XI. BÖLÜM

REGRESYON ÇÖZÜMLEMESİ

XI.1 Giriş

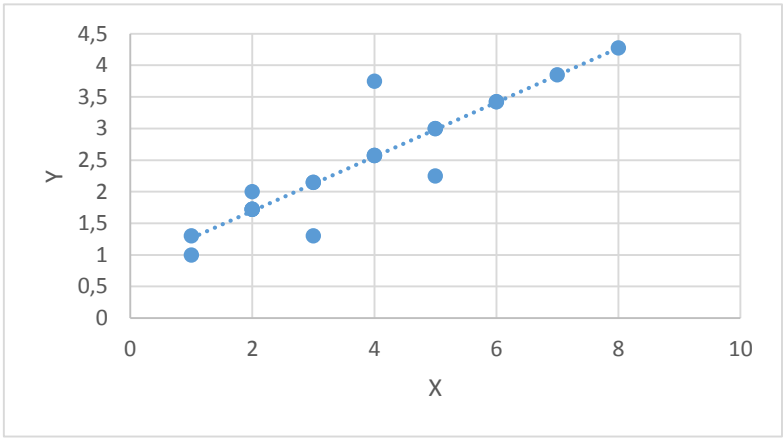
Herhangi iki oranlama ve ya eşit aralıklı değişken arasındaki ilişkinin yönünü ve derecesini belirten korelasyon katsayısı bir önceki bölümde verilmişti. Hatırlanacağı gibi X ve Y değişkenleri arasında korelasyon katsayısı hesaplanırken değişkenler bağımlı değişken ya da bağımsız değişken olarak bir ayırıma tabi tutulmamışlardı. Regresyon çözümlemesinde ise değişkenlerin bağımlı ve bağımsız değişken ya da değişkenler olarak iki gruba ayrılması bir ihtiyaçtır. Bağımlı değişken bağımsız değişken(ler) tarafından açıklanmaya çalışılan değişkendir ve tektir. Regresyonda bağımlı değişken Y ve bağımsız (değişken(ler) de X ile gösterilir.

Regresyon çözümlemesinde amaçlardan biri bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkinin ortaya çıkartılmasıdır. Örneğin X ile Y arasında $Y = \beta_0 + \beta_1 X + \varepsilon$ gibi doğrusal bir ilişki düşünülüyor ise modelin bilinmeyen parametreleri β_0 ve β_1'in edilmesidir. Modelin bilinmeyen parametreleri tahmin edildiğinde bağımsız değişken(ler)in farklı değerleri için bağımlı değişkenin alacağı değerleri tahmin etmek ise regresyon çözümlemesinin diğer bir amacıdır.

XI.2 Uygun Modelin Seçimi

Regresyon çözümlemesinde bir bağımlı değişken ve bir ya da daha fazla sayıda bağımsız değişken vardır. Bir bağımlı ve bir bağımsız değişkenli doğrusal regresyona basit doğrusal regresyon, bir bağımlı ve birden fazla bağımsız değişkenli doğrusal regresyona ise çoklu regresyon adı verilir. Örneğin $Y = \beta_0 + \beta_1 X + \varepsilon$ modeli basit doğrusal, $Y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \varepsilon$ modeli ise çoklu doğrusal regresyon modelidir.

n tane birimin her birinden bağımlı değişken Y ve bağımsız değişken X değerleri saptanmış olsun. Bu durumda n tane gözlem geğeri $(y_1, x_1), \cdots, (y_n, x_n)$ sıralı ikililerden oluşmaktadır. Y ile X arasındaki modeli belirlemek için görsel olarak sıralı ikililerin koordinat düzlemde serpiliş görüntüsüne bakılmalıdır. Bu tür gösterime serpme diyagramı adı verilir. Örneğin Y ile X arasındaki serpme diyagramı, Şekil-XI.1'de verildiği gibi olsun.



Şekil-XI.1 Y ile X arasındaki serpme diyagramı.

Şekil-XI.1'deki gibi noktalar kümesi yaklaşık bir doğru gösteriyorsa seçilecek regresyon modeli, $Y = \beta_0 + \beta_1 X + \varepsilon$ olacaktır.

Basit doğrusal regresyon modelinde bilinmeyen β_0 ve β_1 parametrelerinin hesaplanması için kitledeki $(y_i, x_i), i = 1, \cdots, N$ sıralı ikililerinin hepsi elde edilmiş olmalıdır. Ancak bu genelde mümkün değildir. Bu nedenle kitleden n adet tesadüfî örneklem seçilerek $(y_1, x_1), \cdots, (y_n, x_n)$ değerleri tespit edilerek modelin parametreleri tahmin edilir. Parametreleri tahmin edilen model, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ şeklinde yazılabilir. $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistikleri β_0 ve β_1 parametrelerinin tahminleridir.

XI.3 Parametrelerin Tahmini

Uygulamada $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistiklerinin hesaplanmasında farklı yöntemler kullanılmaktadır. Burada sadece en küçük kareker (EKK) yöntemi olarak bilinen yöntem verilecektir. $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ ifadesini en küçük yapacak $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistikleri EKK tahmin ediciler olarak bilinir. $\hat{\beta}_0$ ve $\hat{\beta}_1$'in EKK tahmin edicileri $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ ve $\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$ eşitlikleri ile verilir.

♦ **Örnek:** Sekiz birimlik bir örnekleme ilişkin dekara verim miktarı (Y) ve dekara sepilen gübre miktarı (X) değerleri takipteki tabloda verildiği gibi tespit edilmiş olsun. Y ile X değerleri arasında doğrusal bir model düşünülmektedir. Modeli denklemini tahmin ediniz.

Gözlem No	Verim (y, 10 Kg)	Gübre (x; 10 Kg)	xy	x^2
1	15	8	120	64
2	18	12	216	144
3	25	14	350	196
4	35	16	560	256
5	48	22	1056	484
6	55	28	1540	784
7	60	35	2100	1225
8	64	41	2624	1681
Toplam	320	176	8566	4834
Ortalama	40	22		

♦ **Çözüm:** Tahmin edilmek istenen denklem; $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ olup tahmin değerleri, $\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{8566 - 8(22)40}{4834 - 8(22)^2} = \frac{1526}{962} \cong 1.5863$ ve $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \cong 40 - 1.5863(22) \cong 5.1019$ bulunur. Öyle ise tahmin denklemi; $\hat{y} = 5.1019 + 1.5863x$ ile verilir. Bu tahmin denklemi kullanılarak çeşitli tahminler yapılabilir. Örneğin $x_4 = 16$ iken y_4 değerini tahmin edelim. $\hat{y}_4 = 5.1019 + 1.5863(16) = 30.4827$ olup 304.827 Kg olarak tahmin edilir (gerçek üretim ise 350Kg olduğu açıktır).

XI.4 Belirleme (Açıklayıcılık) Katsayısı

Belirleme katsayısı regresyon çözümlemesinde önemli bir kavramdır. $R^2_{Y,X}$ ile gösterilir ve $R^2_{Y,X} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ tahmin edicisi ile verilir. Burada $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ yazılır. Bu matematiksel ifade sözel olarak; Kareler Toplamı Y= Regresyon Kareler Toplamı+Hata kareler Toplamı şeklindedir. Kısaca $KT_Y = RKT + HKT$ şeklinde de ifade edilebilir. Belirleme katsayısı, bağımsız değişken(ler)in bağımlı değişken varyansının

%'de kaçını açıkladığının göstergesine denir. Belirleme katsayısı ayrıca $R_{Y,X}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{HKT}{KT_Y}$ tahmin edicisi ile de ifade edilir. $R_{Y,X}^2$'nin bire yakın olması her araştırmacının arzuladığı bir durumdur. Ne kadar bire yakın olursa modelin geçerliliği o kadar kuvvetli olur.

♦ **Örnek:** Bir önceki örnek için belirleme katsayısını hesaplayınız ve sonucu yorumlayınız.

♦ **Çözüm:** $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2624,\ \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \cong 2420.632 => R_{y,x}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cong \frac{2420.632}{2624} \cong 0.9225$ bulunur. Yorum: Bağımlı değişkendeki varyansın % 92.25'i modeldeki bağımsız değişken tarafından açıklanmaktadır, denilebilir.

Belirleme katsayısının tanım aralığı, $0 \leq R_{y,x}^2 \leq 1$'dir. Belirleme katsayısı ile korelasyon katsayısı arasında bir ilişki vardır ve bu ilişki $R_{y,x}^2 = r_{xy}^2$ ile verilir. Son örneğimiz için korelasyon katsayısının değeri, $r_{xy} = \sqrt{0.9225} \cong 0.9605$ bulunur. Bu nedenle bir bağımsız değişkenli model oluştururken bağımlı değişken ile sıkı ilişkiler olan bağımsız değişkenler tercih edilmelidir.

XI.5 Model Parametreleri İçin Hipotez Testi ve Güven Aralığı

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ basit doğrusal regresyon denkleminde $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistikleri β_0 ve β_1 parametrelerinin tahmin edicileridir. Bu nedenle $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistikleri β_0 ve β_1 parametrelerine ilişkin hipotezlerin testlerinde ve güven aralıklarının oluşturulmasında delil olarak kullanılırlar.

XI.5.1 Eğim Parametresi (β_1) İçin Hipotez Testi ve Güven Aralığı

XI.5.1.1 Eğim Parametresi (β_1) İçin Hipotez Testi

Test algoritması şöyle verilebilir.

a) Hipotezler: $H_0: \beta_1 = \beta_{10}$ ve $H_A: \beta_1 \neq \beta_{10}$ şeklinde kurulur. β_{10} genelde sıfır değerini alır.

b) Test İstatistiği: H_0 hipotezinin doğruluğu altında test istatistiği, $t_H = \frac{\hat{\beta}_1 - \beta_{10}}{sh(\hat{\beta}_1)} \sim t_{(n-2)}$'dir. Burada $\hat{\beta}_1$ istatistik değeri, β_{10} bilinen bir değer (sabit) ve $sh(\hat{\beta}_1)$, $\hat{\beta}_1$'in standart hatasıdır ve $sh(\hat{\beta}_1) = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{HKO}{KT_X}}$ eşitliği ile hesaplanır. Ayrıca bu son eşitlikte HKO ve tahmin hatası varyansı, $S_e^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ tahmin edicisi ve buradan $S_e = \sqrt{S_e^2}$ ile verilir.

c) **Karar:** $(n-2)$ s.d. ve $\alpha/2$ önem seviyesinde $t_T = t_{(n-2);\,\alpha/2}$ tablo değeri olmak üzere eğer $|t_H| \leq |t_T| => H_0$ hipotezi kabul ve eğer $|t_H| > |t_T| => H_0$ hipotezi ret edilir.

d) **Yorum:** Karara göre yorum yapılır. Eğer H_0 hipotezi ret edilir ise hem β_1 parametresinin önemli ve hem de modelin anlamlı olduğu söylenir. Aksi durumda ise (yani H_0 hipotezinin kabul edilmesi) β_1'in de modelin de önemsiz olduğuna karar verilir.

♦ **Örnek:** Gübre-Verim örneğimize ilişkin modelin % 5 önem seviyesinde önemli olduğu söylenebilir mi?

♦ **Çözüm:**

a) Hipotezler: $H_0: \beta_1 = 0$ ve $H_A: \beta_1 \neq 0$ şeklinde kurulur.

b) Test İstatistiği: $t_H \cong \frac{1.5863-0}{0.1877} \cong 8.651$ bulunur. Önceki örneklerden $\hat{\beta}_1 \cong 1.5863$ bulunmuştu. $\beta_{10} = 0$ biliniyor ve $sh(\hat{\beta}_1) \cong \frac{5.8219}{\sqrt{962}} \cong 0.1877$, $S_e^2 = \frac{1}{(8-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \cong \frac{203.368}{6} \cong 33.8947$ ve $S_e \cong \sqrt{33.8947} \cong 5.8219$ bulunur.

c) Karar: $(8-2) = 6$ s.d. ve $\alpha/2 = 0.025$ önem seviyesinde $t_T = t_{6;\,0.025} = 2.447$ (t-tablosumdan) bulunur. $|t_H| > |t_T|$ olduğundan H_0 hipotezi ret edilir.

d) Yorum: % 95 güvenilirlikle hem β_1 parametresinin ve hem de doğrusal regresyon modelinin önemli olduğu söylenebilir.

XI.5.1.2 Eğim Parametresi (β_1) İçin Güven Aralığı

β_1 parametresi için $\%(1-\alpha)\times 100$ güvenilirlikle güven aralığı tahmin edicisi, $P\{\hat{\beta}_1 - T_T sh(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + T_T sh(\hat{\beta}_1)\} = (1-\alpha)$ ile verilir. Bu tahmin edicinin ürettiği aralık tahnini ise $P\{\hat{\beta}_1 - t_T sh(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_T sh(\hat{\beta}_1)\} = (1-\alpha)$ olasılık beyanı ile verilir. Aralık sıfır içerdiğinde β_1 parametresinin önemli olmadığı; diğer bir ifade ile modelin önemli olmadığı söylenir.

♦ **Örnek:** Bir önceki örnek verileri ışığı altında β_1 parametresi için % 95 güven aralığı tahmin ediniz ve sonucu yorumlayınız.

♦ **Çözüm:** $P\{1.5863 - 2.447(0.1877) \leq \beta_1 \leq 1.5863 - 2.447(0.1877)\} = 0.95 \Rightarrow P\{1.127 \leq \beta_1 \leq 2.046\} = 0.95$ bulunur. Yorum: $(1.127; 2.046)$ aralığının β_1 parametresini içeriyor olama olasılığı % 95'dir. Dolayısı ile bu sonuçtan anlaşılacağı üzere β_1 parametresi önemlidir, çünkü aralık sıfırı içermemektedir. Modelin de geçerli olduğu söylenir.

XI.5.2 Kesim Parametresi (β_0) İçin Hipotez Testi ve Güven Aralığı

XI.5.2.1 Kesim Parametresi (β_0) İçin Hipotez Testi

Test algoritması şöyle verilebilir.

a) Hipotezler: $H_0: \beta_0 = \beta_{00}$ ve $H_A: \beta_0 \neq \beta_{00}$ şeklinde kurulur. β_{00} genelde sıfır değerini alır. β_0 parametresinin önemli olup olmaması β_1 parametresi kadar önemli değildir.

b) Test İstatistiği: H_0 hipotezinin doğruluğu altında test istatistiği, $t_H = \frac{\hat{\beta}_0 - \beta_{00}}{sh(\hat{\beta}_0)} \sim t_{(n-2)}$'dir. Burada $\hat{\beta}_0$ istatistik değeri, β_{00} bilinen bir değer (sabit) ve $sh(\hat{\beta}_0)$, $\hat{\beta}_0$'ın standart hatasıdır ve $sh(\hat{\beta}_0) = \frac{S_e\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\ \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{HKO \times \sum_{i=1}^{n} x_i^2}{n \times KT_x}}$ eşitliği ile hesaplanır. Ayrıca bu son eşitlikte HKO ve tahmin hatası varyansı, $S_e^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ tahmin edicisi ve buradan $S_e = \sqrt{S_e^2}$ ile verilir.

c) Karar: $(n-2)$ s.d. ve $\alpha/2$ önem seviyesinde $t_T = t_{(n-2);\,\alpha/2}$ tablo değeri olmak üzere eğer $|t_H| \leq |t_T| \Rightarrow H_0$ hipotezi kabul ve eğer $|t_H| > |t_T| \Rightarrow H_0$ hipotezi ret edilir.

d) Yorum: Karara göre yorum yapılır. Eğer H_0 hipotezi ret edilir ise β_0 parametresinin önemli olduğu söylenir. Aksi durumda ise (yani H_0 hipotezinin kabul edilmesi) β_0'in önemsiz olduğuna karar verilir.

♦ **Örnek:** Gübre-Verim örneğimize ilişkin modelin % 5 önem seviyesinde önemli olduğu söylenebilir mi?

♦ **Çözüm:**

a) Hipotezler: $H_0: \beta_0 = 0$ ve $H_A: \beta_0 \neq 0$ şeklinde kurulur.

b) Test İstatistiği: $t_H \cong \frac{5.1019-0}{4.6137} \cong 1.106$ bulunur. Önceki örneklerden $\hat{\beta}_0 \cong 5.1019$ bulunmuştu. $\beta_{00} = 0$ biliniyor ve $sh(\hat{\beta}_0) \cong \frac{5.8219\sqrt{4834}}{\sqrt{8(962)}} \cong 4.6137$, $S_e^2 = \frac{1}{(8-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \cong \frac{203.368}{6} \cong 33.8947$ ve $S_e \cong \sqrt{33.8947} \cong 5.8219$ bulunur.

c) Karar: $(8-2) = 6$ s.d. ve $\alpha/2 = 0.025$ önem seviyesinde $t_T = t_{6;\,0.025} = 2.447$ (t-tablosundan) bulunur. $|t_H| < |t_T|$ olduğundan H_0 hipotezi kabul edilir.

d) Yorum: % 95 güvenilirlikle β_0 parametresi önemli değildir.

XI.5.2.2 Kesim Parametresi (β_0) İçin Güven Aralığı

β_0 parametresi için $\%(1-\alpha)\times 100$ güvenilirlikle güven aralığı tahmin edicisi, $P\{\hat{\beta}_0 - T_T sh(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + T_T sh(\hat{\beta}_0)\} = (1-\alpha)$ ile verilir. Bu tahmin edicinin ürettiği aralık tahnini ise $P\{\hat{\beta}_0 - t_T sh(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_T sh(\hat{\beta}_0)\} = (1-\alpha)$ olasılık beyanı ile verilir. Aralık sıfır içerdiğinde β_1 parametresinin önemli olmadığı; diğer bir ifade ile modelin önemli olmadığı söylenir.

♦ **Örnek:** Bir önceki örnek verileri ışığı altında β_0 parametresi için % 95 güven aralığı tahmin ediniz ve sonucu yorumlayınız.

♦ **Çözüm:** $P\{5.1019 - 2.447(4.6137) \leq \beta_0 \leq 5.1019 - 2.447(4.6137)\} = 0.95 \Rightarrow P\{-6.1878 \leq \beta_0 \leq 16.3914\} = 0.95$ bulunur. Yorum: $(-6.1878; 16.3914)$ aralığının β_0 parametresini kapsıyor olama olasılığı % 95'dir. Yani % 95 güvenilirlikle aralık sıfırı içerdiğinden β_0 parametresi önemli değildir.

XI.5.3 Uyum Parametresi İçin Güven Aralığı

Veri kümesinde var olan belli bir x_i gözlem değerine karşılık tahmin denklemi kullanılarak elde edilen değere uyum tahmini adı verilir. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ tahmin denkleminde X tesadüfi değişkeninin k. gözlem değerine (x_k) karşılık gelen uyum değeri ($\hat{y}_k$) tahmin edicisinin ($\hat{Y}_k$) ortalama ve varyansı sırası ile $E(\hat{Y}_k) = Y_k$ ve $V(\hat{Y}_k) = S_e^2\left[\frac{1}{n} + \frac{(X_k-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]$ ile verilir. Buradan $\hat{y}_k$'nın standart hatası ise $sh(\hat{Y}_k) = \sqrt{V(\hat{Y}_k)}$ ile verilir. Y_k parametresi için $\%(1-\alpha)\times 100$ güvenilirlikle güven aralığı tahmin edicisi, $P\{\hat{Y}_k - T_T sh(\hat{Y}_k) \leq Y_k \leq \hat{Y}_k + T_T sh(\hat{Y}_k)\} = (1-\alpha)$ ile verilir. Bu tahmin edicinin ürettiği aralık tahnini ise $P\{\hat{y}_k - t_T sh(\hat{y}_k) \leq Y_k \leq \hat{y}_k + t_T sh(\hat{y}_k)\} = (1-\alpha)$ olasılık beyanı ile verilir.

♦ **Örnek:** Klasik Gübre-Verim örneğimizde $x_5 = 22$ iken Y_5 için % 95 güven aralığı tahmin ediniz.

♦ **Çözüm:** $\hat{y}_5 \cong 5.1019 + 1.5863(22) \cong 40.0005$ Kg bulunur. Bu değeri 10 Kg ile genişletirsek gerçek tahmin değeri, $\hat{y}_5 \cong 400.005$ Kg olur. $V(\hat{y}_k) = s_e^2\left[\frac{1}{n} + \frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] \cong$

$5.8214^2\left[\frac{1}{8}+\frac{(22-22)^2}{962}\right]\cong\frac{5.8214^2}{8}\cong 4.236$ ve $sh(\hat{y}_5)\cong\sqrt{4.236}=2.0582$ Kg olarak bulunur. Ayrıca $(8-2)=6$ s.d. ve $\alpha/2=0.025$ önem seviyesinde $t_T=t_{6;\,0.025}=2.447$ (t-tablosundan) olduğundan aranan güven aralığı, $P\{40.0005-2.447(2.0582)\leq Y_5\leq 40.0005+2.447(2.0582)\}=0.95 \;\Rightarrow\; P\{34.9641\leq\beta_0\leq 45.0369\}=0.95$ bulunur. Gerçek uzaya dönüştürdüğümüzde (yani, değerleri 10 ile genişlettiğimizde) ise aynı güven aralığı, $P\{349.641\leq\beta_0\leq 450.369\}=0.95$ bulunur. Yorum: % 95 güvenilirlikle Y_5 değeri 349.641 Kg ila 450.369 Kg arasındadır. Ancak bu aralık Y_5'in gerçek değerini içermemiştir (Y_5'in gerçek değeri 480 Kg). Bunun sebebi ise örneklemin oldukça küçük olmasındandır, diyebiliriz.