Problem 3.51

Part a: \( y_i = \beta_0 + \beta_1 x_i + \epsilon \), with \( \epsilon \sim N(0, \sigma^2) \) and observations independent.

Part b: From JMP (or SAS), estimators of model parameters are:

\[
\begin{align*}
\hat{b}_0 &= 155.9 \\
\hat{b}_1 &= -1.09 \\
\end{align*}
\]

Then the LS prediction equation is \( \hat{y} = 155.9 - 1.09x \)

Part c: Test the null hypothesis that \( \beta_1 = 0 \) against the alternative that \( \beta_1 \) is different from 0.

The critical value for the test with \( \alpha = 0.05 \) is \( \pm t(0.025, n-2) \approx \pm 2.262 \).

The test statistic is

\[
t = \frac{\hat{b}_1}{\text{standard error of } b_1} = \frac{-1.09}{0.53} = -2.05
\]

Conclusion: Fail to reject H0 at the 0.05 level. There seems to be no linear association between the time Japanese adults spend practicing a sport and the frequency with which they practice it.

Part d: We want a confidence interval for the true mean time Japanese adults spend on a sport that they practice \( x = 25 \) times per year:

\[
\text{CI} = \hat{y} \pm t_{(\alpha/2, n-2)} \hat{\sigma}_y
\]

Here:

\[
\begin{align*}
\hat{y} &= 155.9 - 1.09 (25) = 128.7 \text{ minutes} \\
t(0.025,9) &= 2.262 \\
\hat{\sigma}_y &= \text{RMSE} [1/n + (25-35.8)^2 / \text{SS}_{xx}]^{1/2} = 19.95
\end{align*}
\]

where RMSE = 63.4, \( n = 11 \), \((25-35.8)^2 = 116.6\) and \( \text{SS}_{xx} = 14,336 \).

**New trick**: you can get \( \text{SS}_{xx} \) out of the JMP or SAS output as follows:

\[
\text{SS}_{xx} = \frac{\text{Model Sum of Squares}}{b_1^2}.
\]

In this problem, Model Sum of Squares is 16,907.24. Then

\[
\text{SS}_{xx} = 16,9907.24 / (-1.09)^2 = 14,336.
\]
Then, the 95% CI for the true mean time spent on a sport that is practiced 25 times a year is (85.6, 173.8). We are reasonably certain that when Japanese adults practice a sport 25 times per year, they spend between 86 and 174 on it each time.

Problem 3.62

Parts a and b: You know how to do this.

Part c: \( \beta_1 \) is the expected change in the length of stay when the number of factors increases by one.

Part d: Set up the null and alternative hypotheses as usual. For \( \alpha = 0.05 \), the critical value for the test is \( \pm t(0.025, 48) = \pm 2.021 \) and the test statistic is

\[
t = \frac{b_1}{\hat{\sigma}_{\beta_1}} = \frac{0.015}{0.00276} = 5.36
\]

Since \( t \) falls in the critical region, we reject \( H_0 \): We conclude that the number of factors per patient does contribute information about the length of stay of patients in the hospital.

Part e: A 95% CI for \( \beta_1 \) is given by

\[
b_1 \pm t(0.025, 48) \cdot \hat{\sigma}_{\beta_1} = 0.015 \pm 2.021 (0.00276) = (0.009, 0.02)
\]

Part f: In the case of simple linear regression, the correlation coefficient is equal to the square root of the coefficient of determination. From Part g: \( R^2 = 0.3740 \), and therefore \( r = 0.61 \). The coefficient of correlation measures the linear association between \( x \) and \( y \).

Part g: see above. About 37% of the sample variability in length of stay can be attributed to the number of factors received by patients.

Part h: We find the confidence interval for the length of stay of a single patient whose number of factors is equal to 231. The CI is computed as usual:

\[
\text{estimator} \pm t(0.025, 48) \cdot \text{(standard error)}
\]

where here, the estimator is just \( \hat{y} \) and the appropriate standard error is \( \hat{\sigma}_{(y-\bar{y})} \). We can get both from JMP easily because \( x = 231 \) happens to occur in our dataset. From JMP:

\[
\hat{y} = 6.71 \\
\hat{\sigma}_{(y-\bar{y})} = 2.12
\]

95% CI: (2.45, 10.98)

Part i: The interval is wide because we are trying to estimate the length of stay of a single patient who receives 231 factors rather than the mean length of stay of patients who receive 231 factors. The ways to decrease the width of the interval, for the same \( x \) are:
• Increase the sample size so that the term $1/n$ in the computation of $\hat{\sigma}_{n-\hat{\gamma}}$ gets smaller. A larger $n$ will also result in a smaller MSE thereby reducing the width of the interval even more.
• Decrease the confidence level: if $\alpha = 0.1$ instead of 0.05, then the interval will be narrower (but our chance of ending up with an interval that does not cover the true beta is also larger).

**Problem 3.64**

Part a: Here we obtain two prediction equations. In one, we regress plasma TCDD on fat TCDD levels and in the other one we do the reverse. From JMP:

When plasma TCDD is the response $y$ and fat TCDD is the predictor $x$:

$$\hat{y} = -0.22 + 0.9 \times FatTCDD$$

When TCDD in fat tissue increases by one unit, TCDD in plasma is expected to increase by about 0.9 units.

When fat TCDD is $y$ and plasma TCDD is $x$:

$$\hat{y} = 1.05 + 0.98 \times PlasmaTCDD$$

When TCDD in plasma increases by one unit, then TCDD in fat tissue is expected to increase by about one unit as well.

Parts b and c: See earlier problems for guidelines on how to set the hypotheses, obtain the critical values and the critical region, compute the test statistic and reach a conclusion. Here I refer to the $t$ ratios and the $p$-values given in the JMP output.

When plasma TCDD is the response: $t$-statistic = 11.03 and $p$-value is below 0.0001. Thus we conclude $H_a$: fat TCDD level is a good predictor of plasma TCDD level, since the $p$-value for the test is below the pre-determined Type I error level $\alpha = 0.05$.

When fat TCDD is the response: $t$-statistic = 11.03 and $p$-value is below 0.0001. Thus we conclude $H_a$: plasma TCDD level is a good predictor of fat TCDD level, since the $p$-value for the test is below the pre-determined Type I error level $\alpha = 0.05$.

Part d: Because if $y$ is associated to $x$ then $x$ must be associated to $y$. In fact, which variable to label a response and which to label a predictor is not always obvious and this may be one of those cases. In real life, we would choose the easier one of the two TCDD levels to measure as the predictor and would try to then predict the level that is harder to measure.

**Problem 3.65**
Parts a, b, c: The t-statistics in all cases are within the critical region; thus we conclude that the predictors contribute information about the responses. If one wishes to get picky, then one might look at the relative magnitude of the t-statistics and decide that the association between predictor and response appears to be stronger in the case of CDF, and least strong in the case of OCDD.

Part d: \[ \hat{y} = 0.9855 + 0.7605(FatCDF) \]

Part e: \[ \hat{y} = 18.1565 + 0.7377(PlasmaCDD) \]

Part f: \[ \hat{y} = 167.723 + 1.5752(FatOCDD) \]

Problem 3.75

Part a: From JMP, we get an estimate for the slope of -$1,783 with a standard error of $1,814. Since we fail to reject the hypothesis that the true slope is equal to zero, we conclude that ratings do not contribute useful information about salary raises.

Part b: If we eliminate observation #3, the new estimate of the slope is -$3,887 with a standard error of $1,699, and this time around we conclude that for every point increase in rating, the employee can expect to see his/her salary decrease by about $3,900. Thus, these results would support the claim that the worst administrators are the ones that get the better raises.

Part c: Several problems:
- At the very low end and at the very upper end of the rating scale, the predicted raises are extrapolations and thus rather suspect. The model might do OK in predicting raises at ratings towards the middle of the scale but will do a poorer job at the extremes.
- Even at the middle ranges of ratings, the table should be used very cautiously. Much more useful would be to obtain a confidence interval (for every rating) of what an administrator could expect for a raise given his/her rating.

Part d: Because of the low response and because non-response is not at random (meaning that the people that tend to respond are those that are mad about the administration) we do not have a representative sample of faculty responding to the survey. Since salary raises are independent of the sample as they are determined by the provost or the president, the errors that creep into our analyses because of the non-representative sample are all in the x’s. Chances are, a representative sample of faculty would have rated the administrators higher and might have even had a very different opinion about them (meaning that administrator 1, for example, might have ended with a higher ranking).

Part e: Not really. Response rate is very low, the predictor cannot be objectively measured (it is a subjective rating) and thus is subject to error, and the sample is not representative. We do not learn much about the association (if any) between real performance and raises in the administration.