The mean grade point averages of students participating in sports programs at 5 area high schools are given in the table below.

	School						
	i=1	*i*=2	*i*=3	*i*=4	*i*=5	Mean	Variance
Baseball (*j*=1)	2.3	2.9	3.1	3.1	3.6	$\bar{y}_1 = 3.0$	$s_1^2 = 0.220$
Football (*j*=2)	2.8	3.3	3.8	3.1	3.5	$\bar{y}_2 = 3.3$	$s_2^2 = 0.145$
Soccer (*j*=3)	1.9	2.6	3.1	2.0	2.4	$\bar{y}_3 = 2.4$	$s_3^2 = 0.235$

Consider the model $Y_{ij} = \mu_j + \varepsilon_{ij}$, where ε_{ij}'s are i.i.d. $N(0, \sigma^2)$.

a. At $\alpha = 0.05$, can one conclude that there is a difference in the mean GPA of the three groups? That is, test H_0: $\mu_B = \mu_F = \mu_S$ at a 5% level of significance. Construct the ANOVA table, report the critical value and state your decision.

b. Use a 95% confidence level and Tukey's pairwise comparison procedure to compare the average GPA for Football with the average GPA for Soccer.

c. Use a 95% confidence level and Scheffe's multiple comparison procedure to compare the average GPA for Football with the average GPA for Soccer.

d. Use a 95% confidence level and Scheffe's multiple comparison procedure to compare the average GPA for Baseball and Football with the average GPA for Soccer.

e. Test H_0: $\mu_B = \mu_F = \mu_S$ at a 10% level of significance using the Kruskal-Wallis test. Report the value of the test statistic, the critical value(s), and your decision.

Sports Programs (without R)

a. Consider the model $Y_{ij} = \mu_j + \varepsilon_{ij}$, where ε_{ij}'s are i.i.d. $N(0, \sigma^2)$.
Test H_0: $\mu_B = \mu_F = \mu_S$ with $\alpha = 0.05$.

$J = 3$.
$N = n_1 + n_2 + \ldots + n_J = 5 + 5 + 5 = 15$.

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + \ldots + n_J \cdot \bar{y}_J}{N} = \frac{5 \cdot 3.0 + 5 \cdot 3.3 + 5 \cdot 2.4}{15} = 2.9.$$

$$\text{SSB} = n_1 \cdot (\bar{y}_1 - \bar{y})^2 + n_2 \cdot (\bar{y}_2 - \bar{y})^2 + \ldots + n_J \cdot (\bar{y}_J - \bar{y})^2$$
$$= 5 \cdot (3.0 - 2.9)^2 + 5 \cdot (3.3 - 2.9)^2 + 5 \cdot (2.4 - 2.9)^2 = 2.1.$$

$$\text{SSW} = (n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2 + \ldots + (n_J - 1) \cdot s_J^2$$
$$= 4 \cdot 0.220 + 4 \cdot 0.145 + 4 \cdot 0.235 = 2.4.$$

SSTotal = SSB + SSW = 2.1 + 2.4 = 4.5.

Completing the ANOVA table,

Source	SS	df	MS	*F*
Between Groups	2.1	$J - 1 = 2$	1.05	**5.25**
Within Groups	2.4	$N - J = 12$	0.2	
Total	4.5	$N - 1 = 14$		

According to the *F*-distribution, the critical region is $F > F_{0.05}(2,12) =$ **3.89**. Since the test statistic lies in the critical region, we reject H_0 and conclude that the model does a significant job of predicting GPA.

b. The 95% confidence level using Tukey's pairwise comparison procedure is

$$(\bar{y}_i - \bar{y}_j) \pm \frac{q_\gamma(J, N-J)}{\sqrt{2}} \cdot s_{pooled} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = (3.3 - 2.4) \pm \frac{q_{0.05}(3,12)}{\sqrt{2}} \cdot \sqrt{0.2} \cdot \sqrt{\frac{1}{5} + \frac{1}{5}}$$
$$= 0.9 \pm \frac{3.77}{\sqrt{2}} \cdot \sqrt{0.2} \cdot \sqrt{\frac{1}{5} + \frac{1}{5}}$$
$$= 0.9 \pm 0.75$$
$$= (0.15, 1.65)$$

With this method, we are 95% confident that the mean difference between the average GPAs of Football and Soccer players is between 0.15 and 1.65.

c. The 95% confidence level using Scheffe's multiple comparison procedure is

$$\sum_{j=1}^{J} c_j \bar{y}_j \pm \sqrt{F_\alpha(J-1, N-J)} \cdot \sqrt{\text{MSW}} \cdot \sqrt{(J-1) \sum_{j=1}^{J} \frac{c_j^2}{n_j}} = (3.3 - 2.4) \pm \sqrt{F_\alpha(2,12)} \cdot \sqrt{0.2} \cdot \sqrt{2 \cdot \left(0 + \frac{1}{5} + \frac{1}{5}\right)}$$
$$= 0.9 \pm \sqrt{3.89} \cdot \sqrt{0.2} \cdot \sqrt{2 \cdot \left(0 + \frac{1}{5} + \frac{1}{5}\right)}$$
$$= 0.9 \pm 0.79$$
$$= (0.11, 1.69)$$

where $c_B = 0$, $c_F = 1$, and $c_S = -1$.

With this method, we are 95% confident that the mean difference between the average GPAs of Football and Soccer players is between 0.11 and 1.69.

d. The 95% confidence level using Scheffe's multiple comparison procedure is

$$\sum_{j=1}^{J} c_j \bar{y}_j \pm \sqrt{F_\alpha(J-1, N-J)} \cdot \sqrt{\text{MSW}} \cdot \sqrt{(J-1) \sum_{j=1}^{J} \frac{c_j^2}{n_j}}$$
$$= \left(\frac{3.0 + 3.3}{2} - 2.4\right) \pm \sqrt{F_\alpha(2,12)} \cdot \sqrt{0.2} \cdot \sqrt{2 \cdot \left(\frac{1}{20} + \frac{1}{20} + \frac{1}{5}\right)}$$
$$= 0.75 \pm \sqrt{3.89} \cdot \sqrt{0.2} \cdot \sqrt{2 \cdot \left(\frac{1}{20} + \frac{1}{20} + \frac{1}{5}\right)}$$
$$= 0.75 \pm 0.68$$
$$= (0.07, 1.43)$$

where $c_B = ½$, $c_F = ½$, and $c_S = -1$.

With this method, we are 95% confident that the difference between the average GPAs of Baseball and Football compared to the Soccer players is between 0.07 and 1.43.

e. First order the GPAs and rank them.

Sport	S	S	B	S	S	F	B	B	B	F	S	F	F	B	F
GPA	1.9	2.0	2.3	2.4	2.6	2.8	2.9	3.1	3.1	3.1	3.1	3.3	3.5	3.6	3.8
Rank	1	2	3	4	5	6	7	9.5	9.5	9.5	9.5	12	13	14	15

Then calculate the rank mean for each group.

$$\bar{r}_B = \frac{3 + 7 + 9.5 + 9.5 + 14}{5} = 8.6. \qquad \bar{r}_F = \frac{6 + 9.5 + 12 + 13 + 15}{5} = 11.1.$$

$$\bar{r}_S = \frac{1 + 2 + 4 + 5 + 9.5}{5} = 4.3. \qquad \bar{r} = \frac{N+1}{2} = 8.$$

Test Statistic:

$$K = \frac{12}{15 \cdot 16}\left[5 \cdot (8.6 - 8)^2 + 5 \cdot (11.1 - 8)^2 + 5 \cdot (4.3 - 8)^2\right] = \mathbf{5.915}.$$

Critical Value: $\chi^2_\alpha(J-1) = \chi^2_{0.10}(2) =$ **4.605**.

Since the test statistic does lie in the critical region (5.915 > 4.605), we reject H_0 and conclude that the model does a significant job of predicting GPA. This is the same result as the "parametric" test back in part a.

In order to rate three brands of a particular product, a consumer agency divided eighteen individuals at random into three groups and asked each one of them to rate one brand of the product on the scale from 0 to 100.

Brand							Mean	Variance
1	66	72	77	81	87	85	$\bar{y}_1 = 78$	$s_1^2 = 64$
2	83	73	69	77	67	87	$\bar{y}_2 = 76$	$s_2^2 = 62$
3	85	74	85	88	89	95	$\bar{y}_3 = 86$	$s_3^2 = 48$

a. Test H_0: $\mu_1 = \mu_2 = \mu_3$ at a 10% level of significance. Construct the ANOVA table, report the critical value, and state your decision.

b. Use a 90% confidence level and Tukey's pairwise comparison procedure to compare the average rating for Brand 3 with the average rating for Brand 2.

c. Use a 90% confidence level and Scheffe's multiple comparison procedure to compare the average rating for Brand 3 with the average rating for Brand 1 and Brand 2.

d. Test H_0: $\mu_1 = \mu_2 = \mu_3$ at a 10% level of significance using the Kruskal-Wallis test. Report the value of the test statistic, the critical value(s), and your decision.

a. Test H_0: $\mu_1 = \mu_2 = \mu_3$ with $\alpha = 0.10$.

$$J = 3.$$
$$N = n_1 + n_2 + \ldots + n_J = 6 + 6 + 6 = 18.$$
$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + \ldots + n_J \cdot \bar{y}_J}{N} = \frac{6 \cdot 78 + 6 \cdot 76 + 6 \cdot 86}{18} = 80.$$

$$\text{SSB} = n_1 \cdot (\bar{y}_1 - \bar{y})^2 + n_2 \cdot (\bar{y}_2 - \bar{y})^2 + \ldots + n_J \cdot (\bar{y}_J - \bar{y})^2$$
$$= 6 \cdot (78 - 80)^2 + 6 \cdot (76 - 80)^2 + 6 \cdot (86 - 80)^2 = 336.$$
$$\text{SSW} = (n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2 + \ldots + (n_J - 1) \cdot s_J^2$$
$$= 5 \cdot 64 + 5 \cdot 62 + 5 \cdot 48 = 870.$$
$$\text{SSTotal} = \text{SSB} + \text{SSW} = 336 + 870 = 1206.$$

Completing the ANOVA table,

Source	SS	df	MS	F
Between Groups	336	$J - 1 = 2$	168	**2.90**
Within Groups	870	$N - J = 15$	58	
Total	1206	$N - 1 = 17$		

According to the F-distribution, the critical region is $F > F_{0.05}(2,15) =$ **3.68**. Since the test statistic does not lie in the critical region, we fail to reject H_0 and conclude that the average ratings of the three brands are not significantly different.

b. The 90% confidence level using Tukey's pairwise comparison procedure is

$$(\bar{y}_i - \bar{y}_j) \pm \frac{q_\gamma(J, N-J)}{\sqrt{2}} \cdot s_{pooled} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = (86 - 76) \pm \frac{q_{0.10}(3,15)}{\sqrt{2}} \cdot \sqrt{58} \cdot \sqrt{\frac{1}{6} + \frac{1}{6}}$$
$$= 10 \pm \frac{3.14}{\sqrt{2}} \cdot \sqrt{58} \cdot \sqrt{\frac{1}{6} + \frac{1}{6}}$$
$$= 10 \pm 9.76$$
$$= (0.24, 19.76)$$

With this method, we are 90% confident that the mean difference between the average rating for Brand 3 and the average rating for Brand 2 is between 0.24 and 19.76.

c. The 90% confidence level using Scheffe's multiple comparison procedure is

$$\sum_{j=1}^{J} c_j \bar{y}_j \pm \sqrt{F_\alpha(J-1, N-J)} \cdot \sqrt{\text{MSW}} \cdot \sqrt{(J-1) \sum_{j=1}^{J} \frac{c_j^2}{n_j}}$$
$$= \left(86 - \frac{78+76}{2}\right) \pm \sqrt{F_\alpha(2,15)} \cdot \sqrt{58} \cdot \sqrt{2 \cdot \left(\frac{1}{24} + \frac{1}{24} + \frac{1}{6}\right)}$$
$$= 9 \pm \sqrt{2.48} \cdot \sqrt{58} \cdot \sqrt{2 \cdot \left(\frac{1}{24} + \frac{1}{24} + \frac{1}{6}\right)}$$
$$= 9 \pm 8.48$$
$$= (0.52, 17.48)$$

where $c_1 = -½$, $c_2 = -½$, and $c_3 = 1$.

With this method, we are 95% confident that the difference between the average rating for Brand 3 as compared to the average rating for Brand 1 and Brand 2 is between 0.52 and 17.48.

d. First order the brand ratings and rank them.

Brand	1	2	2	1	2	3	1	2	1	1
Rating	66	67	69	72	73	74	77	77	81	66
Rank	1	2	3	4	5	6	7.5	7.5	9	1

Brand	2	1	3	3	1	2	3	3	3	2
Rating	83	85	85	85	87	87	88	89	95	83
Rank	10	12	12	12	14.5	14.5	16	17	18	10

Then calculate the rank mean for each group.

$$\bar{r}_1 = \frac{48}{6} = 8.0 \qquad \bar{r}_2 = \frac{42}{6} = 7.0 \qquad \bar{r}_3 = \frac{81}{6} = 13.5 \qquad \bar{r} = \frac{N+1}{2} = 9.5$$

Test Statistic:

$$K = \frac{12}{18 \cdot 19} \cdot [\, 6 \cdot (8 - 9.5)^2 + 6 \cdot (7 - 9.5)^2 + 6 \cdot (13.5 - 9.5)^2 \,] = \mathbf{5.158}$$

Critical Value: $\chi^2_\alpha(J-1) = \chi^2_{0.10}(2) =$ **4.605**.

Since the test statistic does lie in the critical region (5.158 > 4.605), we reject H_0 and conclude that the model does a significant job of product rating.

For this exercise you are not to use R or any other software to solve the exercises. A calculator is allowed.

A students wonders if people of similar heights tend to date each other. She measures herself, her roommate, and some other neighbors in the same apartment complex who are also currently dating. Then she measures the man each woman is dating. Here are the data (heights in inches):

Women (x)	68	64	68	67	70	65
Men (y)	73	68	71	69	73	66

$$\sum x = 402; \quad \sum y = 420; \quad \sum x^2 = 26{,}958; \quad \sum y^2 = 29{,}440; \quad \sum xy = 28{,}167;$$
$$\sum (x-\bar{x})^2 = 24; \quad \sum (y-\bar{y})^2 = 40; \quad \sum (x-\bar{x})(y-\bar{y}) = \sum (x-\bar{x})y = 27$$

Assume that (X, Y) have a bivariate normal distribution.

You are also given that

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 10 & 20 & 5 \\ 20 & 50 & 10 \\ 5 & 10 & 5 \end{bmatrix}; \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix}; \quad \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{bmatrix} 30 \\ 67 \\ 18 \end{bmatrix}$$

$$\sum (y_i - \bar{y})^2 = 13.5; \quad \sum (y_i - \hat{y}_i)^2 = 5$$

a. Find the sample correlation coefficient r between the heights of the women and men.

b. Test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$ at $\alpha = 0.05$. What is the p-value of this test? (You may give a range for the p-value.)

c. Test $H_0: \rho = 0.3$ vs. $H_1: \rho > 0.3$ at $\alpha = 0.05$. What is the p-value of this test?

d. Test $H_0: \rho = 0.5$ vs. $H_1: \rho \neq 0.5$ at $\alpha = 0.05$. What is the p-value of this test?

e. Construct a 95% confidence interval for ρ. The $100(1-\alpha)\%$ confidence interval for ρ is

$$\left(\frac{e^a - 1}{e^a + 1}, \frac{e^b - 1}{e^b + 1} \right), \quad \text{where} \quad a = \ln\frac{1+r}{1-r} - \frac{2z_{\alpha/2}}{\sqrt{n-3}}, \quad b = \ln\frac{1+r}{1-r} + \frac{2z_{\alpha/2}}{\sqrt{n-3}}.$$

f. If every woman wore 2-inch heels when she was measured, what is the correlation between the actual female and male heights? Explain your answer.

g. If every woman dated a man exactly 3 inches taller than herself, what would be the correlation between female and male heights? Explain your answer.

a.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{27}{\sqrt{24}\sqrt{40}} = \mathbf{0.87142}.$$

b. Method 1 begins by calculating the t-test statistic.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.87142\sqrt{6-2}}{\sqrt{1-0.87142^2}} \approx \mathbf{3.553}.$$

There are $n - 2 = 4$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha/2}(n-2) = t_{0.025}(4) = 2.776$. Since the test statistic does lie the critical region, we reject H_0 and conclude that there is a significant correlation between the variables. And since the t-test statistic falls between $t_{0.025}(4) = 2.776$ and $t_{0.01}(4) = 3.747$, the two-sided p-value would be **between 0.02 and 0.05**.

Method 2 begins by calculating the W-test statistic:

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.87142}{1-0.87142}\right) \approx 1.33895.$$

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0}{1-0}\right) = 0$, $\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}$.

We then standardize W under its distribution to create a Z-test statistic:

$$Z = \frac{W - \mu_W}{\sigma_W} = \frac{1.33895 - 0}{\sqrt{1/3}} \approx \mathbf{2.32}.$$

According to the Z-distribution, the critical region is $|z| > z_{\alpha/2} = z_{0.025} = 1.96$. Since the test statistic does lie the critical region, we reject H_0 and conclude that there is a significant correlation between the variables. The two-sided p-value would be $2 \times P(Z > 2.32) = 2 \times 0.0102 = \mathbf{0.0204}$.

c. Begin by calculating the W-test statistic, which is the same as in part b because it's not based on the null hypothesis:

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.87142}{1-0.87142}\right) \approx 1.33895.$$

Under this H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.30}{1-0.30}\right) \approx 0.30952$, $\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}$.

We then standardize W under its distribution to create a Z-test statistic:

$$Z = \frac{W - \mu_W}{\sigma_W} = \frac{1.33895 - 0.30952}{\sqrt{1/3}} \approx \mathbf{1.78}.$$

According to the Z-distribution, the critical region is $z > z_\alpha = z_{0.05} = 1.645$. Since the test statistic does lie the critical region, we reject H_0 and conclude that the correlation between the variables is greater than 0.3. The p-value would be $P(Z > 1.78) = \mathbf{0.0375}$.

d. Test $H_0: \rho = 0.5$ vs. $H_1: \rho \neq 0.5$ at $\alpha = 0.05$. What is the p-value of this test?
Begin by calculating the W-test statistic, which is the same as in part b because it's not based on the null hypothesis:

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.87142}{1-0.87142}\right) \approx 1.33895.$$

Under this H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.50}{1-0.50}\right) \approx 0.54931$, $\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}$.

We then standardize W under its distribution to create a Z-test statistic:

$$Z = \frac{W - \mu_W}{\sigma_W} = \frac{1.33895 - 0.54931}{\sqrt{1/3}} \approx \mathbf{1.37}.$$

According to the Z-distribution, the critical region is $|z| > z_{\alpha/2} = z_{0.025} = 1.96$. Since the test statistic does not lie the critical region, we fail to reject H_0 and conclude that the correlation between the variables is not significantly different than 0.5. The two-sided p-value would be $2 \times P(Z > 1.37) = 2 \times 0.0853 = \mathbf{0.1706}$.

e. The $100(1-\alpha)\%$ confidence interval for ρ is

$$\left(\frac{e^a - 1}{e^a + 1}, \frac{e^b - 1}{e^b + 1} \right), \quad \text{where} \quad a = \ln\frac{1+r}{1-r} - \frac{2z_{\alpha/2}}{\sqrt{n-3}}, \quad b = \ln\frac{1+r}{1-r} + \frac{2z_{\alpha/2}}{\sqrt{n-3}}.$$

$$a = \ln\frac{1+r}{1-r} - \frac{2z_{\alpha/2}}{\sqrt{n-3}} = 2.6779 - \frac{2\cdot 1.96}{\sqrt{3}} = 0.4147.$$

$$b = \ln\frac{1+r}{1-r} + \frac{2z_{\alpha/2}}{\sqrt{n-3}} = 2.6779 + \frac{2\cdot 1.96}{\sqrt{3}} = 4.9411.$$

Thus, we are 95% confident that the true value of the correlation coefficient is in the interval

$$\left(\frac{e^{0.4147} - 1}{e^{0.4147} + 1}, \frac{e^{4.9411} - 1}{e^{4.9411} + 1} \right) = \mathbf{(0.2044, 0.9858)}.$$

f. The correlation coefficient is not affected by linear transformations, including adding (or subtracting) the same number to all values of one variable. So, $r = \mathbf{0.87142}$.

g. These two variables have a perfect linear relationship of $y = x + 3$. So, $r = \mathbf{1}$.

A local fast food restaurant tracks certain data with regard to customer orders in order to predict wait time. Among the variables, they measure the number of food (non-beverage) items ordered (x_1), whether the order is "dine-in" ($x_2 = 1$) or "carry-out" ($x_2 = 0$), and how many minutes elapsed from the moment the order is confirmed and the moment it is delivered (y).

The data for ten customer orders are given here.

Food items in the order (x_1)	2	3	1	1	2	4	1	1	3	2
Dine-In Order (x_2)	1	0	0	1	1	1	0	1	0	0
Time elapsed in minutes (y)	3.1	2.6	2.2	2.9	5.1	4.5	0.7	2.4	3.6	2.9

Consider the multiple linear regression model of

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2).$$

You are also given that

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 20 & 5 \\ 20 & 50 & 10 \\ 5 & 10 & 5 \end{bmatrix}; \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix}; \quad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 30 \\ 67 \\ 18 \end{bmatrix}$$

$$\sum (y_i - \bar{y})^2 = 13.5; \quad \sum (y_i - \hat{y}_i)^2 = 5$$

a. Obtain the least-squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

b. Perform the significance of the regression test at a 5% level of significance. Specify the null and the alternative hypotheses. State the value of the test statistic, critical value(s), and a decision.

c. Test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ at the 5% level of significance. State the value of the test statistic, critical value(s), and a decision.

d. Test the claim that each additional food item purchased adds one minute to wait time against the restaurant's alternate claim that it take less than a minute. That is, test $H_0 : \beta_1 = 1$ vs. $H_1 : \beta_1 < 1$ at a 5% significance level. State the value of the test statistic, critical value(s), and a decision.

e. Construct a 95% confidence interval for β_0.

f. Construct a 90% confidence interval for β_2.

g. Construct a 95% confidence interval for the average wait time for a customer who orders 3 food items for dine-in. Include a sentence of what the interval means.

h. Construct a 90% prediction interval for the wait time for a customer who orders 4 food items for carry-out. Include a sentence of what the interval means.

i. What proportion of the observed variation in wait time is explained by the linear relationship that factors in the number of food items ordered and whether or not the order is for dine-in?

Restaurant Wait Times (without R)

a.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} 30 \\ 67 \\ 18 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.7 \\ 1.2 \end{bmatrix}$$

b. The t-test is not an option for the multiple regression model. Use an F-test.

We're already given two of the necessary values for the ANOVA table.

SSTotal = $SYY = \sum (y_i - \bar{y})^2 = 13.5$; SSError = $RSS = \sum (y_i - \hat{y}_i)^2 = 5$

Next, calculate SSRegression: SSReg = SSTotal – SSError = 13.5 – 5 = 8.5.

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	$\sum (\hat{y}_i - \bar{y})^2 = 8.5$	$p - 1 = 2$	4.25	**5.95**
Error	$\sum (y_i - \hat{y}_i)^2 = 5$	$n - p = 7$	0.714	
Total	$\sum (y_i - \bar{y})^2 = 13.5$	$n - 1 = 9$		

According to the F-distribution, the critical region is $F > F_\alpha(2,7) = F_{0.05}(2,7) = \mathbf{4.74}$. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting wait time.

c. Calculate estimated variance of the slope estimate. From the ANOVA table, we'll use MSE = 0.714 as the estimate of the variance of the residuals, and we'll pull C_{33} from $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$\hat{\text{Var}}\left[\hat{\beta}_2\right] = \hat{\sigma} \cdot C_{33} = (0.714)(0.4) = 0.2856$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\text{Var}}\left[\hat{\beta}_2\right]}} = \frac{1.2 - 0}{\sqrt{0.2856}} = 2.245$$

There are $n - p = 7$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha/2}(n-p) = t_{0.025}(7) = 2.365$. Since the test statistic does not lie the critical region (barely), we fail to reject H_0 and conclude that whether the order is for dine-in is not a significant predictor in the model.

d. Calculate estimated variance of the slope estimate. From the ANOVA table, we'll use MSE = 0.714 as the estimate of the variance of the residuals, and we'll pull C_{22} from $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$\hat{\text{Var}}\left[\hat{\beta}_1\right] = \hat{\sigma} \cdot C_{22} = (0.714)(0.1) = 0.0714$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\text{Var}}\left[\hat{\beta}_2\right]}} = \frac{0.7 - 1}{\sqrt{0.0714}} = -1.12$$

There are $n - p = 7$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_\alpha(n-p) = t_{0.05}(7) = 1.895$. Since the (absolute value of the) test statistic does not lie the critical region, we fail to reject H_0 and conclude that it's more plausible that each additional item order adds at least an extra minute to wait time.

e. Calculate estimated variance of the intercept estimate. We'll use MSE = 0.714 as the estimate of the variance of the residuals, and we'll pull C_{11} from $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$\hat{\text{Var}}\left[\hat{\beta}_0\right] = \hat{\sigma} \cdot C_{11} = (0.714)(0.6) = 0.4284$$

The 95% confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\text{Var}}\left[\hat{\beta}_0\right]} = 1.0 \pm t_{0.025}(7) \cdot \sqrt{0.4284} = 1.0 \pm 2.365 \cdot 0.6545 = 1.0 \pm 1.55 = (-0.55, 2.55)$$

f. We already calculated $\hat{\text{Var}}\left[\hat{\beta}_2\right]$ back in part c. So, the 90% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\text{Var}}\left[\hat{\beta}_2\right]} = 1.2 \pm t_{0.05}(7) \cdot \sqrt{0.2856} = 1.2 \pm 1.895 \cdot 0.5344 = 1.2 \pm 1.01 = (0.19, 2.21)$$

g. The vector representing these predictors is $\boldsymbol{x}_0 = [\,1 \quad 3 \quad 1\,]$. The estimate for the average wait time is

$$\hat{y} = [\,1 \quad 3 \quad 1\,] \begin{bmatrix} 1.0 \\ 0.7 \\ 1.2 \end{bmatrix} = 1 + 0.7(3) + 1.2(1) = 4.3.$$

To calculate the estimate for the variance of the estimate, we need

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = [\,1 \quad 3 \quad 1\,] \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = [\,1 \quad 3 \quad 1\,] \begin{bmatrix} -0.2 \\ 0.1 \\ 0.2 \end{bmatrix} = 0.3.$$

So, the 95% confidence interval is

$$\hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\text{Var}}\left[\hat{Y} \mid x\right]} = 4.3 \pm t_{0.025}(7) \cdot \sqrt{\hat{\sigma}^2 \cdot 0.3} = 4.3 \pm 2.365 \cdot \sqrt{0.714 \cdot 0.3}$$

$$= 4.3 \pm 1.09 = (3.21, 5.39)$$

We are 95% that the average wait time for a dine-in order of 3 items is between 3.21 and 5.39 minutes.

h. The vector representing these predictors is $\boldsymbol{x}_0 = [\,1 \quad 4 \quad 0\,]$. The estimate for the average wait time is

$$\hat{y} = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.7 \\ 1.2 \end{bmatrix} = 1 + 0.7(4) + 1.2(0) = 3.8.$$

To calculate the estimate for the variance of the estimate, we need

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} -0.2 \\ 0.2 \\ -0.2 \end{bmatrix} = 0.6.$$

So, the 95% prediction interval is

$$\hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}[Y \mid x]} = 3.8 \pm t_{0.05}(7) \cdot \sqrt{\hat{\sigma}^2 \cdot (1+0.6)} = 3.8 \pm 1.895 \cdot \sqrt{0.714 \cdot 1.6}$$
$$= 3.8 \pm 2.03 = (1.77, 5.83)$$

There is a 90% chance that a person who orders 4 items for carry-out will have to wait between 1.77 and 5.83 minutes.

i. From the ANOVA table in part b, we see that SSTotal = 13.5 and SSReg = 8.5. Thus, $R^2 = \frac{8.5}{13.5} = 0.629$. That is, the model explains about 63% of the variation in wait time.

e. A certain fitness guru claims that "people who don't take care of themselves and engage in no physical activity spend at least 100 hours per week watching TV and the rest of the time sleeping." Test this claim at a 5% significance level. That is, $H_0: \beta_0 \le 100$ vs. $H_1: \beta_0 > 100$ at a 5% significance level. State the value of the test statistic, critical value(s), and a decision.

f. Construct a 90% confidence interval for the mean number of TV viewing hours in a week when a person engages in 20 hours of physical activity. Include a sentence of what the interval means.

g. Construct a 90% prediction interval for the mean number of TV viewing hours in a week when a person engages in 20 hours of physical activity. Include a sentence of what the interval means.

h. One of the researchers claims that a person who engages in only 2 hours of physical activity per day (14 hours per week) will watch more than 40 hours of TV in that week. Test this claim at a 5% significance level. That is, $H_0: \mathrm{E}[Y \mid x = 14] = 40$ vs. $H_1: \mathrm{E}[Y \mid x = 14] > 40$ at a 5% significance level. State the value of the test statistic, critical value(s), and a decision.

. **Time Use (with R)**

Repeat parts a- h from Problem 1 using R or another software. Please include your code as well as your output. For parts a, b, d, e, and h you may use a *p*-value instead of a rejection region. Label results and highlight or circle relevant output where possible.

For this exercise you are not to use R or any other software to solve the exercises. A calculator is allowed.

You are given a random sample of seven participants who participated in a large behavioral study. Among other things, researchers were looking for a relationship between time spent in physical activity (e.g., exercising, labor) and TV watching. The results of the seven participants are given here.

Physical Activity in hours (x)	16	12	25	19	21	15	18
TV Viewing in hours (y)	30	52	7	32	9	56	38

Consider the simple linear regression model of $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. From Homework 2, we know the following.

$$\sum x_i = 126; \quad \sum y_i = 224; \quad \sum x_i^2 = 2376; \quad \sum x_i y_i = 3600; \quad \sum y_i^2 = 9338$$

$$\bar{x} = 18; \quad \bar{y} = 32; \quad \sum (x_i - \bar{x})^2 = 108; \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = -432; \quad \sum (y_i - \bar{y})^2 = 2170$$

$$\hat{y} = 104 - 4x; \quad \sum (y_i - \hat{y})^2 = 442; \quad s_e = 9.402; \quad R^2 = 0.796$$

a. Use an *F*-test to test the hypothesis that the amount of physical activity does not affect the amount of TV viewing in a week at a 5% significance level. That is, test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \ne 0$ at a 5% significance level. State the value of the test statistic, critical value(s), and a decision.

b. Use a *t*-test to test the hypothesis that the amount of physical activity does not affect the amount of TV viewing in a week at a 5% significance level. That is, test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \ne 0$ at a 5% significance level. State the value of the test statistic, critical value(s), and a decision.

c. Construct a 90% confidence interval for β_1.

d. One of the researchers claims that each additional hour of physical activity would result in two fewer hours of TV viewing. Test this claim at a 10% significance level. That is, $H_0: \beta_1 \ge -2$ vs. $H_1: \beta_1 < -2$ at a 10% significance level. State the value of the test statistic, critical value(s), and a decision.

a. We're already given two of the necessary values for the ANOVA table.

SSTotal = $SYY = \sum (y_i - \bar{y})^2 = 2170$; SSError = $RSS = \sum (y_i - \hat{y}_i)^2 = 442$

Next, calculate SSRegression.

SSReg = SSTotal − SSError = 2170 − 442 = 1728, or

SSReg = $\hat{\beta}_1^2 \cdot SXX = (-4)^2 \cdot 108 = 1728$

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	$\sum (\hat{y}_i - \bar{y})^2 = 1728$	$p - 1 = 1$	1728	**19.55**
Error	$\sum (y_i - \hat{y}_i)^2 = 442$	$n - p = 5$	88.4	
Total	$\sum (y_i - \bar{y})^2 = 2170$	$n - 1 = 6$		

According to the *F*-distribution, the critical region is $F > F_\alpha(1,5) = F_{0.05}(1,5) = \mathbf{6.61}$. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting physical activity hours.

b. Calculate the *t*-test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{(-4) - 0}{9.402/\sqrt{108}} = -4.421$$

There are $n - 2 = 5$ degrees of freedom. According to the *t*-distribution, the critical region is $|t| > t_{\alpha/2}(5) = t_{0.025}(5) = 2.571$. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting physical activity hours.

Note: This is the same decision as in part a, and $(t)^2 = (-4.421)^2 = 19.55 = F$.

c. We are 90% confident that the average change in TV viewing due to a one hour increase in physical activity is between –5.8 and –2.2 hours.

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{SXX}} = -4 \pm t_{0.05} \cdot \frac{s_e}{\sqrt{SXX}} = -4 \pm 2.015 \cdot \frac{9.402}{\sqrt{108}} = -4 \pm 1.823 = (-5.8, -2.2)$$

d. Calculate the *t*-test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{(-4) - (-2)}{9.402/\sqrt{108}} = -2.211$$

There are $n - 2 = 5$ degrees of freedom. According to the *t*-distribution, the critical region is $|t| > t_\alpha(5) = t_{0.10}(5) = 1.476$. Since the test statistic does lie the critical region, we reject H_0 and conclude that each additional hour of physical activity would result in at least two fewer hours of TV viewing.

e. Calculate the t-test statistic using the standard deviation of $\hat{\beta}_0$.

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{s_e \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} = \frac{104 - 100}{9.402 \cdot \sqrt{\frac{1}{7} + \frac{18^2}{108}}} = 0.240$$

There are $n - 2 = 5$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_\alpha(5) = t_{0.05}(5) = 2.015$. Since the test statistic does not lie the critical region, we fail to reject H_0 and conclude that there's not enough evidence to support the fitness guru's claim.

f. We are 90% confident that the mean number of TV viewing hours in a week when a person engages in 20 hours of physical activity is between 16 and 32 hours.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}} = (104 - 4 \cdot 20) \pm t_{0.05}(5) \cdot 9.402 \cdot \sqrt{\frac{1}{7} + \frac{(20 - 18)^2}{108}}$$

$$= 24 \pm 2.015 \cdot 9.402 \cdot \sqrt{\frac{1}{7} + \frac{(20 - 18)^2}{108}} = 24 \pm 8 = (16, 32)$$

g. There's a 90% probability that the number of TV viewing hours in a week when a person engages in 20 hours of physical activity will be between 3.4 and 44.6 hours.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}} = (104 - 4 \cdot 20) \pm t_{0.05}(5) \cdot 9.402 \cdot \sqrt{1 + \frac{1}{7} + \frac{(20 - 18)^2}{108}}$$

$$= 24 \pm 2.015 \cdot 9.402 \cdot \sqrt{1 + \frac{1}{7} + \frac{(20 - 18)^2}{108}} = 24 \pm 20.6 = (3.4, 44.6)$$

h. Calculate the t-test statistic using the standard deviation of $\mathrm{E}[Y \mid x = 14] = \mu_{Y|x=14}$ which is the same as when calculating a confidence interval for $\mu_{Y|x}$ as in part f.

$$t = \frac{\hat{y} - \mu_0}{s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}}} = \frac{(104 - 4 \cdot 14) - 40}{9.402 \cdot \sqrt{\frac{1}{7} + \frac{(14 - 18)^2}{108}}} = 1.577$$

There are $n - 2 = 5$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_\alpha(5) = t_{0.05}(5) = 2.015$. Since the test statistic does not lie the critical region, we fail to reject H_0. We conclude that there's not enough evidence to suggest that a person who engages in only 2 hours of physical activity per day (14 hours per week) will watch more than 40 hours of TV in that week.

For this exercise you are not to use R or any other software to solve the exercises. A calculator is allowed.

You are given a random sample of seven participants who participated in a large behavioral study. Among other things, researchers were looking for a relationship between time spent in physical activity (e.g., exercising, labor) and TV watching. The results of the seven participants are given here.

Physical Activity in hours (x)	16	12	25	19	21	15	18
TV Viewing in hours (y)	30	52	7	32	9	56	38

Consider the simple linear regression model of $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.

a. Find the equation of the least-squares regression line.

b. Calculate the fitted values, $\hat{y}_i$.

c. Calculate the residuals, e_i. Does the sum of the residuals equal zero?

d. Give an estimate for σ, the standard deviation of the observations about the true regression line?

e. What proportion of observed variation in TV viewing is explained by a straight-line relationship with physical activity?

f. Predict the number of TV viewing hours for a participant who engaged in 24 hours of physical activity in the same week.

a. Find the equation of the least-squares regression line.

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$	$(y - \bar{y})^2$
16	30	–2	–2	4	4	4
12	52	–6	20	36	–120	400
25	7	7	–25	49	–175	625
19	32	1	0	1	0	0
21	9	3	–23	9	–69	529
15	56	–3	24	9	–72	576
18	38	0	6	0	0	36
126	224	0	0	108	–432	2170
				SXX	*SXY*	*SYY*

or …

x	y	x^2	xy	y^2
16	30	256	480	900
12	52	144	624	2704
25	7	625	175	49
19	32	361	608	1024
21	9	441	189	81
15	56	225	840	3136
18	38	324	684	1444
126	224	2376	3600	9338

$$\bar{x} = \frac{126}{7} = 18 \qquad \bar{y} = \frac{224}{7} = 32$$

$$SXX = 2376 - \frac{1}{7}(126)^2 = 108$$

$$SXY = 3600 - \frac{1}{7}(126)(224) = -432$$

$$SYY = 9338 - \frac{1}{7}(224)^2 = 2170$$

Putting it all together,…

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{-432}{108} = -4 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 32 - (-4)18 = 104$$

The least-squares regression line is $\hat{y} = 104 - 4x$.

b. Calculate the fitted values, $\hat{y}_i$.

c. Calculate the residuals, e_i. Does the sum of the residuals equal zero?

Fitted Values: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Residuals: $e = y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 x)$

The sum of the residuals equals 0 as we would expect.

x	y	(b) $\hat{y}$	(c) e	e^2
16	30	40	– 10	100
12	52	56	– 4	16
25	7	4	3	9
19	32	28	4	16
21	9	20	– 11	121
15	56	44	12	144
18	38	32	6	36
			0	442
				RSS

d. Give an estimate for σ, the standard deviation of the observations about the true regression line?

We need the residual sum of squares (*RSS*). It can be found by totaling the squares of the errors as seen in the table above. Or, recall that the total variation of Y comes from two sources: $SYY = SSReg + RSS$, or sometimes alternately written as $SST = SSR + SSE$.

$$SSR = \hat{\beta}_1^2 \cdot SXX = (-4)^2 \cdot 108 = 1728$$
$$RSS = SYY - SSR = 2170 - 1728 = 442$$

Often, the more common choice for estimating σ is the unbiased estimator,

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{442}{5}} = \sqrt{88.4} = 9.402.$$

Or, the maximum likelihood estimate of σ is another option,

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{442}{7}} = \sqrt{63.1} = 7.946.$$

e. What proportion of observed variation in TV viewing is explained by a straight-line relationship with physical activity?

This is answered by the coefficient of determination,

$$R^2 = \frac{SSR}{SYY} = 1 - \frac{RSS}{SYY} = 1 - \frac{442}{2170} = 0.796 = 79.6\%$$

f. Predict the number of TV viewing hours for a participant who engaged in 24 hours of physical activity in the same week.

The least-squares regression model predicts 8 hours of TV viewing in that week.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 104 - 4(24) = 8$$

a. Find the least-squares estimate, $\hat{\beta}$.

x	y	x^2	xy
3.6	28	12.96	100.8
4.2	24	17.64	100.8
5.4	32	29.16	172.8
3	13.6	9	40.8
4.8	36	23.04	172.8
6	44	36	264
27	177.6	127.8	852

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{852}{127.8} = 6.67$$

b. Calculate the fitted values, $\hat{y}_i$.

c. Calculate the residuals, e_i. Does the sum of the residuals equal zero?

Fitted Values: $\hat{y} = \hat{\beta} x$

Residuals: $e = y - \hat{y} = y - \hat{\beta} x$

		(b)	(c)
x	y	$\hat{y}$	e
3.6	28	24	4
4.2	24	28	–4
5.4	32	36	–4
3	13.6	20	–6.4
4.8	36	32	4
6	44	40	4
			–2.4

Note that the residuals do not add up to zero. Without the intercept (β_0) and the vector of all 1's in the model, the vector $\boldsymbol{e}$ of the residuals does not have to be orthogonal to it, so the residuals do not have to add up to zero.

You are given a random sample of six nights at a concert venue. We are looking for a relationship between the number of people who attended and the revenue earned in thousands of dollars (e.g., ticket sales, food, drinks).

Number of patrons, in thousands (x)	3.6	4.2	5.4	3	4.8	6
Revenue earned, in thousands of dollars (y)	28	24	32	13.6	36	44

Consider the model $Y_i = \beta x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$. Note that if no one shows up for a concert, no money will be made.

For parts a, b, and c, you are not to use R or any other software to solve the exercises. A calculator is allowed.

a. Find the least-squares estimate, $\hat{\beta}$.

b. Calculate the fitted values, $\hat{y}_i$.

c. Calculate the residuals, e_i. Does the sum of the residuals equal zero?

Meerkats

The standing heights of adult male meerkats (not including the tail) is normally distributed with a mean of 11.4 inches with a standard deviation of 2.1 inches. Female adult meerkats have heights which are normally distributed with a mean of 11.2 inches with a standard deviation of 1.9 inches.

a. Suppose that an adult male and an adult female meerkat are chosen independently at random from the wild. What is the probability that the male is taller than the female?

A group of meerkats is known as a mob. Suppose that in one particular mob of meerkats, the correlation between the heights of the adult males and females is 0.38. Suppose also that all of the summary statistics for the population of all meerkats given above also holds here.

b. Suppose that an adult male and an adult female meerkat are chosen at random from this mob. What is the probability that the male is taller than the female?

Stocks

The daily stock prices of two companies in the same industry sector vary randomly according to a bivariate normal distribution. Because they are in the same sector, it's not surprising that their corporate values tend to fluctuate in similar pattern over a long run and have positive correlation. Based on 2014 data, Exxon Mobil Corp. (X) and BP Oil (Y) generated the following (approximate) parameter values for price per share:

$$\mu_X = \$95,\ \sigma_X = \$4.5,\ \mu_Y = \$46,\ \sigma_Y = \$3.0,\ \rho = 0.64$$

a. What is the probability that on a given day the price of stock for BP (Y) exceeds \$50?

b. Suppose that on a given day the price of stock for Exxon (X) is \$100. What is the probability that the price of stock for BP (Y) exceeds \$50?

c. Suppose you bought 4 shares of Exxon stock and 5 shares of BP stock. What is the probability that on a given day the value of your portfolio ($4X + 5Y$) is below \$600?

d. What is the probability that, on a randomly selected day, 1 share of Exxon stock is worth less than 2 shares of BP stock?

Meerkats

a. Let M = height of a random adult male meerkat; F = height of a random adult female meerkat. We want $P(M > F) = P(M - F > 0)$ $P(M - F > 0)$, so find the distribution of $M - F$. Since M, F are both Normal, then $M - F$ is Normal with

$\mathrm{E}[M - F] = 11.4 - 11.2 = 0.2$,
$\mathrm{Var}[M - F] = (2.1)^2 + (1.9)^2 = 8.02$, and
$\mathrm{SD}[M - F] = (2.1)^2 + (1.9)^2 = 2.83$.

Thus,

$$P(M - F > 0) = P\left(Z > \frac{0 - 0.2}{2.83}\right) = P(Z > -0.07) = 1 - \Phi(-0.07) = 1 - 0.4721 = \mathbf{0.5279}.$$

Or, in R,

```
> 1 - pnorm(0, 0.2, 2.83)
[1] 0.5282
```

b. Let M = height of a random adult male meerkat; F = height of a random adult female meerkat. We want $P(M > F) = P(M - F > 0)$ $P(M - F > 0)$, so find the distribution of $M - F$. Since M, F are both Normal, then $M - F$ is Normal with

$\mathrm{E}[M - F] = 11.4 - 11.2 = 0.2$,
$\mathrm{Var}[M - F] = (2.1)^2 + (1.9)^2 + 2(1)(-1)(0.38)(2.1)(1.9) = 4.99$, and
$\mathrm{SD}[M - F] = (2.1)^2 + (1.9)^2 = 2.23$.

Thus,

$$P(M - F > 0) = P\left(Z > \frac{0 - 0.2}{2.23}\right) = P(Z > -0.09) = 1 - \Phi(-0.09) = 1 - 0.4641 = \mathbf{0.5359}.$$

Or, in R,

```
> 1 - pnorm(0, 0.2, 2.23)
[1] 0.5357
```

d. We want $P(X < 2Y) = P(X - 2Y < 0)$. Since $X - 2Y$ is a linear combination of Normal random variables, then $X - 2Y$ is Normal with

$$\mathrm{E}[X - 2Y] = a\mu_X + b\mu_Y = 1(95) - 2(46) = 3$$
$$\mathrm{Var}[X - 2Y] = \mathrm{Var}[X] - 2 \cdot \mathrm{Cov}[X, 2Y] + \mathrm{Var}[2Y]$$
$$= (4.5)^2 - 2 \cdot (1)(2)(0.64)(4.5)(3.0) + 2^2 \cdot (3.0)^2 = 21.69$$
$$\mathrm{SD}[X - 2Y] = 4.66$$

Thus,

$$P(X - 2Y < 0) = P\left(Z < \frac{0 - 3}{4.66}\right) = P(Z < -0.64) = \Phi(-0.64) = \mathbf{0.2611}.$$

Or, in R,

```
> pnorm(0, 3, 4.66)
[1] 0.2599
```

. **Stocks**

a. We want $P(Y > 50)$. Since Y is Normal with $\mathrm{E}[Y] = 46$ and $\mathrm{SD}[Y] = 3$, then

$$P(Y > 50) = P\left(Z > \frac{50 - 46}{3}\right) = P(Z > 1.33) = 1 - \Phi(1.33) = 1 - 0.9082 = \mathbf{0.0918}.$$

Or, in R,

```
> 1 - pnorm(50, 46, 3)
[1] 0.09121
```

b. We want $P(Y > 50 \mid X = 100)$. Y is Normal with

$$\mathrm{E}[Y \mid X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 46 + 0.64 \cdot \frac{3}{4.5}(100 - 95) = 48.13,$$
$$\mathrm{Var}[Y \mid X] = (1 - \rho^2)\sigma_Y^2 = (1 - 0.64^2)3^2 = 5.31, \text{ and}$$
$$\mathrm{SD}[Y \mid X] = 2.31.$$

Thus,

$$P(Y > 50 \mid X = 100) = P\left(Z > \frac{50 - 48.13}{2.31}\right) = P(Z > 0.81) = 1 - 0.7910 = \mathbf{0.2090}.$$

Or, in R,

```
> 1 - pnorm(50, 48.13, 2.31)
[1] 0.2091
```

c. We want $P(4X + 5Y < 600)$. Since $4X + 5Y$ is a linear combination of Normal random variables, then $4X + 5Y$ is Normal with

$$\mathrm{E}[4X + 5Y] = a\mu_X + b\mu_Y = 4(95) + 5(46) = 610$$
$$\mathrm{Var}[4X + 5Y] = \mathrm{Var}[4X] + 2 \cdot \mathrm{Cov}[4X, 5Y] + \mathrm{Var}[5Y]$$
$$= 4^2 \cdot (4.5)^2 + 2 \cdot (4)(5)(0.64)(4.5)(3.0) + 5^2 \cdot (3.0)^2 = 894.6$$
$$\mathrm{SD}[4X + 5Y] = 29.91$$

Thus,

$$P(4X + 5Y < 600) = P\left(Z < \frac{600 - 610}{29.91}\right) = P(Z < -0.33) = \Phi(-0.33) = \mathbf{0.3707}.$$

Or, in R,

```
> pnorm(600, 610, 29.91)
[1] 0.3691
```

3. Multivariate Normal

Suppose $\boldsymbol{X}$ follows a 3-dimensional multivariate normal distribution with mean $\boldsymbol{\mu} = \begin{bmatrix} 20 \\ 30 \\ 25 \end{bmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 12 & 8 & -6 \\ 8 & 9 & -6 \\ -6 & -6 & 25 \end{bmatrix}$.

a. Find $P(X_2 > 32)$.
b. Find $P(X_3 > 32)$.
c. Find $P(X_1 + X_3 > 50)$.
d. Find $P(X_1 - X_3 > 0)$.
e. Find $P(X_1 + 2X_2 + 3X_3 > 200)$.

4. Inference

Assume that the distributions of X and Y are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively. Given $n = 6$ observations of X,

39, 54, 59, 69, 79, 84

and $m = 4$ observations of Y,

23, 48, 53, 68

we would like to test test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, assuming the population variances are equal. Here is some helpful R code to get started:

```
x <- c(39, 54, 59, 69, 79, 84)
y <- c(23, 48, 53, 68)

mean(x); var(x)
# [1] 64
# [1] 280

mean(y); var(y)
# [1] 48
# [1] 350
```

You may complete the problem using R, or by hand using the given summary statistics.
a. Calculate the value of the test statistic.
b. Calculate the degrees of freedom for the test.
c. Calculate the p-value for the test. (You may give a range if not using R.)
d. State your conclusion for the hypothesis test at a 5% significance level.

Multivariate Normal

Suppose $\boldsymbol{X}$ follows a 3-dimensional multivariate normal distribution with mean $\boldsymbol{\mu} = \begin{bmatrix} 20 \\ 30 \\ 25 \end{bmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 12 & 8 & -6 \\ 8 & 9 & -6 \\ -6 & -6 & 25 \end{bmatrix}$.

a. Find $P(X_2 > 32)$.

X_2 ~ N(30, 9)

$$P(X_2 > 32) = P\left(Z > \frac{32-30}{3}\right) = P(Z > 0.67) = 1 - \Phi(0.67) = 1 - 0.7486 = \mathbf{0.2514}$$

Or, in R,

```
> 1 - pnorm(32, 30, 3)
[1] 0.2525
```

b. Find $P(X_3 > 32)$.

X_3 ~ N(25, 25)

$$P(X_3 > 32) = P\left(Z > \frac{32-25}{5}\right) = P(Z > 1.40) = 1 - \Phi(1.40) = 1 - 0.9192 = \mathbf{0.0808}$$

Or, in R,

```
> 1 - pnorm(32, 25, 5)
[1] 0.08076
```

$$P(X_1 + 2X_2 + 3X_3 > 200) = P\left(Z > \frac{200-155}{14.04}\right) = P(Z > 3.21) = 1 - 0.9993 = \mathbf{0.0007}$$

Or, in R,

```
> 1 - pnorm(200, 155, 14.04)
[1] 0.00068
```

Inference

a. First, calculate the pooled variance.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{(6-1)280 + (4-1)350}{6+4-2} = 306.25$$

Then, calculate the test statistic.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{64-48}{\sqrt{306.25\left(\frac{1}{6} + \frac{1}{4}\right)}} = 1.416$$

Or, in R,

```
> tt <- t.test(x, y, alternative=c("two.sided"), var.equal=T)
> tt$statistic
    t
1.416
```

b. $df = n_1 + n_2 - 2 = 6 + 4 - 2 = 8$

Or, in R,

```
> tt$parameter
df
 8
```

c. p-value = $2 \cdot P(T > 1.416)$ since the alternative is two-sided.

From the t-table where $df = 8$, we see that $0.05 < P(T > 1.416) < 0.10$, thus $0.10 < p\text{-value} < 0.20$.

Or, in R,

```
> tt$p.value
[1] 0.1944
```

c. Find $P(X_1 + X_3 > 50)$.

$$\mathrm{E}[X_1 + X_3] = 20 + 25 = 45$$

$$\mathrm{Var}[X_1 + X_3] = \mathrm{Var}[X_1] + 2 \cdot \mathrm{Cov}[X_1, X_3] + \mathrm{Var}[X_3] = 12 + 2(-6) + 25 = 25, \text{ or...}$$

$$\mathrm{Var}[X_1 + X_3] = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 12 & 8 & -6 \\ 8 & 9 & -6 \\ -6 & -6 & 25 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 19 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = 25$$

$$P(X_1 + X_3 > 50) = P\left(Z > \frac{50-45}{5}\right) = P(Z > 1) = 1 - \Phi(1.00) = 1 - 0.8413 = \mathbf{0.1597}$$

Or, in R,

```
> 1 - pnorm(50, 45, 5)
[1] 0.1587
```

d. Find $P(X_1 - X_3 > 0)$.

$$\mathrm{E}[X_1 - X_3] = 20 - 25 = -5$$

$$\mathrm{Var}[X_1 - X_3] = \mathrm{Var}[X_1] - 2 \cdot \mathrm{Cov}[X_1, X_3] + \mathrm{Var}[X_3] = 12 - 2(-6) + 25 = 49, \text{ or...}$$

$$\mathrm{Var}[X_1 - X_3] = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 12 & 8 & -6 \\ 8 & 9 & -6 \\ -6 & -6 & 25 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 18 & 14 & -31 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = 49$$

$$P(X_1 - X_3 > 0) = P\left(Z > \frac{0-(-5)}{7}\right) = P(Z > 0.71) = 1 - \Phi(0.71) = 1 - 0.7611 = \mathbf{0.2389}$$

Or, in R,

```
> 1 - pnorm(0, -5, 7)
[1] 0.2375
```

e. Find $P(X_1 + 2X_2 + 3X_3 > 200)$.

$$\mathrm{E}[X_1 + 2X_2 + 3X_3] = 20 + 2 \cdot 30 + 3 \cdot 25 = 155$$

$$\mathrm{Var}[X_1 + 2X_2 + 3X_3] = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 12 & 8 & -6 \\ 8 & 9 & -6 \\ -6 & -6 & 25 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 & 8 & 57 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 197$$

$$\mathrm{SD}[X_1 + 2X_2 + 3X_3] = 14.04$$

d. Since the p-value > 0.05, our conclusion is to <u>not</u> reject the null hypothesis. There is <u>not</u> enough evidence to suggest that the means are significantly different.

And here is the complete R output.

```
> tt

   Two Sample t-test

data:  x and y
t = 1.416, df = 8, p-value = 0.1944
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -10.05  42.05
sample estimates:
mean of x mean of y
       64        48
```

1. The number of points a student earns on an exam is often thought to be determined by how prepared the student is. For $n = 10$ students, the following values have been recorded.

y = Final Exam points,
x_1 = number of absences,
x_2 = average number of hours spent studying per week

x_1	x_2	y
1	1	20
1	3	29
2	7	43
2	1	6
3	10	75
3	8	66
4	14	89
4	10	66
5	12	71
5	14	95

Consider the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where ε's are i.i.d. $N(0, \sigma^2)$.

Then $\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 30 & 80 \\ 30 & 110 & 300 \\ 80 & 300 & 860 \end{bmatrix}$, $(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix}$,

$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 560 \\ 2{,}020 \\ 5{,}780 \end{bmatrix}$, and $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix}$, $\Sigma(y_i - \hat{y}_i)^2 = 350$, and $\Sigma(y_i - \bar{y})^2 = 8{,}090$.

a) (12) Perform the significance of the regression test at the 10% level of significance.

$H_0: \beta_1 = \beta_2 = 0$ vs. H_1: at least one $\beta_j \neq 0$

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	7740	$p - 1 = 2$	3870	**77.40**
Error	350	$n - p = 7$	50	
Total	8090	$n - 1 = 9$		

The critical region is $F > F_\alpha(2,7) = F_{0.10}(2,7) =$ **3.26**.
With a calculator, you get p-value = **1.68E-05**.
As a result, **we reject H_0**; the model is a significant model for predicting Final Exam score.

1. (continued) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix}$, $(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix}$.

d) (6) Construct a 95% prediction interval for the final exam score of a student who missed 3 days of class and studied an average of 12 hours per week.

The vector representing these predictors is $\boldsymbol{x}_0 = [\,1 \;\; 3 \;\; 12\,]$. The estimate for the average wait time is

$$\hat{y} = [\,1 \;\; 3 \;\; 12\,] \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix} = 12 - 4(3) + 7(12) = 84.$$

To calculate the estimate for the variance of the estimate, we need

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = [\,1 \;\; 3 \;\; 12\,] \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 12 \end{bmatrix} = [\,1 \;\; 3 \;\; 12\,] \begin{bmatrix} 0.2 \\ -0.3 \\ 0.1 \end{bmatrix} = 0.5$$

So, the 95% prediction interval is

$$\hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}[Y \mid x]} = 84 \pm t_{0.025}(7) \cdot \sqrt{\hat{\sigma}^2 \cdot (1 + 0.5)} = 84 \pm 2.365 \cdot \sqrt{50 \cdot 1.5}$$
$$= \mathbf{84 \pm 20.48}$$
$$= \mathbf{(63.52,\ 104.48)}$$

e) (3) Interpret β_0 in the context of the problem.

β_0 represents the average final exam score of students who did not miss any classes but who also did not do any studying.

f) (3) Interpret β_1 in the context of the problem.

β_1 represents the average change in the final exam score for each additional absence from class (while holding study hours constant)

1. (continued) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 12 \\ -4 \\ 7 \end{bmatrix}$, $(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.575 & -0.225 & 0.025 \\ -0.225 & 0.275 & -0.075 \\ 0.025 & -0.075 & 0.025 \end{bmatrix}$.

b) (7) Test $H_0: \beta_2 = 0$ vs. $H_1: \beta_2 \neq 0$ at the 5% level of significance.

Use MSE = 50 from part a as the estimate of the variance of the residuals.

$$\hat{\mathrm{V}}\mathrm{ar}[\hat{\beta}_2] = \hat{\sigma}^2 \cdot C_{33} = (50)(0.025) = 1.25$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}[\hat{\beta}_2]}} = \frac{7 - 0}{\sqrt{1.25}} = \mathbf{6.261}$$

There are $n - p = 7$ degrees of freedom. The critical region is $|t| > t_{\alpha/2}(n-p) = t_{0.025}(7) =$ **2.365**.
With a calculator, you get p-value = **0.00042**.
As a result, **we reject H_0**; β_2 is a significant predictor in the model.

c) (5) Construct a 90% confidence interval for β_1.

Use MSE = 50 from part a as the estimate of the variance of the residuals.

$$\hat{\mathrm{V}}\mathrm{ar}[\hat{\beta}_1] = \hat{\sigma}^2 \cdot C_{22} = (50)(0.275) = 13.75$$

So, the 90% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}[\hat{\beta}_1]} = -4 \pm t_{0.05}(7) \cdot \sqrt{13.75} = -4 \pm 1.895 \cdot 3.708$$
$$= \mathbf{-4 \pm 7.03}$$
$$= \mathbf{(-11.03,\ 3.03)}$$

2. (16) Suppose a complete second-order model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

was fit to $n = 32$ data points.

```
> sum( lm( y ~ 1 )$residuals^2 )
[1]  600

> summary( lm( y ~ x2 + x4 + x6 ) )$r.squared
[1]  0.65

> summary( lm( y ~ x1 + x3 + x5 + x7 ) )$r.squared
[1]  0.72

> sum( lm( y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)$residuals^2 )
[1]  150
```

Test $H_0: \beta_2 = \beta_4 = \beta_6 = 0$ at a 10% level of significance.
State the alternative hypothesis, the value of the test statistic, the critical value(s), and a decision.

$H_0: \beta_2 = \beta_4 = \beta_6 = 0$ is also represented by the Null Model of

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \varepsilon$$

H_1: at least one $\beta_j \neq 0$ is also represented by the Full Model of

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

For the Full Model, $df = n - p = 32 - 8 = 24$ and $\mathrm{SSE}_{\mathrm{Full}} = 150$.

For the Null Model, $df = n - q = 32 - 5 = 27$ but $\mathrm{SSE}_{\mathrm{Null}}$ is not given. However, $R^2 = 0.72$ for that model, so $R^2 = 1 - \frac{\mathrm{SSE}_{\mathrm{Null}}}{SYY} = 1 - \frac{\mathrm{SSE}_{\mathrm{Null}}}{600} = 0.72$. Thus, $\mathrm{SSE}_{\mathrm{Null}} = 168$.

	SS	df	MS	F
Difference	18	$p - q = 3$	6	**0.96**
Full Model	150	$n - p = 24$	6.25	
Null Model	168	$n - q = 27$		

The critical region is $F > F_\alpha(2,7) = F_{0.10}(2,7) =$ **2.33**.
With a calculator, you get p-value = **0.4276**.
As a result, **we fail to reject H_0**; the Null Model is better.

3. A vaccine is shipped by airfreight to medical facilities in cartons, each containing 1,000 vials. The data presented here concerns 10 such shipments.

y = number of broken vials at final destination

x = number of times the carton was transferred from one aircraft to another

x	y
1	16
0	9
2	17
0	12
3	22
1	13
0	8
1	15
2	19
0	11

Consider the model

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε's are i.i.d. $N(0, \sigma^2)$.

$\sum x = 10$; $\sum y = 142$; $\sum x^2 = 20$; $\sum y^2 = 2{,}194$; $\sum xy = 182$;

$\sum (x - \bar{x})^2 = 10$; $\sum (y - \bar{y})^2 = 177.6$; $\sum (x - \bar{x})(y - \bar{y}) = 40$.

a) (6) Find the equation of the least-squares regression line.

$$\bar{x} = \frac{\sum x}{n} = \frac{10}{10} = 1 \qquad \bar{y} = \frac{\sum y}{n} = \frac{142}{10} = 14.2$$

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{40}{10} = 4$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 14.2 - 4\,(1) = 10.2.$$

Least-squares regression line: $\hat{y} = \mathbf{10.2 + 4\,x}$

b) (12) Perform the significance of the regression test at the 5% level of significance.

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

We have to have SSE, so

SSReg = $\hat{\beta}_1^2 \cdot SXX = (4)^2 \cdot 10 = 160$

SSE = SYY – SSReg = 177.6 – 160 = 17.6

Solution A:

Completing the ANOVA table,

Source	SS	df	MS	*F*
Regression	160	$p - 1 = 1$	160	**72.73**
Error	17.6	$n - p = 8$	2.2	
Total	177.6	$n - 1 = 9$		

The critical region is $F > F_\alpha(1,8) = F_{0.05}(1,8) = \mathbf{5.32}$.
With a calculator, you get *p*-value = **2.74E-05**.
As a result, **we reject H_0**; the model is a significant model for predicting the number of broken vials.

Solution B:

The variance of the residuals is $s_e^2 = \frac{SSE}{n-2} = \frac{17.6}{8} = 2.2$.

Calculate the *t*-test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{4 - 0}{\sqrt{2.2}/\sqrt{10}} = \mathbf{8.528}$$

The critical region is $|t| > t_{\alpha/2}(8) = t_{0.025}(8) = \mathbf{2.306}$.
With a calculator, you get *p*-value = **2.74E-05**.
As a result, **we reject H_0**; the model is a significant model for predicting the number of broken vials.

c) (5) Construct a 95% prediction interval for the number of broken vials after a shipment that had 2 aircraft transfers.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}} = (10.2 + 4 \cdot 2) \pm t_{0.025}(8) \cdot 1.48 \cdot \sqrt{1 + \frac{1}{10} + \frac{(2-1)^2}{10}}$$

$$= 18.2 \pm 2.306 \cdot 1.48 \cdot \sqrt{1 + \frac{1}{10} + \frac{1}{10}}$$

$$= \mathbf{18.2 \pm 3.74}$$

$$= \mathbf{(14.46,\ 21.94)}$$

4. (6) Consider the following data set:

x_1	x_2	y
1	1	6
2	1	9
3	0	10
4	0	11
2	1	15
4	0	17
3	1	18
5	0	18

Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$
$$i = 1, \ldots, 8.$$

where ε_i's are i.i.d. $N(0, \sigma_e^2)$.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 8 & 24 & 4 \\ 24 & 84 & 8 \\ 4 & 8 & 4 \end{bmatrix}; \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 4.25 & -1 & -2.25 \\ -1 & 0.25 & 0.5 \\ -2.25 & 0.5 & 1.5 \end{bmatrix}; \quad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 104 \\ 340 \\ 48 \end{bmatrix}$$

.

Obtain the least-squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 4.25 & -1 & -2.25 \\ -1 & 0.25 & 0.5 \\ -2.25 & 0.5 & 1.5 \end{bmatrix} \begin{bmatrix} 104 \\ 340 \\ 48 \end{bmatrix} = \begin{bmatrix} \mathbf{-6} \\ \mathbf{5} \\ \mathbf{8} \end{bmatrix}$$

5. For this problem we will use a random sample of 35 vehicles from a data set provided by the Environmental Protection Agency regarding fuel economy in cars.

- y = mileage (in miles per gallon)
- x_1 = engine horsepower
- x_2 = top speed (in miles per hour)
- x_3 = vehicle weight (in hundreds of lbs.)

Consider the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$, and the following output from R.

```
> summary(fit.mpg)

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 172.7589    41.0736   4.206 0.000205 ***
x1            0.3459     0.1406   2.461 0.019633 *
x2           -1.1219     0.4210  -2.665 0.012112 *
x3           -1.7274     0.3504  -4.930 2.62e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.142 on [redacted] degrees of freedom
Multiple R-squared:  0.8436,    Adjusted R-squared:  0.8285
F-statistic: 55.75 on [redacted] and [redacted] DF,  p-value: [redacted]
```

a) (5) Construct a 95% confidence interval for β_2.

The *df* for the model is $n - p = 35 - 4 = 31$, so the 95% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{\alpha/2}(n-p) \cdot \mathrm{SE}\left[\hat{\beta}_2\right] = -1.1219 \pm t_{0.025}(31) \cdot 0.4210 = -1.1219 \pm 1.96 \cdot 0.4210$$

$$= \mathbf{-1.1219 \pm 0.8252}$$

$$= \mathbf{(-1.95, -0.30)}$$

You could also opt to use the more conservative $t_{0.025}(30) = 2.042$ in which case the margin of error would be 0.8597 and the CI would be (–1.98, –0.26).

b) (6) Perform the significance of the regression test at the 10% level of significance.

The test statistic of $F = \mathbf{55.75}$ is given.
The critical region is $F > F_\alpha(p-1, n-p) = F_{0.10}(3,31) = \mathbf{2.27}$.
With a calculator, you get *p*-value = **1.36E-12**.
As a result, **we reject H_0**; the model is a significant model for predicting mileage.

6. (8) Suppose the number of number of times a carton is transferred from one aircraft to another (X) and the number of broken vials upon delivery (Y) follow a bivariate normal distribution with

$$\mu_X = 1.5, \quad \sigma_X = 1, \quad \mu_Y = 15, \quad \sigma_Y = 4, \quad \rho = 0.60.$$

Suppose that a recent carton shipment makes 3 aircraft transfers. What is the probability that the carton contains no more than 20 broken vials?

We want $P(Y \le 20 \mid X = 3)$. Given that $X = 3$, Y is Normal with

$$\mathrm{E}[Y \mid X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = 15 + 0.60 \cdot \frac{4}{1}(3 - 1.5) = 18.6,$$
$$\mathrm{Var}[Y \mid X] = (1 - \rho^2)\sigma_Y^2 = (1 - 0.60^2)4^2 = 10.24, \text{ and}$$
$$\mathrm{SD}[Y \mid X] = 3.2.$$

Thus,

$$P(Y \le 20 \mid X = 3) = P\left(Z \le \frac{20 - 18.6}{3.2}\right) = P(Z \le 0.44) = \mathbf{0.6700}.$$

1. (continued)

Examine the following sum of squares calculations.

```
> sum( lm( y ~ v1 + v2 + x )$residuals^2 )
[1] 8.0
> sum( lm( y ~ v1 + v2 )$residuals^2)
[1] 14.0
> sum( lm( y ~ x + 0 )$residuals^2)
[1] 18.0
> sum( lm( y ~ x )$residuals^2)
[1] 11.0
> sum( lm( y ~ 1 )$residuals^2)
[1] 20.0
```

d) (14) We wish to test if the relationship between GPA and time spent studying is the same for all three departments. Perform the appropriate test at $\alpha = 0.10$. Based on the result of this test, which model (full or null) is preferred?

$H_0: \beta_1 = \beta_2 = 0$ is also represented by the Null Model of $Y = \beta_0 + \beta_3 x + \varepsilon$

H_1: at least one $\beta_j \neq 0$ is also represented by the Full Model of

$$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \varepsilon$$

For the Full Model, $df = n - p = 24 - 4 = 20$ and $\mathrm{SSE}_{\mathrm{Full}} = 8$.

For the Null Model, $df = n - q = 24 - 2 = 22$ and $\mathrm{SSE}_{\mathrm{Null}} = 11$.

	SS	df	MS	F
Difference	3	$p - q = 2$	1.5	**3.75**
Full Model	8	$n - p = 20$	0.4	
Null Model	11	$n - q = 22$		

The critical region is $F > F_\alpha(2,7) = F_{0.10}(2,20) = \mathbf{2.59}$.
With a calculator, you get p-value = **0.0414**.
As a result, **we reject H_0**; the Full Model is better.

1. In order to compare the average GPA for the students of three distinct departments at a university, eight students were randomly chosen from each of the three departments (Astronomy, Biology, Communication), the students' GPA (y) and the average time spent studying per week (x) in hours was recorded. Consider the model

$$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \varepsilon,$$

where $v_1 = 1$ if a student is from Astronomy, 0 otherwise;
$v_2 = 1$ if a student is from Biology, 0 otherwise.

To answer parts a-c, consider…

Astronomy:	$Y = \beta_0 + \beta_1$		$+ \beta_3 x + \varepsilon$	
Biology:	$Y = \beta_0$	$+ \beta_2$	$+ \beta_3 x + \varepsilon$	
Communication:	$Y = \beta_0$		$+ \beta_3 x + \varepsilon$	← base category

a) (3) Give an interpretation (in the context of the problem) to the regression coefficient β_0.

The average GPA of Communication majors who do not study ($x = 0$).

b) (3) Give an interpretation (in the context of the problem) to the regression coefficient β_1.

The difference between the average GPA of Astronomy majors and the average GPA of Communication majors who spend the same amount of time studying.

c) (3) Give an interpretation (in the context of the problem) to the regression coefficient β_3.

The change in the average GPA for one additional hour of studying per week (regardless of major).

. (continued)

e) (9) Compute the AIC values for the full and the null models from part d. If the AIC model selection criteria is used, which model (full or null) is preferred?

Using the actual formula,…

$$\mathrm{AIC}_{\mathrm{Full}} = 24 + 24\ln(2\pi) + 24\ln\left({}^{8.0}/_{24}\right) + 2 \times 4 = \mathbf{49.74}$$
$$\mathrm{AIC}_{\mathrm{Null}} = 24 + 24\ln(2\pi) + 24\ln\left({}^{11.0}/_{24}\right) + 2 \times 2 = \mathbf{53.39}$$

Or using the reduced formula, since $n + n\ln(2\pi)$ is constant for all models,…

$$\mathrm{AIC}_{\mathrm{Full}} = 24\ln\left({}^{8.0}/_{24}\right) + 2 \times 4 = \mathbf{-18.37}$$
$$\mathrm{AIC}_{\mathrm{Null}} = 24\ln\left({}^{11.0}/_{24}\right) + 2 \times 2 = \mathbf{-14.72}$$

The lower AIC is preferred, so in either case, we choose **the Full Model**.

f) (9) Compute the Adjusted R-squared values for the full and the null models from part d. If the Adjusted R-squared model selection criteria is used, which model (full or null) is preferred?

$$R^2_{\mathrm{Full}} = 1 - \frac{\mathrm{SSResidual}}{\mathrm{SYY}} = 1 - \frac{8.0}{20.0} = 0.60.$$
$$\text{Adjusted } R^2_{\mathrm{Full}} = 1 - \frac{n-1}{n-p}\cdot\left(1 - R^2\right) = 1 - \frac{23}{20}\cdot(1 - 0.60) = \mathbf{0.54}.$$

$$R^2_{\mathrm{Null}} = 1 - \frac{\mathrm{SSResidual}}{\mathrm{SYY}} = 1 - \frac{11.0}{20.0} = 0.45.$$
$$\text{Adjusted } R^2_{\mathrm{Null}} = 1 - \frac{n-1}{n-p}\cdot\left(1 - R^2\right) = 1 - \frac{23}{22}\cdot(1 - 0.45) = \mathbf{0.425}.$$

The higher Adjusted R^2 is preferred, so we choose **the Full Model**.

(continued)

Consider a new variable. Let $v_3 = 1$ if a student is from Communication, 0 otherwise.

g) (3) Find the Residual Sum of Squares for the model

$$Y = \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$$

This new model is the same as $Y = \beta_1 v_1 + \beta_2 v_2 + \varepsilon$ under the original setup. Therefore, the two models

$$Y = \beta_1 v_1 + \beta_2 v_2 + \varepsilon \qquad \text{and} \qquad Y = \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$$

would have the same predicted/fitted values, the same residuals, and the same Residual Sum of Squares.

```
> sum( lm( y ~ v1 + v2 )$residuals^2)
[1] 14.0
```

h) (4) Suppose we suspect that the rate (the slope) of the relationship between GPA and time spent studying may be different for the different departments. Suggest an appropriate model.

If you want different slopes, you need interaction terms.

$Y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 x + \beta_4 v_1 x + \beta_5 v_2 x + \varepsilon,$
Then,

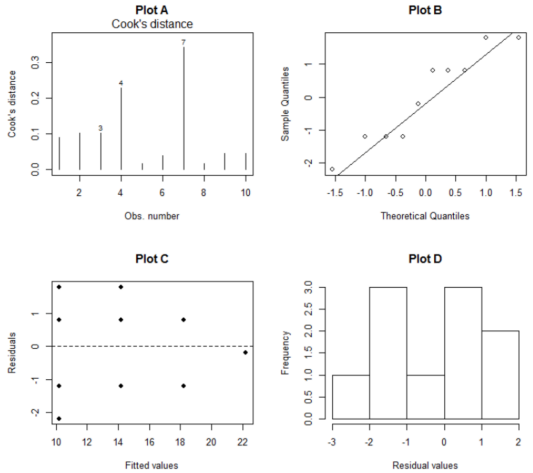
Astronomy:	$Y = \beta_0 + \beta_1 + (\beta_3 + \beta_4)x + \varepsilon$
Biology:	$Y = \beta_0 + \beta_2 + (\beta_3 + \beta_5)x + \varepsilon$
Communication:	$Y = \beta_0 + \beta_3 x + \varepsilon$

OR…
$Y = \mu_1 v_1 + \mu_2 v_2 + \mu_3 v_3 + \gamma_1 v_1 x + \gamma_2 v_2 x + \gamma_3 v_3 x + \varepsilon,$
Then,

Astronomy:	$Y = \mu_1 + \gamma_1 x + \varepsilon$
Biology:	$Y = \mu_2 + \gamma_2 x + \varepsilon$
Communication:	$Y = \mu_3 + \gamma_3 x + \varepsilon$

2. (continued)

b) (16) For each diagnostic plot below, (i) identify what model issues it is used to check and (ii) briefly explain whether this model is okay for each issue listed in (i) or if there's a problem.



Plot A: The Cook's Distance plot checks for any influential points. As long as all observations have Cook's Distance less than $4/n$ (which is $4/10 = 0.4$ here; or some use 0.5), then we're okay. This model does not seem to have any overly influential observations.

Plot B: The Normal Probability Plot (aka QQ Plot) checks the Normality assumption. The points generally follow the line, but we can't get a strong gauge on this because there's only 10 points.

Plot C: The residual plot is used to check the randomness of the error terms and constant variance. The points are generally scattered above and below the mean of 0, but the funnel pattern suggests a decrease in variance as the response values increase.

Plot D: The histogram of the residuals is used to check the Normality assumption. This plot doesn't appear to have much of a bell-shape, but with only 10 observations, it's hard to solidify behavior in the population.

i) (2) For your model in part h, we wish to test if the rate (the slope) of the relationship between GPA and time spent studying is the same for the three departments. Specify the null hypothesis H_0 using the notations of your part h model.

$H_0 : \beta_4 = \beta_5 = 0$ OR… $H_0 : \gamma_1 = \gamma_2 = \gamma_3$

. A vaccine is shipped by airfreight to medical facilities in cartons, each containing 1,000 vials. Consider the model: $Y = \beta_0 + \beta_1 x + \varepsilon,$

where y = number of broken vials at final destination;
x = number of times the carton was transferred from one aircraft to another.

a) (8) List the assumptions about the distribution of the error terms (ε) that must hold in order for the model to be considered valid.

We assume that the error terms are

- independent,
- (identically) Normally distributed with
- mean 0 and
- constant variance, σ^2.

2. (continued)

c) (6) There is something of particular concern regarding one of the model assumptions as seem in the plots of part b. Which of the following tests address that assumption? Do you still have a concern?

```
Shapiro-Wilk test

W = 0.9073, p-value = 0.2632
```

```
Durbin-Watson test

DW = 1.875, p-value = 0.4968
```

```
studentized Breusch-Pagan test

BP = 3.0628, df = 1, p-value = 0.0801
```

```
Levene's Test
      Df F value Pr(>F)
group  3  1.2667 0.3671
       6
```

```
Kolmogorov-Smirnov test

D = 0.2164, p-value = 0.7373
```

```
Bartlett test

Bartlett's K-squared = 0.4765, df = 1, p-value = 0.49
```

There are concerns about the constant variance assumption. The Breusch-Pagan test has a relatively low p-value which suggests that its null hypothesis (constant variance) does not hold.

If you were concerned about the normality test, look at the Shapiro-Wilk test. The p-value is rather large which suggests that the distribution can be explained with the Normal distribution.

[Bonus Knowledge: Levene's Test and the Bartlett Test also both check for constant variance. The Kolmogorov-Smirnov test checks for Normality. These each have such high p-values that we'd say the assumptions are met. However, the small sample size likely means that none of these tests are very reliable.

3. Apex Corporation, which makes corrugated paper products, is currently working to improve its cost control program. The company is analyzing its manufacturing costs to understand more fully the important influences on its costs. Monthly data has been assembled on a group of variables over the course of $n = 27$ months, and regression analysis is to be used to assess how these variables are related to total manufacturing costs. The variables are:

Y = total manufacturing cost per month in thousands of dollars (Cost)
X_1 = total production of paper per month in tons (Paper)
X_2 = total machine hours used per month (Machine)
X_3 = total overhead costs per month in thousands of dollars (Overhead).

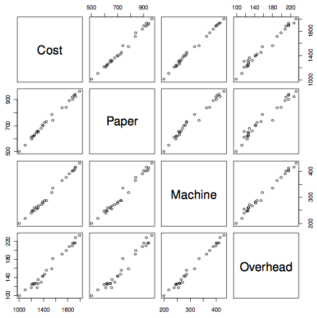
a) (4) Clearly state the first-order, multiple linear regression model for relating Cost to the predictors under consideration.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

b) (6) Consider the following scatterplot matrix and correlation matrix for the predictors and response variable. Comment on the relationship between the predictors and response, and among the predictors

	Cost	Paper	Machine	Overhead
Cost	1.000	0.996	0.997	0.989
Paper	0.996	1.000	0.989	0.978
Machine	0.997	0.989	1.000	0.994
Overhead	0.989	0.978	0.994	1.000

Each of the individual predictors is highly correlated with Cost suggesting that any one of them might be good for simple linear regression. However, the predictors are also highly correlated with one another suggesting collinearity, so it's unlikely that we'll find a significant multiple regression model.



3. (continued)

c) (7) Consider the following output and fill in the missing values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	60.35437	20.84177	2.896	0.0082
Paper	0.95630	0.12133	7.882	0.0000
Machine	2.32128	0.45561	5.095	0.0000
Overhead	0.08603	0.53067	0.162	0.8726

Residual standard error: 11.213
Multiple R-squared: 0.9987 Adjusted R-squared: 0.9986

Analysis of Variance Table

	Sum Sq	Df	Mean Sq	F value	Pr(>F)
Regression	2255666	3	751888.7	5979.75	0.0000
Residuals	2892	23	125.739		
Total	2258558	26			

$$\text{SE(error)} = \sqrt{\text{MSE}} = \sqrt{125.739} = 11.213$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{2892}{2258558} = 0.9987$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST} = 1 - \left(\frac{26}{23}\right)\frac{2892}{2258558} = 0.9986$$

d) (3) Given only the information and analysis contained thus far, clearly state what you feel is the best linear regression model for predicting Cost.

The explanatory variable of Overhead (x_3) is not significant in the full model based on the individual t-test. That coupled with the likelihood of collinearity means that we're justified to remove it from the full model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

1. Listed below are the price quotations of used cars along with their age and odometer mileage. A multiple linear regression analysis was performed using R, the output is given below.

	Age (years) X1	Mileage (thousand miles) X2	Price (thousand dollars) Y
1	1	8.1	9.45
2	2	17	8.4
3	2	12.6	8.6
4	3	18.4	6.8
5	3	19.5	6.5
6	4	29.2	5.6
7	6	40.4	4.75
8	7	51.6	3.89
9	8	62.6	2.7
10	10	80.1	1.47

```
> autos.fit <- lm(Y ~ X1 + X2)
> summary(autos.fit)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7390 -0.2545  0.1114  0.3066  0.5674

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  9.98543    0.34289  29.121  1.45e-08 ***
X1          -1.38474        ⑦        ⑧
X2           0.06481        ⑨        ⑩
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: ① on ② degrees of freedom
Multiple R-Squared: ③ ,     Adjusted R-squared: ④
F-statistic: ⑤ on ⑥1 and ⑥2 DF,  p-value:
```

1. (continued)

```
> X <- cbind(c(rep(1,10)), X1, X2)
> X
      [,1] [,2]  [,3]
 [1,]    1    1   8.1
 [2,]    1    2  17.0
 [3,]    1    2  12.6
 [4,]    1    3  18.4
 [5,]    1    3  19.5
 [6,]    1    4  29.2
 [7,]    1    6  40.4
 [8,]    1    7  51.6
 [9,]    1    8  62.6
[10,]    1   10  80.1

> solve(t(X) %*% X)
            [,1]       [,2]        [,3]
[1,]  0.47166883 -0.3716589  0.03940979
[2,] -0.37165893  0.9238623 -0.11422997
[3,]  0.03940979 -0.1142300  0.01431659

> sum(autos.fit$residuals^2)                    # SSResid
[1] 1.744894

> sum((Y-mean(Y))^2)                            # SYY
[1] 62.55944
```

Answers:

1. a) ②. $= n - p = 10 - 3 = \mathbf{7}$.

$$s^2 = \frac{\text{SSResid}}{n-p} = \frac{1.744894}{7} = 0.24927.$$

① $= s = \sqrt{0.24927} = \mathbf{0.49927}$.

b) ③ $= R^2 = 1 - \frac{\text{SSResid}}{\text{SYY}} = 1 - \frac{1.744894}{62.55944} = \mathbf{0.9721}$.

④ $= R^2_{Adjusted} = 1 - \left(\frac{n-1}{n-p}\right) \cdot \left(1 - R^2\right) = 1 - \frac{9}{7} \cdot (1 - 0.9721) = \mathbf{0.96414}$.

c)

Source	*SS*	*df*	*MS*	*F*
Regression	60.814546	⑥$_1 = 2$	30.407273	⑤ = **121.985**
Residuals	1.744894	⑥$_2 = 7$	0.24927	
Total	62.55944	9		

$F_{0.01}(2, 7) = 9.55$. **Reject** $H_0: \beta_1 = \beta_2 = 0$ at $\alpha = 0.01$.

d) $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.

$\hat{\text{Var}}(\hat{\beta}_1) = 0.24927 \times 0.9238623 = 0.2303$.

⑦ = S.E.$(\hat{\beta}_1) = \sqrt{0.2303} = \mathbf{0.47989}$.

Test Statistic: ⑧ $= t = \frac{-1.38474 - 0}{0.47989} = \mathbf{-2.8855}$.

Rejection Region: $t < -t_{0.025}(7) = -2.365$ or $t > t_{0.025}(7) = 2.365$.

Reject H_0 at $\alpha = 0.05$.

e) $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$.

$\hat{\text{Var}}(\hat{\beta}_2) = 0.24927 \times 0.01431659 = 0.0035687$.

⑨ = S.E.$(\hat{\beta}_2) = \sqrt{0.0035687} = \mathbf{0.05974}$.

Test Statistic: ⑩ $= t = \frac{0.06481 - 0}{0.05974} = \mathbf{1.0849}$.

Rejection Region: $t < -t_{0.025}(7) = -2.365$ or $t > t_{0.025}(7) = 2.365$.

Do NOT Reject H_0 at $\alpha = 0.05$.

f) $H_0: \beta_1 = -2$ vs. $H_a: \beta_1 > -2$.

Test Statistic: $t = \frac{-1.38474 - (-2)}{0.47989} = \mathbf{1.282}$.

Rejection Region: $t > t_{0.05}(7) = 1.895$.

Do NOT Reject H_0 at $\alpha = 0.05$.

a) Adjusted *R*-squared $= 1 - \frac{n-1}{n-p} \cdot \left(1 - R^2\right) = 1 - \frac{19}{15} \cdot (1 - 0.8528) \approx \mathbf{0.813547}$.

b) We remove the predictor (does not include `(Intercept)`) with highest p-value greater than α_{crit}. Therefore, the next step is to **remove `X4` `(HSenglish)`**:

$$\texttt{GPA} = \beta_0 + \beta_1\,\texttt{SATmath} + \beta_2\,\texttt{SATverbal} + \beta_3\,\texttt{HSmath} + \varepsilon.$$

We will then refit the model and remove the remaining least significant predictor provided its p-value is greater than α_{crit}.

c)

```
> drop1(GPA.fit)
Single term deletions

Model:
GPA ~ SATmath + SATverbal + HSmath + HSenglish
          Df Sum of Sq     RSS     AIC
<none>                  1.081 -48.348
SATmath    1     0.853   1.934 -38.718
SATverbal  1     0.372   1.453 -44.440
HSmath     1     0.307   1.388 -45.356
HSenglish  1     0.018   1.099 -50.022       <-- lowest
```

If the AIC model selection criteria is used, can the model be improved? If so, how? Explain.

Want a model with lowest AIC value. Therefore, we can improve the model by **dropping `X4` `(HSenglish)`**.

$$\texttt{GPA} = \beta_0 + \beta_1\,\texttt{SATmath} + \beta_2\,\texttt{SATverbal} + \beta_3\,\texttt{HSmath} + \varepsilon.$$

4. Suppose the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

was fit to $n = 20$ data points, R command `drop1` was applied, and the following results were obtained:

```
> fit = lm(Y ~ X1 + X2 + X3 + X4)
> drop1(fit)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
          Df   Sum of Sq      RSS    AIC
<none>                     25.681   ____
X1         1       5.685   ____    ____
X2         1       7.293   ____    ____
X3         1       1.316   ____    ____
X4         1       8.984   ____    ____
```

a) Fill in the missing AIC values.

b) If the AIC variable selection criteria is used, can the model be improved? If so, how (what is the next step)? Explain.

4. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ $\quad n = 20$

```
> fit = lm(Y ~ X1 + X2 + X3 + X4)
> drop1(fit)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
           Df   Sum of Sq                      RSS     AIC
<none>                                      25.681
X1          1       5.685   25.681+5.685 = 31.366
X2          1       7.293   25.681+7.293 = 32.974
X3          1       1.316   25.681+1.316 = 26.997
X4          1       8.984   25.681+8.984 = 34.665
```

a) R: $\text{AIC} = n \ln\left(\text{RSS}/n\right) + 2p.$

<none> 25.681

None of the variables have been dropped. `Y ~ X1 + X2 + X3 + X4`

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

$p = 5.$

$$\text{AIC} = 20 \ln\left(25.681/20\right) + 2 \times 5 = \mathbf{15}.$$

X1 1 5.685 **31.366**

X1 has been dropped. `Y ~ X2 + X3 + X4`

$Y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

$p = 4.$

$$\text{AIC} = 20 \ln\left(31.366/20\right) + 2 \times 4 = \mathbf{17}.$$

X2 1 7.293 **32.974**

X2 has been dropped. `Y ~ X1 + X3 + X4`

$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

$p = 4.$

$$\text{AIC} = 20 \ln\left(32.974/20\right) + 2 \times 4 = \mathbf{18}.$$

X3 1 1.316 **26.997**

X3 has been dropped. `Y ~ X1 + X2 + X4`

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$

$p = 4.$

$$\text{AIC} = 20 \ln\left(26.997/20\right) + 2 \times 4 = \mathbf{14}.$$

X4 1 8.984 **34.665**

X4 has been dropped. `Y ~ X1 + X2 + X3`

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

$p = 4.$

$$\text{AIC} = 20 \ln\left(34.665/20\right) + 2 \times 4 = \mathbf{19}.$$

```
> drop1(fit)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
           Df   Sum of Sq      RSS    AIC
<none>                      25.681    15
X1          1       5.685   31.366    17
X2          1       7.293   32.974    18
X3          1       1.316   26.997    14   <- lowest
X4          1       8.984   34.665    19
```

b) Want a model with **lowest** AIC value. Therefore, we can improve the model by **dropping X3**:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$$

n) Find the values of $R^2_{Adjusted}$ for both models. Which model is "better"?

n)		linear	**quadratic**
	$R^2_{Adjusted}$	0.56345	**0.92143**

o) Find the values of the Akaike's Information Criterion (AIC) for both models. Which model is "better"?

o)		linear	**quadratic**
	AIC	43.933	**35.331**
	AIC (R)	29.744	**21.142**

1. Full model: $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \beta_4 \vec{\mathbf{x}} + \vec{\mathbf{e}}$

$n = 27$ $\quad \dim(\mathrm{V}) = 4$

$\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$

Null model: $\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}} + \beta_4 \vec{\mathbf{x}} + \vec{\mathbf{e}}$

$\dim(\mathrm{V}_0) = 2$

```
> sum(lm(y ~ 1)$residuals^2)
[1] 240
> sum(lm(y ~ x + 0)$residuals^2)
[1] 1047.374
> sum(lm(y ~ x)$residuals^2)                    <- null model
[1] 117.3333
> sum(lm(y ~ v1 + v2 + v3 + 0)$residuals^2)
[1] 126
> sum(lm(y ~ v1 + v2 + v3 + x + 0)$residuals^2)  <- full model
[1] 27.59630
```

	SS	*DF*	*MS*	*F*
Diff.	$\mathrm{RSS}_{\text{null}} - \mathrm{RSS}_{\text{full}}$	$\dim(\mathrm{V}) - \dim(\mathrm{V}_0)$	…	…
Full	$\mathrm{RSS}_{\text{full}}$	$n - \dim(\mathrm{V})$	…	
Null	$\mathrm{RSS}_{\text{null}}$	$n - \dim(\mathrm{V}_0)$		

	SS	*DF*	*MS*	*F*	
Diff.	89.7370	2	44.8685	**37.3954**	← Test Statistic
Full	27.5963	23	1.19984		
Null	117.3333	25			

Critical Value: $\mathrm{F}_{0.05}(2, 23) = \mathbf{3.42}$. Decision: **Reject H_0.**

c) Find the C_p value for the model

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \vec{\boldsymbol{\varepsilon}}.$$

$$C_p = \frac{\mathrm{SSResid}_{\text{Null}}}{\mathrm{MSResid}_{\text{Full}}} - n + 2 \cdot (\#\text{ of parameters in Null}) = \frac{184}{8} - 25 + 2 \cdot (5) = \mathbf{8}.$$

[Want models with C_p close to or less than (# of parameters in Null).]

d) Compute the AIC values for the full model and the model from part (c). Which model is preferred?

Full: $\mathrm{AIC} = 25 \ln\left({}^{144}/_{25}\right) + 2 \times 7 \approx \mathbf{57.77}$.

Part (c): $\mathrm{AIC} = 25 \ln\left({}^{184}/_{25}\right) + 2 \times 5 \approx \mathbf{59.90}$.

OR

Full: $\mathrm{AIC} = 25 + 25 \ln(2\pi) + 25 \ln\left({}^{144}/_{25}\right) + 2 \times 7 \approx \mathbf{111.39}$.

Part (c): $\mathrm{AIC} = 25 + 25 \ln(2\pi) + 25 \ln\left({}^{184}/_{25}\right) + 2 \times 5 \approx \mathbf{113.52}$.

The **full model** is preferred since the full model has lower AIC value.

2. **Do NOT** use a computer for this problem.

A student is studying for Exam P (the first Actuarial Science exam). He claims that drinking beer has no effect on the amount of time it takes for him to solve a practice problem. The following data show how many seconds he took to solve a problem after consuming various quantities of beer, measured in ounces:

Beer Consumption, x	Time, y
0	141
12	127
24	141
36	163
48	145
60	179
72	161

$\Sigma x = 252$, $\Sigma y = 1{,}057$, $\Sigma x^2 = 13{,}104$, $\Sigma y^2 = 161{,}447$, $\Sigma xy = 40{,}068$,

$\Sigma (x - \bar{x})^2 = 4032$, $\Sigma (y - \bar{y})^2 = 1{,}840$, $\Sigma (x - \bar{x})(y - \bar{y}) = \Sigma (x - \bar{x}) y = 2{,}016$.

Consider the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i's are i.i.d. $\mathrm{N}(0, \sigma^2)$.

a) Find the equation of the least-squares regression line.

b) Calculate the residuals e_i. Does the sum of the residuals equal zero?

c) Give an estimate for σ, the standard deviation of the observations about the true regression line?

d) What proportion of observed variation in time needed to solve a practice problem is explained by a straight-line relationship with the amount of beer consumed?

e) How much time would you expect the student to need to solve a practice problem after consuming 156 ounces of beer.

f) Explain why it may be dangerous to predict the time needed to solve a practice problem for the amount of beer consumed in part (e).

g) Use the F-test to test $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$ (the significance of the regression test) at a 5% significance level. Report the value of the test statistic, critical value(s), and decision.

h) Construct a 95% confidence interval for β_1. Does your answer for part (h) agree with your answer for part (g)?

i) Test the student's claim at a 5% level of significance. That is, test $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 > 0$ at a 5% level of significance. Report the value of the test statistic, critical value(s), and decision. Does the square of your test statistic in part (i) equal your test statistic from part (g)?

j) The student believes that when he is not drinking beer, it takes him on average two minutes to solve a practice problem. What is the p-value (approximately) of the test $\mathrm{H}_0 : \beta_0 = 120$ vs. $\mathrm{H}_1 : \beta_0 \neq 120$?

k) Construct a 95% confidence interval for β_0.

l) Construct a 95% confidence interval for σ^2.

m) Construct 95% limits of prediction for the time the student needs to solve a practice problem after consuming 156 ounces of beer.

n) Construct 90% confidence interval for the average time the student needs to solve a practice problem after consuming 156 ounces of beer.

a) Find the equation of the least-squares regression line.

$$\bar{x} = \frac{\sum x}{n} = \frac{252}{7} = 36. \qquad \bar{y} = \frac{\sum y}{n} = \frac{1057}{7} = 151.$$

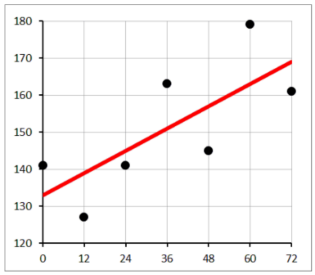
$SXX = 4032, \qquad SXY = 2016.$

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{2016}{4032} = \mathbf{0.5}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 151 - 0.5 \cdot 36 = \mathbf{133}.$$

Least-squares regression line:

$$\hat{y} = \mathbf{133 + 0.5}\,x.$$



b) Calculate the residuals e_i. Does the sum of the residuals equal zero?

x	y	$\hat{y}$	$e = y - \hat{y}$	e^2
0	141	133	**8**	64
12	127	139	**– 12**	144
24	141	145	**– 4**	16
36	163	151	**12**	144
48	145	157	**– 12**	144
60	179	163	**16**	256
72	161	169	**– 8**	64
Sum:			0	832
				SSResid

The sum of the residuals does equal zero.

c) Give an estimate for σ, the standard deviation of the observations about the true regression line?

$$s_e^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 = \frac{SSResid}{n-2} = \frac{832}{5} = 166.4. \qquad s_e = \sqrt{166.4} = \mathbf{12.9}.$$

OR

$$\hat{\sigma}^2 = \frac{1}{n}\sum (y_i - \hat{y}_i)^2 = \frac{SSResid}{n} = \frac{832}{7} = 118.857. \qquad \hat{\sigma} = \sqrt{118.857} = \mathbf{10.9}.$$

Rejection Region:

Reject H_0 if $F > F_{0.05}(1,5)$ $\qquad F_{0.05}(1,5) = \mathbf{6.61}$.

The Test Statistic F is NOT in the Rejection Region.

Do NOT Reject H_0 at $\alpha = 0.05$.

d) What proportion of observed variation in time needed to solve a practice problem is explained by a straight-line relationship with the amount of beer consumed?

Need $R^2 = ?$ (coefficient of determination)

$$R^2 = 1 - \frac{SSResid}{SYY} = 1 - \frac{832}{1840} = \mathbf{0.547826}. \qquad \mathbf{54.7826\%}.$$

e) How much time would you expect the student to need to solve a practice problem after consuming 156 ounces of beer.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 133 + 0.5 \cdot 156 = \mathbf{211} \text{ sec.}$$

f) Explain why it may be dangerous to predict the time needed to solve a practice problem for the amount of beer consumed in part (e).

Extrapolation.

$x = 156$ oz is outside of the observed data range. We do not have any guarantee that the relationship would stay the same for larger values of x.

g) Use the F-test to test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ (the significance of the regression test) at a 5% significance level. Report the value of the test statistic, critical value(s), and decision.

SSRegr = SYY – RSS = 1840 – 832 = 1008.

OR $\qquad SSRegr = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 = \hat{\beta}_1^2 SXX = 0.5^2 \times 4032 = 1008.$

Source	SS	DF	MS	F
Regression	1008	1	1008	**6.0577**
Residuals	832	$n - 2 = 5$	166.4	
Total	1840	$n - 1 = 6$		

h) Construct a 95% confidence interval for β_1. Does your answer for part (h) agree with your answer for part (g)?

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{SXX}} \qquad 7 - 2 = 5 \text{ degrees of freedom,} \qquad t_{0.025} = 2.571.$$

$$0.5 \pm 2.571 \cdot \frac{12.9}{\sqrt{4032}} \qquad \mathbf{0.5 \pm 0.5223} \qquad \mathbf{(-0.0223, 1.0223)}$$

95% confidence interval for β_1 covers 0 $\Leftrightarrow$ Do NOT Reject $H_0: \beta_1 = 0$ at $\alpha = 0.05$.

i) Test the student's claim at a 5% level of significance. That is, test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 > 0$ at a 5% level of significance. Report the value of the test statistic, critical value(s), and decision. Does the square of your test statistic in part (i) equal your test statistic from part (g)?

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{0.5 - 0}{12.9 / \sqrt{4032}} = \mathbf{2.46124}. \qquad n - 2 = 5 \text{ degrees of freedom.}$$

Rejection Region: $T > t_\alpha(n-2)$. $\qquad \alpha = 0.05 \qquad t_{0.05}(5) = \mathbf{2.015}$.

The Test Statistic T IS in the Rejection Region. **Reject H_0** at $\alpha = 0.05$.

p-value = right tail. $\qquad 0.025 < \text{p-value} < 0.05.$

(p-value ≈ 0.02857.)

Indeed, $T^2 = F$.

j) The student believes that when he is not drinking beer, it takes him on average two minutes to solve a practice problem. What is the p-value (approximately) of the test $H_0: \beta_0 = 120$ vs. $H_1: \beta_0 \neq 120$?

$$T = \frac{\hat{\beta}_0 - \beta_{00}}{s_e \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{SXX}}} = \frac{133 - 120}{12.9 \cdot \sqrt{\frac{1}{7} + \frac{(36)^2}{4032}}} = 1.479.$$

$n - 2 = 5$ degrees of freedom.

p-value = 2 tails. $0.05 <$ left tail < 0.10. right tail ≈ 0.10.

$0.10 <$ p-value < 0.20. p-value $\approx$ **0.20**.

(p-value ≈ 0.1992.)

k) Construct a 95% confidence interval for β_0.

$$\hat{\beta}_0 \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}$$ 5 degrees of freedom, $t_{0.025} = 2.571$.

$$133 \pm 2.571 \cdot 12.9 \cdot \sqrt{\frac{1}{7} + \frac{36^2}{4032}}$$ **133 ± 22.6**

(110.4, 155.6)

l) Construct a 95% confidence interval for σ^2.

$$\left(\frac{(n-2)s_e^2}{\chi^2_{\alpha/2}}, \frac{(n-2)s_e^2}{\chi^2_{1-\alpha/2}} \right)$$ $\chi^2_{0.025}(5\text{ df}) = 12.83$, $\chi^2_{0.975}(5\text{ df}) = 0.831$.

$$\left(\frac{5 \cdot 166.4}{12.83}, \frac{5 \cdot 166.4}{0.831} \right)$$ **(64.848, 1001.203)**

m) Construct 95% limits of prediction for the time the student needs to solve a practice problem after consuming 156 ounces of beer.

Limits of prediction for y: $\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$

$\hat{y} = \hat{\alpha} + \hat{\beta} x = 133 + 0.5 \cdot 156 = 211$ sec.

$\alpha = 0.05$. $\alpha/2 = 0.025$. $n - 2 = 5$ degrees of freedom. $t_{0.025}(5) = 2.571$.

$$211 \pm 2.571 \cdot 12.9 \cdot \sqrt{1 + \frac{1}{7} + \frac{(156-36)^2}{4032}}$$ **211 ± 72** **(139, 283)**

n) Construct 90% confidence interval for the average time the student needs to solve a practice problem after consuming 156 ounces of beer.

Confidence interval for $\mu(x)$: $\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$

$\hat{y} = \hat{\alpha} + \hat{\beta} x = 133 + 0.5 \cdot 156 = 211$ sec.

$\alpha = 0.10$. $\alpha/2 = 0.05$. $n - 2 = 5$ degrees of freedom. $t_{0.05}(5) = 2.015$.

$$211 \pm 2.015 \cdot 12.9 \cdot \sqrt{\frac{1}{7} + \frac{(156-36)^2}{4032}}$$ **211 ± 50.1** **(160.9, 261.1)**

1. **Do NOT use a computer for this problem.**

An Anytown State University student wishes to examine the relationship between the monthly rent amount for an apartment (y) (in $) and the number of bedrooms (x). The data are as follows:

x	y
2	567
2	728
2	754
3	808
2	841
2	978
4	1,127
3	1,145
4	1,190
3	1,197
4	1,264
3	1,284
4	1,301
3	1,310
4	1,416
3	1,434

Consider the model

$$Y = \beta_0 + \beta_1 x + \varepsilon, \qquad i = 1, 2, \ldots, 16.$$

where ε's are i.i.d. $N(0, \sigma^2)$.

$$\Sigma x = 48, \quad \Sigma y = 17{,}344,$$
$$\Sigma x^2 = 154, \quad \Sigma y^2 = 19{,}883{,}146,$$
$$\Sigma xy = 54{,}462,$$
$$\Sigma (x - \bar{x})^2 = 10, \quad \Sigma (y - \bar{y})^2 = 1{,}082{,}250,$$
$$\Sigma (x - \bar{x})(y - \bar{y}) = \Sigma (x - \bar{x}) y = 2{,}430.$$

a) Find the equation of the least-squares regression line.

b) Perform the significance of the regression test at a 5% level of significance. Specify the null and the alternative hypotheses. Report the value of the test statistic, the critical value(s), and the decision.

c) Test $H_0: \beta_1 = 202$ vs. $H_1: \beta_1 > 202$ at a 5% level of significance. Report the value of the test statistic, the p-value, and the decision.

d) Construct a 95% confidence interval for the average monthly rent amount for an apartment that has 2 bedrooms.

e) What is the p-value of the test $H_0: \beta_0 = 0$ vs. $H_1: \beta_0 \neq 0$? (You may give a range.)

Find the equation of the least-squares regression line.

$\bar{x} = \frac{\Sigma x}{n} = \frac{48}{16} = 3.$ $\bar{y} = \frac{\Sigma y}{n} = \frac{17{,}344}{16} = 1{,}084.$

SXX = 10, SXY = 2,430. $\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{2{,}430}{10} =$ **243**.

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 1{,}084 - 243 \cdot 3 =$ **355**.

Least-squares regression line: $\hat{y} =$ **355 + 243** x.

Perform the significance of the regression test at a 5% level of significance. Specify the null and the alternative hypotheses. Report the value of the test statistic, the critical value(s), and the decision.

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$.

SSRegression = $\hat{\beta}_1^2 \times SXX = 243^2 \times 10 = 590{,}490$.

RSS = SYY − SSRegression = 1,082,250 − 590,490 = 491,760.

Source	*SS*	*df*	*MS*	*F*
Regression	590,490	2 − 1 = 1	590,490	**16.81**
Residuals	491,760	$n - 2 = 14$	35,125.7143	
Total	1,082,250	$n - 1 = 15$		

Critical Value: $F_{0.05}(1, 14) =$ **4.60**.

Decision: **Reject H_0** at $\alpha = 0.05$.

c) Test $H_0: \beta_1 = 202$ vs. $H_1: \beta_1 > 202$ at a 5% level of significance.
Report the value of the test statistic, the p-value, and the decision.

Test Statistic:

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{s/\sqrt{SXX}} = \frac{243 - 202}{35{,}125.7143/\sqrt{10}} \approx \mathbf{0.692}.$$

Rejection Region:

Reject H_0 if $T > t_{0.05}(14\text{ df}) = 1.761$.

Do NOT Reject H_0 at $\alpha = 0.05$.

OR

P-value $\approx$ **0.25** $> 0.05 = \alpha$.

Do NOT Reject H_0 at $\alpha = 0.05$.

d) Construct a 95% confidence interval for the average monthly rent amount for an apartment that has 2 bedrooms.

Confidence interval for $\mu(x)$: $\hat{y} \pm t_{\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 355 + 243 \cdot 2 = 841$.

$n - 2 = 14$ degrees of freedom. $t_{0.025}(14) = 2.145$.

$$841 \pm 2.145 \cdot \sqrt{35{,}125.7143} \cdot \sqrt{\frac{1}{16} + \frac{(2-3)^2}{10}}$$

841 ± 162

(679, 1,003)

e) What is the p-value of the test $H_0: \beta_0 = 0$ vs. $H_1: \beta_0 \neq 0$?
(You may give a range.)

$$T = \frac{\hat{\beta}_0 - \beta_{00}}{s_e \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{SXX}}} = \frac{355 - 0}{\sqrt{35{,}125.7143} \cdot \sqrt{\frac{1}{16} + \frac{3^2}{10}}} = \mathbf{1.9307}.$$

$n - 2 = 14$ degrees of freedom.

$t_{0.05}(14) = 1.761 < T < t_{0.025}(14) = 2.145$.

$0.025 <$ one tail < 0.05.

Two – tailed. P-value = two tails.

0.05 < p-value < 0.10.

a) Find the sample correlation coefficient r.

1.8	1,284
1.7	1,301
2.0	1,310
2.2	1,416
1.8	1,434

$$r = \frac{295}{\sqrt{0.5}\sqrt{1{,}082{,}250}} \approx \mathbf{0.4010}.$$

```
cor(x1,y)
[1] 0.4010266
```

b) Find the p-value of the test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$.
Use $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ and either EXCEL TDIST or R pt.

Test Statistic: $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.4010 \cdot \sqrt{16-2}}{\sqrt{1-0.4010^2}} \approx \mathbf{1.638}$.

```
2*(1-pt(1.638,16-2))
[1] 0.123693
```

=TDIST(1.638,16-2,2) **0.123693**

c) Find the (approximate) p-value of the test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$.
Use $W = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$.

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.4010}{1-0.4010}\right) \approx 0.42484.$$

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = \frac{1}{2}\cdot\ln\left(\frac{1+0}{1-0}\right) = 0$,

$\sigma_W^2 = \frac{1}{n-3} = \frac{1}{13}$.

Test Statistic: $Z = \frac{W - \mu_W}{\sigma_W} = \frac{0.42484 - 0}{\sqrt{1/13}} \approx \mathbf{1.53}$.

P-value = two tails $= 2 \times P(Z > 1.53) = 2\ 0.0630 = \mathbf{0.1260}$.

d) Find the (approximate) p-value of the test $H_0: \rho = 0.20$ vs. $H_1: \rho > 0.20$.

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = \frac{1}{2}\cdot\ln\left(\frac{1+0.20}{1-0.20}\right) \approx 0.20273$,

$\sigma_W^2 = \frac{1}{n-3} = \frac{1}{13}$.

Test Statistic: $Z = \frac{W - \mu_W}{\sigma_W} = \frac{0.42484 - 0.20273}{\sqrt{1/13}} \approx \mathbf{0.80}$.

P-value = right tail $= P(Z > 0.80) = \mathbf{0.2119}$.

e) Construct a 90% confidence interval for ρ.

$100(1-\alpha)$ % confidence interval for ρ:

$$\left(\frac{e^a - 1}{e^a + 1}, \frac{e^b - 1}{e^b + 1} \right), \quad \text{where} \quad a = \ln\frac{1+r}{1-r} - \frac{2 z_{\alpha/2}}{\sqrt{n-3}}, \quad b = \ln\frac{1+r}{1-r} + \frac{2 z_{\alpha/2}}{\sqrt{n-3}}.$$

$$a = \ln\left(\frac{1+0.4010}{1-0.4010}\right) - \frac{2\times 1.645}{\sqrt{16-3}} \approx -0.06274.$$

$$b = \ln\left(\frac{1+0.4010}{1-0.4010}\right) + \frac{2\times 1.645}{\sqrt{16-3}} \approx 1.76223.$$

$$\left(\frac{e^a - 1}{e^a + 1}, \frac{e^b - 1}{e^b + 1} \right) \approx \mathbf{(-0.03136, 0.70698)}.$$