IST3011 Regresyon Analizi 2020-2021 Güz Dönemi-7. Hafta (26.11.2020)

Örnek : Bir hastalık için uygulanan uygulaması zor, pahalı ve zaman alıcı olan standart bir test yerine yeni bir test geliştiriliyor ve elde edilen yeni test sonuçlarından standart test sonuçlarını bir denklem yardımıyla elde edilmek isteniyor. Bir hasta grubuna grubuna her iki test uygulandıktan sonra elde edilen yeni test sonuçları (x) ve standart test sonuçları (y) aşağıda verilmiştir.

Gözlem	Standart test sonuçları (y)	Yeni test sonuçları (x)
1	49	40
2	51	45
3	61	50
4	62	55
5	71	60
6	71	65
7	80	70
8	76	75
9	90	80
10	102	85
11	98	90
12	100	95
13	112	100

(**Kaynak:** Örnek 7.1, Uygulamalı Çok Değişkenli İstatistiksel Yötntemlere Giriş 1, Reha Alpar, Nobel Yayınevi)

- Bu verileri kullanarak x ve y değişkenleri arasındaki doğrusal ilişkiyi araştırmak için $y = \beta_0 + \beta_1 x + \epsilon$ biçiminde basit doğrusal regresyon modelini oluşturalım.
- İlk olarak verimizden gerekli olan hesaplamaları yapalım.

$$\sum_{i=1}^{13} x_i = 910, \quad \overline{x} = 70, \quad \sum_{i=1}^{13} x_i^2 = 68250,$$

$$\sum_{i=1}^{13} y_i = 1023, \quad \overline{y} = 78.69231, \quad \sum_{i=1}^{13} y_i^2 = 85477, \quad \sum_{i=1}^{13} x_i y_i = 76280.$$

Böylece

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{13}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{13} x_i y_i - n\ \overline{x}\ \overline{y} = 4670, \\ S_{xx} &= \sum_{i=1}^{13}(x_i - \overline{x})^2 = \sum_{i=1}^{13} x_i^2 - n\overline{x}^2 = 4550 \end{aligned}$$

- Regresyon katsayıları için tahmin ediciler

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{4670}{4550} = 1.026374 \quad \text{ve} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 6.846154$$

elde edilir. Böylece tahmin edilen doğrusal regresyon denklemi

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = 6.846154 + 1.026374\ x_i,\ i = 1, ..., 13 \tag{1}$$

olarak elde edilir.

Yorum: Bu regresyon denklemine göre x bağımsız değişkeninde bir birimlik artış olduğunda y bağımlı değişkeninde 1.026374 birimlik artış olmaktadır.

- Şimdi bulduğumuz (1) denklemini kullanarak varyans analizi tablosunu oluşturmak için gerekli hesaplamaları yapalım.

Gözlem	x_i	y_i	$\widehat{y}_i$	$e_i = y_i - \widehat{y}_i$	e_i^2	$(\widehat{y}_i - \overline{y})^2$	$(y_i - \overline{y})^2$
1	40	49	47.90110	1.098901	1.207584	948.09854	881.633136
2	45	51	53.03297	-2.032967	4.132955	658.40176	766.863905
3	50	61	58.16484	2.835165	8.038160	421.37713	313.017751
4	55	62	63.29670	-1.296703	1.681439	237.02463	278.633136
5	60	71	68.42857	2.571429	6.612245	105.34428	59.171598
6	65	71	73.56044	-2.560440	6.555851	26.33607	59.171598
7	70	80	78.69231	1.307692	1.710059	0.00000	1.710059
8	75	76	83.82418	-7.824176	61.21772	26.33607	7.248521
9	80	90	88.95604	1.043956	1.089844	105.34428	127.863905
10	85	102	94.08791	7.912088	62.601135	237.02463	127.863905
11	90	98	99.21978	-1.219780	1.487864	421.37713	543.248521
12	95	100	104.35165	-4.351648	18.936843	658.40176	372.786982
13	100	112	109.48352	2.516484	6.332689	948.09854	454.017751
Toplam	910	1023	1023	0	**181.6044**	**4793.165**	**4974.769**

Bu tabloya göre,

$$\underbrace{\sum_{i=1}^{13}(y_i - \overline{y})^2}_{SST=4974.769} = \underbrace{\sum_{i=1}^{13}(y_i - \widehat{y}_i)^2}_{SSE=181.6044} + \underbrace{\sum_{i=1}^{13}(\widehat{y}_i - \overline{y})^2}_{SSR=4793.165}$$

elde edilir.

- σ^2 için yansız tahmin edici

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{SSE}{11} = 16.50949$$

elde edilir. Böylece regresyonun standart hatası $\sqrt{\widehat{\sigma}^2} = 4.063187$ bulunur.

- $\widehat{\beta}_0$ ve $\widehat{\beta}_1$ için **standart hataları** bulalım.

$$se(\widehat{\beta}_0) = \sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} = \sqrt{\widehat{\sigma}^2\left(\frac{1}{13} + 1.076923\right)} = 4.364563$$

ve

$$se(\widehat{\beta}_1) = \sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}} = 0.06023669.$$

- $\alpha = 0.05$ anlamlılık düzeyinde **güven aralığı ve hipotez testlerini** oluşturalım. Kullanacak olduğumuz tablo değerleri: $t_{11,0.025} = 2.201$, $\chi^2_{11,0.025} = 21.92$, $\chi^2_{11,1-0.025} = 3.82$, $F_{1,11,0.05} = 4.48$ biçimindedir.

- β_0 ve β_1 için %95 güven aralıkları

$$\begin{aligned} \widehat{\beta}_0 - se(\widehat{\beta}_0)t_{11,0.025} &\leq \beta_0 \leq \widehat{\beta}_0 + se(\widehat{\beta}_0)t_{11,0.025} \\ -2.760249 &\leq \beta_0 \leq 16.45256 \end{aligned}$$

  ve

$$\begin{aligned} \widehat{\beta}_1 - se(\widehat{\beta}_1)t_{11,0.025} &\leq \beta_1 \leq \widehat{\beta}_1 + se(\widehat{\beta}_1)t_{11,0.025} \\ 0.8937927 &\leq \beta_1 \leq 1.158955 \end{aligned}$$

  olarak bulunur.

  Yorum: Anakütleden aynı x değerleriyle aynı büyüklükte 100 örneklem alırsak ve herbiri için yukarıdaki gibi β_0 için %95 güvenle oluşturulan aralıkların 95 tanesi gerçek β_0 değerini içerir.

  Anakütleden aynı x değerleriyle aynı büyüklükte 100 örneklem alırsak ve herbiri için yukarıdaki gibi β_1 için %95 güvenle oluşturulan aralıkların 95 tanesi gerçek β_1 değerini içerir.

  Yani, yukarıdaki prosedürü takip ederek elde edilen güven aralıkların %95'i β_0 ve β_1 in gerçek değerlerini içerir.

- σ^2 için %95 güven aralığı

$$\begin{aligned} \frac{(n-2)\,\widehat{\sigma}^2}{\chi^2_{n-2,\alpha/2}} &\leq \sigma^2 \leq \frac{(n-2)\,\widehat{\sigma}^2}{\chi^2_{n-2,1-\alpha/2}} \\ 8.284872 &\leq \sigma^2 \leq 47.54042 \end{aligned}$$

  elde edilir.

- **Varyans analizi (Analysis of Variance, ANOVA)** tablosu aşağıdaki gibi olur.

Değişim kaynağı	Kareler toplamı	Serbestlik derecesi	Kareler ortalaması	F_0 test değeri
Regresyon	4793.165	1	4793.165	$\frac{SSR/1}{SSE/11} = 290.3279$
Artık	181.6044	11	16.50949	
Toplam	4974.769	12		

- $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ hipotezlerini yani r**egresyonun anlamlığını** $\alpha = 0.05$ anlamlılık düzeyinde ANOVA tablosunu ve t testini kullanarak test edelim.

  ANOVA tablosuna göre F_0 istatistiği değeri $F_0 = 290.3279 > F_{1,11,0.05} = 4.48$ olduğundan $H_0 : \beta_1 = 0$ **hipotezi red edilir** yani $\beta_1 \neq 0$ elde edilir. Böylece, bağımlı değişken y ile bağımsız değişken x arasında önerdiğimiz $y = \beta_0 + \beta_1 x + \epsilon$ doğrusal ilişkisi anlamlıdır.

  t **testi kullanarak** $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ r**egresyonun anlamlığını** $\alpha = 0.05$ anlamlılık düzeyinde test edelim.

$$t_0 = \frac{\widehat{\beta}_1 - \beta_1}{se(\widehat{\beta}_1)},\ se(\widehat{\beta}_1) = \sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}$$

test istatistiğinin $H_0 : \beta_1 = 0$ hipotezi altındaki değeri $t_0 = 1.026374/0.06023669 = 17.03902$ bulunur. $t_0 = 17.03902 > t_{11,0.025} = 2.201$ olduğundan $H_0 : \beta_1 = 0$ **hipotezi red edilir**.

Not: $t_0^2 = (17.03902)^2 = 290.3279 = F_0$.

- $H_0 : \beta_0 = 0$, $H_0 : \beta_0 \neq 0$ hipotezlerini $\alpha = 0.05$ anlamlılık düzeyinde test edelim.

$$t_0 = \frac{\widehat{\beta}_0 - \beta_0}{se(\widehat{\beta}_0)},\ se(\widehat{\beta}_0) = \sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$$

test istatistiğinin $H_0 : \beta_0 = 0$, hipotezi altındaki değeri $t_0 = 6.846154/4.364563 = 1.568577$ bulunur. $t_0 = 1.568577 < t_{11,0.025} = 2.201$ olduğundan $H_0 : \beta_0 = 0$ **hipotezi red edilemez** yani $H_0 : \beta_0 = 0$ hipoetezi kabul edilir.

- Oluşturduğumuz model için belirtme (belirlilik) katsayısı

$$R^2 = \frac{SSR}{SST} = 0.9634949$$

olarak bulunur.

Yorum: y bağımlı değişkeni yeni test sonuçlarındaki değişimin %96.3 x bağımsız değişkeni yani eski test sonuçları ile açıklanabilmektedir.

- Oluşturduğumuz regresyon modelini kullanarak yeni test puanı $x_0 = 70$ olan bir hastanın standart test puanlarının ortalaması için %95 güven aralığını bulalım. Ortalama yanıt için %95 güven aralığı

$$\widehat{y} - t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)} \leq E(y|x_0) \leq \widehat{y} + t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)}$$

olduğundan $x_0 = 70$ için $\widehat{y} = 6.846154+1.026374\,(70) = 78.692334$, $\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)} = 1.1269$ ve

$$76.2120 \leq E(y|x_0) \leq 81.1726$$

olarak elde edilir.

- Yeni test puanı $x_0 = 70$ olan bir hastanın standart test puanı y_0 **gelecek gözlemi** için %95 tahmin aralığını bulalım.

$$\widehat{y}_0 - t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)} \leq y_0 \leq \widehat{y}_0 + t_{n-2,\alpha/2}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)}$$

olduğundan $x_0 = 70$ için $\widehat{y}_0 = 78.692334$, $\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)} = 4.2165$ ve

$$69.4118 \leq y_0 \leq 89.9728$$

olarak elde edilir.

- **Ödev 1:** Yukarıda oluşturduğumuz modelde $H_0 : \beta_0 = 0$ hipotezi kabul edilmiştir.

  a) Bu veri için kesim noktasız (orijinden geçen) regresyon modelini oluşturunuz.

  b) ANOVA tablosunu oluşturarak modelin anlamlılığını $\alpha = 0.05$ anlamlılık düzeyinde test ediniz.

  c) Model için belirlilik katsayını bulunuz ve yorumlayınız.

  d) Kesim noktalı ve kesim noktasız modelleri karşılaştırınız. Hangi modeli tercih ederiz, açıklayınız.

- **Ödev 2:** 25-30 yaş grubundan rastgele seçilmiş 26 erkeğe ilşikin kilo (weight) ve sistolik kan basıncı (systolic blood pressure) verileri aşağıda verilmiştir. ($\alpha = 0.05$ anlamlılık düzeyini kullanınız)

  a) Sistolik kan basıncını kilo ile ilişkilendiren bir regresyon modeli oluşturunuz.

  b) Bu model için ANOVA tablosunu oluşturarak modelin anlamlılığını test ediniz.

  c) $H_0 : \beta_0 = 0$, $H_1 : \beta_0 \neq 0$ hipotezini test ediniz.

  d) Alternatif olarak kesim noktasız modeli de oluşturunuz ve bu iki modeli karşılaştırınız.

Subject	Weight	Symbolic BP	Subject	Weight	Systolic BP
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159