DÜZLEŞTİRME SPLAYNLARININ HAYAT DIŞI SİGORTA ÜRÜNLERİ FİYATLAMADA ETKİLERİ

THE EFFECT OF SMOOTHING SPLINES ON PRICING OF NON-LIFE INSURANCE PRODUCTS

HANDAN İLHAN

Dr. Öğr. Üy. UĞUR KARABEY
Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin
Aktüerya Bilimleri Anabilim Dalı için Öngördüğü
YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2018

HANDAN İLHAN' ın hazırladığı **"Düzleştirme Splaynlarının Hayat Dışı Sigorta Ürünleri Fiyatlamada Etkileri"** adlı bu çalışma aşağıdaki jüri tarafından **AKTÜERYA ANABİLİM DALI**' nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Erdem KIRKBEŞOĞLU
Başkan .....

Dr. Öğr. Üy. Uğur KARABEY
Danışman .....

Doç. Dr. Nihal ATA TUTKUN
Üye .....

Dr. Öğr. Üy. Yasemin GENÇTÜRK
Üye .....

Dr. Öğr. Üy. Başak BULUT KARAGEYİK
Üye .....

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof Dr. Menemşe GÜMÜŞDERELİOĞLU
Fen Bilimleri Enstitüsü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin / raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma ama iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **" Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"** kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- o Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- o Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren .... Ay ertelenmiştir. (2)
- o Tezimle ilgili gizlilik kararı verilmiştir. (3)

06 / 09 / 2018

(İmza)

Öğrencinin Adı SOYADI

Handan İLHAN

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. Şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü ve fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir*. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.
Madde 7. 2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

* Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

06./09./2018

HANDAN İLHAN

ÖZET

DÜZLEŞTİRME SPLAYNLARININ HAYAT DIŞI SİGORTA ÜRÜNLERİ FİYATLAMADA ETKİLERİ

Handan İLHAN

Yüksek Lisans, Aktüerya Bilimleri Bölümü

Tez Danışmanı: Dr. Öğr. Üy. Uğur KARABEY

Eylül 2018, 64 Sayfa

Genelleştirilmiş Doğrusal Modeller (GLM) hayat dışı sigorta ürünlerinin fiyatlanmasında en yaygın kullanılan yöntemlerdendir. Hasar dağılımları için üstel dağılım ailesinden dağılımların seçilebiliyor olması, yöntemin sigorta şirketleri tarafından yaygın olarak kullanımına olanak sağlamaktadır. Hasar verileri, genel yapısı itibariyle kalın kuyruklu ve sağa çarpık niteliktedir. Hasar verilerinin bu yapısından dolayı, bu tez çalışmasında GLM kurulumu için hasar sayısı dağılımına Poisson dağılımı, hasar tutarına ise Gamma dağılımı uygunluk göstermiştir.

Hayat dışı sigorta fiyatlama verileri bir ya da birden çok sürekli değişken içermektedir. Fiyatlama uygulamalarında genel yaklaşım, sürekli değişkenlerin belirlenen aralılarla kategorik olarak gruplara ayrılması ve aynı aralıktaki poliçelerin özdeş olarak değerlendirilmesidir. Ancak sürekli değişkenin kategorik hale dönüştürülmesinde genel olarak kabul görülen bir yöntemin olmaması bir dezavantaj oluşturmaktadır. Öte yandan kategoriler için seçilen sınırlar arasındaki geçişlerde risk primleri açısından keskin ayrımların bulunması adil prim ilkesine ters düşmektedir. Bu sorunun ortadan kalkması

için sürekli değişkenleri aralıklara bölmek yerine Genelleştirilmiş Toplamsal Modellerin (GAM) kullanımı, hayat dışı sigorta fiyatlama uygulamalarına bir alternatif sunmaktadır. GAM yöntemi bir düzleştirme splaynı eklentisi ile GLM'nin yarı parametrik bir modele dönüşmüş halidir. Yarı parametrik oluşu ile modele esneklik sağlaması en büyük avantajıdır. Esneklik ile anlatılmak istenen sürekli değişkenlerin düzleştirme splaynları olarak modelde yer almasıdır. Bu sayede sürekli değişkenin her noktasındaki bilgiler modelde yer alır ve bilgi kaybı söz konusu olmaz. Splayn fonksiyonu için çapraz geçerlilik (*Cross-Validation*) yaklaşımı ile düzleştirme parametresi için en uygun değer otomatik olarak seçilmektedir. Bu tezin amacı, sürekli değişkenlerin modeldeki etkisi için B-splayn formunda temsil edilen kübik düzleştirme splaynlarının GAM yöntemindeki kullanımını araştırmaktır. Özel bir sigorta şirketinden alınmış olan kasko veri kümesi üzerine yapılan uygulama ile Genelleştirilmiş Doğrusal Modeller ve Genelleştirilmiş Toplamsal Modellerin karşılaştırılması yapılarak araştırma sorusu cevaplanmıştır.

Anahtar Kelimeler: Genelleştirilmiş Doğrusal Model, Genelleştirilmiş Toplamsal Model, Düzleştirme splaynları, Fiyatlama, Çapraz Geçerlilik

ABSTRACT

THE EFFECT OF SMOOTHING SPLINES ON PRICING OF NON-LIFE INSURANCE PRODUCTS

Handan İLHAN

Master of Science, Departmant of Actuarial Sciences

Supervisor: Assist. Prof. Dr. Uğur KARABEY

September 2018, 64 Pages

The Generalized Linear Models (GLM) is one of the most commonly used methods of pricing non-life insurance products. The method gives a great advantage for selecting loss distribution from exponential family. Typically loss distributions are right-skewed and long-tailed which means the appropriate distributions are Poisson distribution for claim frequency and Gamma distribution for claim severity. Non-life insurance data involves several continuous variables. GLM categorizes the continuous variables into intervals and treats them as identical. However, categorizing the continuous variables method has no common rules and as a result of that it causes some information loss at the breaking points. Instead of categorizing the continuous variables there is an alternative method which is known as Generalized Additive Models. (GAM) provides an alternative modelling without transforming continuous variables into categorical variables. GAM method has same properties with GLM except a semiparametric model with a smoothing spline add-on. In other definition GAM is semiparametric GLM. The biggest advantage of GAM is that the model is flexible with semi-parametric formation. By flexibility we mean that continuous variables are to be included in the model as smoothing splines. In this case, the information on each point of the continuous variable is included in the model. The optimal value for the

smoothing parameter is automatically selected by the cross-validation approach for the spline function.

The aim of this thesis is to study the use of cubic smoothing splines represented in the B-spline form for the effect of continuous variables in the GAM method. Generalized Additive Models and Generalized Linear Models is compared through the insurance loss dataset applications and the research question is answered.

Keywords: Generalized Linear Model, Generalized Additive Model, Smoothing Splines, Cross-Validation

TEŞEKKÜR

Tez çalışmamın her aşamasında desteğini hissettiren sonuna kadar pes etmemem gerektiğine inandıran danışmanım Sayın Dr. Öğr. Üy. Uğur KARABEY'e,

Beni bugünlere hazırlayan her zaman benim için en iyisini isteyen babam Burhan İLHAN'a, her zaman sevgi dolu, şefkatli, anlayışlı, öğretmenim, annem Gülay İLHAN'a, her zaman sorgusuzca yanımda olan küçük eczacı kardeşim Ayşegül İLHAN'a

Beni en iyi anlayan, aynı umutları ve sevinçleri paylaştığım Göksel YÜKSEL'e,

Her geçen gün daha çok özlediğim, her zaman yolundan gideceğim, örnek aldığım canım dayım Şehit Binbaşı Yılmaz TANKÜL'e,

Teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

ÇİZELGELER

Sayfa

ŞEKİLLER

Sayfa

KISALTMALAR

Kısaltmalar

GLM	Genelleştirilmiş Doğrusal Model
GAM	Genelleştirilmiş Toplamsal Model
ÜDA	Üstel Dağılım Ailesi

1. GİRİŞ

Sigorta portföylerinin heterojen bir yapıya sahip olması ve bunun yanı sıra şirketler arasındaki rekabetin zaman içerisinde artıyor olması gibi sebepler dolayısıyla, sigorta şirketleri poliçe sahiplerinin risk profilini analiz edebilecek en doğru yöntemi aramaya yoğunlaşmışlardır. Bu nedenle sigorta şirketleri poliçe sahiplerini farklı risk gruplarına ayırmaktadır. Buradaki amaç; aynı risk grubunda bulunan poliçe sahipleri aynı prim tutarını ödemekle yükümlü olacağı varsayımıdır. Risk gruplarının oluşturulması için sınıflandırılmış değişkenlerin kullanıldığı regresyon yöntemleri kullanılmaktadır. Sınıflandırılmış değişkenler araç sigortası fiyatlamasında oldukça sık kullanılan değişkenlerdir. Sigorta şirketlerinin hayat dışı sigorta branşlarında fiyatlama için risk sınıflandırması ve adil prim ilkesi amacıyla en sık kullandıkları yöntem Genelleştirilmiş Doğrusal Modellerdir. GLM, ilk kez 1972 yılında Nelder ve Wedderburn [1] tarafından tanıtılmıştır. Bu çalışmada en çok olabilirlik yöntemi kullanılarak, parametre tahminleri iteratif algoritma aracılığı ile elde edilmiştir. GLM'nin ilk sigorta uygulaması için hasar sıklığı ve hasar büyüklüğünün modellenmesi McCullagh ve Nelder [2] tarafından 1983 yılında yapılmıştır. Temel çıkarımları McCullagh ve Nelder'ın [3] kitabında yer almaktadır. GLM model kurulumunda, iyi çalışan bir yöntem olması dışında iyi anlaşılır olması ve çeşitli bilgisayar yazılımlarıyla uygulamalara elverişli olması gibi çok sayıda avantajı olan bir modellemedir. Brockman ve Wright [4] tarafından 1992 yılında yazılmış olan makalede, geçmiş hasar verilerinden elde edilen risk ve prim hesabı için istatistiksel detaylara yer verilmiş ve farklı hasar türleri için hasar sıklığının ve şiddetinin iki ayrı model olarak kurulması savunulmuştur. Bu modellemeler için ise GLM teorisini önermişlerdir. Haberman ve Renshaw [5] farklı tip sigorta ürünleri için GLM uygulamıştır. Murphy, Brockman ve Lee [6] GLM kullanarak poliçe sahiplerinin kişisel özelliklerini modele dâhil ederek güçlü bir fiyatlama sistemi üzerinde çalışmışlardır. Smyth, Jørgensen [7] her bir hasar gözlemi için hasar sayısının Poisson, hasar tutarının ise Gamma dağıldığı varsayımı üzerinden Tweedie dağılımını kullanarak direkt olarak risk modellemesinin uygulanmasını önermişlerdir. Arthur ve Renshaw [8] hasar süreci için GLM yöntemini kullanmışlardır. Jong ve Heller [9]; Kaas, Goovaerts, Dhaene ve Denuit [10]; Frees [11] GLM'nin hayat dışı risk modellemesindeki önemini vurgulamışlardır. David [12]; Kafková ve Křivánková [14]; Valecký [15] aktüeryal modelleme için GLM'yi bir sigorta portföyü

üzerinde uygulamışlardır. Rosenlund [16] hayat dışı sigorta çarpımsal fiyatlamada ortalama kareli hata için benzetim ile en iyi nokta tahminini elde ederek Tweedie ve GLM arasında karşılaştırma yapmıştır.

GLM genelleştirilmiş kelimesinden de çağrıştıracağı gibi Normal dağılımın temel alındığı doğrusal modellerin genelleştirilmiş bir versiyonu olup, Üstel dağılım ailesi üyesi olan Poisson, Binom, Gamma gibi dağılımların model içerisinde kullanımına olanak sağlamaktadır. Sigorta uygulamalarında GLM'nin geniş bir yer tutmasının iki temel sebebi vardır. Bunlardan ilki, sigorta verisinin normal dağılıma uyum göstermemesidir. İkinci sebebi ise yanıt ve açıklayıcı değişkenler arasındaki ilişkinin toplamsal olarak değil çarpımsal olarak açıklanabiliyor olmasıdır. Ohlsson ve Johanson [17] hayat dışı veri analizi için çarpımsal modelin yapısını detaylı olarak açıklamıştır.

GLM; rasgele, sistematik ve bağ fonksiyonu olmak üzere 3 bileşenden oluşmaktadır ve açıklayıcı değişkenlerin oluşturduğu sistematik yapı ile yanıt değişkeninin oluşturduğu rasgele parametrik yapıyı bir bağ fonksiyonu sayesinde birleştirir. Ancak bağ fonksiyonu ile kurulan bu yapının doğrusal olmadığı durumlar söz konusu olabilmektedir. Böyle bir sorun ise fiyatlama değişkeninin hasar sıklığı üzerindeki etkisinin yanlış değerlendirilmesine neden olur. Yanıt değişkeni üzerinde doğrusal olmayan sürekli değişkenlerin etkisi için GLM yetersiz kalmaktadır. GLM açıklayıcı değişkenleri bir doğrusal önkestirim formunda modeller. Bu durum kategorik değişkenlere uygulanabilirken sürekli değişkenlerin doğrusal olmayan etkisi için kısıtlı kalmaktadır. Hayat dışı sigorta branşları için sürekli açıklayıcı değişkenlerin fiyatlamadaki problemini çözebilmek ve modele ekleyebilmek için farklı yöntemler aranmalıdır. Bu bağlamda en genel yaklaşım sürekli değişkenleri belirli sayıda sınıflara bölerek modele dâhil etmektir. Dolayısıyla, yeniden kategorik hale dönüşen bu değişkenler yeterli bir şekilde GLM içerisinde yer alabilmektedir.

Hayat dışı sigorta fiyatlanmasında kullanılan genel yaklaşım olan sürekli değişkenlerin kategorik olarak alt sınıflara bölünmesi bazı dezavantajlara neden olmaktadır. Gruplandırma işleminin dezavantajları Altman ve Royston [18] tarafından ele alınmıştır. Fiyatlama modelinde oluşturulan farklı gruplar için ani bir prim değişimi söz konusu olabilir. Ayrıca, açıklayıcı değişkenlerin aralıklara bölünerek gruplar haline getirilirken bilgi kaybına neden olması da veri kullanımında etkisiz bir sonucun ortaya çıkmasına

neden olabilmektedir. Uygun aralıkları belirleme ve gruplara ayırma yöntemi oldukça meşakkatli ve zaman alan bir işlemdir.

GLM dışında kullanım alanı dar olan alternatif yöntemler de mevcuttur. Ruppert [19] yılında esnek regresyon modelleri üzerine bir çalışma yapmıştır. Sürekli değişkenin kategorik hale getirilmesi, o değişkendeki birçok bilginin kaybına neden olsa da, Klein [20] bu şekilde kurulmuş bir modelin daha pratik olacağını savunmuştur. Öte yandan, sürekli değişkeni alt sınıflara bölerken uygun sınırların belirlenebilmesi için herhangi bir kural olmaması sınırları belirleme de zorluk yaratmaktadır. Bu noktada ise Genelleştirilmiş Toplamsal Modeller (GAM), doğrusal olmayan modellemeler için iyi bir seçimdir. GAM, fonksiyonları toplamsal bir yapıya sahiptir ve bileşenleri düzgün olma varsayımı altında GLM'nin yarı parametrik ve genişletilmiş şeklidir. GAM, doğrusal ya da doğrusal olmayan regresyon gibi geleneksel yöntemlere kıyasla çok daha yüksek esneklik özelliğine sahiptir. Parametrik tahmini kolaylaştırır ve yanıt değişkeni ile açıklayıcı değişkenler arasındaki ilişkiyi iyi bir şekilde temellendirmeye yarar. GLM'ye benzer şekilde yanıt değişkeninin ortalaması ile açıklayıcı değişkenlerin düzgünleştirilmiş fonksiyon arasındaki ilişkiyi bir bağ fonksiyonu ile kurmaktadır. GAM'ın güçlü yanlarından biri yanıt değişkeni ile açıklayıcı değişkenler arasındaki yüksek derecede doğrusal olmayan ve monoton olmayan ilişki sorununun üstesinden gelebiliyor olmasıdır. Hastie ve Tibshirani [21] Genelleştirilmiş Toplamsal Modeller teorisi ile GLM başta olmak üzere olabilirlik temeline dayanan tüm regresyon modellerinde uygulanabilen ve $\sum \beta_i X_i$ doğrusal bileşenin yerini düzgün fonksiyonların toplamı olan $\sum s_i(X_i)$ toplamsal kestirimin yer aldığı bir model önermişlerdir. Bu yöntem ile doğrusal olmayan açıklayıcı değişkenlerin model üzerindeki etkileri açığa çıkarılabilmektedir. Hastie ve Tibshirani [22] yayımladıkları makalelerinde toplamsal modelin üstel aileye genişletilmesini örneklendirmişlerdir. GAM formu $g(\mu(x)) = \alpha + \sum_{i=1}^{p} f_i x_i$ olarak belirtilmiştir. Hastie ve Tibshirani [23] tarafından ortaya atılmış olan GAM, Denuit ve Lang [24] ve Wood [25] tarafından yapılan çalışmalarda da yer almıştır.

GAM sürekli değişkenlerin doğrusal olmayan ve veri kümesinden tahmin edilmiş bilinmeyen bir düzgün (*smooth*) fonksiyon ile modele dâhil edilmesine izin verir. Böylece GAM ile modelleme, olası doğrusal olmayan bağımlılığın keşfedilmesini ve sürekli değişken ile yanıt değişkeni arasındaki bağımlılığın temelini düzgün görseller ile elde

edilmesini sağlar. GAM uygulama alanı geniş bir yöntemdir. İstatistiksel olarak etkili düzleştirme parametresi tahmin yöntemi sürecinde modelin bileşen fonksiyonlarının ne derecede düzgünleştirilmiş olması gerektiği konusundaki performansı açısından GAM yönteminin kullanımı sıkça görülmektedir.

GAM sürekli değişkenin model üzerindeki etkisini araştırmak amacıyla düzleştirme splaynlarını kullanmaktadır. Splayn, parçalı polinomların düğüm adı verilen noktalarda birleşerek oluşturduğu bir fonksiyondur. Polinomlar arasında uyum oluşturularak düzgün türevlenebilir bir formda eğri elde edilir. Sürekli değişkenlerin etkisini modele dâhil etmek için splaynların kullanımı GLM teorisinin en önemli gelişimi olmuştur. Bu teori 1994 yılında Green ve Silverman'ın [26] çalışması ile başlangıç yapmıştır. Cezalandırma ve splayn modellerini çeşitlendirerek tartışmışlardır. 1985 yılında Silverman [27] düzleştirme teknikleri için splaynları kullanmıştır. Hastie ve Tibshirani [28] GAM ve düzleştirme yöntemlerinin daha esnek parametrik olmayan fonksiyonlar ile geleneksel parametrik fonksiyonların yerini aldığını savunmuştur. Wahba [29] splayn modellerini matematiksel olarak temellendirmiştir. Ellers ve Marx [30] uygun ceza parametresinin seçimi için B-splaynlarını ele almışlardır. Lee [31] bir benzetim çalışması ile düzleştirme splaynları için düzleştirme parametresinin seçiminde en uygun yöntemi araştırmıştır. Burman [32] GAM için splayn tahminlerini ele almış, model tahmin kriteri için çapraz geçerlilik yöntemini kullanmıştır. Düğüm noktalarının sayısının belirlenmesi konusundaki belirsizlikten söz etmiş ve bundan dolayı veri kümesine bağlı olarak seçilen düğüm sayısının belirlenmesinin model kurulumundaki önemini vurgulamıştır.

Düzleştirme splaynları uygulamalarda başarılı bir şekilde kullanılması dışında, sigorta sektöründe oldukça yeni bir yöntem olup geniş bir uygulama alanı henüz bulamamıştır. Bu tezin amacı, çeşitli fiyatlama sorunlarına ışık tutarak düzleştirme splaynlarının olası kullanımlarını keşfetmektir.

Bu tez çalışmasının birinci bölümünde literatür taramasına ayrıntılı bir şekilde yer verilmiştir.

İkinci bölümde hayat dışı sigorta fiyatlaması ile ilgili açıklamalar, tanımlamalar ve veri kümesinin formundan bahsedilmektedir. Hayat dışı sigorta fiyatlaması uygulamalarında standart yöntem olarak bilinen Genelleştirilmiş Doğrusal Modeller tanıtılmaktadır. Farklı bağ fonksiyonlarıyla herhangi bir üstel dağılım ailesi üyesi doğrusal modelleme için

uygulanabilmektedir. Ancak hasar sıklığı ve hasar şiddeti modellemesi için standart kabul edilen Poisson ve Gamma dağılımlarına ağırlık verilmektedir.

Üçüncü bölümde GLM teorisinin genişletilmiş bir versiyonu olan Genelleştirilmiş Toplamsal Modeller tanıtılmaktadır. GAM parametrik ve parametrik olmayan terimlerin aynı modelde yer almasına izin vermektedir ve bu özellik sayesinde model büyük derecede esneklik kazanmaktadır. Burada sürekli değişkenlerin etkisini saptayabilmek için düzleştirme splaynları yöntemi kullanılmaktadır. Sonrasında tahmin sürecinde oldukça önemli yeri olan kübik ve B-splaynların temel özellikleri verilmektedir.

Dördüncü bölümde Rstudio programı kullanılarak hayat dışı sigorta fiyatlama uygulaması verilmektedir. Uygulama içeriği olarak, GLM ve GAM ile hasar şiddeti ve hasar sıklığı için kurulan modeller ile ilgili sonuçlar kıyaslanmaktadır.

Son olarak, sonuç bölümünde ise tez çalışmasının genel anlatımına ve uygulama ile ilgili sonuçların yorumlamasına yer verilmektedir. GLM ve GAM'ın sigorta sektöründeki yeri, gelişimi ve öneminden bahsedilmektedir.

2. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER

Genelleştirilmiş doğrusal modellerin temel teorisi, 1970'lerin başında Nelder ve Wedderburn tarafından tanıtılmıştır, 1990'ların ikinci yarısında ise kullanımı oldukça yaygınlaşmıştır.

Hayat dışı sigorta matematiği, bir sigorta şirketine ulaşan ve şirketin iflasından kaçınmak için ne kadar primin ödenmesi gerektiği konusunda -tarife analizi olarak bilinen- modellerin analizi ile ilgilidir. Tarife analizi, bir sigorta şirketinin poliçelerinden elde ettiği sigorta verilerine ve portföyden gelen hasarlara dayanarak bir tarife (hesaplanmış prim) elde etmek için yapılan aktüeryal bir çalışmadır.

Hasar- Sigortaya konu olan bir riskin gerçekleşmesi sonucunda oluşan hasarın tazmin edilmesi için, sigorta şirketine poliçe sahibi tarafından yapılan tazminat talebine hasar denir.

Poliçe süresi- Bir sigorta poliçesinin yürürlükte olduğu zaman dilimine poliçe süresi denir. Poliçe süresi, genellikle yıl bazındadır ve poliçe yılı olarak ifade edilir.

Hasar sıklığı- Belirli bir zaman dilimi içerisinde yürürlükte olan grup poliçelerinin hasar sayısının poliçe yılına bölünmesi ile elde edilen ortalama hasar sayısına hasar sıklığı denir.

Hasar şiddeti- Toplam hasar tutarının, hasar sayısına bölünmesi ile hasar başına düşen ortalama maliyet hasar şiddetini vermektedir.

Net risk primi- Belirli bir sigorta dönemine ait muhtemel hasar miktarı ve hasar masraflarını karşılamak üzere, sigorta şirketine ait veriler ile istatistiksel yöntemler kullanılarak hesaplanmış prime net risk primi denir.

Çizelge 2.1. Fiyatlama Oranları

Exposure Ağırlığı: ω	Yanıt: X	Fiyatlama Oranları: $Y = X / \omega$
Poliçe süresi	Hasar sayısı	Hasar sıklığı
Hasar sayısı	Hasar tutarı	(ortalama) Hasar şiddeti
Poliçe süresi	Hasar tutarı	Risk primi

Hasar sıklığı, hasar şiddeti ve risk primi fiyatlama oranlarıdır. Bu oranlar rasgele değişkenlerin sonucu ile exposure ağırlığı arasındadır. Fiyatlama analizi doğrudan yanıt

değişkenleri ile değil bu fiyatlama oranlarına dayanır. Exposure ağırlığı fiyatlama analizinde önemli bir yer almaktadır.

Prim, her bir poliçe için fiyatlama değişkenleri olarak adlandırılan değişkenlerin değerleri ile belirlenmektedir. Poliçe sahibinin yaşı, araç tipi ve araç yaşı gibi değişkenler fiyatlama değişkenlerine örnek gösterilebilir. Her bir fiyatlama değişkeni için aynı sınıfa ait olan poliçeler aynı tarife sınıfına aittir.

Modelin oluşumu için üç temel varsayım vardır:

1. *Zaman bağımsızlığı-* n farklı zaman aralığı olsun. Çizelge 2.1.'deki X_i *i.* zaman aralığındaki yanıt değişkeni olsun. Öyleyse $X_1, \dots, X_n$ bağımsızdır.

2. *Poliçelerin bağımsızlığı-* Aynı zaman dilimine ait n tane farklı poliçe birbirinden bağımsızdır.

3. *Homojenlik-* Aynı exposure değerine ve tarife hücresine ait iki farklı poliçe aynı dağılıma sahiptir.

Çizelge 2.2. r Adet Fiyatlama Değişkeni ile Sigorta Verisi

i	ω_i	x_{1i}	...	x_{ri}	y_i
1	ω_1	x_{11}		x_{r1}	y_1
2	ω_2	x_{12}		x_{r2}	y_2
.	.	.		.	.

Fiyatlama analizinde kullanılan veri setinin formu Çizelge 2.2.'de verildiği gibidir. Sütunlarda bulunan x_{1i}, ..., x_{ri} fiyatlama değişkenleri her i gözlemi için birer değer içerir. Hasar sıklığı analizinde ağırlık ya da exposure olarak kullanılan ω_i poliçe yıl sayısı, y_i ise hasar sıklığıdır. Hasar şiddeti analiz edilirken ise ω_i hasar sayısı, y_i hasar şiddetidir.

2.1. Doğrusal Modeller

Doğrusal modeller yanıt değişkeni Y_i ve açıklayıcı değişkenler x_{1i}, ..., x_{ri} arasındaki doğrusal ilişkinin açıklanması bu ilişkinin bir modelle ifade edilerek tahminlerin yapılması amacıyla kullanılmaktadır.

Y_i raslantı değişkeni $y_1,\ldots,y_n$ n adet bağımsız gözlemler olsun. Klasik doğrusal modellerin temel varsayımları aşağıdaki gibidir:

- Y_i yanıt değişkeni normal dağılıma uymalıdır.

$$Y_i \sim N(\mu_i, \sigma^2) \tag{2.1}$$

- β_0 keşisim noktası $\beta_1,\ldots,\beta_r$ bilinmeyen parametreleri ve x_{ij}'ler bilinen açıklayıcı değişkenler r adet açıklayıcı değişken doğrusal önkestiriciyi verir.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}, \tag{2.2}$$

- Modelin beklenen değerleri ile doğrusal önkestirici arasında doğrudan bir ilişki vardır. (Birim bağ)

$$E(Y_i) = \mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} \tag{2.3}$$

Çoklu doğrusal regresyon gibi doğrusal modellere dayalı teknikler, uygulamalı ekonometride en çok kullanılan istatistiksel yöntemlerdir. Ancak aktüeryal uygulamalarda doğrusal modellerdeki tüm varsayımlar bir sigorta verisi uygulamaları için uygun olmayabilir. İstatistiksel teori ve bilgisayar yazılımlarının ilerlemesi Genelleştirilmiş Doğrusal Modeller teorisi ile yanıt değişkeninin sürekli ya da kategorik olmasına bağlı olmayarak; yanıt değişkeni normal dağılım dışında dağılımlara sahip olabilmektedir ve yanıt değişkeni ile açıklayıcı değişkenler arasındaki ilişkinin basit doğrusal formda ifade edilme zorunluluğu yoktur.

2.2. Genelleştirilmiş Doğrusal Modellerin Yapısı

Genelleştirilmiş Doğrusal Modeller yanıt değişkeninin normal dağılımdan farklı olarak, doğrusal olmayan bir yapı kullanımına ve üstel dağılım ailesi üyesi dağılımların kullanımına olanak sağlamaktadır. Dağılım ve bağ fonksiyonu belirlenerek model kurulması dışındaki özellikler itibariyle Klasik Doğrusal Model ile temel formülizasyonu hemen hemen aynıdır. Genelleştirilmiş Doğrusal Modellerde, olabilirlik maksimizasyonu için iteratif en küçük kareler yaklaşımını gerekli kılıyor olsa da tahmin ve çıkarsamalar En Çok Olabilirlik Tahmini teorisine dayanmaktadır [25].

GLM genel yapı itibariyle (2.4) eşitliğindeki gibidir.

$$g(\mu_i) = \mathbf{X_i}\boldsymbol{\beta} \tag{2.4}$$

Genelleştirilmiş doğrusal modellerin yapısı üç ayrı bileşen ile incelenmektedir. Bu bileşenler; Rasgele bileşen olarak da bilinen yanıtın dağılımı bileşeni, her bir gözlem için atanan doğrusal önkestiricilerin bulunduğu sistematik bileşen ve bağ fonksiyonudur. Bu bileşenler, aşağıdaki gibi tanımlanabilir.

1. *Rasgele Bileşen*: Her bir değişkeninin rasgele ve bağımsız yanıt gözlemlere dayandığı ve üstel dağılım ailesine ait dağılımlardan biri olmak üzere aynı tür olasılık dağılımına sahip olan yapıdır.

2. *Sistematik Bileşen*: Her bir gözlem için bir doğrusal önkestirici atar ve x_{i1}, x_{i2}…x_{ir} açıklayıcı değişkenleri içerir.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_0 x_{ir}\,, \tag{2.5}$$

$\beta_1, \ldots, \beta_r$ bilinmeyen parametreler, β_0 katsayı ve x_{ij}'ler açıklayıcı değişkenlerdir. GLM' nin bir diğer avantajı ise doğrusal önkestiricinin yapısı, doğrusal modelin yapısı ile benzer olmasıdır. Doğrusal önkestirici seçiminde uygun olabilen herhangi bir yapı kullanılabilmektedir [25].

3. *Bağ Fonksiyonu*: Doğrusal modellerde ortalama, açıklayıcı değişkenlerin doğrusal bir fonksiyonudur. GLM için ise, ortalamanın monoton bir dönüşümü açıklayıcı değişkenlerin doğrusal bir fonksiyonudur. Esnek olan ve tersi alınabilen g(.) fonksiyonu doğrusal önkestirici ile Y_i dağılım fonksiyonunun beklenen değeri μ_i ile bir bağ kurmasından dolayı bağ fonksiyonu adını alır.

$$g(\mu_i) \;=\; \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_0 x_{ir}\,. \tag{2.6}$$

Bağ fonksiyonu tersi alınabilir özelliğe sahip olmasından dolayı (2.7) eşitliğindeki gibi yazılabilmektedir.

$$\mu_i = \mathrm{g}^{-1}(\eta_i) = \mathrm{g}^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_0 x_{ir}) \tag{2.7}$$

2.3. Üstel Dağılım Ailesi

Genelleştirilmiş doğrusal modeller, doğrusal regresyon teorisine kıyasla önemli derecede geliştirilmiş bir teoridir. Bu ilerlemenin en büyük göstergesi, doğrusal model ile kıyaslandığında yalnızca normal dağılımla sınırlı kalmayıp daha geniş dağılım ailelerinin kullanımına olanak sağlamasıdır. GLM üstel dağılım ailesinden herhangi bir dağılımla çalışmaya imkân vermektedir ve bu da sigorta verileri üzerindeki uygulamalarda büyük bir esneklik sağlamaktadır.

Üstel dağılım ailesinin genel formu (2.8) eşitliğindeki gibidir.

$$f_i(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}, \tag{2.8}$$

$f_i(y_i;\theta_i,\phi)$, Y_i kesikli ya da sürekli raslantı değişkeninin olasılık yoğunluk fonksiyonunu verir. $a(.)$, $b(.)$, $c(.)$ fonksiyonları üstel dağılım ailesindeki dağılımların genel gösterimini verir. $\theta_i = g(\mu_i)$ ise üstel dağılım ailesi için kanonik parametreyi ifade eder aynı zamanda Y_i raslantı değişkeninin beklenen değerinin fonksiyonudur. $\phi > 0$ ise yayılım (ölçek) parametresidir ve bazı ailelerde sabit ve bilinen bir değeri alırken diğer ailelerde θ_i ile birlikte veriden tahmin edilecek olup bilinmeyen bir parametredir.

$a_i(\phi)$, $b(\theta_i)$ ve $c(y_i, \phi)$ için farklı seçenekler farklı bir dağılım sınıfı ve GLM problemine farklı bir çözüm tanımlamaktadır.

2.3.1. Ortalama ve Varyans

θ_i ve ϕ parametreleri, Y_i raslantı değişkeni ile ilgili ortalama ve varyans bilgilerini verir. Log-olabilirlik fonksiyonu $l(\theta_i, \phi; y_i) = \log f(y_i; \theta_i, \phi)$ şeklinde verilsin. Üstel dağılım ailesi için ortalama ve varyans (2.9) ve (2.10) eşitliğindeki gibi türetilebilmektedir.

$$E\left(\frac{\partial l}{\partial \theta_i}\right) = \frac{\{\mu - b'(\theta)\}}{a(\phi)} = 0 \tag{2.9}$$

$$E\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) + E\left(\frac{\partial l}{\partial \theta_i}\right)^2 = 0. \tag{2.10}$$

2.3.2. Normal Dağılım

μ ve σ^2 olmak üzere iki parametreli olan normal dağılım üstel dağılım ailesinin bir üyesidir. Normal dağılım sigorta verileri modellemesinde kullanımı yaygın bir dağılım değildir. Olasılık fonksiyonu (2.11) eşitliğindeki gibidir.

$$f(y) = f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \tag{2.11}$$

İlk olarak (2.11) eşitliğinin her iki yanının logaritması alınır;

$$\log f(y;\mu,\sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2} \tag{2.12}$$

Elde edilen (2.12) eşitliğini tekrar üstel olarak yazılır;

$$\begin{aligned} f(y;\mu,\sigma^2) &= exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}\right\} \\ f(y;\mu,\sigma^2) &= exp\left\{\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= exp\left\{\frac{2y\mu-\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= exp\left\{\frac{y\mu-\frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)\right\} \end{aligned} \tag{2.13}$$

Buradan, $\theta = \mu,\ \ \phi = \sigma^2,\ \ a(\phi) = \sigma^2,\ \ b(\theta) = \frac{1}{2}\theta^2,\ c(\mathrm{y},\phi) = -\frac{y^2}{2a(\phi)} - \frac{1}{2}\log\big(2\pi a(\phi)\big)$ eşitlikleri ile Normal Dağılımın Üstel Dağılım Ailesinin bir üyesi olduğu görülmektedir.

Beklenen değer ve varyans (2.14) ve (2.15) eşitliklerinde olduğu biçimdedir:

$$E(Y) = \frac{\partial}{\partial\theta} b(\theta) = \frac{1}{2}\frac{\partial}{\partial\theta}\theta^2 = \theta \equiv \mu, \tag{2.14}$$

$$Var(Y) = a(\phi)\frac{\partial^2}{\partial\theta^2} = \sigma^2\frac{\partial}{\partial\theta}\theta = \sigma^2. \tag{2.15}$$

2.3.3. Poisson Dağılımı

Poisson dağılımı üstel dağılım ailesi üyesi olan bir diğer dağılımdır. λ parametreli Poisson dağılımının fonksiyonu (2.16) eşitliğindeki gibidir.

$$f(y) = f(y;\,\lambda) = \; e^{-\lambda}\frac{\lambda^{y}}{y!}. \tag{2.16}$$

Poisson Dağılımının Üstel Dağılım Ailesi üyesi bir dağılım olduğunu göstermek için ilk olarak (2.16) eşitliğinin her iki tarafının logaritması alınır.

$$\log f(y;\,\lambda) = \; y\log\lambda - \lambda - \log(y!). \tag{2.17}$$

Daha sonra elde edilen (2.17) eşitliğin her iki tarafı üstel olarak yazılır.

$$f(y;\,\lambda) = \; exp\{y\log\lambda - \lambda - \log(y!)\} \tag{2.18}$$

$\theta \; = \; \log\lambda$ eşitliği kullanılarak (2.19) eşitliğinde yerine yazılır.

$$f(y;\,\lambda) = \; exp\{y\theta - e^{\theta} - \log(y!)\} \tag{2.19}$$

Buradan elde edilen ϕ = 1, $a(\phi)$ = 1, $b(\theta) = e^{\theta}$, $c(y,\phi)$ =$-\log(y!)$ eşitlikleri ile Poisson Dağılımının Üstel Dağılım Ailesinin bir üyesi olduğu görülmektedir.

Beklenen değer ve varyans (2.20) ve (2.21) eşitliklerinde olduğu biçimdedir:

$$E(Y) = \frac{\partial}{\partial\theta}b(\theta) = \frac{\partial}{\partial\theta}e^{\theta} = e^{\theta} = \lambda, \tag{2.20}$$

$$Var(Y) = a(\phi)\frac{\partial^{2}}{\partial\theta^{2}}b(\theta) = \frac{\partial}{\partial\theta}e^{\theta} = \lambda. \tag{2.21}$$

2.3.4. Gamma Dağılımı

Gamma dağılımına sahip *Y* raslantı değişkeninin α şekil parametresi, β oran parametresi ile olasılık fonksiyonu,

$$f(y;\ \alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \qquad y,\alpha,\beta > 0 \tag{2.22}$$

İlk olarak (2.22) eşitliğinin logaritması alınır.

$$\log f(y;\ \alpha,\beta) = \alpha\log\beta - \log\big(\Gamma(\alpha)\big) + (\alpha-1)\log y - \beta y, \tag{2.23}$$

Daha sonra Logaritması alınmış olan (2.23) eşitliği üstel olarak yazılır.

$$\begin{aligned} f(y;\ \alpha,\beta) &= exp\{-\beta y + \alpha\log\beta + (\alpha-1)\log y - \log\big(\Gamma(\alpha)\big)\} \\ &= exp\left\{\frac{\frac{\beta}{\alpha}y - \log\beta}{-\frac{1}{\alpha}} + (\alpha-1)\log y - \log\Gamma(\alpha)\right\} \end{aligned} \tag{2.24}$$

Buradan, $\theta = \frac{\beta}{\alpha}$, $\phi = \frac{1}{\alpha}$, $a(\phi) = -\frac{1}{\alpha}$, $\beta = \theta\alpha = \frac{\theta}{\phi}$, $\log\beta = \log\theta - \log\phi$ eşitlikleri elde edilir.

Elde edilen eşitlikler olasılık fonksiyonunda yerine koyulduğunda,

$$f(y;\ \theta,\phi) = exp\left\{\frac{\theta y - \log\theta}{-\phi} + \frac{\log\phi}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\big(\Gamma(\alpha)\big)\right\}, \tag{2.25}$$

Buradan $a(\phi) = -\phi, \quad b(\theta) = -\log\theta, \quad c(y,\phi) = \frac{\log\phi}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right)$ eşitlikleri ile Gamma Dağılımının Üstel Dağılım Ailesi üyesi olduğu görülmektedir.

Ortalama ve varyans sırasıyla (2.26) ve (2.27) eşitliklerinde gösterildiği gibidir.

$$E(Y) = \frac{\partial}{\partial\theta} b(\theta) = \frac{\partial}{\partial\theta}\log\theta = \frac{1}{\theta} = \frac{\alpha}{\beta}. \tag{2.26}$$

$$Var(Y) = a(\phi)\frac{\partial^2}{\partial\theta^2}\ b(\theta) = -\phi\frac{\partial}{\partial\theta}\frac{1}{\theta} = \frac{\phi}{\theta^2} = \frac{1}{\alpha}\left(\frac{\alpha^2}{\beta^2}\right) = \frac{\alpha}{\beta^2}. \tag{2.27}$$

Hasar Şiddeti Dağılımı: Gamma Dağılımı

Hasar tutarlarının dağılımları genellikle sağa çarpıktır ve negatif olmayan değerler içerir. Bu sebeple Gamma dağılımı hasar şiddeti için GLM modellemesinde en yaygın olarak kullanılan dağılımlardan biridir. Birbirlerinden bağımsız ω tane her biri Gamma dağılımına uyan raslantı değişkenleri aynı β parametresi ve α parametrelerinin toplamı ile Gamma dağılımına uyum göstermektedir. ω adet raslantı değişkenlerinin toplamı X olarak gösterilsin, $X \sim Gamma(\omega\alpha, \beta)$. Hasar şiddeti için sıklık fonksiyonu $Y = X / \omega$ verildiğinde olasılık fonksiyonu aşağıdaki gibidir:

$$f(y) = \omega f(\omega y) = \frac{(\omega\beta)^{\omega\alpha}}{\Gamma(\omega\alpha)} y^{\omega\alpha-1} e^{-\omega\beta y}; \quad y > 0, \tag{2.28}$$

$Y \sim Gamma(\omega\alpha, \beta)$ olduğuna göre $\mu = \alpha/\beta$ ve $\phi = 1/\alpha$ olarak parametreler (2.29) eşitliğinde yer verildiği gibi gösterilir.

$$f(y) = f(y; \mu, \phi) = \frac{1}{\Gamma(\omega/\phi)} \left(\frac{\omega}{\mu\phi}\right)^{\omega/\phi} y^{(\omega/\phi)-1} e^{-\omega y/(\mu\phi)}. \tag{2.29}$$

Eşitliğin her iki tarafının önce logaritması alınıp daha sonra üstel formda yazıldığında (2.30) eşitliği elde edilir.

$$f(y; \mu, \phi) = exp\left\{\frac{-y/\mu - \log(\mu)}{\phi/\omega} + \log(\omega y/\phi)\omega/\phi - \log(y) - \log\Gamma(\omega/\phi)\right\}. \tag{2.30}$$

Ortalama ve varyans, $E(Y) = \omega\alpha / (\omega\beta) = \mu$ ve $Var(Y) = \omega\alpha / (\omega\beta)^2 = \phi\mu^2/\omega$ şeklinde elde edilir. Son olarak (2.30) eşitliğinde $\theta = -1/\mu$ yazıldığında hasar şiddetinin yoğunluk fonksiyonu (2.31) eşitliğindeki gibi olur.

$$\begin{aligned} f(y_i; \theta_i, \phi) = exp\Bigg\{&\frac{-y_i/\theta_i + \log(-\theta_i)}{\phi/\omega_i} \\ &+ \log(\omega_i y_i/\phi)\omega_i/\phi - \log(y_i) - \log\Gamma(\omega_i/\phi)\Bigg\}, \end{aligned} \tag{2.31}$$

$$a_i(\phi) = \omega_i/\theta, b(\theta_i) = -\log\theta_i, c(y_i, \phi, \omega_i) = \log(y_i\omega_i/\phi)\omega_i/\phi - \log(y_i) - \log\Gamma(\omega_i/\phi).$$

Buradan hasar şiddeti için Gamma dağılımının üstel dağılım ailesinin bir üyesi olduğu görülmektedir.

2.4. Varyans Fonksiyonu

Varyans fonksiyonu her bir üstel dağılım ailesi üyesi için karakteristik bir fonksiyondur. $\mu_i = E(Y_i) = b'(\theta_i)$ fonksiyonun tersi yazılabilen bir fonksiyon olması ile $\theta_i = b'^{-1}(\mu_i)$ yazılabilen fonksiyonu $b''(\theta_i)$ fonksiyonunda yerine yazıldığında (2.32) eşitliği elde edilmektedir.

$$V(\mu_i) = b''\big(b'^{-1}(\mu_i)\big). \tag{2.32}$$

Varyans yapısının ikinci özelliği ise $Var\,(Y_i)$ ile verildiğinde (2.33) eşitliğindeki gibidir.

$$Var(Y_i) = a_i(\phi)V(\mu_i) \tag{2.33}$$

Her üstel dağılım ailesi fonksiyonun o dağılıma ait bir varyans fonksiyonu vardır ve GLM modellemesi uygulandığında ilk olarak varyans fonksiyonu belirlenir. Hasar sıklığı modellemesinde Y_i raslantı değişkeni genellikle Poisson dağılımı $V(\mu_i) = \mu_i$ olarak tahmin edilir. Hasar şiddeti modellemesinde ise Gamma dağılımı $V(\mu_i) = \mu^2{}_i$ genellikle kullanılan üstel dağılım ailesi üyesidir.

Çizelge 2.3. Varyans Fonksiyonları

Dağılım	Normal	Poisson	Gamma
$V(\mu)$	1	μ	μ^2

2.5. Bağ Fonksiyonu

Bağ fonksiyonun kullanımı, ortalamanın doğrusal yapısı göz önüne alındığında normal doğrusal modelin genelleştirilmesi için bir diğer yöntemdir.

$$\eta_i = \sum_{j=1}^{r} x_{ij}\beta_j \qquad i = 1, 2, \dots, n \tag{2.34}$$

Klasik doğrusal modelde önkestirici ve ortalama arasındaki bağ $\mu_i = \eta_i$ birim bağ fonksiyonu ile sağlanırken, GLM için bu bağ fonksiyonu $g(.)$ monoton olma ve türevlenebilme kısıtlaması ile $g(\mu_i) = \eta_i$ bu şekilde tanımlanmaktadır.

$$g(\mu_i) = \eta_i = \sum_{j=1}^{r} x_{ij}\beta_j. \tag{2.35}$$

GLM'de bağ fonksiyonlarının, modelin belirlenmesinde önemli bir yeri vardır ve bunun avantajı ise bağ fonksiyonu seçimi ile yanıt değişkeninin dağılımının birbirinden ayrı şekilde olmasıdır. Sıklıkla kullanılan birçok bağ fonksiyonu çeşidi vardır ve hangi bağ fonksiyonun kullanılacağının seçimi oldukça önem taşır ve dikkatli karar verilmelidir. Hayat dışı sigorta fiyatlamasında log-bağ en çok kullanılan bağ fonksiyonu çeşididir. Açıklayıcı değişkenlerin etkisinin çarpımsal olması özelliği ile $g(\mu_i) = log(\mu_i)$ log-bağ fonksiyonunun tersi $g^{-1}(\mu_i) = e^{\mu_i}$ biçiminde yazılabilir. Bir başka deyişle log-bağ fonksiyonun kullanımı sayesinde toplamsal etki yerine GLM çarpımsal etkinin logaritmasını tahmin eder.

Çizelge 2.4. Bağ Fonksiyonları

	$g(\mu_i)$	$g^{-1}(\mu_i)$
Birim	μ_i	μ_i
Log	$\ln(\mu_i)$	e^{μ_i}
Logit	$\ln(\mu_i/(1-\mu_i))$	$e^{\mu_i}/(1+e^{\mu_i})$
Ters	$1/\mu_i$	$1/\mu_i$

2.6. Yayılım Parametresi

Yayılım parametresi standart hatalar gibi bazı istatistiklerin belirlenmesinde kullanılır. ϕ yayılım parametresi burada eklenti bir parametre olarak düşünülür. ϕ yayılım parametresinin tahmin değeri standart hata ile ilişkilidir ve aralarında doğrusal bir ilişki vardır. ϕ yayılım parametresi büyükse standart hata da büyük değer alır. Bunun sebebi ise ϕ parametresinin büyük değerlerinin varyansı arttırıyor oluşudur. Genel anlamda diğer üstel dağılım ailesi üyesi dağılımlarda en çok olabilirlik yöntemi ile veri setinden

hesaplanmalıdır. Bu yöntemin dezavantajı ise ϕ yayılım parametresi için belirli bir açık formül olmamasından dolayı türevlenemiyor olmasıdır. Bu nedenle de en çok olabilirlik tahmin yöntemi uzun sürebilir. Yayılım parametresi ϕ' nın tahmini için alternatif olarak:

Pearson χ^2 istatistiği moment tahmin edicisi şu şekilde tanımlanır:

$$\hat{\phi}_X = \frac{1}{n-r}\sum_i \frac{\omega_i(y_i-\mu_i)^2}{V(\mu_i)} \tag{2.36}$$

Toplam sapma tahmin edicisi ise:

$$\hat{\phi}_D = \frac{D(y,\hat{\mu})}{n-r} \tag{2.37}$$

r, tahmin edilen β parametrelerinin sayısını temsil eder. D ise sapmayı temsil eder.

2.7. Tweedie Dağılımı

Risk primi modellemesi zorlayıcı bir modelleme türüdür. Bu zorluğun sebebi hangi olasılık dağılımının sigortaya konu olan riski yaklaşık olarak açıklayabilir olduğuna karar vermektir. Çoğu poliçe için hasar oluşmamaktayken hasar oluştuğunda ise hasar dağılımında uzun kuyruklu bir yapıya sebep olmaktadır. Böyle bir kayıp dağılımının karşılığı sıfır yığılmalı aynı zamanda sağa çarpık ve uzun kuyruklu olmalıdır. Tweedie dağılımı bu özellikleri kapsayan bir dağılımdır.

Tweedie dağılımlarında standart üstel dağılım ailesi parametreleri olan μ ve ϕ 'ye ek olarak üçüncü bir p parametresi mevcuttur. p parametresi (0,1) aralığı dışında tüm sayı değerlerini alabilmektedir. Tweedie dağılımının varyans fonksiyonu (2.38) eşitliğinde verildiği gibidir.

$$V(\mu) = \mu^p \tag{2.38}$$

Bazı p parametrelerine göre Tweedie dağılımının özel durumları aşağıda verildiği gibidir.

p=0	Normal dağılım	-
p=1	Poisson dağılımı	Hasar sıklığı
p=2	Gamma dağılımı	Gamma dağılımı
1<p< 2	Bileşik Poisson dağılımı	Risk primi

2.8. Kanonik Bağ

Genelleştirilmiş Doğrusal Modellerde kanonik bağ ile rasgele bileşen arasındaki geçiş (2.39) eşitliğinde verildiği gibidir.

$$\theta \xrightarrow{b'(.)} \mu \xrightarrow{g(.)} \eta \tag{2.39}$$

b(.) ve $b'(.)$ üstel dağılım ailesinin rasgele bileşen yapısı ile belirlenmektedir böylece $g(.) = b'^{-1}(.)$ özel seçimi kanonik bağ olmaktadır. (2.39) eşitliğinde görüldüğü üzere kanonik bağ, θ parametresini η önkesitiricisine eşitler. Bağ fonksiyonu *g*(.) önkestiriciyi kanonik parametre ile aynı yapar.

$$\theta = (b')^{-1}\left(g^{-1}(\eta)\right) = \eta \tag{2.40}$$

Kanonik bağ seçimi dağılımı sadeleştirme anlamına gelmektedir. Dağılımlara ait kanonik bağların örnekleri Çizelge 2.5.'da verildiği gibidir.

Çizelge 2.5. Kanonik Bağ

	Kanonik Bağ
Normal	μ
Poisson	$ln(\mu)$
Gamma	$1/\mu$
Binom	$ln(\mu / (1-\mu))$
Ters Normal	$1/\mu^2$

2.9. Offset

Fiyatlama analizinde μ beklenen değer önceden biliniyor olabilmektedir. Bir diğer deyişle bütün fiyatlama analiz planını tamamen güncellemek yerine bazı unsurlar değiştirilirken bazı unsurlar sabit kalmaktadır. Bu duruma örnek olarak fiyatlama algoritması GLM yani tahmini analizden hariç olarak bir sınıf ya da bir bölge için kendine özgü bir hasar bazı değeri ile başlayabilmektedir. Bu gibi durumlarda sabit değişken (hasar bazı değeri) GLM

ile hesaplanmış katsayılar ile belirlenmemektedir. GLM ile fiyatlama modelinin bir parçası olan bu sabit değişkenin modeldeki varlığı diğer değişkenler için tahmin edilen katsayıların ideal sonuçlar vermesini sağlamaktadır. GLM böyle bir özellikten faydalanabilmek için offset terimine modelde yer verir. Offset terimi belirli bir fiyatlama değişkeninin kademeleri ya da GLM içindeki değişkenler arasındaki bilinen ilişkileri belirlemek için kullanılmaktadır. Offset terimi genellikle Poisson regresyon için bir riske maruz (exposure) ölçü olarak kullanılmaktadır. Verilen önkestirici gösterimi (2.41) eşitliğinde olduğu gibi bir eklenti olarak yer almaktadır.

$$\eta_i = \sum_{j=1}^{r} x_{ij}\beta_j + \xi_i \tag{2.41}$$

Buradan (2.42) eşitliği elde edilmektedir.

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\sum_{j=1}^{r} x_{ij}\beta_j + \xi_i) \tag{2.42}$$

2.10. Parametre Tahmini

GLM'de β parametrelerinin tahmini için en çok olabilirlik yöntemi kullanılmaktadır. Log-olabilirlik fonksiyonu temel basamak olarak parametrelerin tahminleri için her bir β_j parametresine göre türevlenir. Log-olabilirlik fonksiyonunun genel formu (2.43) eşitliğinde verildiği gibidir.

$$\mathcal{L}(\theta, \phi; y) = \sum_{i=1}^{n} \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \tag{2.43}$$

En çok olabilirlik eşitliklerinin açık çözümlemeleri olmamasından dolayı bu eşitlikler iteratif yöntemler kullanılarak çözülmektedir. Doğrusal olmayan eşitlikler için parametre tahmininde kullanılan iki yöntemden biri Newton-Raphson diğeri ise yeniden ağırlıklandırılmış en küçük kareler algoritmasıdır. Bu algoritmalar bilinmeyen β parametrelerinin vektörünü $\frac{\partial_{\mathcal{L}}}{\partial_{\beta_j}} = 0$ çözümü ile iteratif olarak bulmaktadır.

2.10.1. Newton-Raphson

Newton-Raphson yöntemi doğrusal olmayan denklemlerin çözümünde en sık başvurulan yöntemdir. $\frac{\partial_{\mathcal{L}}}{\partial_{\beta_j}}$ skor vektörü ve $\frac{\partial^2{}_{\mathcal{L}}}{\partial_{\beta_j}\partial_{\beta_k}}$ log-olabilirlik ikinci türevinin (gözlem) matrisi olan Hessian matrisi en çok olabilirlik tahmini için iteratif eşitlikler ile hesaplanmalıdır. İterasyon süreci (2.44) eşitliğinde verildiği gibi $\boldsymbol{\beta}$ vektörünün elemanları arasında yaklaşık olarak değişiklik olmayana dek devam eder.

$$\boldsymbol{\beta}^{(p)} \approx \boldsymbol{\beta}^{(p-1)}. \tag{2.44}$$

Newton-Raphson algoritması üstel dağılım ailesi üyesi olan parametrelerin tamamının hesaplanmasını sağlamaktadır.

2.10.2. Yeniden Ağırlıklandırılmış En Küçük Kareler

Fisher-skoru olarak da bilinen algoritma $\boldsymbol{\beta}$ parametre vektörünün tahmin için ağırlıkların iteratif olarak belirlenmesidir. Newton-Raphson yöntemi ile arasındaki fark Hessian matrisi yerine Fisher bilgi matrisi kullanılmasıdır. Fisher bilgi matrisi (2.45) eşitliğinde verilmiştir.

$$\boldsymbol{I} = -E\left(\frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_k}\right) = -E(\boldsymbol{H}) \tag{2.45}$$

Fisher bilgi matrisi parametre tahminlerinin tekrarlı adım sayısı bilgisini vermektedir. Ayrıca Newton-Rapson yönteminde olduğu gibi üstel dağılım ailesi için her bir parametre katsayının tahminini sağlamaktadır. Bu yöntemin uygulanması oldukça kolaydır. Hessian matrisine karşı Fisher bilgi matrisi her zaman pozitiftir. Bundan dolayı Newton -Raphson yöntemine göre daha güvenilir yakınsamalar beklenir.

2.11. Sapma

$\mathcal{L}(\theta, \phi; y)$ fonksiyonu θ_i kanonik parametresi ile *i*. gözlem için GLM'nin log-olabilirlik fonksiyonu olarak tanımlanmıştır. Kanonik bağ özelliği doğrultusunda yanıt değişkeninin beklenen değeri log-olabilirlik fonksiyonu olarak kabul edilmektedir ve buradan yola çıkarak $\widehat{\boldsymbol{\mu}}$ tahmin edici ortalama vektörü olmak üzere $\mathcal{L}(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$ fonksiyonu

yazılabilmektedir. Her bir gözlem için parametre atanması durumunda $\widehat{\boldsymbol{\mu}} = y_i$ eşitliği ile tam uyum sağlanmaktadır ve olabilirlik fonksiyonu $\mathcal{L}(\boldsymbol{y}, \phi; \boldsymbol{y})$ şeklinde yazılabilmektedir. Bu şekilde düzenlenmiş olan model veri seti ile tam bir uyum göstermesi sebebiyle tam model olarak bilinmektedir ve ölçek sapması D^* olarak adlandırılır. Ölçek sapma (2.46) eşitliğinde verildiği gibidir.

$$\begin{aligned} D^* &= D^*(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = 2[\mathcal{L}(\boldsymbol{y}, \phi; \boldsymbol{y}) - \mathcal{L}(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})] \\ &= 2\sum_{i=1}^{n}[\mathcal{L}(y_i, \phi; y_i) - \mathcal{L}(\hat{\mu}_i, \phi; y_i)] \\ &2\sum_{i=1}^{n}\frac{y_i(g(y_i) - g(\hat{\mu}_i)) - (b\big(g(y_i)\big) + b\big(g(\hat{\mu}_i)\big))}{a_i(\phi)} \end{aligned} \tag{2.46}$$

ϕ ve ölçek sapmanın çarpımı (ölçeksiz) sapmayı vermektedir.

$$D = \phi D^* \tag{2.47}$$

Poisson dağılımında $\phi = 1$ eşitliği söz konusu olmasından dolayı ölçek sapma ve sapma $D = D^*$ şeklindedir. Normal, Poisson ve Gamma için sapmalar aşağıda olduğu gibidir:

- Normal: $D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = \sum_i \omega_i \, (y_i - \hat{\mu}_i)^2$
- Poisson: $D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = 2\sum_i \omega_i \, (y_i \log y_i - y_i \log \hat{\mu}_i - y_i + \hat{\mu}_i)$
- Gamma: $D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = 2\sum_i \omega_i \, (y_i/\hat{\mu}_i - 1 - \log(y_i/\hat{\mu}_i))$

Sapma gözlemler ve tahmin değerleri arasındaki farkı vermektedir. Ayrıca (2.48) ve (2.49) eşitliklerinde verildiği şekilde de yazılabilmektedir.

$$d(y_i, \mu_i) = 2\sum_{i=1}^{n} y_i\big(g(y_i) - g(\hat{\mu}_i)\big) - \Big(b\big(g(y_i)\big) + b\big(g(\hat{\mu}_i)\big)\Big), \tag{2.48}$$

$$D = \sum\nolimits_i \omega_i d(y_i, \mu_i). \tag{2.49}$$

2.12. Artıklar ve Model Geçerliliği

Artıklar GLM'de modelin veri setine uyumu, aykırı değerlerin belirlenmesi ve varyans varsayımlarının kontrolü için kullanılmaktadır. Hata terimi ve tahmin değerlerinin gözlem değerlerinden ne derece farklı olduğunu artıkları ile test edilir. GLM'deki artıklar ile doğrusal modellerdeki artıklar genellikle benzeşmektedir. En çok kullanılan Pearson artıkları (2.50) eşitliğinde verildiği gibidir.

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/\omega_i}}. \tag{2.50}$$

Standartlaştırılmış Pearson artıkları (2.51) eşitliğindeki gibidir.

$$r_{SPi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)a_i(\phi)(1-h_i)}}, \tag{2.51}$$

Şapka değeri olarak bilinen $h_i = h_{ii}$ eşitliği doğrultusunda elemanı olduğu şapka matrisi (2.52) eşitliğinde verildiği gibidir.

$$\mathbf{D}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{1/2}. \tag{2.52}$$

Şapka matrisi gözlem değerlerinin vektörü tahmin değerlerinin vektörü ile eşleşmektedir.

$$r_{Di} = sign(y_i - \hat{\mu}_i)\sqrt{\omega_i d(y_i, \hat{\mu}_i)} \tag{2.53}$$

Sapma artıkları ise (2.54) eşitliğindeki gibi tanımlanmaktadır.

$$\sum_i r_{Di}^2 = \phi D^* \tag{2.54}$$

Sapma artıklarının oldukça önemli özellikleri mevcuttur. Genel anlamda normal dağılıma daha yakındır. Gözlemler ve GLM ile tahmin edilmiş beklenen değerler arasındaki fark olarak tanımlanmaktadır. Sapma hesaplaması dağılımın çarpıklığını doğrulamaktadır. Normal dağılımdan büyük bir sapma söz konusu olduğunda model için dağılım varsayımlarının uyum sağlamadığını yorumlayabilmek için iyi bir göstergedir. Standartlaştırılmış sapma artıkları Pearson artıklarında olduğu şekilde elde edilmiştir.

$$r_{SDi} = sign(y_i - \hat{\mu}_i)\frac{\sqrt{d(y_i, \hat{\mu}_i)}}{a_i(\phi)(1 - h_i)} \tag{2.55}$$

Sapma artıkları model tahminleri için varsayımlar geçerli ise varyansın 1'e eşit olabilmesi için standartlaştırılmaktadır.

3. GENELLEŞTİRİLMİŞ TOPLAMSAL MODELLER

Hastie ve Tibshirani 1980'lerde sürekli değişkenin model üzerindeki etkisini çalışmış ve Genelleştirilmiş Toplamsal Modelleri tanıtmışlardır.

Genelleştirilmiş Toplamsal Modeller, Genelleştirilmiş Doğrusal Modellerden farklı olarak hayat dışı sigorta ürünlerinde hemen hemen bir ya da birden çok olan sürekli değişkenleri belirli aralıklara bölerek kategorik bir hale getirmek yerine düzleştirme splaynı uygulaması ile sürekli değişkenin model üzerindeki etkisini analiz edebilmektedir.

Genelleştirilmiş Doğrusal Modeller sürekli değişkenleri aralıklara bölerek gruplandırma işlemi aynı aralıktaki tüm poliçelerin aynı değerleri taşıdığını varsayar. Bu yöntem yeterince iyi çalışmaktadır ancak aynı aralıkta bulunan iki farklı poliçenin prim ödemeleri birbirinden farklı olması gerekebilirken bir sonraki grup aralığındaki poliçe ile de yakın ödemeler olması gerekebilir. Buna örnek olarak poliçe sahibinin yaşı değişkeni üzerinden 18-25 ve 26-30 olarak gruplandırıldığını varsayıldığında 25 ve 26 yaşları arasında çok keskin bir geçiş olması prim ödemelerinde bir adaletsizlik oluşturabilir. Aynı şekilde 26 ve 30 yaşındaki poliçe sahiplerine de aynı derecede risk grubuna ait olarak yaklaşılması da çelişkilidir. Bu da yöntemin dezavantajı olarak görülebilir.

Genelleştirilmiş Doğrusal Modellerin genişletilmiş bir şekli olan Genelleştirilmiş Toplamsal Modeller hem parametrik hem parametrik olmayan terimleri aynı modelde toplayabiliyor olması ile model açısından büyük bir esneklik sağlamaktadır.

3.1. Giriş

x_{ij}, j değişken için i. gözlem değeri olmak üzere $x_{i1}, \dots, x_{ij}$'ler Çizelge 2.1.'deki gibi bir liste formu veri setinde yer alan sürekli ya da kategorik fiyatlama değişkenleri olsun.

$$\eta_i = \sum_{j=1}^{r} x'_{ij}\beta_j, \tag{3.1}$$

β_j'ler bilinmeyen parametrelerdir. x_{ij} yerine x'_{ij} kullanımının nedeni η_i önkestiricisindeki açıklayıcı değişkenlerin x_{ij} orijinal değişkenlerin bir dönüşümü olduğunu vurgulamak içindir. Model açıklayıcı değişkenlerin ilişkili olduğu β parametrelerini gösteren daha açık bir şekilde yazılacak olursa;

$$\eta_i = \beta_0 + \sum_{k_1=1}^{K_1} \beta_{1k_1}\phi_{1k_1}(x_{i1}) + \sum_{k_J=1}^{K_J} \beta_{jk_j}\phi_{jk_j}(x_{ij}). \quad (3.2)$$

β parametreleri (3.2) eşitliğinde yeniden numaralandırılmıştır. j değişkeni kategorik ise $\{z_1, \dots, z_k\}$ arasındaki sınırlı bir değer kümesi varsayımı ile x_{ij}, z_k değerine eşit olup olmamasına göre $\phi_{jk_j}(x_{ij})$, 0 ya da 1 değerini alır. j değişkeni sürekli K_j aralıklarına bölünür sonrasında x_{ij}'nin k aralığında olup olmamasına göre $\phi_{jk_j}(x_{ij})$, 0 ya da 1 değerini alır.

Genelleştirilmiş Toplamsal Modeller ilk kez Hastie ve Tibshirani tarafından 1986 ve 1990 yıllarında tanıtılmıştır. Genelleştirilmiş Toplamsal Modellerin genel yapısı (3.3) eşitliğindeki gibidir.

$$\eta_i = \mathrm{g}(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_j(x_{ij}), \quad (3.3)$$

$\mathrm{g}(\mu_i)$ bağ fonksiyonudur; $\mu_i = E(Y_i)$, Y_i yanıt değişkeninin beklenen değeridir; Y_i yanıt değişkeni üstel dağılım ailesine ait bir dağılıma uymaktadır; f_j fonksiyonları ise düzgün (smooth) fonksiyondur.

Genelleştirilmiş Toplamsal Modeller, GLM'nin genişletilmiş bir hali olduğu için bazı f_j fonksiyonları doğrusal formda varsayılabilmektedir. Doğrusal olmayan f_j fonksiyonları ise düzleştirme splaynları gibi, düzleştirme yöntemleri kullanılarak modellenebilir.

Genelleştirilmiş Toplamsal Modeller, GLM'lere göre daha esnek bir model oluşumu sağlamaktadır. Açıklayıcı değişkenler ile yanıt değişkeninin beklenen değeri arasında doğrusallık şartı aranmamaktadır. Genelleştirilmiş Toplamsal Modellerin temelinde f_j düzgün fonksiyonlarının tahmini yatmaktadır.

3.2. Splayn Fonksiyonları

Splayn, parçalı polinom eğrilerinin kesiştiği yerde birbirine bağlanarak ve bağlandıkları bu noktalarda belli derecelerden türevlenebilen yüksek dereceli pürüzsüz fonksiyonlardır. Parçalı polinomların birbirine bağlandığı $u_1 < \cdots < u_m$ aralığında olacak şekilde sıralanan $u_1, \dots, u_m$ noktaları *düğüm noktaları* olarak adlandırılır. En basit yapılı splayn *doğrusal splayn* olarak tanımlanır.

$$p_k(x) = a_k + b_k x \qquad k = 1, \dots, m-1 \tag{3.4}$$

Doğrusal splaynın bitiş noktalarındaki kesişim (3.5)'te verilmiştir.

$$p_{k-1}(u_k) = p_k(u_k) \qquad k = 2, \dots, m-1. \tag{3.5}$$

$[u_1,\ u_m]$ aralığında sürekli ve herhangi bir u_k ve u_{k+1} aralığında doğrusal bir *s(x)* fonksiyonu tanımlanırsa;

$$s(x) = p_k(x), \quad u_k \le x \le u_{k+1}; \quad k = 1, \dots, m-1. \tag{3.6}$$

Genel bir ifade kullanılacak olursa $[u_1,\ u_m]$ aralığında *j* dereceden bir splayn, her bir $[u_k,\ u_{k+1}]$ aralığındaki *j-1* kez türevlenebiliyor ise *j.* dereceden bir polinom olarak tanımlanabilmektedir.

3.2.1. Kübik Splayn

Kullanımı en yaygın olan kübik splaynlardır. Kübik polinom fonksiyonu (3.7) eşitliğinde verildiği gibidir.

$$p_k(x) = a_k + b_k x + c_k x^2 + d_k x^3 \quad k = 1, \dots, m-1, \tag{3.7}$$

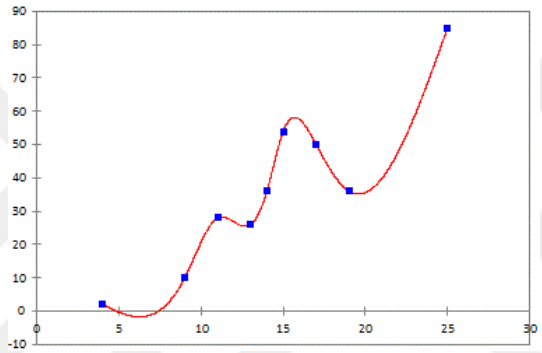
Her bir düğüm noktası (3.8) eşitliğinde ifade edildiği gibidir. Eşitlikte verilen u_k *m* dereceden düğüm sayısını, eşitlik ise ardışık polinomlar için düğüm sayılarının eşitliğini vermektedir.

$$p^{(q)}{}_{k-1}(u_k) = p^{(q)}{}_k(u_k) \qquad k = 2, \dots, m-1, \quad q = 0,1,2, \tag{3.8}$$

$s(x)$ kübik splaynı (3.9) eşitliğindeki gibi tanımlanabilmektedir.

$$s(x) = \begin{cases} p^{(j)}{}_{k,} & u_k \text{ ve } u_{k+1} \\ 0, & diğer \end{cases} \tag{3.9}$$

Bir kübik splayn a_k, b_k, c_k, d_k gibi 4 parametreye sahip m-1 kübik polinom içermektedir. İki kez türevlenebilen sürekli bir *s* fonksiyonu için $4(m-1)$ olarak genel parametre sayısı ifadesi her bir iç düğüm noktasında şartları olan ardışık polinomların düğüm noktalarının birinci ve ikinci türevlerinde eşit olduğu belirterek gerçek parametre sayısını $m+2$'ye düşürülmektedir. Sonrasında ise kübik splayn $[u_1, u_m]$ aralığını dahil ederek $[a, b]$ aralığına genişletilmektedir. $[a, b]$ aralığındaki bir kübik splaynın a ve b noktalarında ikinci ve üçüncü türevleri 0 ise bu splayna doğal splayn denir ve $d_1 = c_1 = d_m = c_m = 0$ olarak gösterilir. Böylece *s*, $[a, u_1]$ ve $[u_m, b]$ uç aralıklarında doğrusaldır.



Şekil 3.1. Kübik Splayn Örneği

Doğal kübik splayn hesaplaması için (3.7) eşitliği uygun bir gösterim değildir. Her bir düğüm noktasında ikinci türevi ve verilen değerleri ile düğüm sayısı kadar parametre sayısı kullanılan doğal kübik splayn tanımlanabilmektedir. $s(x)$ $u_1 < \cdots < u_m$ düğümleri ile bir doğal kübik splayn olsun. O halde (3.10) eşitliği verildiği gibidir.

$$\begin{aligned} \boldsymbol{s} &= (s_1, \ldots, s_m)', \quad s_k = s(u_k) \quad k = 1, \ldots, m \\ \boldsymbol{\gamma} &= (\gamma_2, \ldots, \gamma_{m-1})', \quad \gamma_k = s''(u_k) \quad k = 2, \ldots, m-1 \end{aligned} \tag{3.10}$$

$\boldsymbol{s}$ ve $\boldsymbol{\gamma}$ vektörleri $s(x)$ eğrisini tamamen belirtmektedir.

3.2.2. B-splayn

Kübik splaynları birçok şekilde parametrize etmek mümkündür. Temel splaylanların (basis-splines) doğrusal kombinasyonları ile *B-splaynlar* oluşturulabilmektedir.

B-splayn yaygın olarak kullanılan bir splayn türüdür ve düğüm noktalarından serbestlik özelliği vardır. $k = 1, \ldots, m-2$ için düğüm noktalarında sıçramalar ile basamak fonksiyonları için bir baz tanımlanacak olursa :

$$B_{0,k}(x) = \begin{cases} 1, & x \in [u_k, u_{k+1}]; \\ 0, & \text{öteki değerler için,} \end{cases} \tag{3.11}$$

$$B_{0,m-1}(x) = \begin{cases} 1, & x \in [u_{m-1}, u_m]; \\ 0, & \text{öteki değerler için,} \end{cases} \tag{3.12}$$

Elde edilen (3.11) ve (3.12) sonrasında (3.13) eşitliğinde *B-splaynlar* yinelemeli olarak $j \geq 0$ ve $k = 1, \ldots, m + j$ için tanımlanabilmektedir.

$$B_{j+1,k}(x) = \frac{x - u_{k-j-1}}{u_k - u_{k-j-1}} B_{j,k-1}(x) + \frac{u_{k+1} - x}{u_{k+1} - u_{k-j}} B_{j,k}(x) \tag{3.13}$$

Bu yineleme formülü genellikle Boor's recursion olarak bilinmektedir. $k \leq 0$ ya da $k \geq m + j$ şartları için $B_{j,k}(x) = 0$'dır. Sonrasında $k \leq 0$ için $u_k = u_1$ ve $k \geq m + 1$ için $u_k = u_m$'dir. Buradan anlaşılıyor ki $B_{j,k}(x)$, $(u_{k-j} = u_{j+1})$ aralığında pozitifken diğer noktalarda sıfırdır.

Teorem 1. Verilmiş olan splayn düğümleri için *j* dereceden *B-splaynların* doğrusal bir kombinasyonu olarak $\beta_1, \ldots, \beta_{m+j-1}$ sabitleri için aynı dereceden bir *s* splayn 3.14'teki gibi yazılabilmektedir.

$$s(x) = \sum_{k=1}^{m+j-1} \beta_k B_{j,k}, \tag{3.14}$$

$B_1(x), \ldots, B_{m+2}(x)$'ler kübik B-splaynlar olmak üzere (3.15) eşitliğinde ise *s* kübik splaynı teorem 1'e göre yazılmıştır.

$$s(x) = \sum_{j=1}^{m+2} \beta_k B_j(x) \tag{3.15}$$

3.3. Cezalı Sapma

Birinci bölümde parametre tahminlerinin log-olabilirlik maksimizasyonu ile hesaplandığı gösterilmişti. Ortalama tahmini söz konusu olduğunda sapma minimizasyonu ile tahmin elde edilmektedir. Normal, Poisson ve Gamma dağılımları için $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ ölçeksiz sapmalar verilmiştir.

- Normal: $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_i \omega_i \, (y_i - \hat{\mu}_i)^2$
- Poisson: $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i \omega_i \, (y_i \log y_i - y_i \log \hat{\mu}_i - y_i + \hat{\mu}_i)$
- Gamma: $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i \omega_i \, (y_i / \hat{\mu}_i - 1 - \log(y_i / \hat{\mu}_i))$

Temel fikir sapmaya bir ceza terimi ekleyerek f fonksiyonunu en küçük yapan fonksiyonu bulmaktır.

$$\Delta(f) = D(y, \mu) + \lambda \int_a^b (f''(x))^2 \, dx \tag{3.16}$$

λ parametresi, veriye tam uyum ölçüsü olan sapma ve f fonksiyonunun değişkenliği arasında bir denge kurar. Sapmanın $D(y, \mu)$, $\eta_i = \mathrm{g}(\mu_i) = \beta_0 + f_1(x_{i1})$ eşitliği doğrultusunda f fonksiyonuna bağlı olduğu görülmektedir.

3.4. Tahmin- Tek Fiyatlama Değişkeni

Bu bölümde Poisson ve Gamma dağılımları hasar sıklığı ve hasar şiddeti modellemeleri için Ohlsson, Johansn [17] prensibi kullanılacaktır. Hasar sıklığı ve hasar şiddeti için temel modellerin gösteriminden önce sigorta uygulamalarında nadiren görülen Normal dağılım durumundaki gözlemler için sürekli değişken etkisi incelenecektir.

3.4.1. Normal Dağılım Durumunda

Çizelge 3.1.'deki gibi tek değişkenli bir sigorta verisi üzerinde çalışıldığı varsayılsın. y_i anahtar oran, ω_i exposure ağırlığı, x_i ise herhangi bir sürekli değişken olmak üzere;

Çizelge 3.1. Tek Değişkenli Sigorta Verisi

i	x_i	ω_i	y_i
1			
2			
3			
⋮	⋮	⋮	⋮

Yalnızca bir adet sürekli değişkenin olmasından dolayı f düzgün (pürüzsüz) fonksiyon için μ_i ortalamanın x_i üzerinde bağımlılığı olduğunu gösteren model (3.17) eşitliğindeki gibidir.

$$\eta_i = \mathrm{g}(\mu_i) = f(x_i), \tag{3.17}$$

Genelleştirilmiş toplamsal modellerin en basit örneği bu şekildedir. x_i sürekli değişkeninin muhtemel değerleri $z_1 < \cdots < z_m$ olsun ve $x_i = z_k$ eşitliğindeki tüm i'leri I_k olarak gösterilsin. Tüm ağırlıkların ve gözlemlerin toplamı (3.18) eşitliğindeki gibidir.

$$\widetilde{\omega}_k = \sum_{i \in I_k} \omega_i, \qquad \tilde{y}_k = \frac{1}{\widetilde{\omega}_k} \sum_{i \in I_k} \omega_i y_i. \tag{3.18}$$

Normal dağılım için bağ fonksiyonu olan birim bağ fonksiyonunun özelliği göz önünde bulundurulduğunda $\eta_i = \mu_i$ eşitliği yazılabilmektedir. Normal dağılım için sapma (3.19) eşitliğinde verildiği gibidir.

$$\sum_i \widetilde{\omega}_i (\tilde{y}_i - f(z_k))^2 \tag{3.19}$$

İki kez türevlenebilir sürekli herhangi bir f fonksiyon için (3.16) eşitliğindeki cezalı sapmayı en küçük yapacak olan s doğal kübik splayn;

$$\Delta(s) = \sum_{k=1}^{m} \widetilde{\omega}_k (\tilde{y}_k - s(z_k))^2 + \lambda \int_a^b (s''(x))^2 \, dx \tag{3.20}$$

Doğal kübik splayn $[z_1, z_m]$ aralığı dışında doğrusaldır. Bundan dolayı cezalı sapma eşitliği içinde yalnızca $x \in [z_1, z_m]$ şartı altındaki $s(x)$ doğal kübik splayn fonksiyonu kullanılmaktadır.

$$\int_a^b (s''(x))^2 \, dx = \int_{z_1}^{z_m} (s''(x))^2 \, dx. \tag{3.21}$$

Teorem 1'de görüldüğü gibi $s(x)$ fonksiyonu $[z_1, z_m]$ aralığında (3.22) eşitliğindeki gibi yazılabilir.

$$s(x) = \sum_{j=1}^{m+2} \beta_j B_j(x), \tag{3.22}$$

$B_1(x), \dots, B_{m+2}(x)$'ler $z_1, \dots, z_m$ düğüm noktaları ile kübik B-splaynlar olmak üzere $\beta_1, \dots, \beta_{m+2}$ parametrelerinin fonksiyonu olacak şekilde cezalı sapma (3.23) eşitliğinde yeniden düzenlenmiştir.

$$\Delta(\boldsymbol{\beta}) = \sum_{k=1}^{m} \widetilde{\omega}_k \left(\tilde{y}_k - \sum_{j=1}^{m+2} \beta_j B_j(z_k) \right)^2 + \lambda \sum_{j=1}^{m+2} \sum_{k=1}^{m+2} \beta_j B_k \Omega_{jk}, \tag{3.23}$$

$$\Omega_{jk} = \int_{z_1}^{z_m} B''_j(x) B''_k(x)\, dx. \tag{3.24}$$

$\beta_1, \dots, \beta_{m+2}$ parametreleri (3.25) ve (3.26) eşitliklerinde $\ell = 1, \dots, m+2$ için sırayla, önce kısmi türev ile minimize edilmiş sonra kısmi türevler sıfıra eşitlenmiştir.

$$\frac{\partial \Delta}{\partial \beta_\ell} = -2 \sum_{k=1}^{m} \widetilde{\omega}_k \left(\tilde{y}_k - \sum_{j=1}^{m+2} \beta_j B_j(z_k) \right) B_\ell(z_k) + 2\lambda \sum_{j=1}^{m+2} \beta_j \Omega_{j\ell}\,. \tag{3.25}$$

$$\sum_{k=1}^{m} \sum_{j=1}^{m+2} \widetilde{\omega}_k\, \beta_j B_j(z_k) B_\ell(z_k) + \lambda \sum_{j=1}^{m+2} \beta_j \Omega_{j\ell} = \sum_{k=1}^{m} \widetilde{\omega}_k \tilde{y}_k B_\ell(z_k), \tag{3.26}$$

(3.26) eşitliğini matris formunda yazmak mümkündür. $\widetilde{\omega}_k$ esas köşegen ile $m \times m$'lik $\boldsymbol{W}$ matrisi ve $(m+2) \times (m+2)$'lik Ω_{jk} elemanlarından oluşan $\boldsymbol{\Omega}$ simetrik matris olsun. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{m+2})'$ ve $\boldsymbol{y} = (y_1, \dots, y_m)'$ sütun vektörleri olmak üzere $m \times (m+2)$'lik $\boldsymbol{B}$ matrisi;

$$B = \begin{pmatrix} B_1(z_1) & B_2(z_1) & \dots & B_{m+2}(z_1) \\ B_1(z_2) & B_2(z_2) & \dots & B_{m+2}(z_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(z_m) & B_2(z_m) & \dots & B_{m+2}(z_m) \end{pmatrix},$$

(3.26) eşitliğinin daha öz bir şekilde ifadesi (3.27) eşitliğinde verilmiştir.

$$(\boldsymbol{B'WB} + \lambda \boldsymbol{\Omega})\boldsymbol{\beta} = \boldsymbol{B'Wy}. \tag{3.27}$$

3.4.2. Poisson Dağılımı Durumunda

Hasar sıklığı modellemesi Poisson durumu için tek değişken ile tahmin bu bölümün konusu olacaktır. Hasar sıklığı modellemesi yapılırken logaritmik bağ fonksiyonu $\boldsymbol{g(\mu) = log\mu}$ olarak varsayılır. Poisson dağılımı için sapma kullanılırken $\boldsymbol{\mu_i = exp\{s(x_i)\}}$

eşitliğinden s doğal kübik splayn $\widetilde{\boldsymbol{\omega}}_{\boldsymbol{i}}$ birleşmiş ağırlıklar, $\widetilde{\boldsymbol{y}}_{\boldsymbol{i}}$ birleşmiş gözlemler olsun. Poisson durumu için cezalı sapma ise (3.28) eşitliğinde verildiği gibidir.

$$\Delta(s) = 2\sum_{k=1}^{m} \widetilde{\omega}_k(\tilde{y}_k \log \tilde{y}_k - \tilde{y}_k s(z_k) - \tilde{y}_k + exp\{s(z_k)\} + \lambda \int_{z_1}^{z_m} (s''(x))^2\, dx \tag{3.28}$$

(3.29) eşitliğinde verilen $\beta_1, \ldots, \beta_{m+2}$ parametelerinin en küçük değerleri bulunur.

$$\Delta(\boldsymbol{\beta}) = 2\sum_{k=1}^{m} \widetilde{\omega}_k \left(\tilde{y}_k \log \tilde{y}_k - \tilde{y}_k \sum_{j=1}^{m+2} \beta_j B_j(z_k) \right)^2 - \tilde{y}_k + exp \sum_{j=1}^{m+2} \beta_j B_j(z_k) + \lambda \sum_{j=1}^{m+2} \sum_{k=1}^{m+2} \beta_j B_k \Omega_{jk}. \tag{3.29}$$

β_ℓ'ye göre türevi alındığında (3.30) eşitliğinde verildiği gibidir.

$$-\sum_{k=1}^{m} \widetilde{\omega}_k \tilde{y}_k B_\ell(z_k) + \sum_{k=1}^{m} \widetilde{\omega}_k \gamma(z_k) B_\ell(z_k) + \lambda \sum_{j=1}^{m+2} \beta_j \Omega_{j\ell} = 0, \tag{3.30}$$

$\boldsymbol{\ell = 1, \ldots, m+2}$ için (3.31) eşitliğinde ortalama verilmiştir.

$$\gamma(z_k) = exp\left\{ \sum_{j=1}^{m+2} \beta_j B_j(z_k) \right\}, \tag{3.31}$$

Logaritmik bağ fonksiyonu kullanılarak hasar sıklığı ve hasar şiddeti modellemesi esnasında ortalamanın çarpımsal yapısı için karşılık gelen eşitlik doğrusal olmaktan çıkar ve iteratif olarak çözümlenir. Bu durumda $\gamma(z_k)$ doğrusal olmayan bir şekilde $\beta_1, \ldots, \beta_{m+2}$ parametrelerine bağlıdır. Böylece $\boldsymbol{\beta}$ minimizasyonun belirlenmesi için Newton-Raphson iterasyonu yöntemi kullanılır.

$$h_\ell(\beta_1, \ldots, \beta_{m+2}) = -\sum_{k=1}^{m} \widetilde{\omega}_k \tilde{y}_k B_\ell(z_k) + \sum_{k=1}^{m} \widetilde{\omega}_k \gamma(z_k) B_\ell(z_k) + \lambda \sum_{j=1}^{m+2} \beta_j \Omega_{j\ell} \tag{3.32}$$

$\boldsymbol{\ell = 1, \ldots, m+2}$ için bilinmeyen $\beta_1, \ldots, \beta_{m+2}$ parametrelerini bulmak için (3.33) eşitliği çözülür.

$$h_\ell(\beta_1, \ldots, \beta_{m+2}) = 0, \quad \ell = 1, \ldots, m+2, \tag{3.33}$$

(3.34) eşitliğindeki doğrusal eşitlik sistemleri ile iteratif çözüm yapılır.

$$h_\ell\left({\beta_1}^{(n)},\dots,{\beta_{m+2}}^{(n)}\right)+\sum_{j=1}^{m+2}\left({\beta_j}^{(n+1)}-{\beta_j}^{(n)}\right)\frac{\partial h_\ell}{\partial \beta_j}\left({\beta_1}^{(n)},\dots,{\beta_{m+2}}^{(n)}\right)=0, \tag{3.34}$$

$\ell = 1,\dots,m+2$ için,

$$\frac{\partial h_\ell}{\partial \beta_j}=\sum_{k=1}^{m}\ \widetilde{\omega}_k\gamma(z_k)B_j(z_k)B_\ell(z_k)+\lambda\Omega_{j\ell}. \tag{3.35}$$

$$\gamma_k^{(n)}=exp\sum_{j=1}^{m+2}\left\{{\beta_j}^{(n)}\beta_j(z_k)\right\}, \tag{3.36}$$

(3.36) eşitlinde ortalamadaki n. iterasyonu göstermektedir. Bazı cebirsel düzenlemeler sonucunda doğrusal eşitlik sistemi (3.37) eşitliğinde verildiği gibi yeniden düzenlenmiştir.

$$\begin{aligned}&\sum_{j=1}^{m+2}\sum_{k=1}^{m}\widetilde{\omega}_k\,\gamma_k^{(n)}B_j(z_k)B_\ell(z_k)\beta_j^{(n+1)}+\lambda\sum_{j=1}^{m+2}\beta_j^{(n+1)}\Omega_{j\ell}\\&=\sum_{k=1}^{m}\widetilde{\omega}_k\,\gamma_k^{(n)}\left(\tilde{\gamma}_k/\gamma_k^{(n)}-1+\sum_{j=1}^{m+2}\beta_j^{(n)}\,\beta_j(z_k)\right)\beta_\ell(z_k),\end{aligned} \tag{3.37}$$

$\boldsymbol{\ell = 1,\dots,m+2}$ için $m\times m$ köşegen matrisi $\boldsymbol{W}^{(n)}$ olarak gösterilmektedir.

$$\left(\boldsymbol{W}^{(n)}\right)_{kk}=\widetilde{\omega}_k\,\gamma_k^{(n)} \tag{3.38}$$

$\beta^{(n)}$ ve $y^{(n)}$ vektörleri (3.39) ile (3.40) eşitliğinde verildiği gibidir.

$$\beta^{(n)}=\left(\beta_1^{(n)},\dots,\beta_{m+2}^{(n)}\right)', \tag{3.39}$$

$$\left(y^{(n)}\right)_k=\tilde{\gamma}_k/\gamma_k^{(n)}-1+\sum_{j=1}^{m+2}\beta_j^{(n)}\,\beta_j(z_k) \tag{3.40}$$

Doğrusal eşitlik sistemi normal dağılım durumundaki matris formuna benzer olarak (3.41) eşitliğinde verildiği gibidir.

$$(\boldsymbol{B}'\boldsymbol{W}^{(n)}\boldsymbol{B}+\lambda\boldsymbol{\Omega})\boldsymbol{\beta}^{(n)}=\boldsymbol{B}'\boldsymbol{W}^{(n)}\boldsymbol{y}^{(n)}. \tag{3.41}$$

3.4.3. Gamma Dağılımı Durumunda

Hasar şiddeti modeli için Gamma dağılımı hesaplamaları Poisson durumundaki hesaplamalara çok benzerdir. Poisson durumunda logaritmik bağ fonksiyonu ile kullanılan notasyonlar sonucunda Gamma dağılımı için sapma (3.42) eşitliğinde verildiği gibidir.

$$2\sum_{k=1}^{m}\widetilde{\omega}_k\left(\frac{\tilde{\gamma}_k}{\exp\{s(z_k)\}}-1-\log\tilde{\gamma}_k-s(z_k)\right). \tag{3.42}$$

Poisson durumundaki argümanlar yeniden kullanılarak (3.30) eşitliğine karşılık Gamma için (3.43) eşitliğinde verildiği gibidir.

$$-\sum_{k=1}^{m}\widetilde{\omega}_k\,\frac{\tilde{\gamma}_k}{s(z_k)}B_\ell(z_k)+\sum_{k=1}^{m}\widetilde{\omega}_k B_\ell(z_k)+\lambda\sum_{j=1}^{m+2}\beta_j\Omega_{j\ell} \tag{3.43}$$

$\boldsymbol{\ell=1,\ldots,m+2}$ için bilinmeyen $\beta_1,\ldots,\beta_{m+2}$ yeniden doğrusal olmayan parametrelerdir ve (3.37) eşitliğine karşılık olarak Newton-Raphson yöntemi kullanıldığında (3.44) eşitliği elde edilir.

$$\begin{aligned}&\sum_{j=1}^{m+2}\sum_{k=1}^{m}\widetilde{\omega}_k\,(\tilde{\gamma}_k/\gamma_k^{(n)})B_j(z_k)B_\ell(z_k)\beta_j^{(n+1)}+\lambda\sum_{j=1}^{m+2}\beta_j^{(n+1)}\Omega_{j\ell}\\&\sum_{k=1}^{m}\widetilde{\omega}_k\,(\tilde{\gamma}_k/\gamma_k^{(n)})\left(1-\gamma_k^{(n)}/\tilde{\gamma}_k+\sum_{j=1}^{m+2}\beta_j^{(n)}B_j(z_k)\right)B_\ell(z_k),\end{aligned} \tag{3.44}$$

$\boldsymbol{\ell=1,\ldots,m+2}$ için (3.27) eşitliğindeki normal durumundaki matrisin aynısı elde edilmiştir. Ağırlık matrisi ve gözlem vektörü verildiğinde (3.45) eşitliği ile elde edilir.

$$\left(\boldsymbol{W}^{(n)}\right)_{kk}=\widetilde{\omega}_k\left(\frac{\tilde{\gamma}_k}{\gamma_k^{(n)}}\right),\quad\left(\boldsymbol{y}^{(n)}\right)_k=1-\left(\frac{\tilde{\gamma}_k}{\gamma_k^{(n)}}\right)+\sum_{j=1}^{m+2}\beta_j^{(n)}B_j(z_k) \tag{3.45}$$

3.5. Çoklu Fiyatlama Değişkeni ile Tahmin

Hayat dışı sigorta fiyatlamasında yalnızca tek bir fiyatlama değişkeni ile modelleme nadir görülen bir durumdur. Bu durumda çoklu değişkenin ve en az bir adet sürekli değişkenin olduğu bir model oluşturulmalıdır. (3.46) eşitliğinde görüldüğü üzere iki adet sürekli değişken ile kalan değişkenlerin kategorik olduğunu varsayan bir model oluşturulmuştur.

$$\eta_i=\mathrm{g}(\mu_i)=\sum_{j=0}^{r}\beta_j x'_{ij}+\sum_{k=1}^{m_1+2}\beta_{1k}B_{1k}(x_{1i})+\sum_{\ell=1}^{m_2+2}\beta_{2\ell}B_{2\ell}(x_{2i}), \tag{3.46}$$

(3.46) eşitliğinde x_{1i} modeldeki birinci sürekli değişken olup alacağı muhtemel değerler $z_{11},\ldots,z_{1m_1}$ ve x_{2i} ikinci sürekli değişken olup alacağı muhtemel değerler $z_{21},\ldots,z_{2m_2}$ değerleridir. Kategorik değişkenler için $x'_{i0}=1$'dir ve β_0 katsayıdır. Ortalamanın

çarpımsal yapısı $\eta_i = \log(g(\mu_i))$ olduğunda cezalı olabilirlik poisson dağılımı ile bir hasar sıklığı modeli (3.47) eşitliğindeki gibi oluşturulmuştur.

$$\Delta = 2\sum_i \omega_i(y_i \log y_i - y_i \log \mu_i - y_i + \mu_i) + \lambda_1 \sum_{j=1}^{m_1+2}\sum_{k=1}^{m_1+2} \beta_{1j}\beta_{1k}\Omega_{jk}^{(1)} \qquad (3.47)$$
$$+\lambda_2 \sum_{j=1}^{m_2+2}\sum_{k=1}^{m_2+2} \beta_{2j}\beta_{2k}\Omega_{jk}^{(2)}.$$

$$\Omega_{jk}^{(1)} = \int_{z_{11}}^{z_{1m_1}} B_{1j}''(x)B_{1k}''(x)\,dx, \quad \Omega_{jk}^{(2)} = \int_{z_{21}}^{z_{2m_2}} B_{2j}''(x)B_{2k}''(x)\,dx. \qquad (3.48)$$

Buradan sapmaya iki ceza terimi eklenerek iki sürekli değişkene karşılık olacaktır. (3.49) eşitliğinde olduğu gibidir.

$$\gamma_{0i} = exp\left\{\sum_{j=0}^{r} \beta_j x_{ij}'\right\},$$
$$\gamma_{1i} = exp\left\{\sum_{j=1}^{m_1+2} \beta_{1j}B_{1j}(x_{1i})\right\}, \qquad (3.49)$$
$$\gamma_{2i} = exp\left\{\sum_{j=0}^{m_2+2} \beta_{2j}B_{2j}(x_{2i})\right\}.$$

Burada amaç β parametrelerinin tahmin edilmesidir. $\beta_0, \dots, \beta_r$ ve $\beta_{21}, \dots, \beta_{2,m_2+2}$ parametrelerinin bilindiği varsayımında Poisson dağılımı durumu için sapma (3.50) eşitliğinde olduğu gibidir.

$$2\sum_i \omega_i(y_i \log y_i - y_i \log \mu_i - y_i + \mu_i)$$
$$= 2\sum_i \omega_i(y_i \log y_i - y_i \log(\gamma_{0i}\gamma_{1i}\gamma_{2i}) - y_i + \gamma_{0i}\gamma_{1i}\gamma_{2i}) \qquad (3.50)$$
$$= 2\sum_i \omega_i\gamma_{0i}\gamma_{2i}\left(\frac{y_i}{\gamma_{0i}\gamma_{2i}}\log\frac{y_i}{\gamma_{0i}\gamma_{2i}} - \frac{y_i}{\gamma_{0i}\gamma_{2i}}\log(\gamma_{1i}) - \frac{y_i}{\gamma_{0i}\gamma_{2i}} + \gamma_{1i}\right).$$

(3.50) eşitliğinden yola çıkarak;

$$\omega_i' = \omega_i\gamma_{0i}\gamma_{2i}, \qquad y_i' = \frac{y_i}{\gamma_{0i}\gamma_{2i}}, \qquad (3.51)$$

Cezalı sapma (3.52) eşitliği gibidir.

$$\Delta = 2\sum_{i} \omega_i'(y_i' \log y_i' - y_i' \log \gamma_{1i} - y_i' + \gamma_{1i}) + \lambda_1 \sum_{j=1}^{m_1+2} \sum_{k=1}^{m_1+2} \beta_{1j}\beta_{1k}\Omega_{jk}^{(1)}, \quad (3.52)$$

$\beta_{11}, \dots, \beta_{1,m_1+2}$ parametrelerinin tahmini yapılırken eşitlikte yalnızca katsayı olarak bulunacağından ikinci ceza terimine (3.52) eşitliğinde yer verilmemiştir. (3.4.2) alt başlığında da aynı düzleştirme splayn yöntemi ile bilinmeyen $\beta_{11}, \dots, \beta_{1,m_1+2}$ parametreleri tahmin edilebilmektedir. Sapmanın $\gamma_{0i}\gamma_{1i}$ ve $\gamma_{0i}\gamma_{2i}$ çarpanlarında simetrik oluşundan dolayı çarpanlarına $\gamma_{0i}, \gamma_{1i}, \gamma_{2i}$ aynı şekilde ayrılabilmektedir.

Cezalı terimler modelde yalnızca sabit durumunda olacağı için $\beta_{11}, \dots, \beta_{1,m_1+2}$ ve $\beta_{21}, \dots, \beta_{2,m_2+2}$ parametreleri verilmiş varsayılarak yalnızca kategorik değişkenler için aynı şekilde $\beta_0, \dots, \beta_r$ parametrelerinin tahmini yapılmaktadır.

Backfitting Algoritma

Poisson dağılımı ile çok değişkenli tahmin süreci bu bölümde özetlenecektir. Öncelikle başlangıç parametresi tahmin edilecektir. Örneğin, $\beta_{11} = \cdots = \beta_{1,m_1+2} = \beta_{21} = \dots = \beta_{2,m_2+2} = 0$ olarak alındığında sürekli değişkenlerin çıkarıldığı veri setinin analizi ile $\hat{\beta}_0, \dots, \hat{\beta}_r$ türetilebilir.

$$\begin{aligned} \hat{\gamma}_{0i} &= exp\left\{\sum_{j=0}^{r} \hat{\beta}_j x_{ij}'\right\}, \\ \hat{\gamma}_{1i} &= exp\left\{\sum_{j=1}^{m_1+2} \hat{\beta}_{1j} B_{1j}(x_{1i})\right\}, \\ \hat{\gamma}_{2i} &= exp\left\{\sum_{j=0}^{m_2+2} \hat{\beta}_{2j} B_{2j}(x_{2i})\right\}. \end{aligned} \quad (3.53)$$

Backfitting algoritma iterasyonu üç adım içerir:

1- Bölüm (3.4.2)'de açıklandığı gibi gözlemler ve ağırlıklar kullanılarak splayn yöntemi ile tek sürekli x_{1i} değişkeni için yeni bir set $\hat{\beta}_{11}, \dots, \hat{\beta}_{1,m_1+2}$ tahmin edici parametreler hesaplanır.

$$\omega_i' = \omega_i \gamma_{0i}\gamma_{2i}, \quad y_i' = \frac{y_i}{\gamma_{0i}\gamma_{2i}}. \quad (3.54)$$

2- Gözlemler ve ağırlıklar kullanılarak splayn yöntemi ile sürekli x_{2i} değişkeni için yeni bir set $\hat{\beta}_{21}, \dots, \hat{\beta}_{2,m_2+2}$ tahmin edici parametreler hesaplanır.

$$\omega_i' = \omega_i \gamma_{0i} \gamma_{1i}, \qquad y_i' = \frac{y_i}{\gamma_{0i}\gamma_{1i}}. \tag{3.55}$$

3- Gözlemler ve ağırlıklar kullanılarak kategorik değişkenler için $\hat{\beta}_0, \dots, \hat{\beta}_r$ tahmin edici parametrelerin hesaplanabilmesi için standart GLM yöntemi kullanılır.

$$\omega_i' = \omega_i \gamma_{1i} \gamma_{2i}, \qquad y_i' = \frac{y_i}{\gamma_{1i}\gamma_{2i}}. \tag{3.56}$$

Tahminler yakınsayana dek bu işlem devam ettirilir.

Hasar şiddeti için Gamma dağılımı üzerinden parametre tahminleri ise hasar sıklığı için Poisson dağılımından yalnızca ağırlıkları farklı olmak üzere kalan kısımları aynı şekilde bırakarak tahmin yapılmaktadır.

3.6. λ Düzleştirme Parametresi

λ düzleştirme parametresi verilmiş ise, cezalı sapma yalnızca modeldeki β katsayılarının tahminini yapmaktadır. Düzleştirme splaynları için $\lambda = 0$ ise bu durumda interpolasyon düzleştirici splayn ya da $\lambda = \infty$ ise düz bir doğrusal regresyon eğrisi elde edilir. Düzleştirme parametresine değer seçimi modeli kuran kişiye bağlı bir durumdur. Modelleme sürecinde istatistiksel bilginin yanı sıra veri setinin de iyi bir şekilde anlaşılması oldukça önemlidir. λ düzleştirme parametresinin seçimi için her zaman uygun otomatik bir yöntem vardır. Bu otomatik yöntem oldukça doğru bir uyum yakalayabildiği gibi, zayıf seçimler güvensiz tahminlere ya da aşırı uyum (overfitting) gibi bir durumlara da yol açabilmektedir.

Bu bölümde λ düzleştirme parametresi seçimi için oldukça fazla kullanılan veri temelli ve otomatik bir yöntem olan Çapraz Geçerlilik (Cross-Validation) yöntemi kullanılacaktır. Normal, Poisson ve Gamma dağılımları için yöntem sırasıyla verilecektir.

Normal dağılıma uyan bir veri setinden $\tilde{y}_k$'ya karşılık gelen z_k'nın çıkarıldığı varsayılarak; çıkarılmış olan k gözlemi için ve uygun bir λ değeri için küçültülmüş veri seti ile $s_k^{\lambda}(z_k)$ minimize eden splayn fonksiyonu bulunur. Bu şekilde kalan $m-1$ gözlem ile ilerlenebilir.

Sonrası hata tahmini kareler toplamını en küçük yapan en iyi λ değeri (3.57) eşitliğindeki gibidir.

$$C(\lambda) = \sum_{k=1}^{m} \tilde{\omega}_k(\tilde{y}_k - s_k^{\lambda}(z_k))^2. \tag{3.57}$$

Silinmiş her bir $\hat{y}_k$ veri noktası için m minimizasyon sorunu çözülmelidir. Normal dağılım durumunda $C(\lambda)$ hesaplaması, bütün veri seti için tek bir splayn çözümlemesi durumunda sadeleşebilir.

(3.27) eşitliğinde olduğu gibi λ ve her bir m elemanlı $\tilde{y}$ gözlem vektörü verilmiş olup doğrusal eşitlik sistemi ile bilinmeyen parametreler hesaplanmaktadır.

$$(\boldsymbol{B'WB} + \lambda\boldsymbol{\Omega})\boldsymbol{\beta} = \boldsymbol{B'Wy}. \tag{3.58}$$

$\hat{\mu}$, z_k'da $\hat{\mu}_k$ tahmin edici ortalamaların elemanları ile sütun vektörü ise;

$$\hat{\mu}_k = \sum_{j=1}^{m+2} \hat{\beta}_j B_j(z_k). \tag{3.59}$$

(3.59) eşitliğinin matris formu $\widehat{\boldsymbol{\mu}} = \boldsymbol{B}\widehat{\boldsymbol{\beta}}$ şeklindedir. Bir y vektöründeki gözlemler her ne olursa olsun, tahmin edici ortalamalar $\widehat{\boldsymbol{\mu}} = \boldsymbol{Ay}$ şeklinde yazılabilir. $\boldsymbol{A}$, $m \times m$ boyutlu düzleştirme matrisi olmak üzere (3.60) eşitliğinde verildiği gibidir.

$$\boldsymbol{A} = \boldsymbol{B}(\boldsymbol{B'WB} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B'W}. \tag{3.60}$$

k. gözlemin silinmiş olduğu verilmiş herhangi bir verilmiş λ değeri için en küçük kübik splayn $s_k(x) = s_k^{\lambda}(x)$ şeklindedir. y^* vektörü m uzunluğunda y_j^* elemanları ile (3.61) eşitliğindeki gibidir.

$$y_j^* = \begin{cases} \hat{y}_j & j \neq k \\ s_k^{\lambda}(z_k) & j = k \end{cases} \tag{3.61}$$

$A = \{a_{kj}\}$ ve $\tilde{y}_1, \ldots, \tilde{y}_m$ gözlem değerleri için $s(x)$ en küçük yapan kübik splayn olsun.

$$s_k^{\lambda}(z_k) = \sum_{j=1}^{m} a_{kj} y_j^* = \sum_{\substack{1 \le j \le m \\ j \ne k}} a_{kj} \tilde{y}_j + a_{kk} y_k^*$$

$$= \sum_{j=1}^{m} a_{kj} y_j^* - a_{kk} \tilde{y}_k + a_{kk} s_k^{\lambda}(z_k) \tag{3.62}$$

$$= s(z_k) - a_{kk} \tilde{y}_k + a_{kk} s_k^{\lambda}(z_k).$$

$$\tilde{y}_k - s_k^{\lambda}(z_k) = \tilde{y}_k - s(z_k) + a_{kk}(\tilde{y}_k - s_k^{\lambda}(z_k))$$

$$= \frac{\tilde{y}_k - s(z_k)}{1 - a_{kk}}. \tag{3.63}$$

(3.63) eşitliğindeki ifadeler (3.57) eşitliğinde yerine konulduğunda (3.64) eşitliği elde edilir.

$$C(\lambda) = \sum_{k=1}^{m} \widetilde{\omega}_k \left(\frac{\tilde{y}_k - s(z_k)}{1 - a_{kk}} \right)^2 \tag{3.64}$$

$s(z_k)$ bütün veri seti için en küçük splayndır ve a_{kk}, $\boldsymbol{A}$ matrisinin köşegen elemanlarını oluşturmaktadır. k. gözlemin kaldırılmasına ve düzleştirmeyi yeniden çözümlemeye gerek yoktur. Yalnızca bir kez bütün veri seti ile model fit edilmeli ve düzleştirici matrisin köşegen elemanları hesaplanmalıdır. (3.64) eşitliğindeki sadeleşme normal dağılımın doğrusal yapısından dolayı olan bir durumdur. Normal dağılım için $C(\lambda)$ çözümlemesinin sadeleşmesi Poisson ve Gamma dağılımları için uygulanmamaktadır. Uygun gözlemleri ve ağırlıkları yerine koyarak (3.41) eşitliğindeki mantık ile yaklaşık $C(\lambda)$ hesaplanabilmektedir. Poisson dağılımı için $s^n(z_k)$ splayn mininimizasyonu ile yaklaşık $C(\lambda)$ çözümlemesi (3.65) ve (3.66) eşitliğinde örnek olarak verilmiştir.

$$C^{(n)}(\lambda) = \sum_{k=1}^{m} \left(\boldsymbol{W}^{(n)}\right)_{kk} \left(\frac{\left(y^{(n)}\right)_k - s^{(n)}(z_k)}{1 - A_{kk}^{(n)}} \right)^2 \tag{3.65}$$

$$= \sum_{k=1}^{m} \widetilde{\omega}_k \gamma_k^{(n)} \left(\frac{(\tilde{y}_k / \gamma_k^{(n)} - 1 + \sum_{j=1}^{m+2} \beta_j^{(n)} B_j(z_k)) - s^{(n)}(z_k)}{1 - A_{kk}^{(n)}} \right)^2$$

Matris formu (3.66) eşitliğindeki gibi yazılabilmektedir.

$$\boldsymbol{A}^{(n)} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{W}^{(n)}\boldsymbol{B} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B}'\boldsymbol{W}^{(n)}. \tag{3.66}$$

4. UYGULAMA

Tezin bu bölümünde, sigorta portföyünde yer alan her bir risk profili için ne kadar prim alınmalı sorusu tarife analizi esas alınarak cevaplanmıştır. Önceki bölümlerde bahsedilmiş olan fiyatlama oranlarından hasar sıklığı ve hasar şiddeti modellerinin kurulumu temel odak noktası olmuştur. Veri setini modellemeye hazırlamak amacıyla hasar tutarları için uygun minimum ve maksimum değerler belirlenerek veri setinde düzeltmeler yapılmıştır. Daha sonra, hasar sıklığı ve hasar şiddeti için uygun dağılımlar belirlenmiştir. Öncelikle modellemeye dahil edilecek diğer açıklayıcı değişkenler ile GLM ve GAM modelleri kurulmuş, modellerin kurulumundan sonra ise hasar sıklığı ve hasar şiddeti için yapılan tahminler birleştirilerek örnek bir risk primi hesaplanmıştır. Sonrasında veri kümesinde tek sürekli değişken olan yaş değişkeni ile GLM ve GAM modellemeleri kurulmuştur. Bu modellerle ilgili sonuçlar alınmış görsellerle yorumlamalar yapılmıştır. Son olarak model analizlerinin süreçleri ve kıyaslamalarına yer verilmiştir.

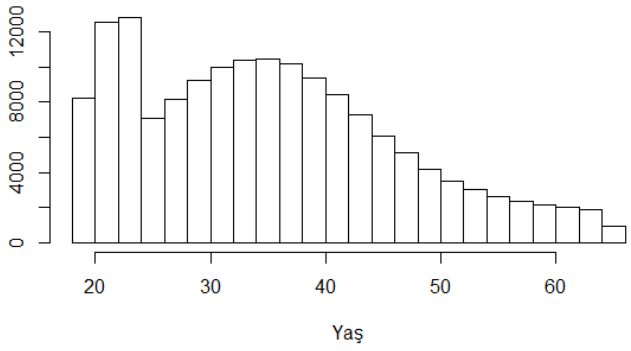
4.1. Veri Açıklaması

Bu tez çalışması için kullanılan veri seti özel bir sigorta şirketinin kasko poliçelerinden oluşturulmuş olup 2015, 2016 ve 2017 yıllarına ait 753.100 adet gözlem, 18 değişken içermektedir. Veri kümesinde, açıklayıcı değişkenler sürekli ve kategorik olmak üzere: İl adı, araç markası, araç bedeli gibi değişkenler yer almaktadır. Her bir satır için poliçe yılı, hasar sayısı ve toplam hasar tutarı değerleri yer almaktadır. Veri seti ile ilgili analizler R.3.4.3 versiyonlu Rstudio programı ile yapılmıştır.

İl adı, hasarsızlık indirimi, marka, araç bedeli, yaş değişkenleri fiyatlama modellemesinde sırasıyla yer verilmiş değişkenlerdir. Bu değişkenlerle ilgili grafikler ve istatistiksel bulgular tezin bu bölümünde açıklanmıştır.

İl adı değişkeni İstanbul, Ankara, İzmir ve diğer olarak yeniden düzenlenmiş ve gruplandırılmıştır. Portföyde sayıca en fazla olan grup diğer grubundaki iller olmuştur. *Hasarsızlık indirimi* değişkeni ise 11 seviyeden oluşmaktadır. Her bir poliçe için hasarsızlık kademesini ifade etmektedir. Örneğin; 3 seviyesindeki bir araç, 3 yıldır hasar getirmemiş olarak açıklanır. Ancak bir üst seviyeden de düşmüş olabileceği için ardı ardına olan yıllarda hasarsızlık olup olmadığı belirsizdir. *Marka* değişkeni çok sayıda farklı marka içermesinden dolayı analizi kolaylaştırması açısından belirli gruplara ayrılarak yeniden düzenlenmiştir. GLM model kurulumu öncesinde sürekli değişkenler kategorik hale

getirilmiştir. Veri setindeki tek sürekli değişken *araç bedeli* değişkenidir. *Yaş* değişkeni GAM modellemesi için oldukça önemli bir değişkendir. Bundan dolayı veri setinde yalnızca tek bir değişkenin sürekli olması nedeniyle yaş değişkeni 18-65 aralığında simülasyon ile üretilmiştir. Simülasyon, Türkiye İstatistik Kurumu'nun (TÜİK) "Trafik Kaza İstatistikleri Karayolu 2013" yılına ait veriler temel alınarak oluşturulmuştur. Bununla birlikte GAM yönteminde modelin sistematik yapısına dahil edilmiş olan sürekli değişkenin splayn fonksiyonlarının etkisini net bir biçimde elde etmek amaçlanmıştır. Yaş değişkenine ilişkin histogram grafiği Şekil 4.1.'de verilmiştir. Ortalama yaş 35,6 medyan 35 olmak üzere iki tepeli bir dağılım elde edilmiştir. Diğer değişkenlerde de olduğu gibi yaş değişkeni GLM analizinde kategorik olarak yer alırken GAM analizinde splayn fonksiyonu olarak yer alacaktır.



Şekil 4.1. Yaş Simülasyon Histogram Grafiği

Yanıt değişkeni için hangi dağılımın kullanılacağına ise sapma artıkları analizine dayanarak karar verilmelidir. Dağılımın veri setine tam olarak uymuyor olabileceği göz önünde bulundurulmalıdır. Buradaki amaç veri setinin modellenmesine en uygun dağılımın kullanılmasıdır.

Sapma artıkları, tahmin değerlerinin gözlem değerlerinden uzaklığının görsel olarak bir ölçüsü olup modele uyumunu belirlemede önemli yöntemlerden biridir. GLM'de modelin rasgele bileşeni ile tahmin edilmektedir. Böylece modelin rasgeleliği de araştırılmış olur. Basit anlamda artık, *i*. gözlem değerinden *i.* ortalama değerinin farkının alınması ile elde

edilir. GLM için ise tahmin ve gerçek değerler arasındaki sapmanın daha kullanışlı ölçümü, model uyumu incelemek için pek çok faydalı özelliği olan sapma artıklarıdır.

Sapma artıkları tahmin edilen GLM dağılımının şekli için düzenlenmiş artıklar olarak düşünülebilir. Model uyumu başarılı bir model için sapma artıkları rasgele bir yapıya sahip ve sabit varyans ile normal dağılımlı olmalıdır. Sapma artıklarından farklı olarak, basit artıkların model kurulumunda normal dağılım haricinde bir dağılımın seçildiği varsayımında normal dağılma zorunluluğu yoktur. Kesikli dağılımlar için sapma artıklarının normal dağılmalı kuralı tam olarak sağlanamamaktadır. Bunun nedeni ise artıkların kümelenerek bir araya gelmesi çok sayıdaki gözlemlerin aynı sonucu almasından kaynaklıdır. Bu durumda sapma artıklarının kullanılabilirliği daralmaktadır.

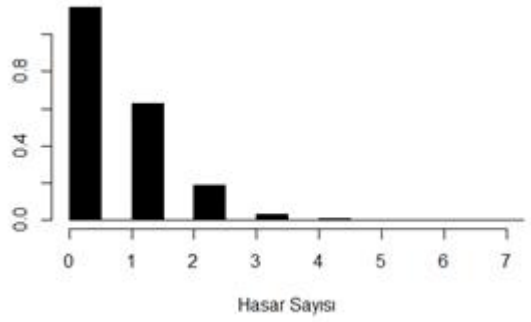
4.1.1. Hasar Sayısı Dağılımının Belirlenmesi

GLM analizinde ilk adım yanıt değişkeninin dağılımının belirlenmesidir. Hasar sayısı dağılımına karar verilebilmesi için öncelikli kriterler bulunmaktadır. Hasar sayısı değişkeninin kesikli bir değişken olması sebebiyle kesikli dağılımlardan Poisson ve Negatif binom ya da lojistik regresyon durumunda binom dağılımlarının kullanımları oldukça yaygındır. Bu noktada değişkenin ortalama ve varyans değerleri kontrol edilmektedir. Poisson dağılımının bir özelliği olan ortalama ve varyans eşitliğinin söz konusu olup olmadığı incelenir. Varyans değerinin ortalamadan büyük olduğu durumda aşırı yayılım olduğu görülmektedir. Böyle bir durum söz konusu olduğunda aşırı yayılım testi yapılır ve aşırı yayılım yeterince yüksek ise negatif binom dağılımının hasar sayısı değişkeni için daha iyi sonuçlar verip vermediği model kıyaslamaları ile açığa çıkarılmaktadır.

753.100 adet gözlemin yer aldığı veri kümesinde yapılan gerekli düzenlemeler sonucunda veri seti 157.960 gözlem sayısına düşürülmüştür. Hasar sayılarının her bir değeri için karşılık gelen poliçe sayıları ve poliçe yıllarının toplamları aşağıda verildiği gibidir. Görülüyor ki toplam poliçe yılı, toplam poliçe sayısına eşit değildir. Bunun sebebi bazı poliçelerin bir poliçe yılını doldurmamış olmasıdır.

Hasar Sayısı	Poliçe Sayısı	Toplam Poliçe Yılı
0	90.655	64.642,41
1	49.709	43.596,47
2	14.817	13.770,95
3	2.352	2.145,03
4	369	336,1151
5	47	42,8384
6	10	8,1151
7	1	1,0027
Toplam	157.960	124.543,2

Bu düzenlemelerin sonucunda hasar sayısı için ortalama ve varyans değerleri incelenerek 0,56 olarak hesaplanmıştır. Hasar sayısı değişkenine ilişkin histogram grafiği Şekil 4.2.'de verildiği gibidir.



Şekil 4.2. Hasar Sayısı Histogram Grafiği

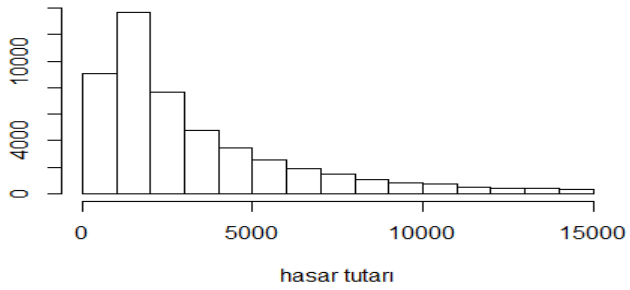
4.1.2. Hasar Tutarı Dağılımının Belirlenmesi

Hasar tutarı verileri negatif olmayan genellikle sağa çarpık dağılımlara uyum sağlayan değerler içerir. Hasar şiddeti modellendiğinde yaygın kullanım Gamma ve Ters Normal dağılımlarıdır. Hasar tutarı verileri için betimsel istatistikler aşağıdaki tabloda verildiği gibidir. Hasar tutarı gözlemleri minimum değeri 500, maksimum değeri 15.000 olarak belirlenmiştir. Tabloda görüldüğü gibi ilk çeyreklik hasar tutarları 1.178'in altındadır. Gözlemlerin yarısı 2.186,9'dan ve %90'ı 7.599,61'den daha küçüktür. Çeyreklikler arası

genişlik 3.209,753'tür. Hasar tutarı gözlemleri için 1,6377 çarpıklık katsayısı ile dağılım asimetriktir.

İstatistik	Değer
Gözlemler	48.886
Min	500
Max	15.000
Ortalama	3.289,7
Standart Sapma	2.951,946
25. yüzdelik	1.178
Medyan	2.186,9
75. yüzdelik	4.387,8
90. yüzdelik	7.599,61
95. yüzdelik	9.913,27
99. yüzdelik	13.485,49
Çarpıklık	1,6377

Hasar tutarı için histogram grafiği Şekil 4.3.'te verildiği gibidir.



Şekil 4.3. Hasar Tutarı Histogram Grafiği

4.2. Genelleştirilmiş Doğrusal Modeller ile Hasar Sıklığı Modellemesi

Genelleştirilmiş Doğrusal Model yöntemiyle hasar sayısı modellemesi, ÜDA'dan kesikli bir dağılımın seçilerek poliçe yılı ağırlıklandırılması ve anlamlı açıklayıcı değişkenlerin belirlenmesi ile ifade edilen model kurulum sürecidir. GLM'de her bir riskin sonucunu

yansıtmak yerine benzer grup risklerin sonuçlarının ortalamalarını yansıtan veri seti etkisi görülür. Bundan dolayı veri setinde daha riskli satırların model katsayılarının tahminlerde daha yüksek ağırlıklarla etkisinin yer alması gerekmektedir. GLM her bir gözlemin modeldeki risk ağırlığını bir ağırlıklandırma değişkeniyle modele dahil eder. Tahmin sürecinde her bir gözlem için ağırlık belirler.

Hasar sıklığı modellemesi aşamasında kullanımı en yaygın olan Poisson dağılımı kullanılmıştır. Model kurulumu için bağ foksiyonu logaritmik bağ fonksiyonu olarak seçilmiştir. Modelde poliçe yılı offset olarak belirlenmiştir. Bunun sebebi modelde poliçe yılının fiyatlama modelinin bir parçası olması istenmekte ancak diğer değişkenlerin katsayılarını etkilemeyecek bir terime ihtiyaç duyulmaktadır. Böyle bir durum modele offset terimi dahil edilerek sağlanır. Modelde bağ fonksiyonu logaritmik bağ fonksiyonu olarak seçilmiş ise offset terimi logaritması alınarak modele dahil edilir. Açıklayıcı değişkenlerin belirlenmesi model kurulumundaki en önemli adımdır. Bu aşamadan önce veri kümesindeki araç bedeli ve simülasyon ile üretilmiş olan yaş değişkenleri GLM analizinde kategorik hale getirilerek modelde yer almıştır. Sonrasında hasar sıklığı modeli kurulumu için model sabit terim ile başlatılmış açıklayıcı değişkenler teker teker eklenerek model kurulmuştur. Çizelge 4.1.'de, modelde yer alan açıklayıcı değişkenlerin açıklamaları yer almaktadır.

Çizelge 4.1. Modelde Yer Alan Değişkenlerin Sınıf Bilgisi

Değişkenin Açıklaması	Kademe	Değerler
Sigortalı yaşı	8	18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-65
Araç bedeli	5	974-30860, 30860-48069, 48069-56677, 56677-69189, 69189-1649053
Hasarsızlık indirimi	11	0,1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
İl adı bilgisi	4	İstanbul, Ankara, İzmir, Diğer
Araç markaları	11	A/OPEL, B/VOLKSWAGEN, C/TOYOTA, …, K/TOFAS-FIAT, L/Diğer

Modele eklenen her bir açıklayıcı değişkenin, başarılı model ile arasındaki fark sapma değeri ile ölçülmektedir. Çizelge 4.2.'de olduğu gibi eklenen her bir açıklayıcı değişken anlamlı sonuç çıkararak başarılı modelde yer almaktadır.

Çizelge 4.2. Hasar Sıklığı GLM Anova Sonuçları

Değişkenin Açıklaması	Sapma	Df	Pr> Ki-Kare
Sigortalı yaşı	2174,2	7	<0,001
Araç bedeli	307,5	4	<0,001
Hasarsızlık indirimi	1444,2	12	<0,001
İl adı bilgisi	15627,9	3	<0,001
Araç markaları	418,2	10	<0,001

Tahmin edilen bütün değerlerin tablosunu vermek yerine sonuçlara ilişkin bir örnek verilecektir. Örneğe ilişkin poliçe sahibinin karakteristik özellikleri Çizelge 4.3.'te verildiği gibidir.

Çizelge 4.3. En Çok Olabilirlik Parametre Tahmin Analizi

		Tahmin	Std. Hata	p
sabit		-0,7961	0,0140	0,001
Sigortalı yaşı	35-40	0,2876	0,01283	0,001
Araç bedeli	974-30860	0,0852	0,01088	0,001
Hasarsızlık indirimi	2	-0,1259	0,01214	0,001
İl adı bilgisi	İstanbul	0,8762	0,00762	0,001
Araç markaları	Nissan	-0,0513	0,01735	0,0031
Ölçek Parametresi		1	0	

Örnek hasar sıklığı tahmini şu şekildedir:

$$exp\{-0{,}7961 + 0{,}2876 + 0{,}0852 - 0{,}1259 + 0{,}8762 \ - 0{,}0513\} = 1{,}3175$$

4.3. Genelleştirilmiş Doğrusal Modeller ile Hasar Şiddeti Modellemesi

Hasar şiddeti modellemesinde izlenilen yöntem hasar sıklığında olduğu gibidir. Burada GLM'nin rasgele bileşeni için Gamma dağılımı kullanılmıştır. Hasar şiddeti modellemesinde hasar sıklığı modelindekinden farklı olarak bir offset teriminin eklenmemiş olmasıdır. Ancak modelde hasar sayısının bir ağırlık terimi kullanılmıştır. Hasar şiddeti modelinde yer alan açıklayıcı değişkenler Çizelge 4.4.'te verildiği gibidir.

Çizelge 4.4. Modelde Yer Alan Değişkenlerin Sınıf Bilgisi

Değişkenin Açıklaması	Kademe	Değerler
Sigortalı yaşı	8	18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-65
Araç bedeli	5	974-30860, 30860-48069, 48069-56677, 56677-69189, 69189-1649053
Hasarsızlık indirimi	11	0,1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
İl adı bilgisi	4	İstanbul, Ankara, İzmir, Diğer
Araç markaları	11	A/OPEL, B/VOLKSWAGEN, C/TOYOTA, …, K/TOFAS-FIAT, L/Diğer

Çizelge 4.5.'teki p değerleri göz önünde bulundurularak tüm açıklayıcı değişkenlerin anlamlı olduğu söylenebilmektedir.

Çizelge 4.5. Hasar Şiddeti GLM Anova Sonuçları

Değişkenin Açıklaması	Sapma	Df	Pr> Ki-Kare
Sigortalı yaşı	63,1	7	<0,000
Araç bedeli	677,79	4	<0,001
Hasarsızlık indirimi	92,76	11	<0,000
İl adı bilgisi	382,16	3	<0,001
Araç markaları	144,38	10	<0,001

Hasar sıklığı modelinde olduğu gibi tahmin değerlerinin tamamına tabloda yer vermek yerine hasar şiddeti modeli üzerinden örnek verilmiştir.

Çizelge 4.6. En Çok Olabilirlik Parametre Tahmin Analizi

		Tahmin	Std. Hata	p
sabit		7,8050	0,0178	0,001
Sigortalı yaşı	35-40	0,0928	0,0166	0,001
Araç bedeli	974-30860	-0,2226	0,0139	0,001
Hasarsızlık indirimi	2	0,0146	0,0156	0,001
İl adı bilgisi	İstanbul	0,0910	0,0098	0,001
Araç markaları	Nissan	0,0605	0,0224	0,0070
Ölçek Parametresi		1,2686	0	

Çizelge 4.6.'da verilmiş olan özelliklere bağlı olarak örnek hasar şiddeti

$$exp\{7{,}8050 + 0{,}0928 - 0{,}2226 + 0{,}0146 + 0{,}0910 + 0{,}0605\} = 2543{,}509$$

gibidir.

4.4. Genelleştirilmiş Toplamsal Modeller ile Hasar Sıklığı Modellemesi

GAM modelleme süreci GLM modelleme süreci ile hemen hemen aynı şekilde işlemektedir. Daha öncede belirtildiği gibi iki model kurulumu arasındaki fark GAM'ın düzleştirme parametresi ile bir splayn fonksiyonunun modele dahil edilerek modeli yarı parametrik hale getirmesidir. Hasar sıklığı GAM analizi için modelde yer alan açıklayıcı değişkenler Çizelge 4.7.'de verildiği gibidir.

Çizelge 4.7. Modelde Yer Alan Değişkenlerin Sınıf Bilgisi

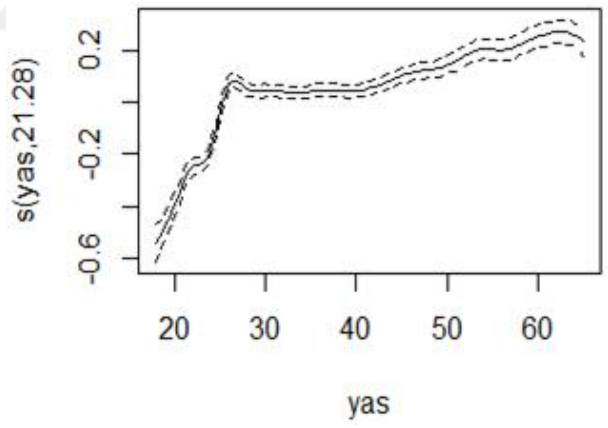
Değişkenin Açıklaması	Kademe	Değerler
Sigortalı yaşı	sürekli değişken	18-65
Araç bedeli	5	974-30860, 30860-48069, 48069-56677, 56677-69189,
Hasarsızlık indirimi	11	0,1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
İl adı bilgisi	4	İstanbul, Ankara, İzmir, Diğer
Araç markaları	11	A/OPEL, B/VOLKSWAGEN, C/TOYOTA, …, K/TOFAS-FIAT, L/Diğer

Hasar sayısı dağılımı için veri setini Poisson dağılımının uyum göstermesi ve çarpımsal model kurulumunun sağlanabilmesi durumundan dolayı model için logaritmik bağ fonksiyonu seçilmiştir. Poliçe yılı, modelin offseti olarak yer almaktadır Düzleştirme parametresi ise Çapraz Geçerlilik ile belirlenmektedir. Model kurulumu ile ilgili model sonuçları Çizelge 4.8.'de verildiği gibidir. Çizelgedeki sonuçlara göre modele Splayn fonksiyonu olarak katılan yaş değişkeninin anlamlı olduğu söylenebilmektedir. Aynı şekilde diğer kategorik değişkenlerin de modelde anlamlı olarak yer alabildiği söylenebilmektedir.

Çizelge 4.8. Model Anova Sonuçları

	sd	F	p-değeri
Araç bedeli	4	25,68	<0,000
Hasarsızlık indirimi	11	138,23	<0,000
İl adı bilgisi	3	5.189,63	<0,000
Araç markaları	10	42,85	<0,000
s(yas)	21,28	65,97	<0,000

Model kurulumunda GCV (*Generalized Cross-Validation*) çapraz geçerlilik metodu kullanılarak model sonuçları alınmıştır. Böylece düzleştirme parametreleri çapraz geçerlilik ile tahmin edilmiştir. Düğüm sayısı 25 olarak belirlenen yaş değişkeninin splayn fonksiyonu için λ düzleştirme parametresi 0,0056 serbestlik derecesi 21,28 olarak hesaplanmıştır. Yaş değişkeni için splayn fonksiyonunun grafiği Şekil 4.6.'da verildiği gibidir.



Şekil 4.6. Hasar Sıklığı Modellemesinde Yaş Değişkeninin Splayn Grafiği

Şekil 4.6.' ya göre yaş değişkeni için 20-30 değerleri arasında bir sıçrama söz konusudur. Riskin bu yaş aralıklarında birden yükseldiğini ve sonrasında azalan bir artışla devam ettiği söylenebilmektedir. Bunun sebebi yaş simülasyonu yapılırken bu yaş grubunun diğer yaş gruplarına göre az sayıda veri içermesi ve yüksek riskli olarak düzenlenmesidir.

4.5. Genelleştirilmiş Toplamsal Modeller ile Hasar Şiddeti Modellemesi

Hasar şiddeti dağılımı için Gamma dağılımının varsayılan dağılım olması ve çarpımsal model kurulumunun sağlanabilmesi durumundan dolayı model için 'log' bağ fonksiyonu seçilmiştir. Hasar şiddeti GAM analizi için modelde yer alan açıklayıcı değişkenler Çizelge 4.9.'da verildiği gibidir.

Çizelge 4.9. Modelde Yer Alan Değişkenlerin Sınıf Bilgisi

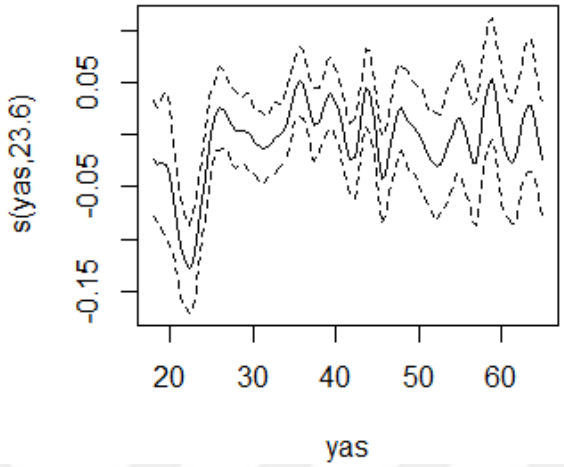
Değişkenin Açıklaması	Kademe	Değerler
Sigortalı yaşı	sürekli değişken	18-65
Araç bedeli	5	974-30860, 30860-48069, 48069-56677, 56677-69189,
Hasarsızlık indirimi	11	0,1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
İl adı bilgisi	4	İstanbul, Ankara, İzmir, Diğer
Araç markaları	11	A/OPEL, B/VOLKSWAGEN, C/TOYOTA, …, K/TOFAS-FIAT, L/Diğer

Hasar şiddeti modelinde hasar sayısı ağırlık olarak modele eklenmiştir. Düzleştirme parametresi ise Çapraz Geçerlilik ile otomatik olarak belirlenmektedir. Model kurulumu ile ilgili model sonuçları Çizelge 4.10.'da verildiği gibidir. Çizelgedeki sonuçlara göre modele Splayn fonksiyonu olarak katılan yaş değişkeninin anlamlı olduğu söylenebilmektedir. Aynı şekilde diğer kategorik değişkenlerin de modelde anlamlı olarak yer alabildiği söylenebilmektedir.

Çizelge 4.10. Model Anova Sonuçları

	sd	F	p-değeri
Araç bedeli	4	85,939	<0,000
Hasarsızlık indirimi	11	4,668	<0,000
İl adı bilgisi	3	111,27	<0,000
Araç markaları	10	12,562	<0,000
s(yas)	23,6	2,828	<0,000

Düğüm sayısı 30 olarak belirlenen yaş değişkeninin splayn fonksiyonu için λ düzleştirme parametresi 0,0109 serbestlik derecesi 23,6 olarak hesaplanmıştır. Yaş değişkeni için splayn fonksiyonunun grafiği Şekil 4.8.'de verildiği gibidir.



Şekil 4.8. Hasar Şiddeti Modellemesinde Yaş Değişkeninin Splayn Grafiği

Şekil 4.8.'e göre yaş değişkeni için tüm değerlerde aşırı salınım söz konusudur. Bu durumun bir nedeni düğüm sayısının yüksek belirlenmesidir. Düğüm sayısı model ile hesaplanabildiği gibi kullanıcı tarafından da seçilebilmektedir. Burada önemli olan düğüm sayısı ve serbestlik derecesi değerlerinin çok yüksek değerler olmayacak şekilde belirlenmesidir. Hasar şiddeti modeli olduğu değerlerdeki bu salınım beklenen bir durumdur. Değişkendeki doğrusal olmayan etkinin model ile yansıması bu şekildedir. Hasar sıklığında olduğu gibi bu değerlerde ani sıçramalar olduğu söylenebilmektedir. Bu durumun bir diğer sebebi ise serbestlik derecesinin yüksek λ düzleştirme parametresinin ise küçük bir değer olmasıdır. Düğüm sayısıyla bu sorun dengelenmeye çalışılsa da daha düşük düğüm sayıları için serbestlik derecesinin düşmesine rağmen düzleştirme parametresinde düşüşe neden olmaktadır.

4.6. Yaş Değişkeni ile Tek Değişkenli GLM ve GAM Modellemesi

Veri kümesinde tek sürekli değişken olarak yer alan yaş değişkenin GLM ve GAM analizlerinde süreklilik etkisinin daha anlaşılır olması amacıyla tek değişkenli analizleri ve hesaplanmış risk primleri karşılaştırılması amaçlanmıştır.

Yaş Değişkeni için Tek Değişkenli GLM modeli

Yaş değişkeni tek değişken olarak GLM için hem hasar sıklığı hem de hasar şiddeti modellemelerinde anlamlı bir değişkendir.

Çizelge 4.11. Yaş Değişkeni GLM Sıklık ve Şiddet Modeli Anova Sonuçları

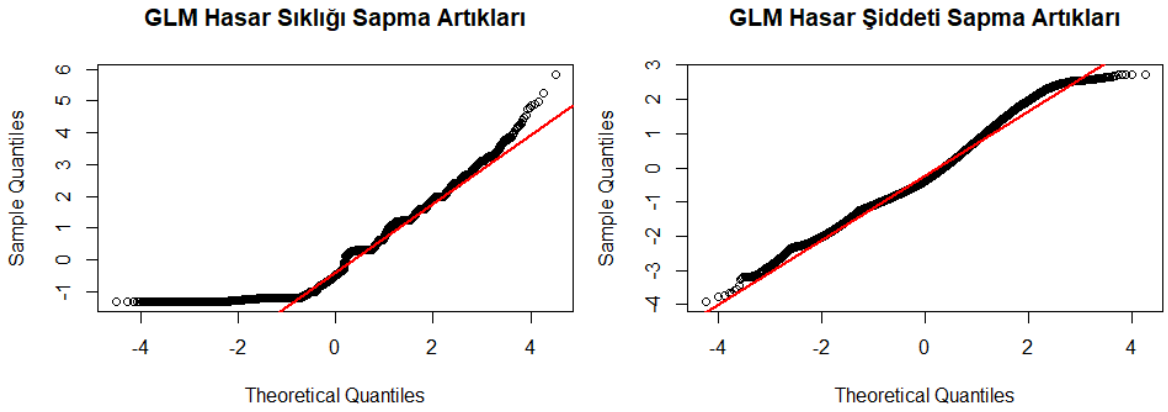
Hasar Sıklığı Anova		
LR Ki-kare	Sd	Pr(>Kikare)
2174.1	7	0
Hasar Şiddeti Anova		
LR Ki-kare	Sd	Pr(>Kikare)
2174.1	7	0

Yaş değişkeni için hasar sıklığı ve hasar şiddeti modelleri için ayrı ayrı model uyumu testleri Çizelge 4.12.'de olduğu gibi p> 0,05 değerleri ile model anlamlı çıkmıştır.

Çizelge 4.12. Yaş Değişkeni GLM Sıklık ve Şiddet Model Uyum İyiliği

Hasar Sıklığı Model İyiliği Testi		
Sapma Artıkları	Sd	p
148633	157952	1
Hasar Şiddeti Model İyiliği Testi		
Sapma Artıkları	Sd	p
48518.53	48878	0,87

Hasar sıklığı için seçilen Poisson ve hasar şiddeti için seçilen Gamma dağılımlarının model sonuçları ve veri setinin şekline uygun olup olmadığının bir ölçüsü olan sapma artıklarının normal dağılıma yakın bir uyum göstermesi gerekmektedir.



Şekil 4.9. Hasar Sıklığı ve Hasar Şiddeti için Sapma Artıklarının QQ-plot Grafikleri

Şekil 4.8.'de görüldüğü üzere yaş değişkeni ile oluşturulan tek değişkenli hasar sıklığı ve hasar şiddeti modelleri yeterince uyumlu olduğu söylenebilmektedir.

Yaş Değişkeni için Tek Değişkenli GAM modeli

Yaş değişkeni tek değişken olarak GAM için hem hasar sıklığı hem de hasar şiddeti modellemelerinde anlamlı bir değişkendir.

Çizelge 4.13. Yaş Değişkeni GAM Sıklık ve Şiddet Modeli Anova Sonuçları

GAM Hasar Sıklığı			
	sd	F	p-değeri
s(yas)	22,36	99.49	<0,000
GAM Hasar Şiddeti			
	sd	F	p-değeri
s(yas)	23,38	3,479	0

Yaş değişkeni için hasar sıklığı ve hasar şiddeti modelleri için ayrı ayrı model uyumu testleri Çizelge 4.13.'te olduğu gibi p> 0,05 değerleri ile model anlamlı çıkmıştır.

Çizelge 4.14. Yaş Değişkeni GAM Sıklık ve Şiddet Model Uyum İyiliği

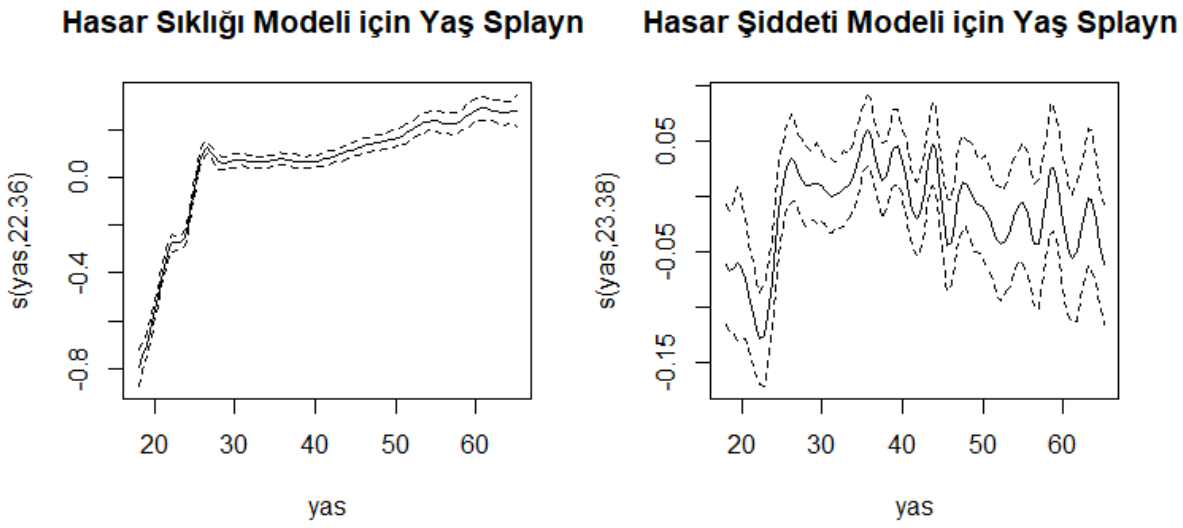
GAM Hasar Sıklığı Model İyiliği Testi		
Sapma Artıkları	Sd	p
148050	157936	1
GAM Hasar Şiddeti Model İyiliği Testi		
Sapma Artıkları	Sd	p
48464	48861	0.8986

Yaş değişkeni için hasar sıklığı ve hasar şiddeti modelleri için ayrı ayrı model uyumu testleri Çizelge 4.14.'de olduğu gibi p> 0,05 değerleri ile model anlamlı çıkmıştır. GAM sürekli değişkene modelde splayn fonksiyonu olarak yer vermesinden dolayı hasar sıklığı ve hasar şiddeti modelleri için seçilen düğüm sayılarının uygunluğu Çizelge 4.15.'te verildiği gibidir. Hasar sıklığı ve hasar şiddeti modelleri için sırasıyla λ düzleştirme parametresi 0,01083 ve 0,010 değerlerini almıştır. Düğüm sayıları ise hasar sıklığı için 25 hasar şiddeti için 29 olarak belirlenmiştir.

Çizelge 4.15. Yaş Değişkeni GAM Sıklık ve Şiddet Modelleri için Optimal Düğüm Sayısı

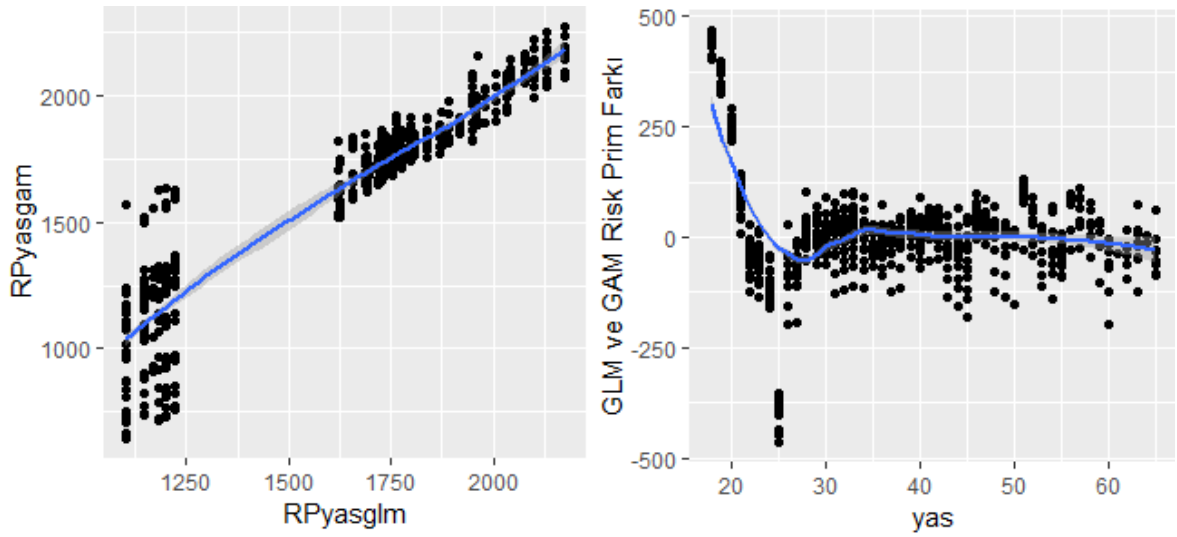
Sıklık				
	k	k-indeks	p-değeri	λ
s(yas)	25	0,95	0,82	0,01083
Şiddet				
	k	k-indeks	p-değeri	λ
s(yas)	29	0,97	0,56	0,0100

Şekil 4.9.'da elde edilen splayn grafikleri ile hasar sıklığı ve hasar şiddeti için çoklu değişkenlerin oluşturduğu splayn grafikleri oldukça benzer görseller vermiştir. Bunun sebebi ise aynı yanıt değişkenleri ile aynı modellerin kurulması ve modellerdeki yaş değişkeninin sürekli etkisinin aynı olmasından kaynaklıdır.



Şekil 4.10. Hasar Sıklığı ve Hasar Şiddeti için Yaş Splaynları

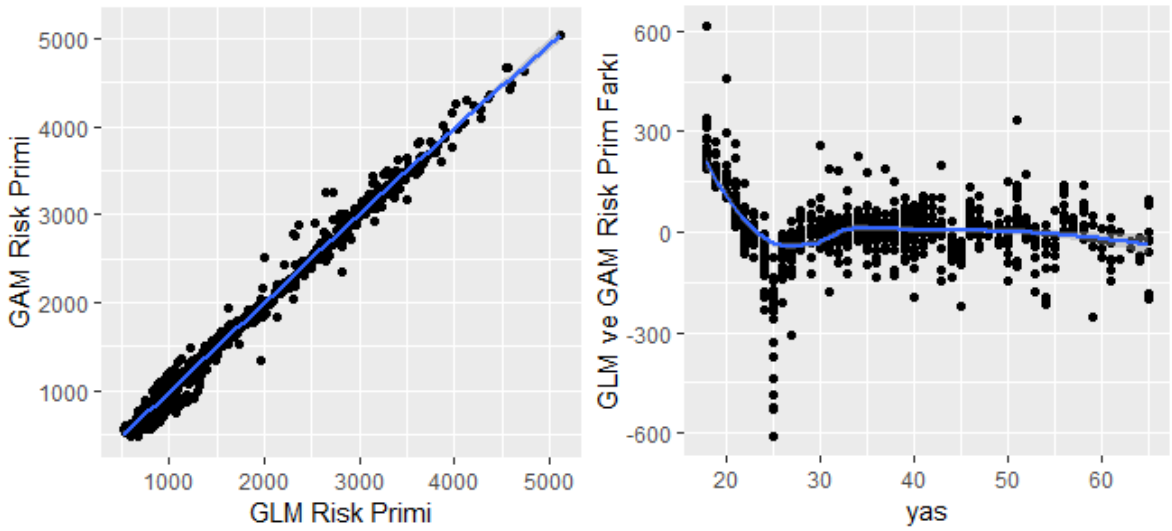
Son olarak tek değişkenli GLM ve GAM hasar sıklığı ve hasar şiddeti modellerinin tahmin değerlerinin çarpımı ile risk primleri hesaplanmıştır. Şekil 4.10.'da solda verilen grafik GAM ve GLM risk prim tahmin değerlerini üzerinde dağılımını göstermektedir. Sağda verilen grafik ise GLM ve GAM risk primlerinin fark grafiğidir. Sağdaki grafiği GLM ile GAM risk primleri arasında yaş değişkenine göre 20-30 yaş aralığı arasında yaklaşık 250 birimlik bir farkın oluştuğu diğer yaş grupları arasında ise daha yakın sonuçlar alındığı yorumlanabilmektedir. GAM modelde yaş değişkeninin sürekli etkisini kullandığı için her yaş değeri için risk primi sonucunu verebilmektedir.



Şekil 4.11. Risk Primlerinin GLM ve GAM Hesaplama Karşılaştırmaları

4.7. Genelleştirilmiş Doğrusal Modeller ile Genelleştirilmiş Toplamsal Modellerin Karşılaştırılması

GAM ve GLM yöntemleri kullanılarak hasar sıklığı ve hasar şiddeti modellerinin tahmin değerleri elde edilmiştir. Bu modellerin sonucundan tahmin edilen risk primleri GLM ve GAM yöntemlerini karşılaştırmak amacıyla Şekil 4.11.'daki görsel elde edilmiştir. Solda verilen grafik GAM ve GLM risk prim tahmin değerlerini üzerinde dağılımını göstermektedir. Sağda verilen grafik ise GLM ve GAM risk primlerinin fark grafiğidir. Çoklu değişken GLM ve GAM modellerinin hesaplamış olduğu risk primlerine göre iki model risk prim sonuçlarını çok yakın vermiştir. Bunun nedeni modeldeki sürekli değişken sayısının daha fazla olabildiği modellerde sürekliliğin etkisi daha açık görülebilmektedir. Sağdaki grafik ile GAM ve GLM risk prim farkları daha net görülmektedir. Tek değişkenli modelde olduğu gibi 20-30 yaş aralığında iki model arasında fark oluşmuş diğer yaş gruplarında bu fark kapanmıştır.



Şekil 4.12. Tahmin Edilen Risk Primlerinin GLM ve GAM Karşılaştırması

İki modelin yakın sonuçlar vermeleri beklenen bir durumdur. Aynı dağılımlarla modele giren hasar sayısı ve hasar tutarı değişkenleri için yaş değişkeninin GAM'da splayn fonksiyonu olarak yer alması dışında iki model kurulumları benzerdir. Böyle bir çalışmada sonuçların daha belirli çıkabilmesi için daha önce de bahseldildiği gibi modele daha çok sürekli değişkenin katılmış olması gerekmektedir.

5. SONUÇ

Bu tez çalışmasında GLM ve GLM'nin geliştirilmiş yarı parametrik bir hali olan GAM yöntemleri incelenmiştir. GLM ve GAM yöntemleri için belirli tanımlamalar yapılmış, her iki yöntemin de güçlü ve zayıf yanlarından bahsedilmiştir. GAM, düzleştirme splaynı eklentisi sayesinde GLM'nin daha esnek bir hale gelmiş şeklidir. İki yöntem arasında keskin çizgiler olmamasına rağmen sigorta sektöründe GLM yönteminin yaygın olması sebebiyle GAM henüz kullanım alanı bulamamıştır. Buna rağmen GAM yöntemin kullanım rahatlığı ve modele uyumu açısından artıları oldukça fazladır.

Tezin uygulama bölümünde özel bir sigorta şirketinden alınan 2015-2016-2017 yıllarının kasko verileri ile GLM ve GAM yöntemleri için hem hasar sıklığı hem hasar şiddeti olmak üzere 4 ayrı model kurulmuştur. Model kurulumu öncesinde yanıt değişkenleri için dağılımlar araştırılmış ve veri setinde düzenlemeler yapılmıştır Orijinal veri setindeki tek sürekli değişken olan araç bedeli değişkeni düzleştirme splaynı olarak modelde yeterince anlamlı sonuç vermemesinden dolayı her 4 model için de kategorik olarak modelde yer almıştır. Bu sebepten dolayı veri setine bir yaş değişkeni simülasyon ile eklenmiş ve GLM model kurulumu için kategorik hale getirilmiştir. GAM modelleri için modelde splayn fonksiyonu olarak yer alan tek değişkendir. GLM ve GAM model kurulumunda hasar sıklığı ve hasar şiddeti için yaygın kullanım biçimi olan logaritmik bağ fonksiyonu seçilmiştir. GLM tahmin modelinden alınan sonuçlar ile hasar sıklığı ve hasar şiddeti modelleri için örnekler sunulmuştur. GAM yönteminde GLM'den farklı olarak splayn düzleştirmesi için uygun düğüm sayısı aranmıştır. Uygun düğüm sayısının belirlenmesi splayn fonksiyonu için aşırı dalgalanma durumu ya da aşırı düzgünlükte bir eğri olması problemine karşı oldukça önemlidir. Model anlamlılıkları ile ilgili tüm sonuçlar yorumlanmış artık grafikleri ile görseller sunulmuştur. Sürekli değişkenin etkisini araştımak amacıyla yalnızca yaş değişkeninin bulunduğu GLM ve GAM modelleri kurulmuştur. Model kurulumları sonrasında GLM ve GAM yöntemleri için kurulan hasar şiddeti ile hasar sayısı modellerinin tahminleri çarpımı ile risk primleri hesaplanmıştır. Risk primlerinin karşılaştırılması sonucunda tek değişkenli risk primleri tahminleri arasında fark görülebilirken çok değişkenli modellerin risk primi tahminleri yakın sonuçlar verdiği görülmüştür.

GAM model kurulumu açısından GLM'e göre oldukça kullanışlı ve esnek bir yöntemdir. Sürekli değişkenler ile alınan sonuçların esnekliği ve modele etkisi açısından avantajlı bir yöntemdir. Ancak düğüm sayısının belirlenmesi ve düzleştirme parametresinin alacağı değerlere göre aşırı uyum sorunu ile karşılaşılabilmektedir.

Sigorta sektöründe henüz kullanım alanı bulamamış olan GAM yöntemi sektörde kolaylık sağlayacak iyi çalışan bir yöntemdir. GLM'in genişletilmiş ve doğrusal olmayan veri yapıları ile kolayca başa çıkabilmesi gibi artılar gelecek çalışmalara yansıtılabilir.

KAYNAKLAR

[1] Nelder, J. A., Wedderburn, R. W. M., "Generalized Linear Models", *Journal of the Royal Statistical Society Series A*, 135, 3, 370–384, **1972**.

[2] McCullagh, P., Nelder, J. A., *Generalized Linear Models*, Chapman Hall, **1983**.

[3] McCullagh, P., Nelder, J. A., *Generalized Linear Models,* 2. Basım, London Chapman Hall, **1989**.

[4] Brockman, M. J., Wright, T. S., "Statistical Motor Rating: Making Effective Use of Your Data", *Journal of the Institute of Actuaries.*, 03, 457–543, **1992**.

[5] Haberman, S., Renshaw, A. E., "Generalized Linear Models and Actuarial Science", *Journal of the Royal Statistical Society*, 4, 407–436, **1996**.

[6] Murphy, K.P., Brockman M.J., Lee, P.K.W., "Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance", *Casualty Actuarial. Forum,* 107–140, **2000**.

[7] Smyth , G. K., Jørgensen, B., "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling", *ASTIN Bulletin: the Journal of the International Actuarial Association.*, 1, 143–157, **2002**.

[8] Renshaw, A. E., "Modelling the Claims Process in the Presence of Covariates", *Astin Buletin: the Journal of the International Actuarial Association.*, 2, 265–285, **1994**.

[9] Heller, G. Z., de Jong, P., *Generalized Linear Models for Insurance Data*. Cambridge University Press, **2008**.

[10] Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M., *Modern Actuarial Risk Theory Using R*, Springer, **2008**.

[11] Frees, E.W., *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press, **2010**.

[12] David, M., "Auto Insurance Premium Calculation Using Generalized Linear Models", *Procedia Economics and Finance.*, 15, 147–156, **2013**.

[13] David, M., "Automobile Insurance Pricing with Generalized Linear Models", *3rd*

Global Virtual Conference., 4, 32–39, **2015**.

[14] Kafková, S., Krivánková, L., "Generalized linear models in vehicle insurance", *Acta Universitatis Agriculturae Silviculturae Mendelianae Brunensis*, 2, 383–388, **2014**.

[15] Valecký, J., "Modelling Claim Frequency in Vehicle Insurance", *Acta Universitatis Agriculturae Silviculturae Mendelianae Brunensis*, 2, 683–689, **2016**.

[16] Rosenlund, S., "Inference in multiplicative pricing", *Scandinavian Actuarial Journal*, 8, 690–713, **2014**.

[17] Ohlsson, D. E., Johansson, D. B., *Non-Life Insurance Pricing with Generalized Linear Models*, Springer, **2010**.

[18] Royston, P., Altman, D. G., "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious", *Journal of the Royal Statistical Society,* 3, 429–467, **1994**.

[19] Ruppert, D., Wand, M. P., Carroll, R. J., *Semiparametric Regression*, Cambridge University Press, **2003**.

[20] Klein, N., Denuit, M., Lang, S., Kneib, T., "Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape", *Insurance Mathematics and Economics*, 1, 225–249, **2014**.

[21] Hastie, T., Tibshirani, R., "Generalized additive models", *Statistical Science*, 3, 297–310, **1986**.

[22] Hastie, T., Tibshirani, R., "Generalized Additive Models: Some Applications", *Journal of the American Statistical Association*, 398, 371–386, **1987**.

[23] Hastie, T., Tibshirani, R., *Generalized Additive Models*, Chapman Hall, , **1990**.

[24] Denuit, M., Lang, S., "Non-life rate-making with Bayesian GAMs", *Insurance Mathematics and Economics*, 3, 627–647, **2004**.

[25] Wood, S. N., *Generalized Additive Models: An Introduction with R*, Chapman Hall, **2006**.

[26] Green, P., Silverman, B., *Nonparametric Regression and Generalized Linear Models*. **1994**.

[27] Silverman, B., “Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting”, *Journal of the Royal Statistical Society*, 1, 1-52, **1985**.

[28] Hastie, T., Tibshirani, R., “Generalized additive models for medical research.”, *Statistical Methods Medical Research*, 3, 187–196, **1995**.

[29] Wahba, G., *Spline Models for Observational Data*, **1990**.

[30] Eilers, P. H. C., Marx, B. D., “Flexible Smoothing with B-splines and Penalties”, *Statistical Science*, 2, 89–102, **1996**.

[31] Lee, T. C. M., “Smoothing Parameter Selection for Smoothing Splines: A Simulation Study”, *Computational Statistics and Data Analysis*, 1–2, 139–148, **2002**.

[32] Burman, P., “Estimation of Generalized Additive Models”, *Journal of Multivariate Analysis*, 230–255, **1990**.

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı: Handan İLHAN

Doğum Yeri: Konya/ Akşehir

Medeni Hali: Bekar

E-posta: handan.ilhan11@hacettepe.edu.tr

Adresi:

Eğitim

Lise: 2007-2011 Eskişehir Ahmet Kanatlı Anadolu Lisesi

Lisans: 2011-2015 Hacettepe Üniversitesi Aktüerya Bilimleri

Yüksek Lisans: 2016-2018 Hacettepe Üniversitesi Aktüerya Bilimleri

Yabancı Dil ve Düzeyi

İş Deneyimi

Deneyim Alanları

Tezden Üretilmiş Projeler ve Bütçesi

Tezden Üretilmiş Yayınlar

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
Aktüerya Bilimleri ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 27/09/2018

Tez Başlığı / Konusu: **Düzleştirme Splaynlarının Hayat Dışı Sigorta Ürünleri Fiyatlamada Etkileri**
Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 73 sayfalık kısmına ilişkin, 27/09/2018 tarihinde ~~şahsım~~/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 4 'tür.

Uygulanan filtrelemeler:
1- Kaynakça hariç
2- Alıntılar ~~hariç~~/dâhil
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

27/09/2018
Tarih ve İmza

Adı Soyadı:	Handan İlhan
Öğrenci No:	N14328563
Anabilim Dalı:	Aktüerya Bilimleri
Programı:	Yüksek Lisans
Statüsü:	☑ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Dr. Öğr. Üyesi Uğur Karabey
(Unvan, Ad Soyad, İmza)