Sigara Kullanımı ve Akciğer Kanserine Yakalanma Durumunun Kategorik Veri Çözümlemesi ile İncelenmesi

Büşra Şencan, Mert Yanık, Şükriye Nur Şencan

Öz

Birçok uygulamalı bilim dalında araştırmacılar, değişkenler arasındaki ilişki veya değişken düzeyleri arasındaki farklılığı incelemeyi amaçlar. Düzeyleri sözel olarak ifade edilebilen değişkenler kategorik veri olmakla beraber, düzeyleri sayısal olarak ifade edilebilen değişkenlerin alabileceği değerler sınıflandırılarak kategorik veri haline getirilebilir. Kategorik değişken, her bir gözlemin belirli bir kategoriye, yani sınıfa ait olduğu sınıflanabilir ve sıralanabilir özelliğe sahip değişkenlerdir. Kategorik değişkenlerin birleşik dağılımı olumsallık tabloları ile özetlenir. Olumsallık tabloları iki ya da daha fazla kategorik değişkenin, kategorilerine göre nasıl dağıldığını frekanslarla gösteren tablolardır. Olumsallık tablolarında her değişken belirli sayıda düzeye sahiptir. Bu çalışmada Çin'de farklı şehirlerde yaşayan sigara kullanıcıları ve akciğer kanserine yakalanma durumları çok boyutlu tablo olarak verilmiş ve incelenmiştir. Üç yönlü olumsallık tablosu için model oluşturulmuş, log doğrusal modeller, logit model, lojistik regresyon modeli ve meta analizi yapılmıştır. Analizler sonucunda sigara içenler ile içmeyenler arasında kansere yakalanma yönünden ciddi farklılıklar bulunmuştur.

Anahtar Kelimeler: Kategorik veri analizi, Çok boyutlu tablolar, Olumsallık tabloları, Log-doğrusal modeller, Logit model, Lojistik regresyon, Meta analizi

1.Giriş

Sigara içmek, akciğer kanseri için ana risk faktörü olarak iyi bir şekilde belirlenmiştir. Çin'de, bir dizi epidemiyolojik çalışma, akciğer kanseri ve sigara içimi arasındaki ilişkiyi araştırmıştır ve mevcut makalede, bu konu incelenmiştir. Yedi vaka kontrol çalışması Pekin, Şanghay, Shenyang, Nanjing, Harbin, Zhengzhou ve Taiyuan'da gerçekleştirilmiş ve 8169 kişi içerisinden toplam 3956 akciğer kanseri vakası ortaya çıkmıştır. Belli analizlerle sonuçlar yorumlanmış ve tartışılmıştır.

Kategorik verilerde, standart normal dağılım varsayımı altında ki istatistiksel analiz yöntemlerinin kullanılması uygun değildir. Kategorik verilerin analizi için çeşitli istatistiksel metotların geliştirilmesi gerekmektedir. Kategorik verilerin analizi uygulamalarında kontenjans tabloları temel alınmaktadır. Bu tabloların düzenlenme amacı genel olarak kategorik değişkenlerin kategorilerine göre dağılımlarını göstermek ve bu değişkenler arasındaki ilişkileri ortaya çıkarmaktır. İki yönlü olumsallık tablolarında istatistiksel çıkarsamalar için Pearson'ın ki-kare istatistiği yeterli olmaktadır ancak daha büyük boyutlu olumsallık tablolarında bu yöntem kullanılamamaktadır. Üç veya daha çok boyutlu olumsallık tablolarında ilişki yapılarının belirlenmesi için değişkenler arasında bağımlı – bağımsız ayrımı yapmayan Logaritmik Doğrusal Modeller kullanılabilir. Logaritmik Doğrusal Modeller yardımıyla daha çok değişken arasındaki etkileşimler sorgulanabilmektedir.

Denekler; sosyoekonomik durum karakteristiklerine bağlı olarak şehirler arasında farklılıklar gösterebilmektedirler. Buna bağlı olarak sigara içme oranları ve akciğer kanseri bakımından şehirler arasında heterojenliğe neden olabilmektedir. Bu durumdan dolayı örneğin R değişkeni kontrol altına alınarak, C ve K değişkenleri arasındaki birliktelik araştırılır. Bu çalışmada da bu ve buna benzer 8 farklı bağımsızlık modeli logaritmik doğrusal modellere ek olarak incelenmiştir.

Olumsallık tablosunda yer alan kategorik değişkenlerden bir tanesi yanıt değişkeni ya da bağımlı değişken olabilir, bu durumda log doğrusal model yerine bağımlı değişken alınarak logit model kurulabilir. Logit terimi, odds değerinin doğal logaritma değeridir. Ek olarak logit modelin bir çeşidi olan lojistik regresyon modeli, varsayımlara ihtiyaç duymadan ve nispeten esnek bir yolla regresyon modeli kurmaya imkan tanıdığı için analizlerde buna da yer verilmiştir.

Aynı konuda yapılmış farklı çalışmaların tutarlı ve uyumlu bir şekilde birleştirilmesi ile elde edilen ortak çıkarsama işlemine meta analiz denir. Birçok konuda meta analizi yapılırken bu çalışmada iller bazında odds oranı ile ilgili meta analizi verilmiştir.

2. Veri Tanıtımı

İkiden çok değişkene göre düzenlenen olumsallık tabloları, değişken sayısına göre 3 boyutlu, 4 boyutlu ya da çok boyutlu olarak isimlendirilir. Tabloda sigara kullanımı ve akciğer kanseri konusunda Çin'de yapılan sekiz ayrı incelemenin özetleri yer almaktadır.

- Satır değişkeni (R) sigara kullanım (1 = Evet, 2 = Hayır)
- Sütun değişkeni (C) akciğer kanseri (1 = Evet, 0 = Hayır)
- Tabaka değişkeni (K) şehir (1 = Pekin, 2 = Shanghai, 3 = Shenyang, 4 = Nanjing , 5 = Harbin, 6 = Zhengzhou, 7 = Taiyuan)

Tablo 1. Çalışmada Kullanılan Veri Tablosu

		Akciğer Kanseri	
Şehirler	**Sigara Kullanımı**	**Evet**	**Hayır**
Pekin	Evet	126	100
	Hayır	35	61
Shanghai	Evet	908	688
	Hayır	497	807
Shenyang	Evet	913	747
	Hayır	336	598
Nanjing	Evet	235	172
	Hayır	58	121
Harbin	Evet	402	308
	Hayır	121	215
Zhengzhou	Evet	182	156
	Hayır	72	98
Taiyuan	Evet	60	99
	Hayır	11	43

3. Yöntemler

3.1 Üç Yönlü Olumsallık Tablosu İçin Model Oluşturma

Çok boyutlu tablolarda birden çok hipotez test edilir. Bu hipotezler;

- Tam Bağımsızlık
- Kısmi Bağımsızlık
- Koşullu Bağımsızlık
- Karşılıklı Bağımsızlık

Üç değişken tam bağımsız ise, her değişkenin marjinal olasılığı çarpımı ortak olasılığa eşit olur. Eğer iki değişkenin ortak olasılığı ile üçüncü değişkenin marjinal olasılığı çarpımı üç değişkenin ortak olasılığına eşit ise bu durumda kısmi bağımsızlıktan söz edilir. Eğer bir değişkenin tüm düzeylerinde diğer iki değişken birbirinden bağımsız ise koşullu bağımsızlıktan söz edilebilir.

3.1.1 M_0 Tam Bağımsızlık Modeli

Satır, sütun ve tabaka değişkenlerinin birbirinden bağımsız olup olmadığını inceleyen modeldir, dolayısıyla bizim verimiz için şehirler, sigara kullanımı ve akciğer kanseri değişkenlerinin bağımsız olup olmadığı incelenmiştir. Analiz sonucunda serbestlik derecesi 19 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için tam bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde istatistiksel olarak söylenebilir. Dolayısıyla şehir, sigara kullanımı ve akciğer kanseri değişkenlerinin bağımsız olmadığı görülmüştür.

3.1.2 M_1 Kısmi Bağımsızlık Modeli

Satır değişkeninin, sütun ve tabaka değişkenlerinden bağımsız olup olmadığını inceleyen modeldir, dolayısıyla bizim verimiz için sigara değişkeninin, akciğer kanseri ve şehir değişkenlerinden bağımsız olup olmadığı incelenmiştir. Analiz sonucunda serbestlik derecesi 13 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için M_1 kısmi bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde söylenebilir. Dolayısıyla sigara değişkeninin, akciğer kanseri ve şehir değişkenlerinden bağımsız olmadığı görülmüştür.

3.1.3 M_2 Kısmi Bağımsızlık Modeli

Sütun değişkeninin, satır ve tabaka değişkenlerinden bağımsız olup olmadığını inceleyen modeldir. Dolayısıyla kanser değişkeninin, şehir ve sigara kullanımından bağımsız olup olmadığı incelenmiştir. Analiz sonucunda serbestlik derecesi 13 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için M_2 kısmi bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde söylenebilir. Dolayısıyla kanser değişkeninin, şehir ve sigara kullanımından bağımsız olmadığı görülmüştür.

3.1.4 M_3 Kısmi Bağımsızlık Modeli

Tabaka değişkeninin, satır ve sütun değişkeninden bağımsız olup olmadığını inceleyen modeldir. Dolayısıyla şehir değişkeninin, sigara kullanımı ve akciğer kanserinden bağımsız olup olmadığı incelenmiştir. Analiz sonucunda serbestlik derecesi 18 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için M_3 kısmi bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde söylenebilir. Dolayısıyla şehir değişkeninin, sigara kullanımı ve akciğer kanserinden bağımsız olmadığı görülmüştür.

3.1.5 M_4 Koşullu Bağımsızlık Modeli

Tabakanın her bir düzeyinde, satır ve sütun değişkeninin bağımsız olup olmadığını inceleyen modeldir. Dolayısıyla şehir değişkeninin her bir düzeyinde sigara kullanımı ve akciğer kanserinin bağımsızlığı incelenmiştir. Analiz sonucunda serbestlik derecesi 7 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için M_4 koşullu bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde söylenebilir. Dolayısıyla şehir değişkeninin her bir düzeyinde sigara kullanımı ve akciğer kanserinin bağımsız olmadığı görülmüştür.

3.1.6 M_5 Koşullu Bağımsızlık Modeli

Sütun değişkeninin her bir düzeyinde, satır ve tabaka değişkenlerinin bağımsız olup olmadığını inceleyen modeldir. Dolayısıyla kanser değişkeninin her bir düzeyinde sigara ve şehir değişkeninin bağımsızlığı incelenmiştir. Analiz sonucunda serbestlik derecesi 12 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için M_5 koşullu bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde söylenebilir. Dolayısıyla kanser değişkeninin her bir düzeyinde sigara ve şehir değişkeninin bağımsız olmadığı görülmüştür.

3.1.7 M_6 Koşullu Bağımsızlık Modeli

Satır değişkeninin her bir düzeyinde, sütun ve tabaka değişkenlerinin bağımsız olup olmadığını inceleyen modeldir. Dolayısıyla şehir ve sigara değişkenlerinin birbiri ile ilişkili ve sigara ve kanser değişkenlerinin birbiri ile ilişkili olup olduğu incelenmiştir. Analiz sonucunda serbestlik derecesi 12 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.00 olduğu görülmektedir. P-value değeri α=0.05 değerinden küçük olduğu için M_6 koşullu bağımsızlık modeline uyum olmadığı 0.05 anlamlılık düzeyinde söylenebilir.

3.1.8 M_7 Karşılıklı Bağımsızlık Model

Tüm ikili etkileşimlerin bağımsız olup olmadığı incelenmiştir. Analiz sonucunda serbestlik derecesi 7 olarak bulunmuştur. Olabilirlik oranına bakıldığında p-value 0.528 olduğu görülmektedir. P-value değeri α=0.05 değerinden büyük olduğu için karşılıklı bağımsızlık modeline uyum olduğu 0.05 anlamlılık düzeyinde söylenebilir. Dolayısıyla tüm ikili etkileşimlerin bağımsız olduğu görülmüştür.

3.1.9 M_8 Doygun Model

Doygun model olası tüm terimleri içerir. Tamamen uyum sağlayan bir modeldir fakat elinizdeki verinin özelliklerine ve kalitesine bağlıdır. Buna benzer iteratif model yaklaşımı log-doğrusal modelde incelenecektir.

Aşağıdaki tabloda üç yönlü olumsallık tablosu için oluşturulan modellerin özet bilgileri verilmiştir.

Tablo 2. Sigara, Akciğer Kanseri ve Şehir değişkenleri için Çok Boyutlu Tablo Sonuçları

Model	**Model Adı**	**Serbestlik Derecesi**	**G^2**	**P-değeri**	**Sonuç**
M_0	Tam Bağımsızlık	19	427.1909	0.00	Uyum yok
M_1	Kısmi Bağımsızlık	13	404.8047	0.00	Uyum yok
M_2	Kısmi Bağımsızlık	13	305.4937	0.00	Uyum yok
M_3	Kısmi Bağımsızlık	18	156.8229	0.00	Uyum yok
M_4	Koşullu Bağımsızlık	7	283.1075	0.00	Uyum yok
M_5	Koşullu Bağımsızlık	12	134.4367	0.00	Uyum yok
M_6	Koşullu Bağımsızlık	12	35.12564	0.00	Uyum yok
M_7	**Karşılıklı Bağımsızlık**	**6**	**5.124**	**0.528**	**Uyum var**
M_8	Doygun	0	0	-	

Verinin modellere uyumu incelendikten sonra uyum sağlayan modeller içinden en iyi modeli seçmek için bilgi kriterlerinden yararlanılır. Burada uyum sağlayan karşılıklı bağımsızlık modeline ait AIC bilgi kriteri değeri (G^2-2*sd) formülünden "-6.876" olarak bulunmuştur.

3.2. Çok Boyutlu Tablolarda Log-Doğrusal Modeller

Olumsallık tablolarında iki ya da daha çok boyutlu kategorik değişken arasındaki ilişkiyi araştıran modellerdir. Çok boyutlu olumsallık tablolarında tercih edilir ve aynı anda çok sayıda hipotezin test edilmesine olanak sağlar. Bu modelde "ana etki" olarak bilinen tek bir değişkenin etkisi ve "etkileşim" olarak bilinen bileşik değişkenlerin etkisi logaritmik doğrusal modellerin terimleriyle ifade edilir. Böylece değişkenler arasındaki ilişkinin modellenmesinin yanı sıra, bu etkilerin anlamlılıklarının sınanması da sağlanmış olur.

3.2.1 K-Yönlü Etkiler

Tablo 3. K-Yönlü Etkiler Tablosu

K		**sd**	**Olabilirlik Oran**		**Pearson**	
			Ki-Kare	**P-değeri**	**Ki-Kare**	**P-değeri**
K-Yönlü	1	27	6929,899	,000	7586,726	,000
ve daha	2	19	427,191	,000	418,661	,000
Yüksek Etkiler	3	6	5,124	,528	5,129	,527
K-Yönlü	1	8	6502,708	,000	7168,065	,000
Etkiler	2	13	422,067	,000	413,531	,000
	3	6	5,124	,528	5,129	,527

Tablo 3'te ana etkilerin, ikili etkileşimlerin ve üçlü etkileşimin istatistiksel olarak anlamlı olup olmadığı incelenmiştir.

1. Pearson sig. değerine bakıldığında $\alpha = 0.05$ değerinden küçük olduğu için sigara kullanımı, akciğer kanseri ve şehir ana etkileşimlerinin istatistiksel olarak önemli olduğu 0.05 anlamlılık düzeyinde söylenebilir.
2. Pearson sig. değerine bakıldığında $\alpha = 0.05$ değerinden küçük olduğu için sigara kullanımı ve şehrin, sigara kullanımı ve akciğer kanserinin, şehir ve akciğer kanserinin ikili etkileşimlerinin istatistiksel olarak önemli olduğu 0.05 anlamlılık düzeyinde söylenebilir.
3. Pearson sig. değerine bakıldığında $\alpha = 0.05$ değerinden büyük olduğu için sigara, şehir ve akciğer kanseri üçlü etkileşiminin istatistiksel olarak önemli olmadığı 0.05 anlamlılık düzeyinde söylenebilir.

Tablo 4. Ana Etkiler ve Etkileşimlerin Anlamlılık Tablosu

Etki	**sd**	**Ki-Kare**	**P-değeri**
Şehir*Sigara	6	129,313	,000
Şehir*Kanser	6	30,002	,000
Sigara*Kanser	1	277,983	,000
Şehir	6	5988,388	,000
Sigara	1	506,233	,000
Kanser	1	8,087	,004

Tablo 4'e bakıldığında p değerinin hepsi $\alpha = 0.05$ değerinden küçük olduğu için ana etkilerin ve ikili etkileşimlerin;

- Şehir ve sigara değişkenlerinin
- Şehir ve kanser değişkenlerinin
- Sigara ve kanser değişkenlerinin

Birlikte etkileşimlerinin önemli olduğu 0.05 anlamlılık düzeyinde söylenebilir.

3.2.2 Geriye Doğru Seçim İşlemi

Log-doğrusal modellerde doygun modelden başlayarak geriye doğru seçim (backward selection) işlemi ile en iyi model bulunur. Bulunan en iyi model üzerinden beklenen sıklıklar ve parametre tahminleri hesaplanarak tablo yorumlanır.

Tablo 5. Geriye Doğru Seçim İşlemi Tablo Sonuçları

Adım		Etki	Ki-Kare	sd	p-değeri
0	Üretilmiş Sınıf	Şehir*Sigara*Kanser	,000	0	.
	Silinen Etki 1	Şehir*Sigara*Kanser	5,124	6	,528
1	Üretilmiş Sınıf	Şehir*Sigara, Şehir*Kanser, Sigara*Kanser	5,124	6	,528
	Silinen Etki 1	Şehir*Sigara	129,313	6	,000
	2	Şehir*Kanser	30,002	6	,000
	3	Sigara*Kanser	277,983	1	,000
2	Üretilmiş Sınıf	Şehir*Sigara, Şehir*Kanser, Sigara*Kanser	5,124	6	,528

Tablo 5'te hem etkileşimlerin anlamlılıkları hem de uyumları test edilmiştir.
Sıfırıncı adımda doygun modelin p-değeri 0.528, α = 0.05 değerinden büyük olduğu için üçlü etkileşim önemsizdir ve modelden çıkarılır. Ana etkilerin ve ikili etkileşimlerin yer aldığı modellerin p-değerlerine bakıldığında α = 0.05 değerinden küçük oldukları görülmektedir.
Böylece en iyi model;

$$log_{Eijk} = u + u_i^{\text{Sigara}} + u_j^{\text{Kanser}} + u_k^{\text{Şehir}} + u_{ij}^{\text{Sigara*Kanser}} + u_{jk}^{\text{Kanser*Şehir}} + u_{ik}^{\text{Sigara*Şehir}}$$

Şeklinde bulunmuştur. Bu model uyum iyiliği testi ile test edilmiştir.

Tablo 6. Uyum İyiliği Testi Sonuçları

	Ki-Kare	sd	P-value
Olabilirlik Oran	5,124	6	,528
Pearson Ki-Kare	5,129	6	,527

Tablo 6 'daki p-value değerine bakıldığında 0.528 değeri α = 0.05 değerinden büyük olduğu için modele uyum olduğu istatistiksel olarak 0.05 anlamlılık düzeyinde söylenebilir.

3.2.3 Model Parametre Tahminleri

Tablo 7.Parametre Tahminleri Tablosu

Parametre	Tahmin	Standart Hata	Z	Sig.	%95 Güven Aralığı	
					Alt Sınır	Üst Sınır
Sabit	3,747	,140	26,782	,000	3,473	4,021
[Sigara = 1]	,854	,160	5,351	,000	,541	1,167
[Kanser = 1]	-1,296	,152	-8,539	,000	-1,593	-,998
[Şehir = 1]	,362	,178	2,033	,042	,013	,711
[Şehir = 2]	2,950	,143	20,610	,000	2,669	3,230
[Şehir = 3]	2,647	,144	18,355	,000	2,364	2,930
[Şehir = 4]	,983	,161	6,087	,000	,666	1,299
[Şehir = 5]	1,608	,152	10,592	,000	1,310	1,905
[Şehir = 6]	,923	,163	5,664	,000	,603	1,242
[Sigara = 1] * [Kanser = 1]	,779	,047	16,464	,000	,686	,872
[Kanser = 1] * [Şehir = 1]	,746	,186	4,020	,000	,382	1,109
[Kanser = 1] * [Şehir = 2]	,802	,152	5,269	,000	,504	1,100
[Kanser = 1] * [Şehir = 3]	,718	,152	4,712	,000	,419	1,017
[Kanser = 1] * [Şehir = 4]	,752	,169	4,440	,000	,420	1,083
[Kanser = 1] * [Şehir = 5]	,764	,160	4,779	,000	,451	1,077
[Kanser = 1] * [Şehir = 6]	,775	,173	4,491	,000	,437	1,113
[Sigara = 1] * [Şehir = 1]	-,357	,202	-1,767	,077	-,752	,039
[Sigara = 1] * [Şehir = 2]	-1,021	,164	-6,228	,000	-1,343	-,700
[Sigara = 1] * [Şehir = 3]	-,632	,165	-3,838	,000	-,955	-,309
[Sigara = 1] * [Şehir = 4]	-,392	,184	-2,137	,033	-,752	-,033
[Sigara = 1] * [Şehir = 5]	-,468	,173	-2,705	,007	-,807	-,129
[Sigara = 1] * [Şehir = 6]	-,531	,186	-2,856	,004	-,895	-,167

1. Sigara değişkeninde, sigara içmeyenler referans alınmıştır.

Sig. değeri 0.00 $\alpha = 0.05$ değerinden küçük olduğu için sigara içenlerin modele katkısının istatistiksel olarak anlamlı olduğu %95 güvenle söylenebilir.

2. Kanser değişkeninde kanser olmayanlar referans alınmıştır.

Sig. değeri 0.00, $\alpha = 0.05$ değerinden küçük olduğu için kanser olanların modele katkısının istatistiksel olarak anlamlı olduğu %95 güvenle söylenebilir.

3. Şehir değişkeninde Taiyuan şehri(7) referans alınmıştır.

Bütün şehirlere ait Sig. değerleri $\alpha = 0.05$ değerinden küçük olduğu için tüm şehirlerin modele katkısının istatistiksel olarak anlamlı olduğu %95 güvenle söylenebilir.

4. Sigara içmeyenler ve kanser olmayanlar referans alınmıştır.

Sig. değeri 0.00 $\alpha = 0.05$ değerinden küçük olduğu sigara içen ve kanser olanların modele katkısının istatistiksel olarak anlamlı olduğu %95 güvenle söylenebilir.

5. Kanser olmayanlar ve Taiyuan şehrinde(7) yaşayanlar referans alınmıştır.

Sig. değerleri hepsinde 0.00, α = 0.05 değerinden küçük olduğu için kanser olan ve diğer 6 şehirde yaşayanların modele katkısının istatistiksel olarak anlamlı olduğu %95 güvenle söylenebilir.

6. Sigara içmeyenler ve Taiyuan şehrinde(7) yaşayanlar referans alınmıştır.

Sig. değeri 0.077, α = 0.05 değerinden büyük olduğu için, sigara içen ve Pekin şehrinde yaşayanların modele katkısının istatistiksel olarak anlamlı olmadığı %95 güvenle söylenebilir.

3.3. Logit Model

Logit modeller, log-linear modellerin özel bir halidir. Logit modeller ile analize başlamadan önce log-linear modeller analizi yardımıyla bu sorudaki çapraz tablo için en iyi model aşağıdaki gibi bulunmuştur.

$$log_{Eijk} = u + u_i^{\text{Sigara}} + u_j^{K\text{anser}} + u_k^{\text{Şehir}} + u_{ij}^{\text{Sigara*}K\text{anser}} + u_{jk}^{K\text{anser*Şehir}} + u_{ik}^{\text{Sigara*Şehir}}$$

Tablo 8. Logit Model İçin Uyum İyiliği Tablosu

	Ki-Kare	**sd**	**P-value**
Olabilirlik Oran	5,124	6	,528
Pearson Ki-Kare	5,129	6	,527

G^2 değeri 5.124 olarak bulunmuştur. P-değerine bakıldığında 0.528 değeri, $\alpha = 0.05$ değerinden büyük olduğu için modele uyum olduğu %95 güven düzeyinde istatistiksel olarak söylenebilir.

Tablo 9. Gözlemlenen ve Beklenen Sıklık Değerleri Tablosu

			Gözlemlenen		**Beklenen**	
Sigara	**Şehir**	**Kanser**	**Değer**	**%**	**Değer**	**%**
Evet	Pekin	Evet	126	55,8%	125,878	55,7%
		Hayır	100	44,2%	100,122	44,3%
	Shanghai	Evet	908	56,9%	910,886	57,1%
		Hayır	688	43,1%	685,114	42,9%
	Shenyang	Evet	913	55,0%	913,256	55,0%
		Hayır	747	45,0%	746,744	45,0%
	Nanjing	Evet	235	57,7%	227,272	55,8%
		Hayır	172	42,3%	179,728	44,2%
	Harbin	Evet	402	56,6%	398,648	56,1%
		Hayır	308	43,4%	311,352	43,9%
	Zhengzhou	Evet	182	53,8%	190,662	56,4%
		Hayır	156	46,2%	147,338	43,6%

	Taiyuan	Evet	60	37,7%	59,398	37,4%
		Hayır	99	62,3%	99,602	62,6%
Hayır	Pekin	Evet	35	36,5%	35,122	36,6%
		Hayır	61	63,5%	60,878	63,4%
	Shanghai	Evet	497	38,1%	494,114	37,9%
		Hayır	807	61,9%	809,886	62,1%
	Shenyang	Evet	336	36,0%	335,744	35,9%
		Hayır	598	64,0%	598,256	64,1%
	Nanjing	Evet	58	32,4%	65,728	36,7%
		Hayır	121	67,6%	113,272	63,3%
	Harbin	Evet	121	36,0%	124,352	37,0%
		Hayır	215	64,0%	211,648	63,0%
	Zhengzhou	Evet	72	42,4%	63,338	37,3%
		Hayır	98	57,6%	106,662	62,7%
	Taiyuan	Evet	11	20,4%	11,602	21,5%
		Hayır	43	79,6%	42,398	78,5%

- $\frac{125,878}{100,122}$= 1,2572 yani Pekin'de yaşayan ve sigara içenlerin kanser olma olasılığı, kanser olmama olasılığına göre yaklaşık 1.25 kat daha fazladır.
- $\frac{227,272}{179,728}$= 1,264 yani Nanjing'de yaşayan ve sigara içenlerin kanser olma olasılığı, kanser olmama olasılığına göre yaklaşık 1.26 kat daha fazladır.
- $\frac{35,122}{60,878}$= 0.57 ters çevirirsek 1.736 yani Pekin'de yaşayan ve sigara içenlerin kanser olmama olasılığı, kanser olma olasılığına göre yaklaşık 1.736 kat daha fazladır.

Tablo 10. Parametre Tahminleri Tablosu

Parametre	Tahmin	Std. Hata	Z	P-value	95% Güven Aralığı	
					Alt Sınır	Üst Sınır
[Kanser = 1]	-1,296	,152	-8,539	,000	-1,593	-,998
[Kanser = 1] * [Sigara = 1]	,779	,047	16,464	,000	,686	,872
[Kanser = 1] * [Şehir = 1]	,746	,186	4,020	,000	,382	1,109
[Kanser = 1] * [Şehir = 2]	,802	,152	5,269	,000	,504	1,100
[Kanser = 1] * [Şehir = 3]	,718	,152	4,712	,000	,419	1,017
[Kanser = 1] * [Şehir = 4]	,752	,169	4,440	,000	,420	1,083
[Kanser = 1] * [Şehir = 5]	,764	,160	4,779	,000	,451	1,077
[Kanser = 1] * [Şehir = 6]	,775	,173	4,491	,000	,437	1,113

Kanser olma durumu şehirlere göre farklılık göstermemektedir. Çünkü tablo 10'a bakıldığında olasılıklar birbirine yaklaşık değerler olarak gelmiştir.

3.4. Lojistik Regresyon

Lojistik regresyon, sınıflandırma problemi için bağımlı değişken ve bağımsız değişkenler arasındaki ilişkiyi tanımlayan doğrusal bir model kurar. Bağımlı değişken kategoriktir ve adını bağımlı değişkene uygulanan logit dönüşümünden alır. Doğrusal regresyonda aranan varsayımlar burada aranmadığı için daha esnek kullanılabilirliği vardır. Lojistik regresyon, ikili(binary) 1 veya 0 olarak kodlanmış verileri içerir.

Bu çalışmada veri kümesinin analizi sonucu kişinin kanser olup olmadığını bulacağı bir lojistik regresyon analizi için, sonuç 1 ise kanser, 0 ise kanser değildir, diyebiliriz.

Bağımlı değişken olarak incelenen kanser değişkenine ait kayıp gözlem bulunmamıştır.

Tablo 11. Katsayı Anlamlılığı için Omnibus Testi

	Ki- Kare	sd	P-değeri
Adım	300,370	7	,000
Blok	300,370	7	,000
Model	300,370	7	,000

Tablo 11'deki p-değerlerine bakıldığında $\alpha = 0.05$ değerinden küçük olduğu için model katsayılarının istatistiksel olarak anlamlı olduğu 0.05 anlamlılık düzeyinde söylenebilir.

Tablo 12. Hosmer- Lemeshow Testi

Ki-kare	sd	P-value
1,455	6	,962

Tablo 12'deki p-value değerine bakıldığında $\alpha = 0.05$ değerinden büyük olduğu için modele uyum olduğu 0.05 anlamlılık düzeyinde söylenebilir.

Tablo 13. Karışıklık Matrisi (Confusion Matrix)

		Tahmin		
		Kanser		
Gözlenen		**Hayır**	**Evet**	**Doğru Sınıflandırma Oranı**
Kanser	Hayır	2042	2171	48,5
	Evet	1190	2766	69,9
Toplam Yüzde				58,9

Modelin doğru sınıflandırma oranı %58.9'dur. Ne kadar yüksekse analiz o kadar doğrudur.

Tablo 14. Denklemdeki Değişkenler Tablosu

	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Sigara(1)	,779	,047	271,060	1	,000	2,179
Şehir			28,446	6	,000	
Şehir(1)	,746	,186	16,161	1	,000	2,108
Şehir(2)	,802	,152	27,762	1	,000	2,229
Şehir(3)	,718	,152	22,200	1	,000	2,051
Şehir(4)	,752	,169	19,715	1	,000	2,120
Şehir(5)	,764	,160	22,834	1	,000	2,147
Şehir(6)	,775	,173	20,163	1	,000	2,170

Constant	-1,296	,152	72,905	1	,000	,274

- Sigara değişkeni için hayır (2) grubu referans alınmıştır. Sigara içenlerde kanser görülme riski içmeyenlere göre 2.179 kat daha fazladır.
- Şehir değişkeni için Taiyuan (7) şehri referans alınmıştır. Pekin(1) şehrinde yaşayanlarda kanser görülme riski Taiyuan(7) şehrinde yaşayanlara göre 2.10 kat daha fazladır.

3.5 Meta Analizi

Aynı konuda yazılmış farklı çalışmaların tutarlı ve uyumlu bir şekilde birleştirilmesi ile elde edilen ortak çıkarsama işlemine meta analiz denir. Burada Odds oranı ile meta analizi yapılmıştır.

Meta analiz yapılan çalışmaların iller düzeyinde birleştirilmesi ile gerçekleştirilmiştir ve ortak bir odds oranı hesaplanmıştır.

Tablo 15. Odds Oranı için Homojenlik Testi Tablosu

	Ki-kare	**sd**	**P-value (2-sided)**
Breslow-Day	5,129	6	,527
Tarone's	5,129	6	,527

Öncelikle iller düzeyinde odds oranlarının eşit olup olmadığı incelenmiştir. P değeri $\alpha = 0.05$ değerinden büyük olduğu için iller üzerinden odds oranlarının benzer olduğu %95 güven düzeyinde söylenebilir.

Tablo 16. Ortak Odds Oranı Anlamlılık Testi Tablosu

	Ki-kare	**sd**	**P-value (2-sided)**
Cochran's	275,344	1	,000
Mantel-Haenszel	274,359	1	,000

P değeri $\alpha = 0.05$ değerinden küçük olduğu için ortak odds oranının anlamlı olduğu 0.05 anlamlılık düzeyinde söylenebilir.

Tablo 17. Mantel-Haenszel Ortak Odds Oranı Tahmini Tablosu

Kestirim	ln(Kestirim)	Std. Hata ln(Kestirim)	P-değeri (2-sided)	Ortak Odds Oranı		ln(Ortak Odds Oranı)	
				Alt Sınır	Üst Sınır	Alt Sınır	Üst Sınır
2,179	0,779	0,047	0,000	1,986	2,390	0,686	0,871

Ortak Odds oranı 2.179 olarak elde edilmiştir. Ortak odds'un %95 güven aralığına bakıldığında da [1.986 ; 2.390] <1> içermediği de görülmektedir. Buraya bakarak da ortak odds oranının anlamlı olduğu görülmektedir.

Tüm şehirler için sigara içenlerde akciğer kanseri görülme odds'u içmeyenlere göre 2.179 kat daha fazladır yorumu yapılabilir.

4. Sonuç ve Tartışma

Sigara tüm dünyada kanserin en önemli nedenlerinden biridir. Sigara dumanı 70'den fazla, kansere neden olan madde içermektedir. Dumanın solunmasıyla, bu kimyasallar akciğerlerinize girer ve vücudunuzun kalanına yayılır. Bilim adamları bu kimyasalların DNA'ya zarar verdiğini ve önemli genlerde değişikliğe neden olduğunu göstermiştir. Bu durum ise hücrelerinizin gelişmesi ve kontrol dışı çoğalması sonucu kansere neden olmaktadır. Her beş akciğer kanserinden dördü sigaradan kaynaklanmaktadır. Akciğer kanseri tüm kanserler içinde hayatta kalma oranı en düşük olandır. İyi haber ise şudur: bu ölümlerin çoğu sigaranın vaktinde bırakılması ile önlenebilir.

Kategorik veri analiz yöntemlerinin incelenmesi amacıyla yapılan çalışma kapsamında farklı alt başlıklarda bu alandaki analiz yöntemlerinin bazılarına yer verilmiştir. Çalışmanın uygulama kısmında analiz yöntemlerinin kullanım alanları, kullanım şekilleri, değerlendirme kriterleri ifade edilmiştir. İlk olarak üç boyutlu olumsallık tabloları uygulanmıştır. Verinin modellere uyumu incelendikten sonra uyum sağlayan modeller içinden en iyi modeli seçmek için bilgi kriterlerinden yararlanılabilir. Uygulama sonucunda sadece M7 Karşılıklı bağımsızlık modeli anlamlı bulunmuştur ve [Kanser, Sigara], [Kanser, Şehir], [Şehir, Sigara] etkileşimlerinin birbirinden bağımsız olduğu sonucuna ulaşılmıştır. Geriye doğru seçilim işlemi uygulanarak yapılan log doğrusal modelin de karşılıklı bağımsızlık modeliyle uyuştuğu görülmüştür.

Logit model ve lojistik model uygulanarak belli odds oranları yorumlanmış ve genel olarak sigara kullanımının kanseri tetiklediği görülmüştür.

Elimizde bulunan değişkenlerin tamamı nominal yani sınıflanabilir olduğu için sıralanabilir log doğrusal model uygulanmamıştır. Uyum analizi yapılarak ilişkilerin çok boyutlu uzayda grafiksel gösterimleri elde edilmek istenmiştir ancak veri setinde değişkenler 2 ve 7 boyutlu olduğu için, uyum analizinin yapısı da çalışmamıza uygun değildir.

Genel bir çerçevede sigara kullanımının akciğer kanserini tetiklediğini ve yapılan alt çalışmalar doğrultusunda kullanılan sigara, kanser ve şehir değişkenlerinin yapılan analizler için anlamlı olduğunu söyleyebiliriz. Sigara kullanımının ve buna bağlı olarak akciğer kanseri olma durumunun iller bazında farklılık göstermediğini homojenlik testi ile söyleyebiliriz. Buna bağlı olarak iller bazında yapılan birleştirmenin faydalı olduğunu ve genel çerçevede bize sigara içenlerde akciğer kanseri görülme olasılığının içmeyenlere göre 2 kat daha fazla olduğunu göstermiştir.

Kaynakça

Liu Z. Smoking and lung cancer in China: combined analysis of eight case-control studies. *Int J Epidemiol*. 1992;21(2):197-201. doi:10.1093/ije/21.2.197

YALÇINKAYA, A. E. (n.d.). KATEGORİK VERİ ANALİZİNİN İSTATİSTİKSEL VERİ ANALİZİ İÇERİSİNDEKİ YERİ VE ÖNEMİ. Retrieved from http://acikerisim.deu.edu.tr:8080/xmlui/bitstream/handle/20.500.12397/11235/226854.pdf?sequence=1&isAllowed=y

ŞENEL, S., & ALATLI , B. (n.d.). Lojistik Regresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme. Retrieved from http://static.dergipark.org.tr/article-download/imported/1040000015/1040000013.pdf

(n.d.). Sigara ve Kanser - Türkiye Kanserle Savaş Vakfı. Retrieved from http://www.kanservakfi.com/sigara-ve-kanser-132.html