1. (Problem 6.3 on p.115).
The formula I gave in class for the variance of the estimated intercept $b_0$ looks a little different than the one given in the book, but they are equal.

$$V(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

By examining the formula above for the sampling variance of $b_0$ in simple linear regression, we can see that the variance of $b_0$ is larger when the mean of the x’s is far from 0. But tell why it is intuitively sensible that this is the case. Illustrate with a graph (you can draw this by hand).

2. (Problem 6.4 on p.116).
The formula for the sampling variance of the estimated slope $b_1$ in simple linear regression

$$V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

shows that it helps to have spread-out x’s if you want to estimate $\beta_1$ precisely. Explain why this result is intuitively sensible, illustrating your explanation with a graph. What happens to the variance when there is NO variation in the x’s?
3. Use the *Robey* data set on fertility and contraception in developing countries in the *car* library to perform a least-squares regression of fertility on contraception. Plot the data and the least squares line. Does the line adequately summarize the relationship between these variables? Examine and interpret the values of A, B, $S_E$, and $r$ (or $r^2$), which is the book notation. Recall that the book uses the following notation for simple linear regression:

- $A \equiv$ intercept
- $B \equiv$ slope
- $S_E \equiv$ standard error of the regression
- $r \equiv$ correlation coefficient

To interpret the slope, provide information on how the mean of $Y$ changes as $x$ changes by 1 unit (put in context of the problem, use specific units).

To interpret the intercept, provide the mean $Y$ value when $x = 0$ (put in context of the problem). Depending on context of the data, this interpretation may be meaningful, but sometimes it is not.

4. Continuing to use the *Robey* data set as in problem 3...

(a) Test for a significant linear relationship between the variables. Provide the table used to make your decision.

(b) Give the estimated standard errors for the estimated slope and intercept.

(c) Use the ‘confint’ function to compute 95% confidence intervals for the population slope and intercept.

(d) Use the estimated slope and intercept (from output) to compute by hand the predicted Total Fertility Rate for a country with 25% contraceptors among married women of childbearing age.

(e) Use the ‘predict’ function in R to predict the TFR for countries with contraceptive values of 10% and 40%, the code is shown below... R will be picky and you need to have it just as shown:

```r
predict(lm.out,newdata=data.frame(contraceptors=c(10,40)),se.fit=TRUE)
```

i) What are the predicted conditional means, or $\hat{Y}$’s?

ii) What are the standard errors for the estimated $\hat{Y}$’s?

iii) Compare the size of the standard errors and how this relates to the x-values at which the predictions were made.
5. Data on PCB levels in trout fish in Cayuga Lake, NY, are available from our class website under the *Data sets* link titled “PCB_in_fish.csv” (see the info file, too).

(a) Plot PCB vs Age. Does it look like there is a linear relationship? Or a non-linear relationship? Provide the plot.

(b) Regress PCB on Age. Provide the plots for checking for constant variance and normality. Comment on what the plots suggest about these two assumptions.

(c) Apply the $\log_e$ transformation to PCB and re-fit the model with this transformed $Y$ as the response. Provide the plots for checking for constant variance and normality. Do the assumptions seem better met?

(d) Test for a significant linear relationship between log(PCB) and Age. Provide the results of your test and interpret the meaning.

(e) Interpret the slope of the fitted model in (c).