Juan Villegas

Quiz: Chapter 9

1. Indicate whether the following statements are true or false. If the statement is false, alter it so that the statement becomes true.

   a. If we would like to approximate the relationship between a response variable and two continuous predictors, we would need a plane. True

   b. In linear regression, while the response variable is typically continuous, it may be categorical as well. False, the response variable has to be continouous The predictor variable can be categorical.

   c. In general, for a multiple regression with m predictor variables, we would interpret the coefficient b_i as follows: "the estimated change in the response variable for a unit increase in variable x_i is b_i." False, It forgets to mention that the other predictor variables have to be held constant.

   d. In multiple regression, the residual is represented by the vertical distance between the data point and the repression plane or hyperplane. True

   e. Whenever a new predictor variable is added to the model, the value of R^2 always goes up. True, the model is trying to become perfect.

   f. Whenever a new predictor variable is added to the model, the value of the standard error of the estimate s_e always goes down. False, it is not guaranteed that the standard error goes down but, it is highly likely.

   g. For use in regression, a categorical variable with k categories must be transformed into a set of k indicator variables.

   h. The first sequential sum of squares is exactly the value for the SSR form the simple linear regression of the response on the first predictor. True

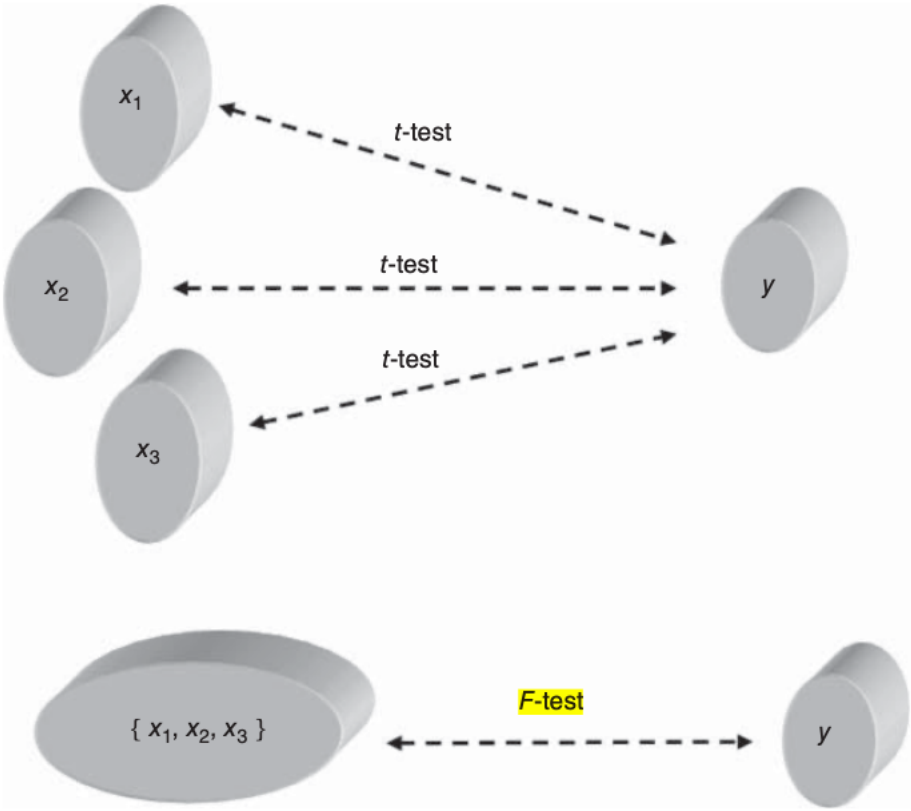
   i. A variable that has been entered into the model early in the forward selection process will remain significant once other variables have been entered into the model. True

2. Clearly explain why s_e and R^2_{adj} are preferable to R^2 as measures for model building.

R^2 doesn't consider the # of predictor variables, this means that the more variables you add, the bigger R^2 giving us misinformation about how good the fit of our model is. Adding more variables can lead to over-fitting the data. R^2 adjusted considers the extra variables and penalizes us for adding extra variables. Standard error tells us far residuals are from the regression line, so s_e is a good indicator as to how far our model is from the average residuals.

3. Explain the difference between the t-test and the F-test for assessing the significance of the predictors.

T-test considers the significance of each predictor variable as independent when comparing it to the target variable. The F-test considers all predictors variables in a model and assess their significance.



4. Construct the indicator variables for the categorical variable class, which takes four values: first-year, sophomore, junior, senior.

Sophomore {1 if sophomore = 1, 0 if sophomore != 1 (not 1)

Junior {1 if junior = 1, 0 if junior != 1 (not 1)

Senior {1 if senior =1, 0 if senior !=1 (not 1)

5. Explain what it means when R^2_{adj} is much less than R^2.

There's a high number of insignificant variables and R^2 adjusted is taking them into account.

6. Explain some of the drawbacks of a set of predictors with high multicollinearity.

It creates redundancy and the R^2 adjusted decreased because of the more redundant variables.

7. Return to the model for predicting nutritional rating (response) from the predictor variables sugars, fiber, and sodium:

$$\widehat{rating} = 61.009 - 2.163 sugars + 2.793 fiber - 0.056 sodium$$

a. How do we interpret the value of the constant term? Note in the output from the summary command (see below), the *t*-test for intercept indicates it is significantly different from 0. Explain how this makes sense.
When the sugars, fibers, and sodium of a cereal from this data set is 0, the nutritional rating is 61.009190.
The t-test indicates that this is significantly different than being = to 0 because if a cereal doesn't have sodium, fiver, or sugar, it does not mean that the nutritional rating is = 0.

```
> summary(M_su_fi_so)

Call:
lm(formula = rating ~ sugars + fiber + sodium, data = cereal)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4634  -1.1559   0.1652   2.2281   9.2223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.009190   1.379679  44.220  < 2e-16 ***
sugars      -2.162825   0.110378 -19.595  < 2e-16 ***
fiber        2.793019   0.200727  13.915  < 2e-16 ***
sodium      -0.056370   0.005744  -9.813 1.01e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.06 on 69 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9233,    Adjusted R-squared:  0.9199
F-statistic: 276.7 on 3 and 69 DF,  p-value: < 2.2e-16
```

b. What is the conclusion regarding the overall significance of the overall regression?
The overall significance of the regression model is <2.2e16. The p-value is a lot smaller than .05 so we reject the null hypothesis. There is evidence for a linear relationship between nutritional rating, sugar, fiber, and sodium.

c. What is the typical error in prediction?

The typical error for this model is 4.06 on 69 degrees of freedom. This means that residuals are typically about 4.06 points (-/+) away from our regression model.