Theory:

8.21  
   a. Hard Hat: \( Y = \beta_0 + \beta_1 X_1 + \beta_2 \)  
   Bump Cap: \( Y = \beta_0 + \beta_2 X_1 + \beta_3 \)  
   None: \( Y = \beta_0 + \beta_1 X_1 \)  
   b. (1) \( H_0: \beta_3 = 0, \ H_A: \beta_3 < 0 \)  
   (2) \( H_0: \beta_2 = \beta_3, \ H_A: \beta_2 \neq \beta_3 \)  

8.22  
   (1) \( \beta_3 \) is the change of tool wear when tool model M3 is used, holding the tool speed at a constant rate.  
   (2) \( \beta_4 - \beta_3 \) is the difference in tool wear when using model M4 versus model M3, holding the tool speed constant.  
   (3) \( \beta_1 \) is the change in tool wear for every 1 unit increase in tool speed, when using the M1 tool model. It is the change for every unit increase since \( X_1 \) is not a categorical variable and it is the model M1 since all other variables are 0.  

8.23  
   No I disagree with the results. The result of \( \beta_1 = 0 \) represents that there is no difference in sales between winter and summer. Since \( \beta_0 \neq 0 \), the summer season does have an impact on sales. Because \( \beta_2 \neq 0 \) and \( \beta_3 \neq 0 \), spring and fall have different effects on sales than the summer season has. Therefore the report should have said that the effects of summer and winter on sales are very similar, however spring and fall exhibit a much different effect on sales.
The relationship seems to be very similar for both corner and non-corner lots. Both seem to have a linear relationship with selling price.

b. R Code:
   ```r
   full <- lm(y~x1+x2)
   reduced <- lm(y~x1)
   anova(reduced,full)
   ```

After performing the General F-test the results of $F^* = 27.04$ and $p$-value $< 0.05$ show that there is a difference between corner and non-corner lots. So even though the graph does not strongly support it, the general F-test has concluded that the lots affect the selling price.
C. Selling Price vs. Assessed Valuation By Lot Location

The have similar slopes with just different intercepts.

5. a. After conducting the Box Cox analysis, it showed that there needed to be a transformation of Y where $\lambda = 10/99$.

b. The residual plots formed by fitting the model was
As shown by the graphs observations 1 and 2 seem to be a little off compared to the other observations. Observation 1 especially has an extremely large Cook’s distance. I removed both 1 and 2 since those were the two counties with the highest population, having 2-4 times the number of people as the other counties. This may have contributed for the numbers to be so far away from the other observations. After removing those observations, there were some additional observations that were skewing the data as well because of their large population sizes. Observations 3, 4, and 5 were also very large however I did not remove them since their Cook’s distances were not very large and the residuals were fairly normal.

c. R code:
   ```r
   full <- lm(physicians~income+population+region,subset=c(-1,-2))
   reduced <- lm(physicians~income+population,subset=c(-1,-2))
   anova(reduced,full)
   ```
   The General F-test produced $F^* = 1.37$ and p-value > 0.05 which means that geographic region is not significant in determining the number of physicians.

d. The average income was 7,231 (without 1st and 2nd observations) and the average population was 362,915 (without 1st and 2nd observations).
   95% Confidence interval for W Region: (399, 541)
   95% Confidence interval for NC Region: (396, 511)
   95% Confidence interval for S Region: (466, 576)
   95% Confidence interval for NE Region: (457, 594)
e. The General F-test produced $F^* = 0.93$ and p-value $> 0.05$ so state does not have an effect on the number of physicians.

6. a.

![Pulse vs. Weight by Gender](image)

b. The model fitted for males:

|            | Estimate | Std. Error | t value  | Pr(>|t|) |
|------------|----------|------------|----------|----------|
| (Intercept)| 82.58503 | 5.16884    | 15.977   | <2e-16 ***|
| weight     | -0.10398 | 0.08664    | -1.200   | 0.233    |
| gendermale | -2.13471 | 2.57527    | -0.829   | 0.409    |

The model fitted for females:

|            | Estimate | Std. Error | t value  | Pr(>|t|) |
|------------|----------|------------|----------|----------|
| (Intercept)| 84.30384 | 4.72757    | 17.832   | <2e-16 ***|
| weight     | -0.14700 | 0.06926    | -2.122   | 0.0362 * |

So males have a lower y intercept which in the context of the problem doesn’t matter since weight never equals zero. Weight seems to have more of an impact for females, since it is larger in absolute value. It also appears to be more significant in the female model. The following plot displays the regression functions with the data.
The plots seem to be consistent with the results from part b. The males seem to be evenly spread out. The females have some slight problems in the lower fitted values, which indicates that they may be affected strongly by something, in this case it is weight. The female plot seems to have nonconstant variance because of the cone-like shape of the residuals.

d. Fitting a regression model with exercise level gave the following output:

| Coefficients       | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 76.11994 | 5.94520    | 12.804  | <2e-16   *** |
| weight              | -0.08722 | 0.08639    | -1.010  | 0.3151   |
| gendermale          | -1.57154 | 2.57162    | -0.611  | 0.5425   |
So we see that the low exercise levels produce significantly different pulse rates than high levels of exercise. However, there is not a significant difference between moderate and high levels of exercise.

e. 95% Confidence interval for male, moderate exerciser weighing 56 kg: (69.64, 79.34)
95% Confidence interval for female, low exerciser, weighing 43 kg: (75.27, 84.17)

f. R code:
hm <- rbind(c(0,0,0,1,-1))
rhs <- c(0)
linear.hypothesis(res3,hm,rhs)

H₀: β₃ - β₄ = 0
Hₐ: β₃ - β₄ ≠ 0
p-value = 0.27 > 0.05
Since the p-value is less than 0.05, there is no significant difference between moderate and low exercise levels in terms of the pulse rate.

g. R code:
hm <- rbind(c(0,20,1,0,-1))
rhs <- c(0)
linear.hypothesis(res3,hm,rhs)

H₀: 20β₁ + β₂ - β₄ = 0
Hₐ: 20β₁ + β₂ - β₄ ≠ 0
p-value = 0.02 < 0.05
Since the p-value was less than 0.05 we have enough evidence to show that there is a difference in pulse rate between a 60kg male who exercises at a high level and a 40kg female who exercises at a moderate level.