Bilişimde İstatistiksel Analiz

Ders 12

Ali Mertcan KOSE Msc.
amertcankose@ticaret.edu.tr

İstanbul Ticaret Üniversitesi

Korelasyon Analizi

Korelasyon analizi ölçümle belirtilen iki(ya da daha çok) değişken arasında ilişki olup olmadığını, ilişkinin yönünü ve gücünü inceler.

Değişkenler arasındaki ilişkiler doğrusal olabilir, ya da doğrusal değildir. Değerler grafik üzerinde noktalanırsa ilişkinin yapısı kolay anlaşılır.

Korelasyon Analizi

İki değişken birbirini aynı yönde etkiliyorsa “pozitif ilişki”, ters yönde etkiliyorsa “negatif ilişki” olduğu söylenir. Etkileyen değişkene bağımsız değişken denir ve “x” ile sembolize edilir. Etkilenen değişkene bağımlı değişken denir ve “y” ile sembolize edilir.

Korelasyon Analizi

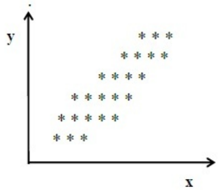
Korelasyon katsayısı “r” ile gösterilir. İlişkinin gücünü gösterir.İlişki pozitif ise işareti “+”, negatif ise işareti “-” olur. r yalnız **-1** ile **+1** arasında değer alabilir. r, her iki yönde **0**’a yaklaşırken ilişkinin gücü azalır. Her iki yönde 1’e yaklaşırken ilişkinin gücü artar.

Korelasyon Analizi

Korelasyon katsayısının (r) yorumu: - 0$<$ r $<$0.19 arasında ise çok zayıf ilişki ya da korelasyon yok - 0.20 $<$ r $<$0.39 arasında ise, zayıf korelasyon - 0.40$<$ r $<$0.59 arasında ise orta şiddette korelasyon - 0.60$<$r$<$0.79 arasında ise yüksek şiddette korelasyon - 0.80 $<$ r $<$ 1 arasında ise çok yüksek korelasyon olduğu yorumu yapılır

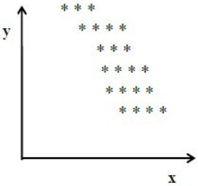
İki değişken arasında pozitif ya da negatif kuvvetli ilişki varsa korelasyon katsayısı 1'e yaklaşır. Eğer iki değişkene ilişkin noktalar bir doğru üzerinde ise, değişkenler arasında tam bir ilişkiden söz edilir.

Korelasyon Analizi



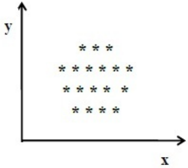
a)Pozitif Yüksek Korelasyon

$(1 > r > 0)$



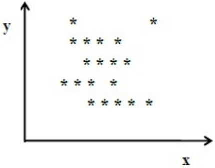
b) Negatif Yüksek Korelasyon

$(-1 < r < 0)$



c) Sıfır Korelasyon

$(r \cong 0)$



d) Sıfır Korelasyon

$(r \cong 0)$

Basit Korelasyon Katsayısının (r) Tahmini

Normal dağılım gösteren, x ve y gibi iki sürekli değişken arasındaki Pearson korelasyon katsayısı,

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum yi}{n}}{\sqrt{\left[\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right) \cdot \left(\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}\right)\right]}}$$

formülle hesaplanır.

Örnek:

12 kişinin ağırlık ve HDL kolesterol değerleri yine aşağıda verilmiştir. Bu özellikler (değişkenler) arasındaki ilişkinin yönü ve ölçüsü nedir? Korelasyon katsayısını hesaplayınız.

Korelasyon Analizi

i	Ağırlık	HDL
1	56	34
2	65	38
3	73	40
4	68	42
5	80	46
6	77	38
7	85	45
8	59	35
9	64	37
10	67	39
11	75	42
12	70	40

Korelasyon Analizi

Çözüm için önce formülde yer alan $x_i y_i$, ${x_i}^2$ *ve* ${y_i}^2$ değerleri hesaplanarak aşağıdaki tablo hazırlanır. Sonra tablodaki her bir sütunun toplamını en alt satıra yazılır. Bu tablo işlem kolaylığı sağlar.

Korelasyon Analizi

i	Ağırlık	HDL	x*y	x^2	x^2
1	56	34	1904	3136	1156
2	65	38	2450	4225	1444
3	73	40	2920	5329	1600
4	68	42	2856	4624	1764
5	80	46	3680	6400	2116
6	77	38	2926	5929	1444
7	85	45	3825	7225	2025
8	59	35	2065	3481	1225
9	64	37	2368	4096	1369
10	67	39	2613	4489	1521
11	75	42	3150	5625	1764
12	70	40	2899	4900	1600
Toplam	**839**	**476**	**33577**	**59459**	**19028**

Korelasyon Analizi

Bulunan bu değerler ilgili korelasyon formülünde yerine yazılarak gerekli hesaplama yapılır. Yani:

$$r = \frac{33577 - \frac{839 * 476}{12}}{\sqrt{\left[\left(59459 - \frac{(839)^2}{12}\right) \cdot \left(19028 - \frac{(476)^2}{12}\right)\right]}} \cong 0.867 = 0.87 \text{ olarak}$$

hesaplanır. Bu katsayı ağırlık ve HDL arasında pozitif ve yüksek bir ilişkinin olduğunu gösterir. Yani ağırlık arttıkça HDL değeri de artmaktadır. Ancak bu katsayının istatistiksel önemliliğinin kontrolü için hipotez testi yapılmalıdır.

Korelasyon Katsayısının Önem Testi

Hipotez testini yaparken aşağıdaki dört aşama uygulanır. Önce hipotezler kurulur. Hipotezler

$H_0 : \rho = 0$ $H_1 : \rho > 0$, $\rho < 0$ veya $\rho \neq 0$ şeklinde kurulur.

Sonra kritik tablo değeri belirlenir. Tablo değeri t tablosundan bulunur.

Tek yönlü test için $t_c = t_{\alpha;n-2}$

Çift yönlü test için $t_c = t_{\alpha/2;n-2}$

Üçüncü aşamada test istatistiği hesaplanır. Test istatistiğinin hesaplanmasında aşağıdaki formül kullanılır.

Korelasyon Analizi

$t_h = \frac{r-\rho}{S_r}$; $S_r = \sqrt{\frac{1-r^2}{n-2}}$

Son aşamada karşılaştırma yapılır. Karşılaştırma işleminde,

$t_h < t_c$ ise H_0 hipotezi kabul edilir,

$t_h > t_c$ ise H_0 hipotezi reddedilir.

veriler için hesaplanan r = 0.87 korelasyon katsayısının önem testini yapınız. Bu katsayının 0'dan farklı olup olmadığını %1 hata ile test ediniz.

Hipotezler; $H_0 : \rho = 0$ ve $H_1 : \rho \neq 0$ şeklindedir.

H_1 hipotezi çift yönlü ve α=0.01 olduğundan kritik tablo değeri; $t_c = t_{0.01/2;12-2} = 3.169$'dur.

Bu aşamada önce r'nin standart hatası S_r hesaplanır. $S_r = \sqrt{\frac{1-0.87^2}{12-2}} = 0.156$

Test istatistiği;

$t_h = \frac{0.87-0}{0.156} = 5.577$ olarak hesaplanır.

Son aşama olarak karşılaştırma yapılır. Bu örnek için, $t_h = 5.577$ ve $t_c = 3.169$ olup test istatistiği t_h, tablo değerinden t_c büyük olduğu için H0 hipotezi reddedilir. H_1 hipotezi kabul edilir. Yani ağırlık ve HDL kolesterol arasında önemli ve pozitif bir ilişkinin olduğuna karar verilir.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Regresyon analizi iki değişken arasındaki sayısal ilişkiyi inceler. Başka bir deyişle, **"x"** değişkeni belirli bir birim değiştiğinde, **"y"** değişkeninin nasıl bir değişime uğradığını inceler. İlişki bir denklemle ifade edilir.

Bir bağımlı, bir bağımsız değişken olması durumunda doğrusal regresyon denklemi;

$$Y_i = \alpha + \beta\ X_i + \epsilon_i$$

a ve b regresyon modelinin bilinmeyenleridir. a, regresyon doğrusunun y eksenini kestiği noktayı gösterir iken, b ise regresyon katsayısı olarak adlandırılır.

Ve bağımsız değişkende bir birimlik değişme (artma yada azalma) olduğunda bağımlı değişkende meydana gelecek değişiklik miktarını verir.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Basit doğrusal regresyon modelinde aşağıdaki varsayımlar geçerlidir:

- Bağımsız değişken X sebep değişkenidir. Yani araştırıcı tarafından peşin hükümlü olarak alınabilir. Bu durumda X şans değişkeni olmayabilir. Fakat X şansa bağlı değişken olsa da regresyon analizi yapılabilir.
- Bağımlı ve bağımsız değişkenler en az hata ile ölçülmelidir. Ölçümlerde hata olabileceğinden en az hatalı ölçüm yapmaya özen gösterilmelidir.
- Bağımsız değişkenlerin her bir değeri için bağımlı değişkenlerin bir alt populasyonu vardır. Hipotez testlerinin ve tahminlerin sağlam yapılabilmesi için alt populasyonların normal dağılıma uygunluk göstermesi gerekir. Yani Y değerleri normal dağılıma uygun olmalıdır.
- Bağımlı değişken Y'nin alt populasyonlarının varyansları eşit ve ortalamaları doğrusal olmalıdır.

Regresyon Analizi İçin Parametrelerin Tahmini

En küçük kareler tahmin metodu

En küçük kareler metoduna göre regresyon parametrelerinin tahmini, doğruyu; noktaların kendisinden ayrılışlarının kareleri toplamını minimum yapacak şekilde tayin etme esasına dayanmaktadır. Yani hataların karelerine ilişkin toplamın en küçük olması temeline dayanır. Populasyonda (X_i, Y_i) gözlemlerine ait doğrusal regresyon modeli aşağıdaki gibi yazılır:

$Y_i = \alpha + \beta\ X_i + \epsilon_i$, i=1,2,. . . ,n

Bu denklemde;

Y_i: i. gözlemin cevap değeri veya bağımlı değişkeni,

X_i: i. gözlemdeki bağımsız değişkenin değerini,

β: Regresyon doğrusunun eğimini,

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

En küçük kareler tahmincisi $\hat{\beta}$,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

şeklinde bulunur. Burada; $\overline{Y}$ ve $\overline{X}$, X ve Y'nin ortalamalarıdır. Bu ortalamalar aritmetik ortalama hesaplama yaklaşımı ile basit olarak aşağıdaki gibi hesaplanır:

$$\overline{Y} = \frac{\sum Y_i}{n}$$

$$\overline{X} = \frac{\sum X_i}{n}$$

Regresyon doğrusunun Y eksenini kestiği noktanın en küçük kareler tahmincisi $\hat{\alpha}$ ise

$\hat{\alpha} = \overline{Y}$ - $\beta\overline{X}$ şeklinde bulunur.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Basit doğrusal regresyonda amaç, bağımsız değişken X ile bağımlı değişken Y arasındaki ilişkiyi doğru bir şekilde belirleyecek eşitliği tespit etmektir. Matematik eşitlik belirlenirken populasyonun tüm elemanları kullanılmaz; çünkü bir populasyonu tümüyle ele alıp incelemek hem zor hem de masraflıdır. Bunun için eşitlik belirlenirken populasyondan şansa bağlı örnek çekilir ve bu örnek üzerinden populasyon için eşitlik tahmin edilir. Örnekte bağımsız değişken X ile bağımlı değişken Y arasındaki basit regresyon eşitliği,

y_i= a+ b x_i + e_i olarak ifade edilir.

İstatistik modeldeki a ve b, parametrik eşitlikte sırasıyla α ve β'nın tahminleridir. Regresyon modelinde,

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

y_i: Bağımlı değişkeni (sonuç değişkenini),

x_i: Bağımsız değişkeni (sebep değişkeni),

a: Doğrunun y eksenini kestiği değer olarak başlangıç değerini,

b: Regresyon katsayısını veya doğrunun eğimini, diğer bir ifade ile de, X'deki 1 birim değişmeye karşılık y'deki değişim miktarını,

e_i: Şansa bağlı hata değerini göstermektedir.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Örnek için hesaplanan a ve b değerleri α ve β'nın sapmasız tahmincileridir.

Örnekler için hesaplanan a değerlerinin ortalaması α'ya ve b değerlerinin ortalaması da β'ya eşittir. Buna göre

$\mu_a = \alpha$

$\sigma_a^2 = \frac{\sigma_{YX}^2 \sum X_i^2}{n \sum (X_i - \overline{X})^2}$

$\mu_b = \beta$

$\sigma_b^2 = \frac{\sigma_{YX}^2}{\sum (X_i - \overline{X})^2}$

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Eşitliklerde σ^2_{YX}, hata kareler ortalamasıdır. a ve b aşağıdaki formüllere göre hesaplanır:

Regresyon denklemindeki b ve a'nın tahmini aşağıdaki formüller ile yapılır.

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

formüldeki a ise, a $= \overline{y}$ - b$\overline{x}$ eşitliği ile tahmin edilir.

Bu değerler y = a + bx denkleminde yerlerine yazılarak doğrusal regresyon denklemi kurulmuş olur. Bu eşitlik (denklem) bir doğrunun denklemi olup dik koordinat sisteminde aşağıdaki gibi gösterilir.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

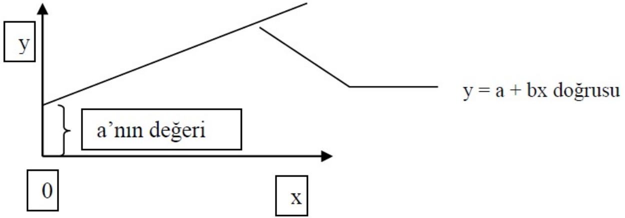


Figure 2: Basit Doğrusal Regresyonun Şekilsel Gösterimi

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Regresyon analizinde y'nin x'e regresyonu byx ile x'in y'ye regresyonu bxy eşit değildir. Farklı katsayılar ve eşitlikler üretirler. En küçük kareler metoduna göre elde edilen regresyon doğrusunun özellikleri aşağıdaki gibidir.

1- Hataların toplamı sıfırdır.

$e_i = Y_i - (a + bX_i$

$\sum_{i=1}^{n} e_i = 0$

2- Hataların kareleri toplamı minimumdur.

$\sum_{i=1}^{n} e_i^2 = minimum$

3- Bağımlı değişkenlerin (Y_i) toplamı elde edilen regresyon modeline göre tahmin edilen değerlerin $(\hat{Y}_i)$ toplamına eşittir.

$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

4- i. değerdeki hatanın ağırlığı i. değerdeki bağımsız değişkene göre alındığında hataların toplamı veya hata değerleri ile bağımsız değerlerin çarpımları toplamı sıfırdır.

$\sum_{i=1}^{n} X_i e_i = 0$

5- Hata değerleri ile elde edilen modele göre tahmin edilen değerlerin çarpımlarının toplamı sıfırdır.

$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

5- Regresyon doğrusu ($\overline{X}, \overline{Y}$) noktasından geçer.

Regresyon katsayısının önem testi

β 'nın önemlilik testi yapılırken sıfır hipotezi H0 ve karşıt (alternatif) hipotez H_1

H_0: $\beta = 0$ H_1: $\beta \neq 0$ veya H_1: $\beta > 0$ veya H_1: $\beta < 0$

şeklinde kurulur. Karşıt hipotezin H_1: $\beta \neq 0$ şeklinde ifade edilmesi durumunda test iki yönlü, H_1: $\beta > 0$ ve H_1: $\beta < 0$ şeklinde ifade edilmesi durumunda test tek yönlüdür.

Regresyon katsayısı β'nın hipotez testinde (n-2) serbestlik dereceli t dağılımı kullanılır ve test istatistiği aşağıdaki formüle göre hesaplanır.

$$t_h = \frac{b-\beta}{S_b}$$

$$S_b = \sqrt{\frac{S_{XY}^2}{\sum(X_i-\overline{X})^2}} = \sqrt{\frac{S_{XY}^2}{\sum X_i^2-(\sum X_i)^2/n}}$$

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Populasyon hata kareler ortalaması olan σ^2_{YX}'nin tahmincisi S^2_{YX} ise aşağıdaki formüle göre hesaplanır.

$$\text{HKO} = S^2_{YX} = \frac{\sum(Y_i - \hat{Y})^2}{n-2} = \frac{n-1}{n-2}(S^2_Y - b^2 S^2_X)$$

$$= \frac{1}{n-2}\left[\sum(Y_i - \overline{Y})^2 - \frac{(\sum(X_i-\overline{X})(Y_i-\overline{Y}))^2}{\sum(X_i-\overline{X})^2}\right]$$

Kritik cetvel değeri (t_c) alternatif hipotez tek yönlü ise $t_{\alpha(n-2)}$, iki yönlü ise $t_{\alpha/2(n-2)}$ dir. İstatistiksel karar verilirken; $t_h > t_c$ ise H_0 hipotezi red, H_1 kabul edilir ve regresyon katsayısının önemli olduğuna karar verilir.

$t_h < t_c$ ise H_0 hipotezi kabul, H_1 reddedilir ve regresyon katsayısının önemsiz olduğuna karar verilir.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

Regresyon doğrusunun Y eksenini kestiği nokta olan a değerinin önemlilik testi yapılırken aşağıdaki eşitlik kullanılır.

$t_h = \frac{\alpha}{S_\alpha}$

S_a $= \sqrt{HKO(\frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2})} = \sqrt{HKO\frac{1}{}[\frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2}]}$

Örnek:

On iki kişinin ağırlık ve HDL kolesterol değerleri aşağıdaki gibi verilmiştir. HDL kolesterol tipinin ağırlığın bir sonucu olduğu varsayılarak bu iki değişken arasındaki fonksiyonel ilişkiyi ifade eden eşitliği kurunuz ve regresyon katsayısının önemliliğini test ediniz. α=0.01.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

i	Ağırlık	HDL
1	56	34
2	65	38
3	73	40
4	68	42
5	80	46
6	77	38
7	85	45
8	59	35
9	64	37
10	67	39
11	75	42
12	70	40

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

	Ağırlık (x)	HDL (y)	x*y	x^2	y^2
1	56	34	1904	3136	1156
2	65	38	2470	4225	1444
3	73	40	2920	5329	1600
4	68	42	2856	4624	1764
5	80	46	3680	6400	2116
6	77	38	2926	5929	1444
7	85	45	3825	7225	2025
8	59	35	2065	3481	1225
9	64	37	2368	4096	1369
10	67	39	2613	4489	1521
11	75	42	3150	5625	1764
12	70	40	2800	4900	1600
toplam	839	476	33577	59459	19028
ortalama	69.92	39.67			

Bulunan bu değerler regresyon katsayısı (b) formülünde yerine yazarak hesaplama yapılırsa,

b= $\frac{33577-\frac{839*476}{12}}{59459-\frac{839^2}{12}} = 0.376$ olarak hesaplanır. Buradan a değeri, a=$\overline{y}-b\overline{x}$ eşitliğinden a=39.67-0.376*69.92=13.4 olarak bulunur. Tahmin edilen a ve b değerleri ($\hat{a}$, $\hat{b}$) regresyon eşitliğinde yerlerine yazıldığında

Tahmin edilen b regresyon katsayısı kullanılarak β'nın hipotez testi için

$H_0 : \beta = 0$ ve

$H_1 : \beta \neq 0$ (İki yönlü, çünkü 0'dan farklı olup olmadığı sorulmuştur.)

t tablosunda iki yönlü testte, $\alpha = 0.01$ önem düzeyinde (test iki yönlü olduğu için $\alpha/2 = 0.005$'tir.) tablo değeri $t_c = t_{0.005(12-2)} = 3.169$ olarak okunur.

Regresyon katsayısının standart hatası S_b, ilgili formüller yardımıyla hesaplandığında $S_b \cong 0.068$ olarak bulunur. Buradan b=0.376 ve $S_b = 0.068$ t formülünde yerlerine yazıldığında

$t_h = \frac{0.376}{0.068} = 5.529$ olarak bulunur.

$t_h > t_c$ yani $5.529 > 3.169$ olduğundan H_0 reddedilir. Sonuç olarak bu çalışmada "Ağırlıktaki her bir kg artışa bağılı olarak HDL değerindeki 0.376 birimlik artış, %1 hata ile istatistiksel olarak çok anlamlıdır" kararına varılır.

Regresyon denkleminin uyum (belirlenme) katsayısının tahmini

Regresyon denkleminin uygunluğunu ifade eden belirleme katsayısı R^2 ile sembolize edilir ve aşağıdaki formüller kullanılarak hesaplanır:

$$R^2 = \frac{\frac{[\sum(X_i-\overline{X})(Y_i-\overline{Y})]^2}{\sum(X_i-\overline{X})^2}}{\sum(Y_i-\overline{Y})^2} = \frac{[\sum(X_i-\overline{X})(Y_i-\overline{Y})]^2}{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}$$

Kareler toplamları yaklaşımı ile

$$R^2 = \frac{RKT}{GKT} = \frac{\sum(\hat{Y}_i-\overline{Y})^2}{\sum(Y_i-\overline{Y})^2}$$

formülü ile hesaplanır.

Belirleme katsayısı R_2 0 ile 1 arasında $(0 \leq \mathrm{R2} \leq 1)$ değişir. Belirleme katsayısının 1'e yaklaşması doğrunun uyumunun iyi olduğunu, 0'a yakın olması ise uyumun olmadığını gösterir.

BASİT DOĞRUSAL REGRESYON ÇÖZÜMLEMESİ

i	Ağırlık (x)	HDL (y)	$\sum(X_i - \overline{X})$	$\sum(Y_i - \overline{Y})$
1	56	34	-13,9167	-5,66667
2	65	38	-4,91667	-1,66667
3	73	40	3,083333	0,333333
4	68	42	-1,91667	2,333333
5	80	46	10,08333	6,333333
6	77	38	7,083333	-1,66667
7	85	45	15,08333	5,333333
8	59	35	-10,9167	-4,66667
9	64	37	-5,91667	-2,66667
10	67	39	-2,91667	-0,66667
11	75	42	5,083333	2,333333
12	70	40	0,083333	0,333333

soruya ilişkin belirleme katsayısı:

$=[\sum(X_i - \overline{X})(Y_i - \overline{Y})]^2 = 88011.11$

$=\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2 = 117174.4$

$=\frac{[\sum(X_i-\overline{X})(Y_i-\overline{Y})]^2}{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2} = \frac{88011.11}{117174.4} = 0.75$

ÖDEV

Aynı yaşta 8 kişinin boy ve ağırlıklarına ait değerler aşağıdaki tabloda verilmiştir (Rakamlar gerçek değildir, örnek için türetilmiştir.)

- a) Ağırlığınn boya göre ilişkisini ifade eden regresyon denklemini bulunuz.
- b) Regresyon katsayısı b ve başlangıç değeri a' nın önemlilik testlerini yapınız.
- c) Boyu 170 olan kişinin ağırlığını tahmin ediniz.
- d) Ağırlığı 75 olan kişinin boyunu tahmin ediniz.

ÖDEV

i	Boy	Ağırlık
1	185	84
2	180	100
3	174	70
4	161	60
5	184	80
6	175	70
7	168	72
8	173	65