Stats for Clinical Trials, Math 150
Jo Hardin
Homework # 2, Due: Thurs, 9/18/10

1. Eisenhofer et al. (1999) investigated the use of plasma normetanephrine and metanephrine for detecting pheochromocytoma in patients with von Hippel-Lindau disease and multiple endocrine neoplasia type 2. The data set (vonHippelLindau.csv) contains data from this study on 26 patients with von Hippel-Lindau disease and nine patients with multiple endocrineneoplasia. The variables in the data set are (PROBLEM FROM DUPONT, CHP 2.22):

   `disease` = 0: patient has von Hippel-Lindau disease;
   1: patient has multiple endocrine neoplasia type 2
   `p_ne` = plasma norepinephrine (pg/ml)
   `tumorvol` = tumor volume (ml)

   Use R for the following problems.

   (a) Regress plasma norepinephrine against tumor volume. Draw a scatter plot of norepinephrine against tumor volume together with the estimated linear regression curve. What is the slope estimate for this regression? What proportion of the total variation in norepinephrine levels is explained by the regression?

   (b) Experiment with different transformations of norepinephrine and tumor volume. Find transformations that provide a good fit to a linear model. Report your new linear model. What is your new R^2? Does the R^2 matter in choosing your transformation? Explain.

   (c) Using the transformed model, what is the predicted average plasma norepinephrine concentration for a patient with a tumor volume of 100 ml? What is the 95% confidence interval for your prediction? Interpret. What would be the 95% prediction interval for a new patient with a 100 ml tumor?

2. The systolic blood pressure (SBP, in mmHg) in 36 women in the age interval 20-82 years is given below (and you don't need to type in the data at all).

```
age <- c(30,25,55,47,27,21,30,27,31,34,35,47,20,52,55,48,52,47,40,48,60,38,
54,59,63,60,62,63,45,38,72,67,81,44,82,60)
```

```
SBP <- c(105,110,112,116,117,119,120,112,122,123,125,122,119, 129,132,134,
136,140,142,142,143,145,145, 151,159,160,163,168,160,170, 170,171,171,172,
205,180)
```

Note:

$$\begin{aligned}
\sum Y_i &= 5110 \\
n &= 36 \\
\overline{X} &= 47.75 \\
\sum (X_i - \overline{X})(Y_i - \overline{X}) &= 10,314.5 \\
\sum (Y_i - \overline{Y})^2 &= 20,575.889 \\
\sum (Y_i - \hat{Y}_i)^2 &= 8836.747 \\
\sum (X_i - \overline{X})^2 &= 9,062.75 \\
\sum (X_i Y_i) &= 254,317
\end{aligned}$$

Recall:

$$\begin{aligned}
b_0 &= \overline{Y} - b_1 \overline{X} \\
b_1 &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \\
r &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}} \\
s_{y|x} &= \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}} \\
SE(b_1) &= \frac{s_{y|x}}{\sqrt{\sum (X_i - \overline{X})^2}} \\
SE(\hat{y}|x^*) &= s_{y|x} \sqrt{\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum (X_i - \overline{X})^2}} \\
SE(\text{ indiv } y|x^*) &= s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum (X_i - \overline{X})^2}}
\end{aligned}$$

(a) BY HAND, find the least squares estimates of β_0 and β_1.

(b) BY HAND, find the estimate of the SE of $\hat{y}$ at x = 24 years old.

(c) BY HAND, find the estimates of the variance (or SE) of b_1.

(d) BY HAND, find a 95% confidence interval for β_1.

(e) BY HAND, find a 90% confidence interval for the mean of Y at X = 24 years old.

(f) BY HAND, find a 93% prediction interval for an individual value of Y at X = 24 years old.

(g) BY HAND, find the proportion of variability of SBP explained by age.

(h) Give some general conclusions about the data set.

3. Which of the following assumptions are required to test hypotheses using simple linear regression. If you think it isn't valid, explain why not.

   (a) The random variable Y is normally distributed.

   (b) The variance of Y depends only upon X.

   (c) The random variable Y is normally distributed for each value of X.

   (d) The mean of Y is a linear function of X.