STA 302 / 1001 (A. Gibbs)
Sketch of Solutions to Exercises in Chapter 2 of Sheather

1. (a) $t_{16,0.025} = 2.12$
   95% CI for β_1: $0.982 \pm 2.12(0.014) = (0.95, 1.01)$
   Since 1 is in the CI, it is a plausible value for β_1.

   (b) Test statistic: $(6805 - 10000)/9929 = -0.32$
   From a t-distribution with 16 degrees of freedom, the p-value is 0.75 (from tables, can estimate that $p > 0.5$) so the data give no evidence against the null hypothesis and we cannot rule out that the intercept is 10000.

   (c) $\hat{y} = 6805 + 0.982 * 400000 = 399605$
   95% PI: $399605 \pm 2.12 * 18008\sqrt{1 + \frac{1}{18} + \frac{(400000-622187)^2}{17*91642100481}} = (359800, 439400)$
   Since \$450,000 is not in the prediction interval, it is not a feasible value.

   (d) The proposed prediction rule is a reasonable estimate for shows with box office results near \$378000 (where the regression line crosses the line $y = x$). But these data illustrate the phenomenon of regression to the mean: shows with low box office results in the previous week on average have higher box office results in the current week and shows with high box office results in the previous week on average have lower box office results in the current week.

2. The answers are in the SAS output. There is evidence of a significant negative linear association and 0% is not a feasible value for $\mathrm{E}(Y|X = 4)$. You should be able to construct the confidence intervals using other numbers on the SAS output and a t-table.

3. (a) $t_{28,.0.025} = 2.048$
   95% CI for β_0: $0.64171 \pm 2.048(0.12227) = (0.391, 0.892)$

   (b) Test statistic: $(0.01129 - 0.01)/0.00081840 = 1.576$
   From a t distribution with 28 degrees of freedom, the p-value is 0.126. (From tables, the estimated p-value is $0.10 < p < 0.20$.)
   So there is no evidence that the slope is different than 0.01, that is there is no evidence that the measured values differ from the benchmark.

   (c) For 130 invoices, the estimated time is $0.64171 + 0.01129(130) = 2.109$
   95% PI: $2.109 \pm 2.048 * 0.32977\sqrt{1 + \frac{1}{30} + \frac{(130-130.033)^2}{29*5598.86092}} = (1.422, 2.796)$

4. (a) Differentiating $\sum(y_i - \hat{\beta}x_i)^2$ with respect to $\hat{\beta}$ and setting the derivative equal to 0 gives the result.

   (b) i. $\mathrm{E}(\hat{\beta}|X) = \frac{\sum x_i \mathrm{E}(Y_i|X=x_i)}{\sum x_i^2} = \frac{\sum x_i\,(\beta x_i)}{\sum x_i^2} = \beta$

   ii. $\mathrm{Var}(\hat{\beta}|X) = \frac{\sum x_i^2 \mathrm{Var}(Y_i|X=x_i)}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2}$

iii. Since $e|X$ has a normal distribution, $Y|X$ also has a normal distribution so $\hat{\beta}$ has a normal distribution (since a linear combination of normally distributed random variables is also normally distributed). Thus, using the results of i. and ii., the distribution of $\hat{\beta}$ is what is given.

5. Statement (d) is correct. Since y is the same in both plots, SST is the same. RSS for model 2 is greater than RSS for model 1 since there is more scatter about the line in model 2. Since SST=RSS+SSreg, SSreg for model 1 must be greater than SSreg for model 2.

6. (a) Plug in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and then $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ to get the result.

   (b) Result follows immediately from part (a).

   (c) Plug in the result from part (b) and $\hat{\beta}_0 + \hat{\beta}_1 x_i$ for $\hat{y}_i$ giving

$$\hat{\beta}_1 \left[\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i - \overline{x}) \right]$$

   which expands and simplifies to give

$$\hat{\beta}_1 \left[SXY - \hat{\beta}_1 SXX \right]$$

   which is 0 since $\hat{\beta}_1 = SXY/SXX$.

7. The plots show confidence intervals for the regression line rather than prediction intervals. We would expect approximately 95% of points to fall within 95% prediction intervals.