Test Corrections Test 1

1. a- This is valid because 31.6 isn't within the 90% confidence interval. So it's safe to say that in this area, the prevelance of lung cancer is likely higher.

b- This isn't valid since the statement used the word chance. Confidence intervals don't say what the chance of something happening is.

c- This is a valid statement since that's how confidence intervals are supposed to be interpreted. It's true that with repeated random samples, the confidence intervals produced will have the actual true mean prevelance of lung cancer in this area 90% of the time.

d- This isn't valid since the statement used the word chance. Confidence intervals don't say what the chance of something happening is.

3. a- This is true since the prediction interval has an extra "1" in the formula when compared to the confidence interval.

confidence interval; $SE_{\hat{\mu}} = \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma(x - \bar{x})^2}}$

prediction interval; $SE_{\hat{y}} = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma(x - \bar{x})^2}}$

b- If the model is really bad, then the sum of squared errors could be very large and larger than SSmod. So, that would make this statement false.

c- This is true since correlation is "r" and the coefficient of determination is r^2. So, if the absolute value of the correlation is greater, then the coefficient of determination has to be greater too.

4. a- You can check linearity in a residual vs fits plot by seeing if there's a pattern in the residuals.

b- You can check constant variance in a residual vs fits plot by seeing if the variance is larger/smaller on certain parts of the plot or if it's not.

c- You cannot check independence using a residual vs fits plot. You find out if the errors are independent using the context of how the data was collected.

4. continued. d- Normality cannot be checked by a residual vs fits plot. It can by checked by a normal quantile plot but not by a residuals vs fits plot.

5. a- We don't need to assess the normality and randomness conditions in this situation since there's no inference involved in this situation. The scientist is just trying to find how much the average weight (g) of a random sample of sparrows changes with each additional mm in wing length

b- We don't need to asses the normality and randomness conditions in this situation because prediction isn't inference. Since the person is only interested to predict the sale price of their house, they're not using inference, so we don't need to asses the normality and randomness conditions.

c - Since in this situation the nutritionist is trying to get an upper and lower bound on the size of the increase of change in calories, they are using inference. They are trying to develop a confidence interval. In inference, you need to check the normality and randomness conditions.

d - Since in this situation the used car salesperson is trying to see if an increase in the number of miles on a car has a statistically significant impact on the sale price of the vehicle, there is use of inference. Trying to see if something is statistically significant is a use of inference. In inference, you need to check the normality and randomness conditions.

6. Step 1: I would model a SLR model with APR being the predictor and payments made on your credit card being the response. I'd use a SLR model because they are trying to explore a linear relationship.

Step 2: The way I'd fit the data to my model is by finding the values of $\hat{\beta}_0$ and $\hat{\beta}_1$. I'd put those into my SLR equation of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. $\hat{y}$ = average APR, x = payments made on credit card, $\hat{\beta}_0$ = intercept, $\hat{\beta}_1$ = slope.

Step 3: I would check for linearity in the scatter plot to see if a linear relationship makes sense. In the residual plot, I would check for residuals having a zero mean, that there's constant variance, and that the residuals are independent. I would use a quantile plot to check the normality condition and I would check check the write up of the problem to see if there was random selection.

Step 4: As long as the first 3 steps are completed and look good, I would create a confidence interval for the mean response (APC). I would do this to get a range of APC's to see where the mean of APC may lie.

8. a- Conditions

random: We can't make any assumptions about randomization using the plots. We would need context about how the data was collected.

linearity: residual plot has no patterns ✓

zero mean: given ✓, we could also check it by finding the mean of the errors and making sure it's zero using the residual plot

constant variance: seems like there's constant variance in the residual plot ✓

independence: We can't make any assumptions about independence of errors using the plots. We would need context about what data was collected

normality: studentized residuals shows a normal looking plot and normality plot looks like most points are near the line ✓

b- No, we would not expect any of the behavior of any plots to change. All you're doing is changing the numbers a little bit for the predictor (size of diamonds). You're not changing it in a way where any relationships change though. The relationship between size of diamonds and cost of diamonds would stay the same since all you're doing is changing units for the size of the diamonds.

9. a- The estimate for standard deviation of the number of calories burned based on this linear model is 30.84 calories. This is the residual standard error at a df of 17. of the SLR model for the relationship between speed of the run (mph) and number of calories burned.

b- Our runner can expect to burn 80.82 calories more for each mph increase in average running speed. This is the β_1 of the SLR model for the relationship between speed of the run (mph) and number of calories burned.

c- A good procedure to determine if our runner's rate of burning calories is different from the average for all people in the age group is a 95% prediction interval for our runner. If 100 is in the interval, then you can't say anything significant about the data. If 100 is outside the interval, then you could say you're 99% confident that the runner has a different rate of burning calories than the average for all people in the age group.

9 continued: d- R^2: .4313: shows how much of the variance from the response can be predicted by the model

p-value: .00225: This is the p-value of the F-statistic and it can show you how probable it is to have a slope (β_1) of 0. Since p-value is very small, we reject H_0 ($\beta_1 = 0$).