DAMT-5301-Spring 2019
HW6- Solution

11.1 We are given: $n = 30$, $\bar{x} = 126$, $s_x = 35$, $\bar{y} = 0.04$, $s_y = 0.01$, and $r = 0.86$.

Compute the least squares estimates,

$$\begin{aligned} b_1 &= r\left(\frac{s_y}{s_x}\right) = (0.86)\left(\frac{0.01}{35}\right) = 0.000246 \\ b_0 &= \bar{y} - b_1\bar{x} = 0.04 - (0.000246)(126) = 0.009. \end{aligned}$$

The fitted regression line has an equation

$$\boxed{y = 0.009 + 0.000246x}$$

The time it takes to transmit a 400 Kbyte file is predicted as

$$\hat{y}_* = 0.009 + 0.000246x_* = 0.009 + (0.000246)(400) = \boxed{0.107 \text{ seconds}}$$

11.2 Here $n = 75$, $\bar{x} = 32.2$, $s_x^2 = 6.4$, $\bar{y} = 8.4$, $s_y^2 = 2.8$, and $s_{xy} = 3.6$.

(a) Compute the least squares estimates

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2} = \frac{3.6}{6.4} = 0.5625 \\ b_0 &= \bar{y} - b_1\bar{x} = 8.4 - (0.5625)(32.2) = -9.7125. \end{aligned}$$

The sample regression line is

$$\boxed{y = -9.7125 + 0.5625x}$$

(b) Compute the sums of squares,

$$\begin{aligned} SS_{\text{TOT}} &= (n-1)s_y^2 = (75-1)(2.8) = 207.2 \\ SS_{\text{REG}} &= b_1^2 S_{xx} = b_1^2 s_x^2 (n-1) = (0.5625)^2(6.4)(75-1) = 149.85 \\ SS_{\text{ERR}} &= SS_{\text{TOT}} - SS_{\text{REG}} = 57.35 \end{aligned}$$

Also, compute degrees of freedom

$$\mathrm{df}_{\mathrm{TOT}} = n - 1 = 74, \quad \mathrm{df}_{\mathrm{REG}} = 1, \quad \text{and} \quad \mathrm{df}_{\mathrm{ERR}} = n - 2 = 73$$

and the mean squares

$$MS_{\mathrm{REG}} = \frac{SS_{\mathrm{REG}}}{\mathrm{df}_{\mathrm{REG}}} = 149.85, \quad MS_{\mathrm{ERR}} = \frac{SS_{\mathrm{ERR}}}{\mathrm{df}_{\mathrm{ERR}}} = \frac{57.35}{73} = 0.7856.$$

Finally, we compute the F-ratio

$$F = \frac{MS_{\mathrm{REG}}}{MS_{\mathrm{ERR}}} = 190.75$$

and find from Table A7 (1 and 73 d.f.) that it is significant at the 0.1% level.
Complete the ANOVA table:

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	149.85	1	149.5	190.75
Error	57.35	73	0.7856	
Total	207.2	74		

Predictor X can explain

$$R^2 = SS_{\mathrm{REG}}/SS_{\mathrm{TOT}} = \boxed{0.7232 \text{ or } 72.32\%}$$

of the total variation.

(c) The 99% confidence interval for β_1 is

$$\begin{aligned} b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}} &= b_1 \pm t_{0.005} \frac{\sqrt{MS_{\mathrm{ERR}}}}{\sqrt{(n-1)s_x^2}} \\ &= 0.5625 \pm (2.648) \frac{\sqrt{0.7856}}{\sqrt{(74)(6.4)}} \\ &= \boxed{0.5625 \pm 0.1078 \text{ or } [0.4547,\ 0.6703]} \end{aligned}$$

where $t_{0.005}$ is obtained from Table A5 with 73 d.f.
This interval does not contain 0, therefore, the slope is significant (significantly different from 0) at a 1% level of significance.

11.5 (a) In this multivariate regression analysis,

$$X = \begin{pmatrix} 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 1 \\ 1 & 6 & 0 \\ 1 & 7 & 1 \\ 1 & 8 & 1 \\ 1 & 9 & 1 \\ 1 & 10 & 1 \\ 1 & 11 & 0 \\ 1 & 12 & 1 \\ 1 & 13 & 1 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} 17 \\ 23 \\ 31 \\ 29 \\ 33 \\ 39 \\ 39 \\ 40 \\ 41 \\ 44 \\ 47 \end{pmatrix}$$

The 3rd column of the design matrix $\boldsymbol{X}$ is the dummy variable Z that equals 1 if the company reports profit and 0 otherwise. In the 2nd column of $\boldsymbol{X}$, we again defined

$$X = \text{year} - 2000,$$

to simplify the calculations.
We then compute

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 11 & 88 & 7 \\ 88 & 814 & 64 \\ 7 & 64 & 7 \end{pmatrix},$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 0.6742 & -0.0707 & -0.0278 \\ -0.0707 & 0.0118 & -0.0370 \\ -0.0278 & -0.0370 & 0.5093 \end{pmatrix},$$

and

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{pmatrix} 13.3586 \\ 2.3569 \\ 4.0926 \end{pmatrix}$$

That is,

$$\boxed{b_0 = 13.3586, \quad b_1 = 2.3569, \quad b_2 = 4.0926}$$

(b) The estimated regression equation is

$$\hat{y} = 13.3586 + 2.3569x + 4.0926z$$

For a company reporting profit ($z_* = 1$) in 2015 (that is, $x_* = 15$), we predict the investment amount as

$$\hat{y}_* = 13.3586 + 2.3569(15) + 4.0926(1) = \boxed{52.8047 \text{ thousand dollars}}$$

(c) The slope β_2 is the change in the response variable when the dummy variable z changes from 0 to 1. Thus, if the company reports a loss during year 2007 instead of a gain, its expected investment amount reduces by β_2. Our prediction will decrease by $b_2 = 4.0926$ thousand dollars.

(d) The total sum of squares

$$SS_{\text{TOT}} = S_{yy} = \sum(Y_i - \bar{Y})^2 = 841.64$$

is computed in the previous exercise.
The error sum of squares can be computed, say, by filling the table,

Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
17	20.4293	−3.4293	11.7600
23	22.7862	0.2138	0.0457
31	29.2357	1.7643	3.1128
29	27.5000	1.5000	2.2500
33	33.9495	−0.9495	0.9015
39	36.3064	2.6936	7.2555
39	38.6633	0.3367	0.1134
40	41.0202	−1.0202	1.0408
41	39.2845	1.7155	2.9429
44	45.7340	−1.7340	3.0068
47	48.0909	−1.0909	1.1901

Alternatively, one can multiply matrices,

$$SS_{\text{ERR}} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

Then

$$SS_{\text{ERR}} = \sum_i (Y_i - \hat{Y}_i)^2 = 33.62$$

and

$$SS_{\text{REG}} = SS_{\text{TOT}} - SS_{\text{ERR}} = 841.64 - 33.62 = 808.02$$

Complete the ANOVA table:

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	808.02	2	404.01	96.14
Error	33.62	8	4.20	
Total	841.64	10		

Comparing $F_{obs} = 96.14$ against Table A7 with 2 and 8 d.f., we find that the model is significant at the 0.1% level (in fact, its P-value is less than 0.0001).

(e) From Exercise 11.4,

$$SS_{\text{ERR}}(\textit{Reduced}) = (MS_{\text{ERR}})(\text{df}_{\text{ERR}}) = (9)(7.39) = 66.51 \quad (9 \text{ d.f.})$$

This error sum of squares is obtained from the reduced model where investment is predicted based on the year only.
For the full model of Exercise 11.5,

$$SS_{\text{ERR}}(\textit{Full}) = 33.62 \quad (8 \text{ d.f.})$$

Hence, the new variable explains additional

$$\frac{SS_{\text{EX}}}{SS_{\text{TOT}}} \cdot 100\% = \frac{6.51 - 33.62}{841.64} \cdot 100\% = \underline{3.9\%}$$

of the total variation.

Significance of the new dummy variable (reporting profit) in addition to the time trend is tested by the partial F-statistic

$$F_{\text{obs}} = \frac{SS_{\text{EX}}/\text{df}_{\text{EX}}}{MS_{\text{ERR}}(Full)} = \frac{(SS_{\text{ERR}}(Reduced) - SS_{\text{ERR}}(Full))/(9-8)}{MS_{\text{ERR}}(Full)}$$

$$= \frac{6.51 - 33.62}{4.20} = \underline{7.83}.$$

From Table A7 with 1 and 8 d.f., addition of the new variable is significant at the 2.5% but not at the 1% level. The P-value of this test is between 0.01 and 0.025.

11.9 From the data in Example 11.10 on p. 390, we can obtain:

$$n = 7,\ \bar{x}_2 = 9.57,\ s^2_{x_2} = 63.29,\ \bar{y} = 35,\ s^2_y = 242,\ r_{x_2 y} = 0.758$$

Compute the least squares estimates

$$\begin{aligned} b_1 &= r\sqrt{\frac{s^2_y}{s^2_{x_2}}} = (0.758)\sqrt{\frac{242}{63.29}} = 1.48 \\ b_0 &= \bar{y} - b_1\bar{x}_2 = 35 - (1.48)(9.57) = 20.84. \end{aligned}$$

The fitted regression line has the equation

$$\boxed{\hat{y} = 20.84 + 1.48x_2}$$

The coefficient of determination is

$$R^2 = r^2 = \boxed{0.575}$$

showing that 57.5% of the total variation of the number of processed requests is explained by the number of tables only.

Next, compute sums of squares,

$$\begin{aligned} SS_{\text{TOT}} &= (n-1)s^2_y = (6)(242) = 1452 \quad (n-1 = 6 \text{ d.f.}) \\ SS_{\text{REG}} &= R^2 SS_{\text{TOT}} = (0.575)(1452) = 834.9 \quad (1 \text{ d.f.}) \\ SS_{\text{ERR}} &= SS_{\text{TOT}} - SS_{\text{REG}} = 1452 - 834.9 = 617.1 \quad (n-2 = 5 \text{ d.f.}) \end{aligned}$$

Based on this, the adjusted R-square is

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/\text{df}_{\text{ERR}}}{SS_{\text{TOT}}/\text{df}_{\text{TOT}}} = 1 - \frac{617.1/5}{1452/6} = \boxed{0.51}$$

which is lower than the adjusted R-square of the full model ($R^2_{\text{adj}} = 0.68$) in Example 11.10. According to the adjusted R-square criterion, the full model is better.

Complete the ANOVA table,

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	834.9	1	834.9	6.77
Error	617.1	5	123.4	
Total	1452	6		

This F-statistic (6.77) is just a little higher than $F_{0.05} = 6.61$ from Table A7 with 1 and 5 d.f. Therefore, the P-value is just below 0.05, and this reduced model predicting the number of processed requests based on the number of tables is significant at the 5% level.