Math 326 – Homework 08 (11.6 – 11.7 and 11.10 – 11.12)
Due (via upload to Canvas) Wednesday, April 13, 2022 at 11:59 PM

1. Suppose the following represents a random sample of points (x, y):

x	-2.0	-1.0	0.0	1.0	2.0
y	3.0	2.0	1.0	1.0	0.5

(a) Find the 90% confidence interval for $E(Y)$ when $x^* = 0$ and again when $x^* = 2$.

Solution: Calculating the confidence intervals first requires computation of the LSR line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The usual suspects are

$$\bar{x} = 0, \bar{y} = 1.5, S_{yy} = 4, S_{xx} = 10, \text{ and } S_{xy} = -6.$$

Then $\hat{\beta}_0 = 1.5$ and $\hat{\beta}_1 = -0.6$ and

$$\hat{y} = 1.5 - 0.6x.$$

To compute the confidence intervals, we need the standard error associated with our estimator $\hat{y}$. Recall $V(\hat{y}) = S^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right) = S^2 \left(\frac{1}{5} + \frac{(x^*)^2}{10}\right)$. As $S^2 = \text{SSE}/(n-2)$ and $\text{SSE} = S_{yy} - \hat{\beta}_1 S_{xy}$, we get $\text{SSE} = 0.4$ and $S^2 = 0.4/3 \approx 0.13333$. At 3 degrees of freedom, the critical value is $t^*_{0.05}(3) = 2.353$ and the confidence interval will be of the form

$$(-0.6 + 1.5x^*) \pm t^*_{0.05}(3)\sqrt{V(\hat{y})}.$$

The intervals for our test points are:

$$\begin{array}{lll} \text{at } x = 0 & 1.5 \pm 0.384 \text{ or} & (1.116, 1.884), \\ \text{at } x = 2 & 0.3 \pm 0.666 \text{ or} & (-0.366, 0.966). \end{array}$$

(b) Find the 90% confidence interval for Y^* when $x^* = 0$ and again when $x^* = 2$.

Solution: This is essentially the exact same computation as above, but the variance for $V(y^*)$ is just a little bit wider; $V(\hat{y}) = S^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right) = S^2 \left(1 + \frac{1}{5} + \frac{(x^*)^2}{10}\right)$. This time the confidence interval will be of the form

$$(-0.6 + 1.5x^*) \pm t^*_{0.05}(3)\sqrt{V(\hat{y})}.$$

The intervals for our test points are:

$$\begin{array}{lll} \text{at } x = 0 & 1.5 \pm 0.941 \text{ or} & (0.559, 2.441), \\ \text{at } x = 2 & 0.3 \pm 1.087 \text{ or} & (-0.787, 1.387). \end{array}$$

(c) Are the intervals for $E(Y)$ or the intervals for Y^* wider? How can this be explained?

Solution:
The intervals of Y^* will always be wider as the extra $+\sigma^2$ in the variance term will make standard error just a bit larger. Recall this is to account for the total spread of Y rather than just the expected value of Y and is coming from the ϵ, the probabilistic component of the model.

The following table contains dietary data (calories and the content of fat, sodium, carbohydrate, and protein) in some standard hamburgers that can be found at local fast food restaurants.

	cal	fat (g)	sodium (mg)	carbs (g)	protein (g)
BK Jr.	310	18	390	27	13
Wendy's Jr.	250	11	420	25	13
McDonald's	250	9	480	31	12
Culvers	390	17	480	38	20
Steak-n-Shake	320	14	830	32	15
Sonic Jr.	330	16	610	32	15

2. Assume the relationship between defining calories is a linear one. That is,

$$\text{cal} = A \cdot (\text{fat}) + B \cdot (\text{carbs}) + C \cdot (\text{protein}) + D.$$

(a) Determine the best-fit hyperplane that predicts calories of a burger based upon the fat, carbohydrate, and protein content of the burger. Explicitly state the terms of the normal forms: X, X^TX, $(X^TX)^{-1}$, and X^TY.

Solution: Using our text's least squares notation, we seek a best-fit hyperplane of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where x_1 corresponds to the fat content, x_2 the carbohydrates, and x_3 the protein. The matrices we need for the normal equations $\hat{\boldsymbol{\beta}} =: (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ are

$$\mathbf{X} = \begin{pmatrix} 1 & 18 & 27 & 13 \\ 1 & 11 & 25 & 13 \\ 1 & 9 & 31 & 12 \\ 1 & 17 & 38 & 20 \\ 1 & 14 & 32 & 15 \\ 1 & 16 & 32 & 15 \end{pmatrix}, \quad \mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 18 & 11 & 9 & 17 & 14 & 16 \\ 27 & 25 & 31 & 38 & 32 & 32 \\ 13 & 13 & 12 & 20 & 15 & 15 \end{pmatrix},$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 6 & 85 & 185 & 88 \\ 85 & 1267 & 2646 & 1275 \\ 185 & 2646 & 5807 & 2768 \\ 88 & 1275 & 2768 & 1332 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 1850 \\ 26970 \\ 57990 \\ 27830 \end{pmatrix},$$

and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 11.6325 & -0.2268 & -0.4660 & 0.4169 \\ -0.2268 & 0.0260 & 0.0109 & -0.0317 \\ -0.4660 & 0.0104 & 0.0370 & -0.0561 \\ -0.4169 & -0.0317 & -0.0561 & 0.1201 \end{pmatrix},$$

.

Then $\hat{\boldsymbol{\beta}} = \langle -15.870, 7.671, 3.974, 6.340 \rangle$ and the hyperplane that best fits the data is

$$\hat{Y} = -15.870 + 7.671x_1 + 3.974x_2 + 6.340x_3.$$

(b) Based upon your least-squares regression analysis, how many calories are in expected in a burger made with 10 grams of fat, 20 grams of carbohydrates, and 15 grams of protein.

Solution: Setting $x_1 = 10$, $x_2 = 20$, and $x_3 = 15$ (or $\mathbf{x}^* = (10, 20, 15)$ in the parlance of our text), we get $\hat{Y} = 235.4$ calories.

(c) Find SSE and S.

Solution: Recall SSE $= \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y}$ and we get SSE $= 55.547$. To compute standard deviation, we need to use the estimator $S^2 = \text{SSE}/(n-(k+1))$. Here we have $n = 6$ and 4 regression coefficients. Thus we have 2 degrees of freedom and this yields $S^2 = 55.547/2 \approx 27.774$. Thus, $S \approx 5.270$

(d) Find a 95% confidence interval for the amount of calories in a burger made with 10 grams of fat, 20 grams of carbohydrates, and 15 grams of protein.

Solution: Note that $\hat{Y} = 235.4 = -15.870 + 7.671(10) + 3.974(20) + 6.340(15)$. To use our multivariate regression formulas, we need to identify $\mathbf{a}$ such that $\hat{Y}(\mathbf{x}^*) = \mathbf{a}^T\boldsymbol{\beta}$. Looking at the previous line as a dot products, we determine that $\mathbf{a} = \langle 1, 10, 20, 15\rangle$. Then $V(\mathbf{a}\boldsymbol{\beta}) = S^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = 178.048$. For our confidence interval, we need $t^*_{0.025}(2) = 4.303$ and the equation $\mathbf{a}^T\boldsymbol{\beta} \pm t^*_{0.025}(2)\sqrt{V(\mathbf{a}\boldsymbol{\beta})}$ yields

$$235.4 \pm 57.4 \text{ calories.}$$

3. A family kept track of its natural gas usage for two heating seasons and the accompanying outdoor temperatures. Gas usage is measured in hundreds of cubic feet, and temperature is average temperature in degrees Fahrenheit.

month:	Oct 1	Nov 1	Dec 1	Jan 1	Feb 1	Mar 1
temp:	53	41	14	22	32	39
gas:	3.0	6.2	12.6	9.2	7.5	5.5
month:	Oct 2	Nov 2	Dec 2	Jan 2	Feb 2	Mar 2
temp:	56	36	33	13	35	43
gas:	2.0	6.5	7.3	12.5	6.9	5.3

(a) Fit he model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ where X is the temperature and Y is the gas usage.

Solution: Let

$$\begin{aligned}\mathbf{x} &= \langle 53, 41, 14, 22, 32, 39, 56, 36, 33, 13, 35, 43\rangle,\\ \mathbf{x}^2 &= \langle 2809, 1681, 196, 484, 1024, 1521, 3136, 1296, 1089, 169, 1225, 1849\rangle.\end{aligned}$$

Then defining $\mathbf{X} = [\mathbf{1} \quad \mathbf{x} \quad \mathbf{x}^2]$ we get

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 12 & 417 & 16479 \\ 417 & 16479 & 706065 \\ 16479 & 706065 & 31998951 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 84.5 \\ 2465.3 \\ 84827.5 \end{pmatrix},$$

and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 2.899 & -0.172 & 0.002 \\ -0.172 & 0.011 & -0.0001 \\ -0.002 & 0.0001 & 2.4 \times 10^{-6} \end{pmatrix},$$

.

Then $\hat{\boldsymbol{\beta}} = \langle 16.194, -0.301, 0.0009\rangle$ and the hyperplane that best fits the data is

$$\hat{Y} = 16.194 - 0.301x + 0.0009x^2.$$

(b) Test if the quadratic term is really necessary. That is, test $H_0 := \beta_2 = 0$.

Solution: We want to test $\beta_2 = 0$ versus $\beta_2 \neq 0$. Since we are isolating on just β_2, we can use $\hat{\beta}_2$ and $V(\hat{\beta}_2) = \sigma^2 c_{22} \approx S^2 c_{22}$ where $c_{22} = 2.4 \times 10^{-6}$. (Equivalently, we could have done the $V(\mathbf{a}\boldsymbol{\beta}) = S^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}$ calculation we used in Problem 2 where $\mathbf{a} = \langle 0, 0, 1 \rangle$.) Note $\text{SSE} = \mathbf{Y}^T \mathbf{Y} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{Y} \approx 1.815$, $S^2 = \text{SSE}/(12-3) \approx 0.202$ and $S = 0.449$. Then the test statistic is

$$T = \frac{\hat{\beta}_2 - 0}{\sqrt{S^2 c_{22}}} = \frac{0.0009}{0.449\sqrt{2.4 \times 10^{-6}}} \approx 1.377.$$

Using R, I calculated the two-sided P-value at 9 degrees of freedom to be 0.2017919. This is not sufficient evidence to reject the null hypothesis.