Stat 371-003, Solutions to Homework #11

1. **12.10 (pg 540)**

   (a) The estimated slope is $b_1 = \hat{\beta}_1 = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sum(x_i - \bar{x})^2 = 5288/1172 = 4.512$
   The estimated y-intercept is $b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 660 - 180.4 \cdot 4.512 = -154.0$
   So the estimate regression line for y on x is

   $$y = -154.0 + 4.512x$$

   (b) For the predicted (aka fitted) values, we calculate $\hat{y} = -154.0 + 4.512x$; see the following table. (I asked you to calculate just the first three.)

Subject	x	y	$\hat{y}$	$y - \hat{y}$
1	174	733	631.1	101.9
2	183	572	671.7	–99.7
3	176	500	640.1	–140.1
4	169	738	608.5	129.5
5	183	616	671.7	–55.7
6	186	787	685.2	101.8
7	178	866	649.1	216.9
8	175	670	635.6	34.4
9	172	550	622.1	–72.1
10	179	660	653.6	6.4
11	171	575	617.5	–42.5
12	184	577	676.2	–99.2
13	200	783	748.4	34.6
14	195	625	725.8	–100.8
15	176	470	640.1	–170.1
16	176	642	640.1	1.9
17	190	856	703.3	152.7

   (c) For the residuals, we calculate $y - \hat{y} = y - (-154.0 + 4.512x)$; see the previous table. (I asked you to calculate just the first three.)

   (d) $s_{Y|X} = \sqrt{\text{SS(resid)}/(n-2)} = \sqrt{198,909/(17-2)} = 115.2.$
   The units are the same as for y (Li/min).

   (e) 12/17 = 71% of the data points are within $\pm s_{Y|X}$ of the regression line.

2. **12.18 (pg 548)**

   The estimated mean peak flow for men 180 cm tall is

   $$-154.0 + 4.512 \cdot 180 = 658.2$$

   The estimated SD of peak flow for men 180 cm tall is $s_{Y|X} = 115.2$.

3. **12.26 (pg 553)**

   We first calculate the estimated standard error of the estimated slope.

$$\hat{\text{SE}}(\hat{\beta}_1) = s_{Y|X}/\sqrt{\sum(x_i - \bar{x})^2} = 115.2/\sqrt{1172} = 3.364$$

   (a) To test $\beta_1 = 0$, we look at $\hat{\beta}_1/\hat{\text{SE}}(\hat{\beta}_1) = 4.512/3.364 = 1.341.$

   For a non-directional test at $\alpha = 0.1$, we compare this to the 95th percentile of a t distribution with 15 degrees of freedom, 2.131. Since 1.341 < 2.131, we fail to reject the null hypothesis: there is insufficient evidence to conclude that there is a relationship between peak flow and height.

   Note that the P-value is very close to 0.2.

   (b) For a directional test at $\alpha = 0.1$, we compare the t statistic, 1.341, to the 90th percentile of a t distribution with 15 degrees of freedom, 1.341. We just reject the null hypothesis of no relationship and conclude that peak flow increases with height.

4. **12.32 (pg 564–565)**

   (a) The correlation coefficient (which I'd prefer to call the *estimated* correlation) is

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{893.689}{\sqrt{1419.82 \cdot 853.396}} = 0.8119$$

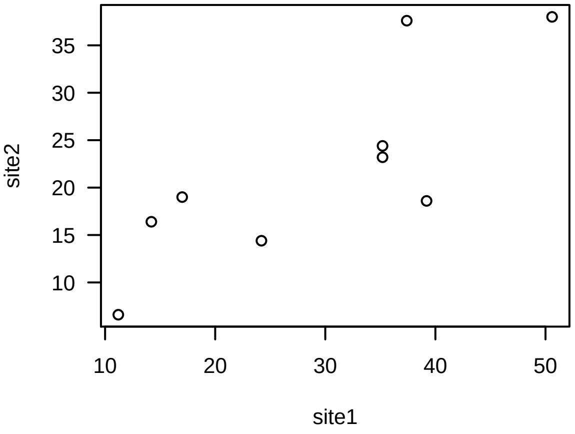
   (b) You could create a scatterplot by hand, or you could read the data into R and plot them as follows:

```
dat <- read.csv("http://www.biostat.wisc.edu/~kbroman/teaching/stat371/data_12-32.csv")
plot(dat[,2], dat[,3])
```

   Here are a couple of alternatives for creating the scatterplot:

```
plot(dat$site1, dat$site2)
plot(site2 ~ site1, data=dat)
```

   Here's the actual plot.



(c) Regarding the four potential sources of errors and their impact on the correlation, it is the variation between horses that is the primary contributor to the correlation between the sites.

   i. *Errors in counting nerve cells*: one would expect this to be independent between the two sites, and so this would weaken the association.

   ii. *Sampling error due to choosing certain slices for counting*: sampling error is similar to counting error and would be expected to be independent between the two sites, and so again this would weaken the association.

   iii. *Variation from one horse to another*: this is the major contributor to the correlation between the sites. If all horses were the same, there would be a single dot in the scatterplot, and so no correlation between sites.

   iv. *Variation from site to site within a horse*: this shouldn't have any influence on the correlation between the sites. It would shift the scatterplot sideways or up-and-down, but wouldn't affect the correlation.

5. **12.33 (pg 565)**

To test whether the true correlation is equal to zero, we could calculate the estimated slope (and estimated standard error) for the regression of site 2 on site 1, and test whether the true slope is 0 or not.

We first get the estimated slope:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = 893.689/1419.82 = 0.6294$$

We then need to calculate the residual sum of squares:

$$\text{SS(resid)} = \sum(y_i - \bar{y})^2 - \hat{\beta}_1 \sum(x_i - \bar{x})(y_i - \bar{y}) = 853.396 - 0.6294 \cdot 893.689 = 290.8740$$

We use this to estimate the residual SD:

$$s_{Y|X} = \sqrt{\text{SS(resid)}/(n-2)} = \sqrt{290.8740/7} = 6.446$$

We then calculate the estimated standard error of $\hat{\beta}_1$:

$$\hat{\text{SE}}(\hat{\beta}_1) = s_{Y|X}/\sqrt{\sum(x_i - \bar{x})^2} = 6.446/\sqrt{1419.82} = 0.1711$$

Finally, we calculate the t statistic: $t = \hat{\beta}_1/\hat{\text{SE}}(\hat{\beta}_1) = 0.6294/0.1711 = 3.679$

We compare this to a t distribution with 7 degrees of freedom. For the *directional* alternative, we look at the 95th percentile of the t distribution with 7 df, which is 1.895. We thus reject the null hypothesis and conclude that the true slope is non-zero and so the true correlation is non-zero. The p-value (from the table) is between 0.001 and 0.01. Calculated with R, it is approximately 0.008.

The simpler approach described in the book is the following.

The test statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.8119\sqrt{\frac{7}{1-0.8119^2}} = 3.679$$

Note that this is exactly the same as the test statistic we got by my convoluted method above. It will always be that way! Our conclusions are, of course, the same.