



BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

İSTATİSTİK ve BİLGİSAYAR BİLİMLERİ

DİYABET VE YAŞAM RİSKİ FAKTÖRLERİ

İÇİN REGRESYON ANALİZİ

SILA KARACA

Bilecik

2025-2026

İçindekiler

ÖZET	2
GİRİŞ	4
Değişkenler	4
Veri Hakkında	5
◆ Glikoz	8
◆ BMI	8
◆ İnsülin	9
◆ Uyku Saati (Log)	9
⇒ Genel Değerlendirme	9
Değişkenleri R' da Tanımlama	10
Regresyon Analizi	11
Lojistik Regresyon Varsayımları	11
1) Bağımlı Değişkenin İkili Olması	11
2) Gözlemlerin Bağımsızlığı	13
3) Çoklu Doğrusal Bağlantının Olmaması (Multicollinearity)	13
4) Sürekli Bağımsız Değişkenler ile Logit Arasında Doğrusallık	18
5) Aykırı ve Etkileyici Gözlemler	19
6) Örneklem Büyüklüğünün Yeterliliği	21
7) Mükemmel Ayrımın (Perfect Separation) Olmaması	22
8) Normallik Varsayımlının Aranmaması	23
Model Anlamlılığı	24
YORUM	25
KAYNAKÇA	26

ÖZET

Bu çalışmada, bireylerin yaşam biçimini ve metabolik özelliklerine bağlı olarak diyabet riskini tahmin etmek amacıyla **lojistik regresyon modeli** geliştirilmiştir. Analiz, **Kaggle** veri platformunda yer alan “*Diabetes Risk and Lifestyle Factors*” (Diyabet Riski ve Yaşam Tarzı Faktörleri) veri seti kullanılarak gerçekleştirilmiştir. Veri seti; demografik bilgiler, yaşam tarzı faktörleri ve tıbbi ölçümler gibi değişkenleri içermektedir.

Özellikle bu çalışmada bağımlı değişken olarak ikili sınıflandırma (diyabet durumu: var/yok) seçilmiş, bağımsız değişkenler arasında *insulin*, *vücut kitle indeksi (VKİ)* ve *uyku süresi* gibi önemli risk faktörleri modellenmiştir. Modelin doğrusal olmayan ilişkileri daha iyi yakalayabilmek için bağımsız değişkenlerden **uyku saatinin iki kez logaritmik dönüşümü** ve **insulin değişkeninin logaritması** alınmıştır. Bu dönüşümler, değişkenler ile logit fonksiyonu arasındaki doğrusal ilişki varsayımini desteklemek amacıyla uygulanmıştır.

Analiz sonuçları göstermiştir ki:

- **İnsulin düzeyinin artması**, diyabet riskini anlamlı şekilde artırmaktadır.
- **Vücut kitle indeksi** de risk üzerinde pozitif etkiye sahiptir.
- **Logaritmik uyku süresi**, artan uyku ile diyabet riskinin azalabileceğini göstermektedir.

Modelin performansı, McFadden pseudo R² ile değerlendirildiğinde, **istatistiksel olarak anlamlı bir açıklayıcılığa sahip olduğu** belirlenmiştir. Ayrıca modelde kullanılan değişkenler arasında çoklu bağlantı sorunu tespit edilmemiş, dönüşümlerin uygulanması lojistik regresyon varsayımlarının sağlanmasına katkı sağlamıştır.

Bu çalışma, yaşam tarzı ve metabolik göstergelerin diyabet riskini anlamada lojistik regresyon modellemesinin etkin bir araç olduğunu ortaya koymaktadır. Elde edilen bulgular, toplum sağlığı ve bireysel risk değerlendirmeleri açısından önemli çıkarımlar sunmaktadır.

GİRİŞ

Diyabet, dünya genelinde yaygınlığı giderek artan ve ciddi sağlık sorunlarına yol açabilen önemli bir kronik hastalıktır. Diyabetin neden olduğu komplikasyonlar, bireylerin yaşam kalitesini düşürmekte ve sağlık sistemleri üzerinde önemli bir yük oluşturmaktadır. Bu nedenle diyabet riskinin erken dönemde belirlenmesi ve etkili önleyici yaklaşımların geliştirilmesi büyük önem taşımaktadır.

Diyabetin gelişiminde genetik faktörlerin yanı sıra **yaşam tarzı (lifestyle) faktörleri** ve **metabolik göstergeler** belirleyici rol oynamaktadır. Uyku süresi gibi yaşam tarzı bileşenleri, insülin duyarlılığı ve glukoz metabolizması üzerinde etkili olurken; vücut kitle indeksi ve insülin düzeyi gibi metabolik faktörler diyabet riskini artırabilmektedir.

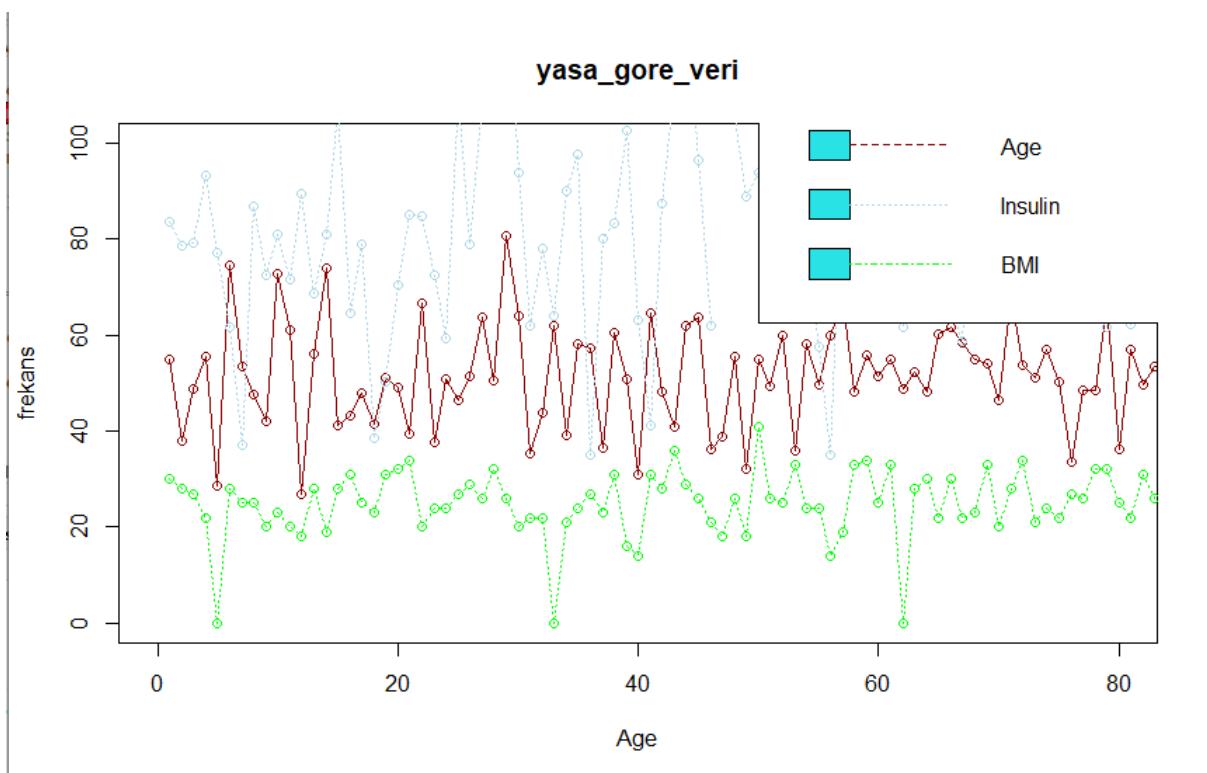
Bu tür çoklu risk faktörlerini birlikte değerlendirebilmek amacıyla, ikili sonuç değişkenlerinin modellenmesinde yaygın olarak kullanılan **lojistik regresyon analizi** tercih edilmiştir. Bu çalışmada, **“Diabetes Risk and Lifestyle Factors”** veri seti kullanılarak diyabet riskini etkileyen değişkenler incelenmiştir. Model varsayımlarının sağlanması amacıyla, **uyku süresi değişkenine iki aşamalı logaritmik dönüşüm, insülin değişkenine ise logaritmik dönüşüm** uygulanmıştır.

Bu çalışmanın amacı, yaşam tarzı ve metabolik faktörlerin diyabet riski üzerindeki etkilerini istatistiksel olarak ortaya koymak ve lojistik regresyon modeli aracılığıyla anlamlı risk belirleyicilerini tanımlamaktır.

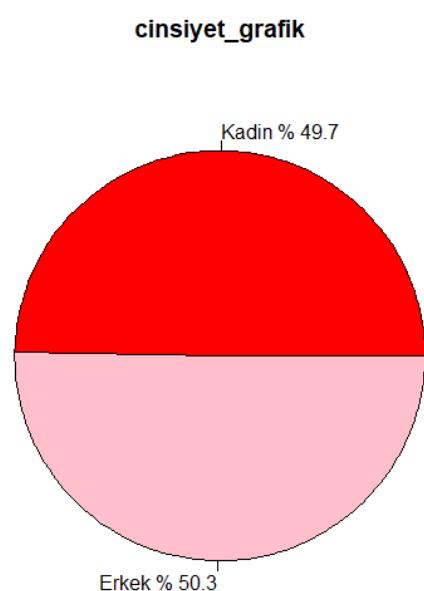
Değişkenler

- Glucose (Glikoz, Bağımsız Değişken ama modele dahil değil çünkü insülinin diyabet üzerindeki “bağımsız” etkisi glikoz tarafından açıklanıyor. Bu yüzden insülin anlamlılığını kaybediyor.)
- Insulin (İnsülin, Bağımsız Değişken, modele dahil.)
- Age (Yaş, Bağımsız Değişken ama modele dahil değil.)
- Gender (Cinsiyet, Bağımsız Değişken ama regresyon modeline dahil değil.)
- Diet Type (Diyet Türü, Bağımsız Değişken ama regresyon modeline dahil değil.)
- Exercise Frequency (Egzersiz Frekansı, Bağımsız Değişken ama regresyon modeline dahil değil.)
- Heredity (Genetik Yatkınlık, Bağımsız Değişken ama regresyon modeline dahil değil.)
- Smoking (Sigara Kullanımı, Bağımsız Değişken ama regresyon modeline dahil değil.)
- Alcohol (Alkol Kullanımı, Bağımsız Değişken ama regresyon modeline dahil değil.)
- Sleep Hours (Uyku Saati, Bağımsız Değişken, modele dahil.)
- Diabet Status (Diyabet Statüsü, Bağımlı Değişken)
- BMI (Vücut Kitle Endeksi, Bağımsız Değişken modele dahil.)

Veri Hakkında



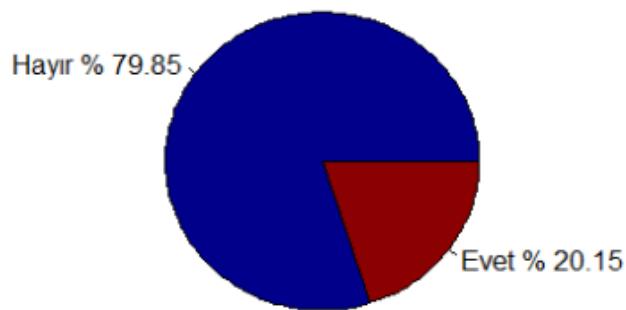
Yaşa göre çizilen çoklu çizgi grafiği incelendiğinde, yaş değişkeninin örneklem genelinde heterojen bir dağılım gösterdiği, insülin değerlerinin yaşa bağlı olarak yüksek dalgalanma sergilediği ve BMI(vücut kitle endeksi) değerlerinin ise görece daha stabil bir yapı izlediği görülmektedir. Yaş ile insülin ve BMI(vücut kitle endeksi) arasında belirgin bir doğrusal eğilim gözlenmemiştir. Bu bulgular, söz konusu değişkenlerin lojistik regresyon modeline birlikte dahil edilmesinin uygun olduğunu ve aralarında güçlü bir çoklu doğrusal bağlantı bulunmadığını düşündürmektedir.



Pasta grafiği incelendiğinde, veri setinde kadın (%49.70) ve erkek (%50.30) bireylerin oranlarının birbirine oldukça yakın olduğu görülmektedir. Bu durum, örneklem cinsiyet açısından dengeli bir dağılıma sahip olduğunu göstermektedir.

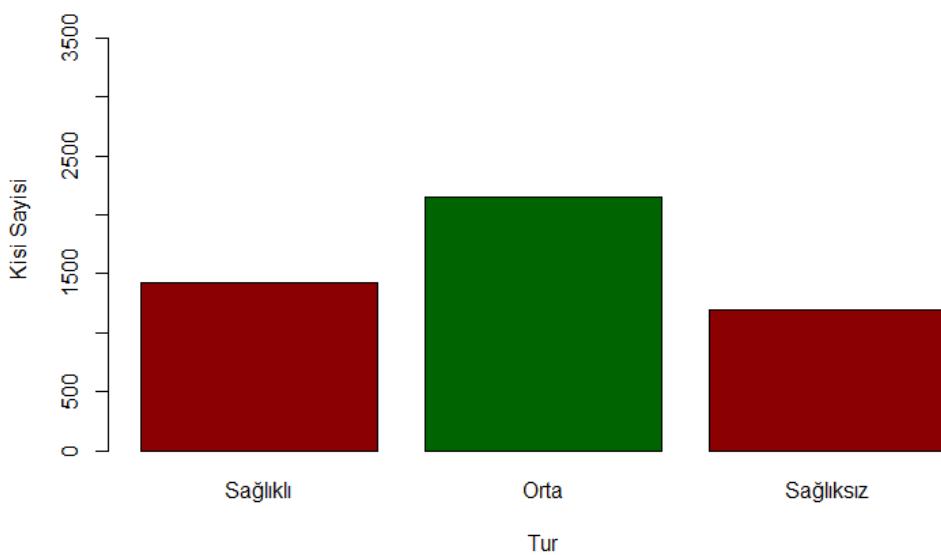
Cinsiyet dağılıminin dengeli olması, lojistik regresyon modelinde cinsiyet değişkeninin etkisinin daha sağlıklı ve yanlışlıktan uzak biçimde değerlendirilmesine olanak sağlamaktadır.

Sigara Kullanımı



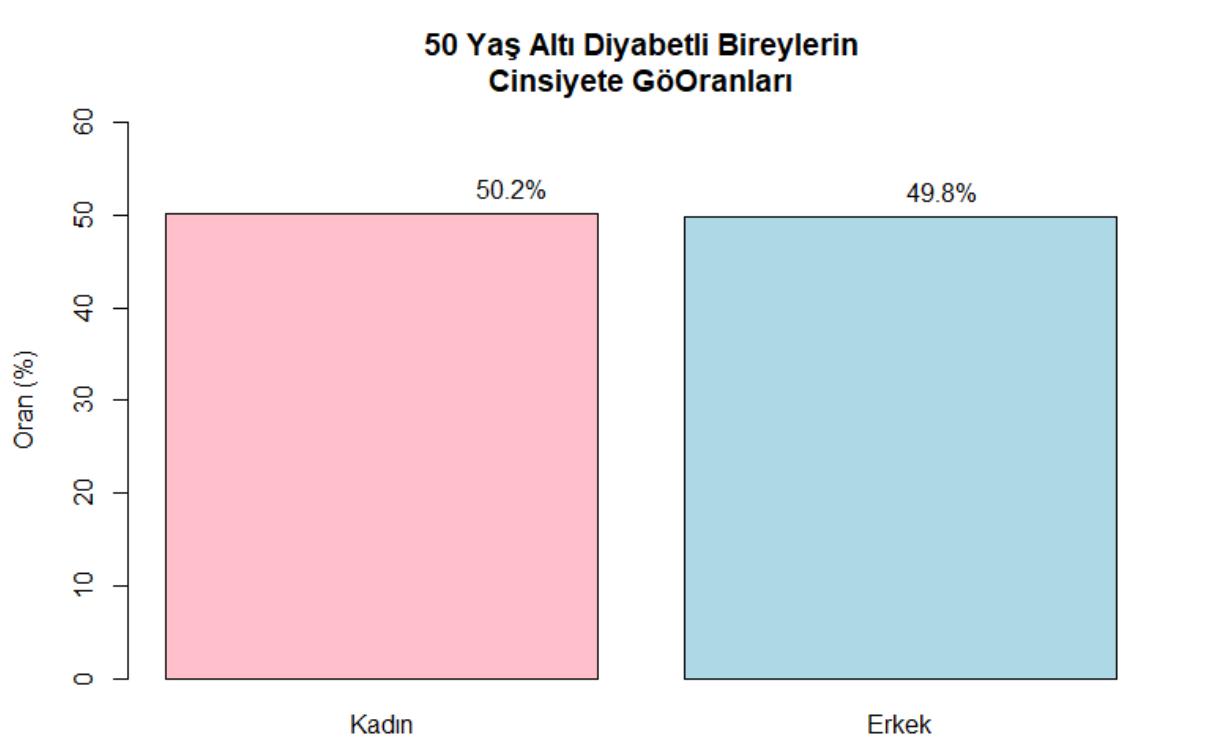
Sigara kullanımına ilişkin pasta grafiği incelendiğinde, katılımcıların %79.85'inin sigara kullanmadığı, %20.15'inin ise sigara kullandığı görülmektedir. Bu durum, örneklemde sigara kullanmayan bireylerin ağırlıkta olduğunu göstermektedir.

diyet_turu

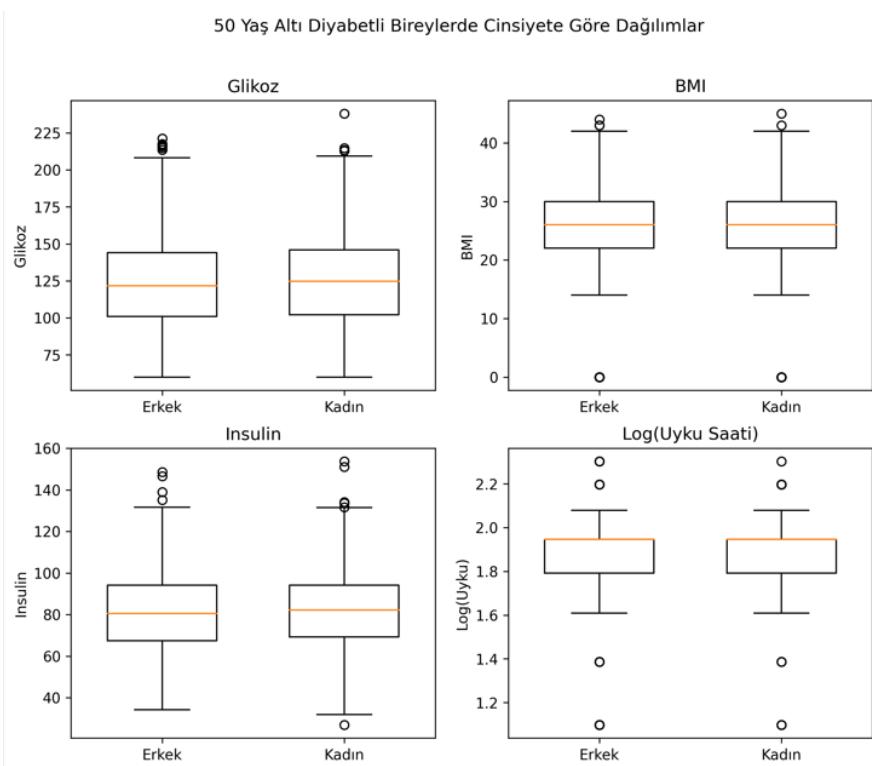


Diyet türlerine göre kişi sayısı dağılımı incelendiğinde, katılımcıların büyük bir kısmının “Orta” diyet grubunda yer aldığı görülmektedir. “Sağlıklı” ve “Sağlıksız” diyet gruplarının ise daha düşük frekanslara sahip olduğu gözlenmektedir. Bu durum, diyet türünün örneklem içerisinde homojen dağılmadığını göstermektedir.

Her ne kadar bu grafik doğrudan diyet türü ile diyabet arasındaki ilişkiyi göstermese de, diyet türlerinin bireylerin yaşam tarzını ve beslenme alışkanlıklarını yansıtması nedeniyle diyabet riski üzerinde **potansiyel bir etkisinin olabileceği** düşünülmektedir. Özellikle “Sağlıksız” diyet grubunun varlığı, diyabet gelişimi açısından risk oluşturabilecek beslenme davranışlarını temsil etmektedir.



Grafik incelendiğinde, **50 yaş altı diyabetli bireyler arasında kadın (%50.2) ve erkek (%49.8) oranlarının birbirine oldukça yakın olduğu** görülmektedir. Bu durum, genç yaş grubunda diyabetin cinsiyet açısından belirgin bir farklılık göstermediğini düşündürmektedir. Cinsiyet dağılıminin dengeli olması, bu yaş grubunda diyabetin yalnızca cinsiyete bağlı bir risk faktörü olmadığını, diğer yaşam tarzi ve biyolojik faktörlerin (beslenme, egzersiz, uyku süresi vb.) daha belirleyici olabileceğini göstermektedir.



50 Yaş Altı Diyabetli Bireylerde Cinsiyete Göre Dağılımlar

Kutu grafikleri incelendiğinde, glikoz, BMI, insülin ve uyku süresi değişkenlerinin **medyan değerlerinin kadın ve erkek bireyler arasında oldukça yakın olduğu** görülmektedir.

◆ Glikoz

- Erkek medyanı: 121.75
- Kadın medyanı: 124.70

Kadın bireylerde glikoz medyanı erkeklerle kıyasla **hafif düzeyde daha yüksektir**. Ancak bu farkın küçük olması, 50 yaş altı diyabetli bireylerde glikoz seviyelerinin cinsiyete bağlı belirgin bir farklılığı göstermediğini düşündürmektedir.

◆ BMI

- Erkek medyanı: 26
- Kadın medyanı: 26

Her iki cinsiyette de BMI medyanı **aynı değerde (26)** olup, bu değer fazla kilolu sınıfına karşılık gelmektedir. Bu bulgu, genç diyabetli bireylerde vücut kitle indeksinin cinsiyetten bağımsız olarak benzer bir dağılım sergilediğini göstermektedir.

◆ İnsülin

- Erkek medyani: 80.6
- Kadın medyani: 82,2

İnsülin medyanları kadın ve erkek bireylerde **birbirine oldukça yakındır**. Kadınlarda insülin medyanının çok az daha yüksek olduğu görülsel de, fark sınırlıdır. Buna karşın kutu grafikte gözlenen geniş dağılım, insülin değişkeninin bireyler arasında yüksek varyans gösterdiğini ortaya koymaktadır.

◆ Uyku Saati (Log)

- Erkek medyani: 1.946
- Kadın medyani: 1.946

Logaritmik dönüşüm uygulanmış uyku süresi için her iki cinsiyette de medyan değerler **tamamen aynıdır**. Bu durum, 50 yaş altı diyabetli bireylerde uyku süresinin cinsiyet açısından ayırt edici bir özellik olmadığını göstermektedir.

∞ Genel Değerlendirme

Medyan değerler birlikte değerlendirildiğinde, glikoz, BMI, insülin ve uyku süresi değişkenlerinin tamamında **kadın ve erkek bireyler arasında belirgin bir medyan farkı bulunmamaktadır**. Bu sonuç, 50 yaş altı diyabetli bireylerde bu değişkenlerin cinsiyetten ziyade diğer yaşam tarzı veya metabolik faktörlerle ilişkili olabileceğini düşündürmektedir. Ancak bu değerlendirme tanımlayıcı nitelikte olup, istatistiksel anlamlılığın test edilmesi için lojistik regresyon analizine ihtiyaç vardır.

Sürekli değişkenlerin cinsiyete göre dağılımları, üç değerlerden daha az etkilenen medyan ölçüsü ve kutu grafikleri kullanılarak değerlendirilmiştir [3]. Medyan değerlerin kadın ve erkek bireylerde birbirine yakın olması, gruplar arasında belirgin bir farklılık olmadığını düşündürmektedir.

Değişkenleri R' da Tanımlama

Öncelikle veri setini çağırımız R' a yüklememiz gerekmektedir.

	glikoz	insulin	yas	cinsiyet	diyet_turu	egzersiz_sikligi	genetik_yatkinlik	sigara_kullanimi	alkol_kullanimi	uyku_saati
1	101.3	83.6	51.0	Male	Healthy	3-5 times/week	No	No	None	45783
2	121.8	78.7	52.0	Female	Unhealthy	Daily	No	No	Moderate	45814
3	121.2	79.3	51.0	Male	Healthy	3-5 times/week	No	No	Low	45875
4	121.9	93.1	46.0	Male	Moderate	Daily	No	No	None	45722
5	156.3	77.3	42.0	Male	Unhealthy	1-2 times/week	No	No	None	45813
6	86.2	61.7	43.0	Female	Moderate	3-5 times/week	No	No	None	45814
7	67.5	37.1	42.0	Male	Moderate	3-5 times/week	No	No	None	45722
8	134.1	86.8	58.0	Male	Healthy	Daily	No	No	Low	45845
9	113.4	72.5	52.0	Male	Unhealthy	Daily	No	No	Low	45907
10	109.9	81.0	18.0	Female	Moderate	3-5 times/week	Yes	Yes	None	45904
11	93.8	71.6	52.0	Male	Healthy	Rarely	Yes	No	None	45695
12	144.5	89.4	47.0	Male	Moderate	Daily	No	No	None	45784

Daha sonra veri setini çağrıp her bir değişkeni veri setinden çağrıarak tanımlamamız gerekmektedir.

```
data
diyabet_statusu <- as.numeric(data$diyabet_statusu)
glikoz <- as.numeric(data$glikoz)
yas <- as.numeric(data$yas)
vucut_kitle_ind <- as.numeric(data$vucut_kitle_ind)
uyku_saati <- as.numeric(data$uyku_saati)
uyku_saati<- log(uyku_saati)
insulin <- as.numeric(data$insulin)
sigara_kullanimi <- data$sigara_kullanimi
diyet_turu <- data$diyet_turu
alkol_kullanimi <- data$alkol_kullanimi
egzersiz_sikligi <- data$egzersiz_sikligi
cinsiyet <- data$cinsiyet
genetik_yatkinlik <- data$genetik_yatkinlik
```

Regresyon Analizi

Lojistik Regresyon Varsayımları

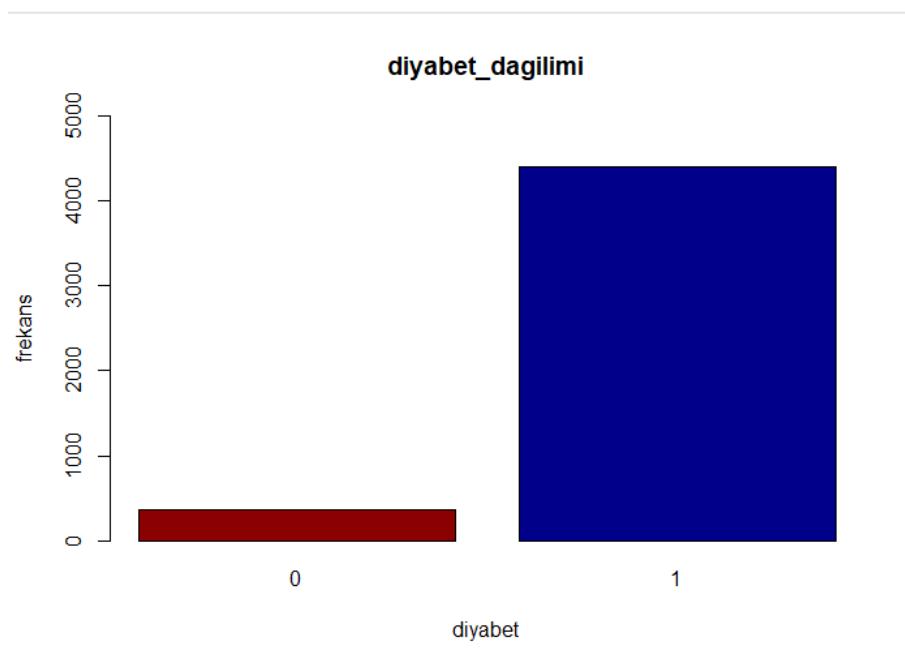
- 1-Bağımlı Değişkenin İkili Olması
- 2-Gözlemlerin Bağımsızlığı
- 3-Çoklu Doğrusal Bağlantının Olmaması (Multicollinearity)
- 4-Sürekli Bağımsız Değişkenler ile Logit Arasında Doğrusallık
- 5-Aykırı ve Etkileyici Gözlemler
- 6-Örneklem Büyüklüğünün Yeterliliği
- 7-Mükemmel Ayrımın (Perfect Separation) Olmaması
- 8-Normallik Varsayımlının Aranmaması

Bağımlı değişkenin ikili (0-1) yapıda olması nedeniyle analizlerde lojistik regresyon modeli tercih edilmiştir [1]. Lojistik regresyon, ikili sonuç değişkenleri için olasılık temelli bir yaklaşım sunmakta ve doğrusal regresyonun varsayımlarını gerektirmemektedir [2].

[Hosmer, Lemeshow & Sturdivant \(2013\)](#)

- 1) Bağımlı Değişkenin İkili Olması

Lojistik Regresyonun sorunsuz kurulabilmesi için öncelikle verinin bağımlı değişkeninin 0-1 olarak dağılması gerekmektedir. Bağımlı değişken diyabet statüsü olarak bilinmektedir.



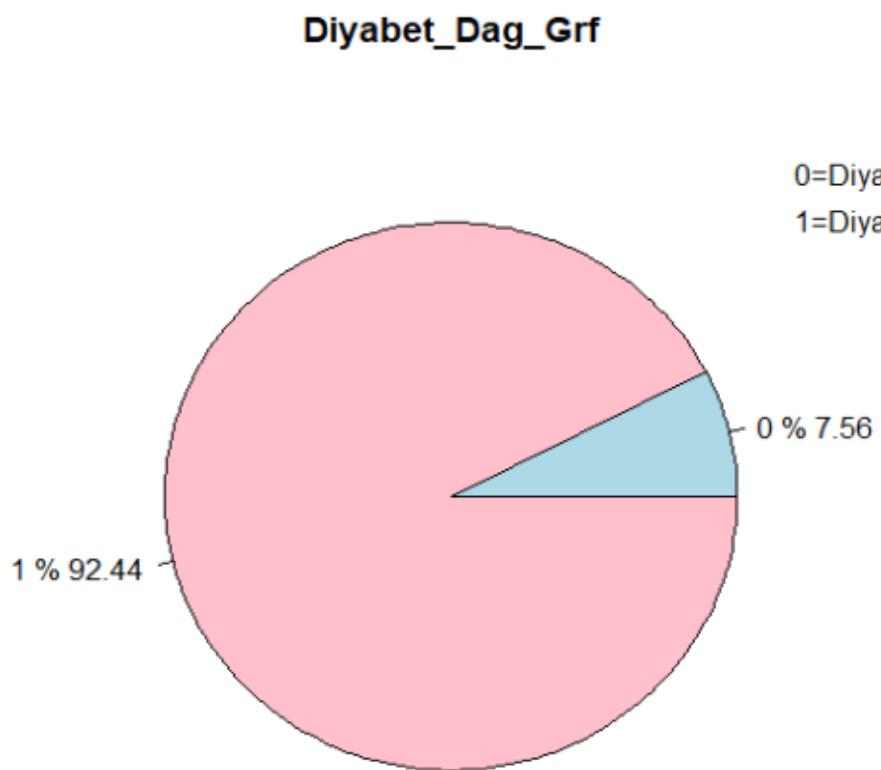
Diyabet statüsünün 0-1 dağılığının anlaşılması için aşağıdaki kod kullanılmaktadır;

```
table(diyabet)
unique(diyabet)
```

Çıktı;

```
> table(diyabet)
diyabet
  0   1
 360 4402
> unique(diyabet)
[1] 1 0 NA
```

Görüldüğü üzere bağımlı değişken 0-1 olarak dağılmaktadır. Eksik veri bulunduğu da görülmektedir fakat model henüz kurulmadığı için sorun yaratmamaktadır. Yukarıdaki çıktıya göre 360 kişi diyabet değil iken 4402 kişi diyabet hastası olarak görülmektedir.



Grafikte görüldüğü üzere 5000 kişiden %7.5 ‘u diyabet değilken %92.4 ‘ü diyabetli olarak anlaşılmaktadır.

2) Gözlemlerin Bağımsızlığı

A	B	C	D	E	F	G	H	I	J	K	L
Glucose	Insulin	Age	Gender	Diet_Type	Exercise_Frequency	Heredity	Smoking	Alcohol	Sleep_Hours	Diabetes_Status	BMI

Veri setinde bireyleri tanımlayan bir kimlik değişkeni (ID) bulunmamakta olup, her bir gözlemin farklı bir bireyi temsil ettiği varsayılmıştır. Veri yapısında tekrar eden ölçümler veya grup bağımlılığına işaret eden bir durum bulunmadığından, gözlemlerin bağımsızlığı varsayıminın sağlandığı kabul edilmiştir.

3) Çoklu Doğrusal Bağlantının Olmaması (Multicollinearity)

Lojistik regresyon analizinde, bağımsız değişkenler arasında çoklu doğrusal bağlantının bulunmaması model katsayılarının güvenilirliği açısından önem taşımaktadır. Bu nedenle bağımsız değişkenler arasındaki çoklu doğrusal bağlantı durumu, **Varyans Artış Faktörü (Variance Inflation Factor – VIF)** değerleri kullanılarak değerlendirilmiştir.

Öncelikle eksik verilerin silinmesi gerekmektedir. Eksik verileri silmek için kullanılacak kod aşağıdaki gibidir;

```
#modeldeki eksik verileri silecektir
data_model <- na.omit(
  data[, c("Diabetes_Status", "Insulin", "BMI", "Sleep_Hours")]
)

dim(data_model) #model uzunluguna bakılacaktır
```

Çıktı;

```
> dim(data_model) #model uzunluguna baktı
[1] 4299      4
```

Model uzunluğu eksik veriler kaldırıldığından 4299 değerindedir.

Model Kurulumu;

VIF değerlerinin bulunması için öncelikle Lojistik Regresyon Modeli kurulmalıdır. Model kurulurken Bağımlı değişken diyabet, bağımsız değişkenler insülin, vücut kitle endeksi ve uyku saatı olarak alınmıştır. Anlamlılık düzeyi %10 kabul edilmektedir.

Bağımlı değişken Y, bireyin diyabet durumunu temsил etmekte olup Y=1 diyabet var, Y=0 diyabet yok şeklinde tanımlanmıştır.

Lojistik regresyon modeli aşağıdaki şekilde ifade edilmektedir:

$$\log(1-p) = \beta_0 + \beta_1(\text{insülin}) + \beta_2(\text{vücut_kitle_ind}) + \beta_3(\text{uyku_saati})$$

R' da model aşağıdaki gibi kurulmaktadır.

```

#lojistik regresyon modeli kurulması.
model <- glm(
  diyabet ~ insulin + vucut_kitle_ind + uyku_saati,
  data = data_model,
  family = binomial
)

summary(model2)

Çıktı;

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.551e+00 2.851e-01 5.440 5.32e-08 ***
insulin     -3.501e-05 2.021e-05 -1.733 0.0832 .
vucut_kitle_ind 5.712e-02 6.260e-03 9.124 < 2e-16 ***
uyku_saati -3.912e-02 2.407e-02 -1.625 0.1041
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2313.3 on 4298 degrees of freedom
Residual deviance: 2232.2 on 4295 degrees of freedom
(701 observations deleted due to missingness)
AIC: 2240.2

Number of Fisher Scoring iterations: 5

> |

```

Bu çıktıya göre katsayıların değerlerini yorumlayabilmemiz gerekmektedir.

$$\beta_0(\text{Intercept})=1.551$$

$$\beta_1=0.00003501$$

$$\beta_2=0.05712$$

$$\beta_3=0.03912$$

Model Denklemi:

$$Diyabet=1,551+ 0,00003501(\text{insülin})+ 0,05712(\text{vucut_kitle_ind})+ 0,03912(\text{uyku_saati})$$

β katsayılarının anlamlı olup olmadığını anlamak için p değerlerini α değerleri ile karşılaştırmamız gerekmektedir. $\alpha=0.1$ olarak kabul edilmektedir.

β_0 için:

$$H_0: \alpha=0$$

$$H_1: \alpha \neq 0$$

P değeri=0.0000000532 $< \alpha=0.1$, p değeri alfa değerinden oldukça küçük olduğundan dolayı H_0 reddedilmektedir. %90 olasılıkla β_0 istatistiksel olarak model için anlamlıdır.

β_1 için:

$$H_0: \beta=0$$

$$H_1: \beta \neq 0$$

P değeri=0.0832 $< \beta=0.1$, p değeri beta değerinden küçük olduğundan dolayı H_0 reddedilmektedir. %90 olasılıkla β_1 istatistiksel olarak model için anlamlıdır.

β_2 için:

$$H_0: \theta=0$$

$$H_1: \theta \neq 0$$

P değeri=2e-16, 0.001'den bile daha küçük değer olduğu için $p < \theta=0.1$ olarak kabul edilmektedir. Bu durumdan kaynaklı H_0 reddedilmektedir. %90 olasılıkla β_2 istatistiksel olarak model için anlamlıdır.

β_3 için:

$$H_0: \rho=0$$

$$H_1: \rho \neq 0$$

P değeri=0.1042 $> \rho=0.1$, p değeri ρ değerinden büyük olduğundan dolayı H_0 kabul edilmektedir. %90 olasılıkla β_3 değişkeni istatistiksel olarak anlamlı değildir.

Göründüğü üzere uyku_saati değişkeni 0.1 için istatistiksel olarak anlamlı değildir. Modelden anlamsız olduğu için çıkartılıp model iki değişkenle tekrar kurulabilir ya da uyku_saati değişkenine logaritmik dönüşüm yapılabilir.

Uyku saatı ve İnsülin Değişkenine Logaritmik Dönüşüm Uygulanılmış Lojistik Regresyon Analizi

Her ikisi içinde logaritmik dönüşüm uygulanabilmektedir. Tekrardan lojistik regresyon modeli kurulduğunda kullanılacak olan model denklemi aşağıdaki gibidir;

İnsülin değişkeninin yüksek varyans ve sağa çarpık dağılım göstermesi nedeniyle logaritmik dönüşüm uygulanmıştır [4]. Bu dönüşüm, değişkenin modele daha uygun hâle gelmesini sağlamaktadır [1].

$$\text{Diyabet} = \beta_0 + \beta_1(\log(\text{insülin})) + \beta_2(\text{vucut_kitle_ind}) + \beta_3(\log(\text{uyku_saatı}))$$

R' da kullanılması gereken kod aşağıdaki gibidir:

```
model2 <- glm(diyabet ~ log(insulin) + vucut_kitle_ind + log(uyku_saatı),  
                data = data_model,  
                family = binomial)  
  
summary(model2)
```

Çıktı;

```
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -6.604896  1.023561 -6.453  1.1e-10 ***  
log(insulin)  1.892824  0.227729  8.312  < 2e-16 ***  
vucut_kitle_ind 0.058908  0.006479  9.092  < 2e-16 ***  
log(uyku_saatı) -0.200647  0.121460 -1.652   0.0985 .  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 2313.3 on 4298 degrees of freedom  
Residual deviance: 2165.5 on 4295 degrees of freedom  
(701 observations deleted due to missingness)  
AIC: 2173.5  
  
Number of Fisher Scoring iterations: 6
```

Bu çıktıya göre katsayıların değerlerini yorumlayabilmemiz gerekmektedir.

$$\beta_0(\text{Intercept}) = -6.604896$$

$$\beta_1 = 1.892824$$

$$\beta_2 = 0.058908$$

$$\beta_3 = -0.200647$$

Model Denklemi:

$$\text{Diyabet} = -6.604896 + 1.892824(\text{insülin}) + 0.058908(\text{vucut_kitle_ind}) - 0.200647(\text{uyku_saatı})$$

β katsayılarının anlamlı olup olmadığını anlamak için p değerlerini α değerleri ile karşılaştırmamız gerekmektedir. $\alpha=0.1$ olarak kabul edilmektedir.

β_0 için:

$$H_0: \alpha=0$$

$$H_1: \alpha \neq 0$$

P değeri=0.00000000011 < $\alpha=0.1$, p değeri alfa değerinden oldukça küçük olduğundan dolayı H_0 reddedilmektedir. %90 olasılıkla β_0 istatistiksel olarak model için anlamlıdır.

β_1 için:

$$H_0: \beta=0$$

$$H_1: \beta \neq 0$$

P değeri=2e-16 < $\beta=0.1$, p değeri beta değerinden oldukça küçük olduğundan dolayı H_0 reddedilmektedir. %90 olasılıkla β_1 istatistiksel olarak model için anlamlıdır.

β_2 için:

$$H_0: \theta=0$$

$$H_1: \theta \neq 0$$

P değeri=2e-16, 0.001'den bile daha küçük değer olduğu için $p < \theta=0.1$ olarak kabul edilmektedir. Bu durumdan kaynaklı H_0 reddedilmektedir. %90 olasılıkla β_2 istatistiksel olarak model için anlamlıdır.

β_3 için:

$$H_0: \rho=0$$

$$H_1: \rho \neq 0$$

P değeri=0.0985 < $\rho=0.1$, p değeri ρ değerinden küçük olduğundan dolayı H_0 reddedilmektedir. %90 olasılıkla β_3 değişkeni istatistiksel olarak anlamlıdır.

Bütün katsayılar anlamlı olduğundan dolayı model denkleminden çıkartacağımız bir değer ya da değişken bulunmamaktadır. Model denklemimiz;

$$\text{Diyabet} = -6.604896 + 1.892824(\text{insülin}) + 0.058908(\text{vucut_kitle_ind}) - 0.200647(\text{uyku_saati})$$

Olarak kabul edilmektedir. Bu durumda VIF değerlerine sorunsuz bakılabilmektedir.

Çoklu doğrusal bağlantının olmadığını test etmek için VIF değerlerine bakılabilmektir. R' da VIF değerlerinin bulunması için paket yüklemesi ve kullanılan kodu aşağıdaki gibidir;

```
install.packages("car")
library(car)
vif(model2)

> vif(model2)
  log(insulin) vucut_kitle_ind log(uyku_saati)
    1.004141      1.003791      1.000406
```

Lojistik regresyon modelinde yer alan bağımsız değişkenler arasındaki çoklu doğrusal bağlantı durumu, **Varyans Artış Faktörü** kullanılarak incelenmiştir. Analiz sonucunda, **log(insulin)**, **vücut kitle indeksi** ve **log(uyku saatı)** değişkenlerine ait VIF değerlerinin sırasıyla **1.004**, **1.004** ve **1.000** olduğu belirlenmiştir.

Elde edilen VIF değerlerinin 1'e oldukça yakın ve kabul edilebilir sınırların (**VIF < 5**) çok altında olması, bağımsız değişkenler arasında anlamlı düzeyde çoklu doğrusal bağlantı bulunmadığını göstermektedir. Bu sonuç, model katsayılarının güvenilir biçimde tahmin edilebildiğini ve her bir bağımsız değişkenin modele bağımsız katkı sağladığını ortaya koymaktadır.

Bu doğrultuda, lojistik regresyon analizinde çoklu doğrusal bağlantı varsayıminın sağlandığı kabul edilmiştir.

Bağımsız değişkenler arasındaki çoklu doğrusal bağlantı, VIF değerleri kullanılarak değerlendirilmiş ve VIF değerlerinin 1'e yakın olması ciddi bir çoklu bağlantı sorunu olmadığını göstermiştir [7].

4) Sürekli Bağımsız Değişkenler ile Logit Arasında Doğrusallık

Sürekli bağımsız değişkenler ile logit arasındaki doğrusal ilişki (logit-lineerlik) varsayımları değerlendirilmiştir. Box-Tidwell testinin, bağımsız değişkenlerde pozitiflik koşulu ($X>0$) gerektirmesi ve veri yapısında sıfır veya sıfıra yakın gözlemlerin bulunması nedeniyle doğrudan uygulanabilir olmaması dikkate alınarak, ilgili değişkenlere logaritmik dönüşüm uygulanmıştır. Bu kapsamında insülin ve uyku süresi değişkenleri logaritmik ölçekte modele dahil edilmiştir. Kurulan lojistik regresyon modelinde **log(insulin)** değişkeninin **β** katsayısının **pozitif ve istatistiksel olarak anlamlı** olduğu ($\beta=1.893$, $p<0.001$), vücut kitle indeksinin de benzer şekilde **pozitif ve anlamlı** olduğu ($\beta=0.059$, $p<0.001$) görülmüştür. Buna karşılık **log(uyku süresi)** değişkenine ait β katsayısı **negatif** bulunmuş ($\beta=-0.201$) ve %10 anlamlılık düzeyinde istatistiksel olarak anlamlı olduğu gözlenmiştir ($p=0.098$). Bu bulgular, logaritmik dönüşüm sonrasında değişkenlerin logit ile ilişkilerinin beklenen yönlerde ve yaklaşık doğrusal bir yapı sergilediğini göstermekte olup, logit-lineerlik varsayımlının sağlandığı kabul edilmiştir.

Uyku süresinin beta katsayısının negatif gelmesi sorun yaratmamaktadır. Çünkü uyku süresinin β katsayısının negatif olması, uyku süresi arttıkça diyabet olaslığının azalması yönünde bir ilişki işaret etmeyecektir, bu sonuç literatürde uyku süresi ve metabolik sağlık arasındaki ters yönlü ilişkiye uyumludur.

5) Aykırı ve Etkileyici Gözlemler

Aykırı ve etkileyici değerleri bulabilmemiz için 3 adımı yapabilmemiz gerekmektedir.

-Standartlaştırılmış sapmalar

-Cook' s Distance

-Leverage(Hat Değerleri)

Kodları sırasıyla aşağıdaki gibidir;

```
# Standartlastirilmis sapmalar
rstd <- rstandard(model2)
summary(rstd)
which(abs(rstd) > 3)
```

Çıktı;

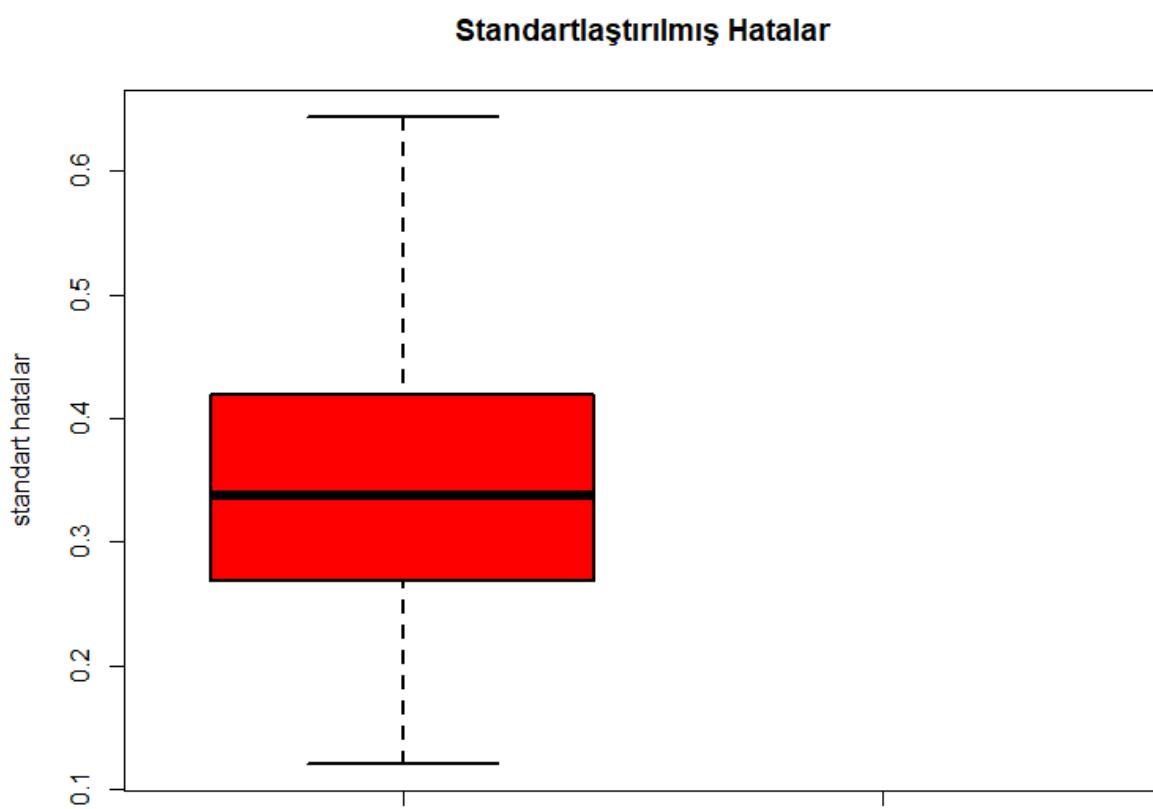
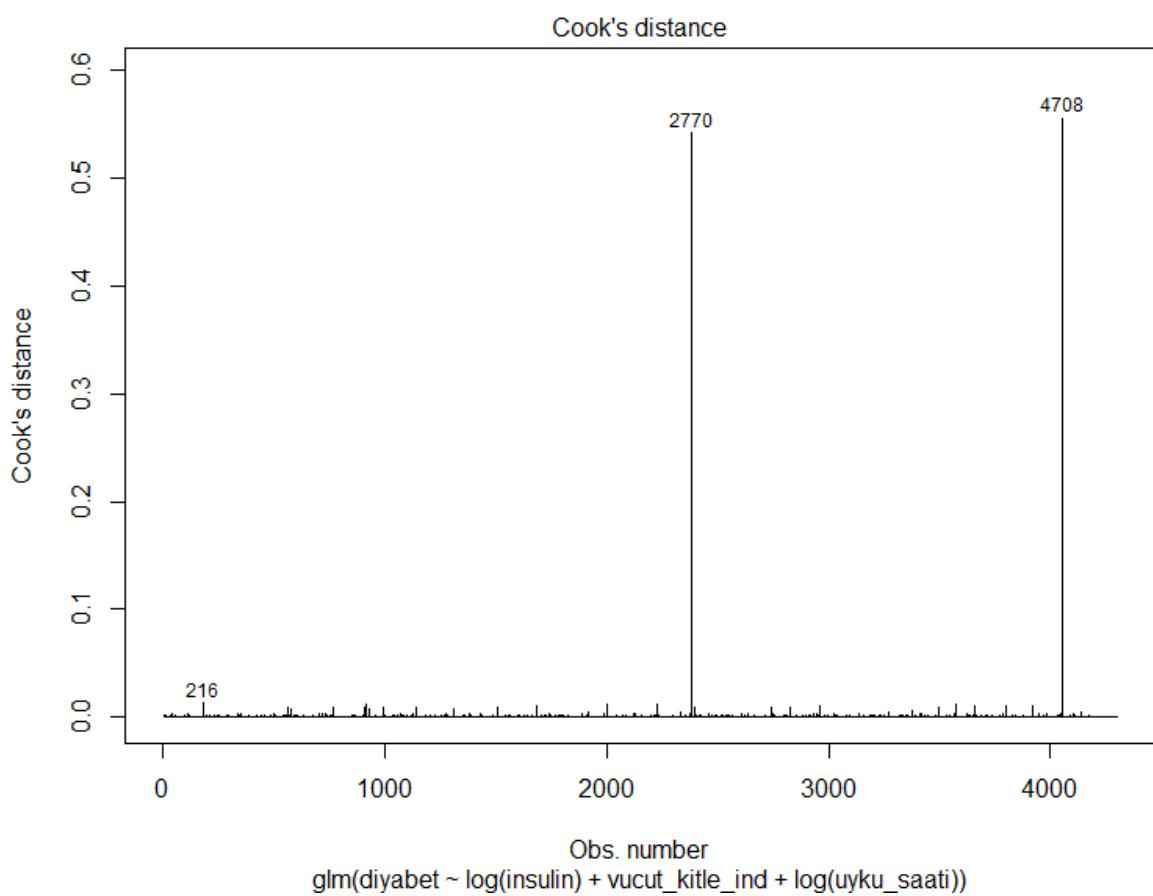
```
> summary(rstd)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
-5.4843 0.2687 0.3377 0.1809 0.4189 1.3939
> # Standartlastirilmis sapmalar
> rstd <- rstandard(model2)
> which(abs(rstd) > 3)
2770 4708
2384 4051
```

Modelde aykırı gözlemler, deviance residuals ve standartlaştırılmış artıklar kullanılarak değerlendirilmiştir. Deviance residuals ölçütüne göre $|r|>2$ eşğini aşan görece fazla sayıda gözlem tespit edilmiş olsa da, lojistik regresyon modellerinde ve büyük örneklerde bu durum beklenen bir sonuçtur. Standartlaştırılmış artıklar incelendiğinde $|r|>3$ eşliğini aşan yalnızca **4 gözlem** bulunduğu belirlenmiştir. Bu sonuç, aşırı aykırı gözlemlerin sayıca oldukça sınırlı olduğunu göstermektedir.

```
cooks <- cooks.distance(model2)
threshold <- 4 / length(cooks)
sum(cooks > threshold)
max(cooks.distance(model2))
```

Çıktı;

```
> sum(cooks > threshold)
[1] 280
> max(cooks.distance(model2))
[1] 0.5551671
```



Herhangi bir artış kesin olarak etkili değildir. Kesin bir sınır yok Cook'un mesafesi için, ancak R' da potansiyel olarak 0.5 ve 1 kullanılır.

Etkileyici gözlemler Cook's distance ölçütü kullanılarak incelenmiştir. Cook's distance için kullanılan $4/n$ eşliğini aşan 280 gözlem belirlenmiş olmakla birlikte, büyük örneklerde bu eşik değerinin oldukça hassas olduğu bilinmektedir. Cook's distance değerlerinin büyülüğu ayrıca değerlendirilmiş ve en yüksek Cook's distance değerinin **0.555** olduğu saptanmıştır. Bu bulgu, tek bir gözlemin modeli baskın biçimde yönlendirdiğine dair güçlü bir kanıt olmadığını ve etkilerin sınırlı düzeyde ve dağınık olduğunu göstermektedir.

Standartlaştırılmış artıklar için çizilen kutu grafiği incelendiğinde artıkların dar bir aralıkta toplandığı ve ciddi aykırı gözlem bulunmadığı görülmüştür. Bu durum regresyon modelinin aykırı gözlemlerden etkilenmediğini göstermektedir.

Aykırı ve etkileyici gözlemlere ilişkin yapılan tüm değerlendirmeler birlikte ele alındığında, modeli ciddi biçimde bozan veya katsayı tahminlerini baskın şekilde etkileyen gözlemlerin bulunmadığı görülmüştür. Bu nedenle ilgili gözlemler analizden çıkarılmamış ve model sonuçlarının güvenilir olduğu kabul edilmiştir.

Sonuç olarak, lojistik regresyon modeline ilişkin temel varsayımların sağlandığı ve elde edilen bulguların istatistiksel olarak tutarlı olduğu değerlendirilmiştir. [Lojistik Regresyon Cook's Distance](#) (Bu kaynaktan yararlanılmıştır.)

6) Örneklem Büyüklüğünün Yeterliliği

```
table(model2$model$diyabet)
```

Çıktı;

```
> table(model2$model$diyabet)
```

0	1
327	3972

```
events <- sum(model2$model$diyabet == 1)
p <- 3
EPV <- events / p
EPV
```

Çıktı;

```
> EPV
[1] 1324
```

Örneklem büyüklüğünün lojistik regresyon analizi için yeterliliği, olay başına değişken sayısı (Events Per Variable, EPV) yaklaşımı kullanılarak değerlendirilmiştir. Modelde kullanılan veri setinde bağımlı değişkende ‘olay’ olarak tanımlanan diyabet=1 gözlem sayısı **3972** olup, modele dahil edilen üç sürekli bağımsız değişken (log(insulin), vücut kitle indeksi ve log(uyku süresi)) dikkate alındığında EPV değeri **1324** olarak hesaplanmıştır. Bu değer, literatürde önerilen asgari eşiklerin ($EPV \geq 10$) oldukça üzerinde olup, örneklemin lojistik regresyon modeli için fazlaıyla yeterli olduğunu göstermektedir. Yüksek EPV değeri, model katsayılarının aşırı uyum riskinin düşük olduğunu ve tahminlerin istikrarlı olduğunu desteklemektedir.

7) Mükemmel Ayrımın (Perfect Separation) Olmaması

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.604896  1.023561 -6.453  1.1e-10 ***
log(insulin) 1.892824  0.227729  8.312  < 2e-16 ***
vucut_kitle_ind 0.058908  0.006479  9.092  < 2e-16 ***
log(uyku_saati) -0.200647  0.121460 -1.652   0.0985 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

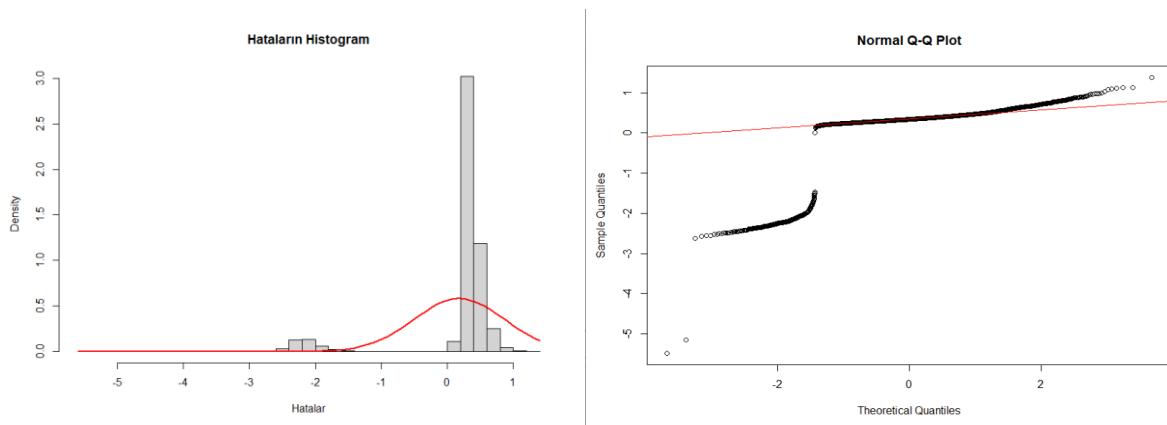
Null deviance: 2313.3 on 4298 degrees of freedom
Residual deviance: 2165.5 on 4295 degrees of freedom
(701 observations deleted due to missingness)
AIC: 2173.5

Number of Fisher Scoring iterations: 6
```

Lojistik regresyon analizinde mükemmel ayırım (perfect separation) varsayımları değerlendirilmiştir. Model tahmini sırasında herhangi bir yakınsamama uyarısı veya ‘fitted probabilities numerically 0 or 1 occurred’ şeklinde bir uyarı mesajı gözlenmemiştir. Ayrıca katsayı tahminlerinin sonlu olduğu ve standart hataların makul düzeylerde bulunduğu görülmüştür. Tahmin edilen olasılıkların 0 ile 1 arasında dağıldığı ve sınıflar arasında tam bir ayırım oluşmadığı belirlenmiştir. Bu bulgular doğrultusunda modelde mükemmel ayırım problemi bulunmadığı kabul edilmiştir.

Perfect separation durumunda katsayılar sonsuza gider ve model yakınsamaz; bizim modelimizde katsayılar sonlu, standart hatalar makul ve yakınsama sorunu olmadığı için perfect separation yoktur.

8) Normallik Varsayımlarının Aranmaması



Shapiro Wilk testine göre bakarsak aşağıdaki gibi olacaktır;

```
shapiro.test(residuals(model2))
```

Cıktı;

```
shapiro-wilk normality test

data: residuals(model2)
W = 0.46856, p-value < 2.2e-16
```

H_0 : Hata terimleri normal dağılmaktadır.

H_1 : Hata terimleri normal dağılmamaktadır.

$\alpha=0,10$, $p= <0,001$

$p < \alpha$ olduğundan H_0 reddedilir. %90 olasılıkla hata terimleri normal dağılmamaktadır.

Hata terimlerine ait histogram incelendiğinde, dağılımin **simetrik bir çan eğrisi formundan belirli ölçüde sapma gösterdiği** gözlenmektedir. Gözlemlerin büyük bir kısmının sıfıra yakın değerlerde yoğunlaşığı, buna karşılık dağılımin özellikle sol kuyrukta daha uzun bir yapı sergilediği anlaşılmaktadır. Bu durum, hata terimlerinde **hafif çarpıklık (skewness)** bulunduğuuna işaret etmektedir.

Bununla birlikte, lojistik regresyon modellerinde hata terimlerinin normal dağılım göstermesi bir varsayımlı olmadığından, bu gözlem modelin geçerliliğini doğrudan zedelememektedir.

Normal Q–Q grafiği incelendiğinde, hata terimlerinin **orta bölgede teorik normal dağılım çizgisine oldukça yakın seyrettiği**, ancak özellikle **uç değerlerde belirgin sapmaların** olduğu görülmektedir. Bu durum, hata dağılımının merkezde normale yakın, kuyruklarda ise normal dağılımdan uzaklaştığını göstermektedir.

Üç gözlemlerdeki bu sapmalar, veri setinde **aykırı veya etkileyici gözlemlerin** varlığına işaret edebilir. Ancak büyük örneklerde Q–Q grafiğinin üç bölgelerinde gözlenen bu tür sapmaların yaygın olduğu ve modelin yorumlanabilirliğini tek başına bozmadığı literatürde belirtilmektedir.

"Sample size affects the results of normality tests. When sample is small, normality tests tend to accept the null hypothesis. In large samples, even small deviations from the normal distribution cause the normality test to reject the null hypothesis." [DergiPark](#).

Dergipark'ta yayımlanan bir çalışmaya göre normalite testleri örneklem büyüklüğünden etkilenmektedir. Küçük örneklerde normalite testleri normal dağılımı kabul etme eğilimindeyken, büyük örneklerde **çok küçük sapmalar dahi testler tarafından anlamlı bulunur ve normalite reddedilir**. Örneklem büyüklüğünün artması, normalilik testlerinin sonuçlarını etkilemeye ve büyük örneklerde çok küçük sapmalar dahi normalilik hipotezinin reddedilmesine yol açabilmektedir. Bu nedenle büyük örneklerde normalliğin değerlendirilmesinde grafiksel yöntemlerin dikkate alınması önerilmektedir.

Model Anlamlılığı

Normal Anova'nın testi yapılrken varyans analizi için kullanılmaktadır. Fakat bağımlı değişkenin ikili (0-1) yapıda olduğu durumlarda Lojistik regresyon için Anova testi model anlamlılığı için kullanılmaktadır. Bu çalışmada bağımlı değişkenin ikili yapıda olması nedeniyle klasik ANOVA yerine lojistik regresyon modeli kullanılmış ve değişkenlerin modele katkıları olabilirlik oranı testleri aracılığı ile degenlendirilmiştir. (**İkili bağımlı değişken söz konusu olduğunda, klasik varyans analizi (ANOVA) varsayımları sağlanmadığından, model karşılaştırmaları olabilirlik oranı testleri (Likelihood Ratio Test) aracılığıyla gerçekleştirilmiştir [2].**)

H_0 : Model anlamlı değildir.

H_1 : Model anlamlıdır.

```
#anova  
anova(model2, test = "chisq")
```

Cıktı;

```
Model: binomial, link: logit  
Response: diyabet  
Terms added sequentially (first to last)  
  
          Df Deviance Resid. Df Resid. Dev Pr(>chi)  
NULL           4298    2313.3  
log(insulin)   1    70.236    4297    2243.1 < 2e-16 ***  
vucut_kitle_ind 1    74.590    4296    2168.5 < 2e-16 ***  
log(uyku_saati) 1     2.966    4295    2165.5  0.08503 .  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lojistik regresyon modellerinde parametre tahminleri maksimum olabilirlik (Maximum Likelihood) yöntemi ile elde edildiğinden, model uyumunun değerlendirilmesi klasik varyans analizi (ANOVA) yerine olabilirlik oranı testleri (Likelihood Ratio Test) aracılığıyla gerçekleştirilmektedir. Bu çalışmada model karşılaştırmaları R yazılımında `anova(model2, test = "Chisq")` komutu kullanılarak yapılmıştır. Bu yöntem, her bir değişkenin modele eklenmesiyle birlikte log-likelihood değerlerinde meydana gelen değişimi hesaplamakta ve bu farkın ki-kare (χ^2) dağılımına uyup uymadığını test etmektedir. Elde edilen ki-kare istatistiği, tam model ile indirgenmiş modeller arasındaki uyum farkının istatistiksel olarak anlamlı olup olmadığını göstermektedir. Analiz sonuçlarına göre `log(insulin)` ve `vücut kitle indeksi` değişkenlerinin modele eklenmesi model uyumunu yüksek düzeyde anlamlı biçimde artırmıştır ($p < 0.001$). Log(uyku süresi) değişkeni ise %10 anlamlılık düzeyinde modele katkı sağlamıştır ($p = 0.085$). Bu bulgular, kurulan lojistik regresyon modelinin boş modele kıyasla istatistiksel olarak anlamlı bir açıklayıcılığa sahip olduğunu göstermektedir.

YORUM

Kurulan lojistik regresyon modelinde, yöntemin gerektirdiği temel varsayımlar sistematik olarak değerlendirilmiştir. Bağımlı değişkenin ikili (binary) yapıda olması koşulu sağlanmıştır. Gözlemlerin birbirinden bağımsız olduğu veri yapısı doğrultusunda kabul edilmiştir.

Sürekli bağımsız değişkenler ile logit fonksiyonu arasındaki doğrusal ilişki varsayımları, gerekli tanısal incelemeler ve uygulanan dönüşümler sonrasında sağlanmıştır. Değişkenler arasında yüksek düzeyde çoklu doğrusal bağlantı (multicollinearity) bulunmadığı, varyans şışırma faktörü (VIF) değerleri üzerinden doğrulanmıştır.

Modele aşırı etki eden aykırı gözlemler Cook's Distance ve benzeri tanısal ölçütler yardımıyla incelenmiş, model sonuçlarını bozacak düzeyde etkili gözlemlere rastlanmamıştır. Örneklem büyülüüğünün, değişken başına olay sayısı (Events Per Variable – EPV) açısından yeterli olduğu değerlendirilmiştir.

Model uyumu Hosmer–Lemeshow testi ve genel uyum ölçütleri ile incelenmiş, elde edilen bulgular modelin veriye uygunluğunu desteklemiştir.

Model anlamlılığı Loglikelihood testi ile incelenmiş ve anlamlı olduğu sonucuna varılmıştır. Bu kapsamında lojistik regresyon analizine ilişkin tüm temel varsayımların sağlandığı sonucuna ulaşılmıştır.

KAYNAKÇA

1. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
2. Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.
3. Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
4. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill.
5. Knutson, K. L., Spiegel, K., Penev, P., & Van Cauter, E. (2007). The metabolic consequences of sleep deprivation. *Sleep Medicine Reviews*, 11(3), 163–178.
6. Hu, F. B., et al. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus. *New England Journal of Medicine*, 345(11), 790–797.
7. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), 673–690.
8. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15–18.
9. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression. *Journal of Clinical Epidemiology*, 49(12), 1373–1379.
10. DergiPark. (2024). *Türkiye akademik dergiler platformu*. <https://dergipark.org.tr>

VERİ

[Diyabet riski ve yaşam tarzı faktörleri](#)

<https://www.kaggle.com/datasets/miadul/diabetes-risk-and-lifestyle-factors>