MATH 356H Assignment #2 Solutions

1. Chapter 12, #32

   The null hypothesis is $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The test statistic has a t distribution with 13 degrees of freedom. The observed value of 22.64 has a P-value of 0.000 under that distribution, so the null hypothesis is rejected for all practical levels of α (note that the precise value of P is 8×10^{-12}, absolutely significant).

   The 95% confidence interval for β_1 is $\hat{\beta}_1 \pm t_{.025,13} S/S_{xx}$, which is $(0.748067661, 0.905878634)$. Again, our conclusion that there is a useful linear relationship is supported because the interval does not include 0.

2. Chapter 12, #37

   (a) $n = 10$, $\sum x_i = 2615$, $\sum y_i 39.20$, $\sum x_i^2 = 860,675$, $\sum y_i^2 = 161.94$. $\sum x_i y_i = 11,453.5$, so

   $$\begin{aligned}\hat{\beta}_1 &= \frac{12,027}{1,768,525} = .00680058 \\ \hat{\beta}_0 &= 2.14164770\end{aligned}$$

   Thus

   $$\begin{aligned}SSE &= .09696713 \\ S^2 &= .09696713/8 = .0121208913 \\ S &= \sqrt{.0121208913} = .1100949193 \\ S_{\hat{\beta}_1} &= .1100949193/\sqrt{176,852} = .000262\end{aligned}$$

   Of course, all this detail is unnecessary when using `R`. From the summary of the linear model for time vs. pressure, we obtain the standard error of the pressure coefficient to be 0.0002618 and the estimate of σ to be (Residual standard error) 0.1101, which agree with the above results.

   (b) The null hypothesis is $H_0 : \beta_1 = .006$ against $H_1 : \beta_1 \neq .006$. The test statistic is

   $$t = \frac{.00680058 - .006}{.000262} = 3.0556.$$

   At $\alpha = .1$, the right critical value is $t_{.05,8} = 1.860$, so we reject H_0. The P-value satisfies $.005 < P < .01$ so although rejection is significant here, it would not be for absolutely all practical levels of α.

3. A student, working on a summer internship in the economic research office of a large corporation, studied the relation between sales of a product (Y, in millions of dollars) and population (X, in millions of persons) in the firm's 50 marketing districts. The normal error regression model was employed. The student first wished to test whether or not a linear association between Y and X existed. The following is part of the output he obtained:

Parameter	Estimated value	95 percent Confidence limits	
Intercept	7.43119	-1.18518	16.0476
Slope	.755048	.452886	1.05721

   (a) The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

   Yes, the conclusion is warranted. The confidence interval for the slope does not contain 0, so at .05 level of significance, the slope is significantly different than 0.

   (b) Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss briefly.

   The value of $X = 0$ is not within the scope of the model so it is not important if the interval contains negative numbers. When 0 is not within the scope of the model, the intercept coefficient has no interpretation.

4. Chapter 12, #48.

(a) $S_{xx} = 18.24 - \dfrac{(12.6)^2}{9} = 0.6, \quad S_{xy} = 40.968 - \dfrac{(12.6)(27.68)}{9} = 2.216;$
$S_{yy} = 93.3448 - \dfrac{(27.68)^2}{9} = 8.213; \quad \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{2.216}{.6} = 3.693;$
$\hat{\beta}_0 = \dfrac{\sum y}{n} - \hat{\beta}_1 \dfrac{\sum x}{n} = \dfrac{27.68 - (3.693)(12.6)}{9} = -2.095$, so the point estimate is
$\hat{y}_{1.5} = -2.095 + 3.693(1.5) = 3.445$. Also, $SSE = 8.213 - 3.693(2.216) = .0293$, which yields $s = \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{.0293}{7}} = .0647$. Thus
$s_{\hat{y}_{1.5}} = .0647\sqrt{\dfrac{1}{9} + \dfrac{(1.5-1.4)^2}{.6}} = .0231$. Hence a 95% confidence interval for $\mu_{Y \cdot 1.5}$ is
$3.445 \pm 2.365(.0231) = 3.445 \pm .055 = (3.390, 3.5)$.

(b) A 95% prediction interval for Y when $X = 1.5$ is
$3.445 \pm 2.365\sqrt{(.0647)^2 + (.0231)^2} = 3.445 \pm .162 = (3.283, 3.607)$. The prediction interval for a future Y value is wider than the confidence interval for an average value of Y when $X = 1.5$.

(c) A prediction interval for Y when $X = 1.2$ would be wider than that for $X = 1.5$ since 1.2 is farther away from the mean of $\bar{x} = 1.4$.

5. Chapter 12, #49.

The estimator of the expected lead content is the midpoint of the interval, which is 529.9. At 95% confidence with 8 degrees of freedom, the critical value is $t_{.025,8} = 2.306$. Hence
$529.9 + (2.306)(s_{\hat{Y}_{15}}) = 597.7$, from where $s_{\hat{Y}_{15}} = 29.402$, and a 99% CI for the same quantity is
$529.9 \pm (3.355)(29.402) = (431.3, 628.5)$.

6. You work for a cellular phone industry analyst and gather the data shown in the following table:

Number of subscribers (in millions) x	Average monthly bill (in dollars) y
1.2	96.83
2.1	98.02
3.5	89.03
5.3	80.9
7.6	72.74
11.0	68.68
16.0	61.48
24.1	56.21
33.8	51
44	47.7
55.3	42.78
69.2	39.43
86.0	41.24

(a) Draw a scatter plot of the cellular phone data. Using the scatter plot, what type of correlation (negative or positive), if any, do you think the data have?

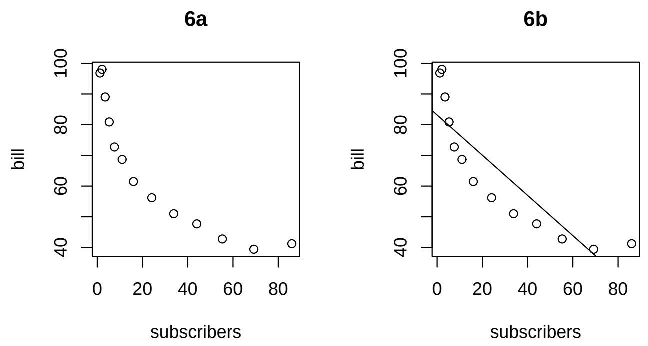
The data seem to have a negative correlation, since as X grows, Y decreases.

(b) Find an equation of the regression line of the data. Graph the regression line together with your scatter plot.

According to the output, the equation is $\hat{y} = 83.1704 - 0.6549x$.

(c) Use the regression line to obtain a 95% confidence interval for the true average monthly bill when $x = 25$ million subscribers.

The point estimate for the average monthly bill at $x = 25$ is $83.17036748 + 25(-0.654900521) = 66.7978$. Moreover,
$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 19382.93 - (359.1)^2/13 = 9463.483077$, so that

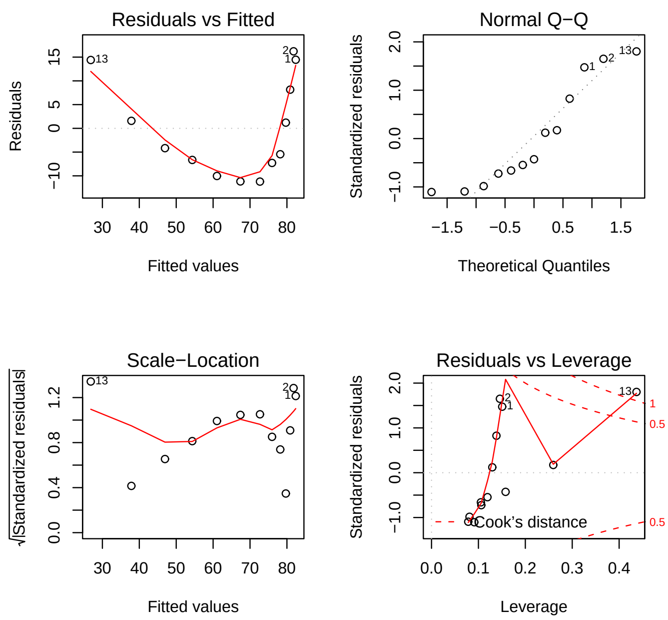


$s_{\hat{Y}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 10.63991\sqrt{\frac{1}{13} + \frac{(25 - 27.6231)^2}{9463.483077}} = 2.96489$ and hence a 95% CI for the true average at $x = 25$ is
$66.7978 \pm 2.201(2.96489) = (60.272077, 73.323522)$.

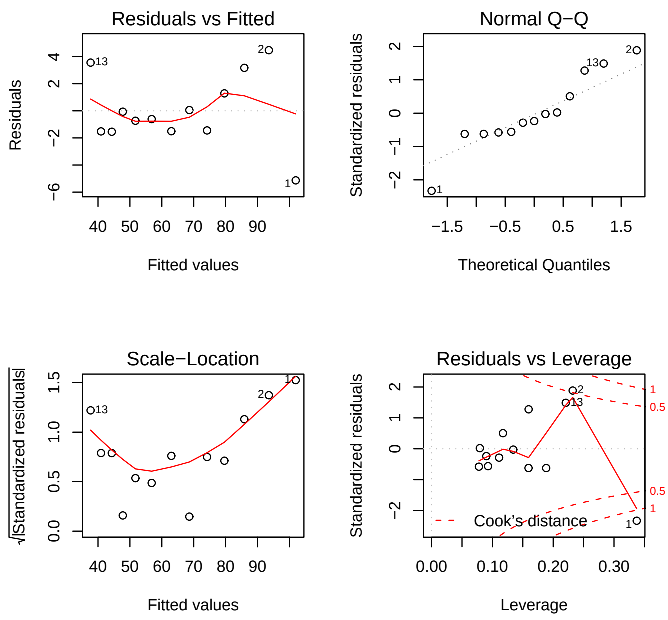
(d) The analyst wants to use the regression line you found to predict the average monthly bill for $x = 140$ million subscribers. Is a 95% prediction interval valid? If so, obtain it.
140 million subscribers is not within the scope of the model, so a prediction interval at that level is not valid.

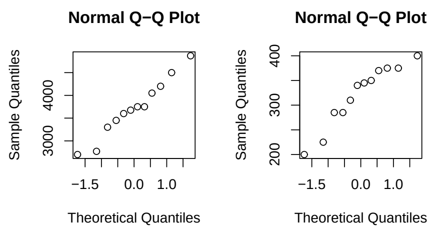
(e) The analyst claims that the data have a significant correlation at $\alpha = .01$. Verify this claim.
A test of significance of correlation is equivalent to the test for model validity on the slope. With a P-value of 9.04×10^{-5}, we can reject the null hypothesis of zero slope, and hence the hypothesis of no correlation. The analyst's claim is supported by the evidence.

(f) Plot the residuals against the independent variable. Does your model seem appropriate? If not, fit a model that does seem appropriate.



The residual plot has a curved pattern. Therefore the model is not appropriate. From the scatter plot it is clear that a linear model is not the correct fit, and the normal probability plot for residuals does not indicate normality either.
We use the transformation $X' = \ln(X)$, which gives a high R^2 value of .9848, with a standard error of 2.71. The diagnostic plots are better than the previous model, indicating a better model.





7. Chapter 12, #58

All of the answers relate to the following `R` output:

```
Pearson's product-moment correlation
data:  TOST and RBOT
t = 7.5939, df = 10, p-value = 1.853e-05
alternative hypothesis:  true correlation is not equal to 0
95 percent confidence interval:  0.7427808 0.9785966
sample estimates:
cor = 0.9231564
```

(a) The correlation coefficient is .9231, which indicates a strong positive linear relationship between TOST time and RBOT time.

(b) Changing the values of x and y does not affect the correlation coefficient.

(c) Changing the scale does not affect the correlation coefficient.

(d) Both variables seem to be normally distributed.

(e) According to the output above, the hypothesis of no correlation must be rejected, so TOST and RBOT are indeed linearly related.