Review Problems for Exam 2

These review problems are not meant to be comprehensive. Make sure to review the lessons and homework too!

Problem 1 (Adirondack High Peaks)**.** A goal for hikers in the Adirondack region of upstate New York is to become a "46er" by scailing each of 46 mountains with elevations near or above 4,000 feet. Consider the data in `HighPeaks` from the `Stat2Data` library. Suppose you want to develop a model to predict `Time` from all available useful variables.

a. Find the best model using best subset variable selection and R^2. State the fitted model and the R^2.

b. Consider the simple linear regression model to predict `Time` from `Length`. Use the following code to split your data:

```
set.seed(2022)
train = sample(46, 36)
```

You will then be able to use `HighPeaks[train,]` when you want to use the training data and `HighPeaks[-train,]` for the test data. Calculate the cross-validation correlation.

Problem 2 (Leafhopper diets.)**.** If you eat nothing but sugar, how long will you live? Experimenters prepared eight petri dishes, two for each diet: control, sucrose, glucose, and fructose. Eight leafhoppers were put into each dish. Diets were randomly assigned to dishes. The response variable was time (in days) until half the leafhoppers in a dish died. The data is in `Leafhoppers` from the `Stat2Data` library.

a. Is this an observational study or an experiment? Briefly justify your answer.

b. Create a boxplot of survival (in `Days`) versus `Diet`.

c. What is the grand mean?

d. How far is each group mean from the grand mean? i.e., What is each group effect?

e. State and check the conditions for ANOVA inference.

f. Regardless of your answer to part e, does diet affect the survival of leafhoppers? Conduct an appropriate test and provide the test statistic used to make this determination.

g. Construct a 95% CI for the mean length of life for leafhoppers on the control diet.

Problem 3 (House Prices.)**.** Suppose you want to predict house prices from their size (in sq ft) and their lot size (in sq ft).

a. Which model is a quadratic regression model, for both predictors?

(a) $Price = \beta_0 + \beta_1 Size^2 + \beta_2 Lot^2 + \varepsilon$

(b) $Price = \beta_0 + \beta_1 Size + \beta_2 Lot + \beta_3 Size^2 + \beta_4 Lot^2 + \varepsilon$

(c) $Price = \beta_0 + \beta_1 Size + \beta_2 Lot + \beta_3 (Lot \times Size) + \varepsilon$

(d) $Price^2 = \beta_0 + \beta_1 Size + \beta_2 Lot + \varepsilon$

b. Which model is a complete second order regression model?

(a) $Price = \beta_0 + \beta_1(Size \times Lot) + \beta_2 Size^2 + \beta_3 Lot^2 + \beta_4(Size^2 \times Lot^2) + \varepsilon$

(b) $Price = \beta_0 + \beta_1 Size^2 + \beta_2 Lot^2 + \beta_3 Size \times Lot + \varepsilon$

(c) $Price = \beta_0 + \beta_1 Size + \beta_2 Lot + \beta_3 Size^2 + \beta_4 Lot^2 + \varepsilon$

(d) $Price = \beta_0 + \beta_1 Size + \beta_2 Lot + \beta_3 Size^2 + \beta_4 Lot^2 + \beta_5(Size \times Lot) + \varepsilon$

c. Suppose you fit a model with only linear terms and get this output.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 34121.649  29716.458   1.148   0.2668
Size           23.232     17.700   1.313   0.2068
Lot             5.657      3.075   -----   0.0834 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47400 on 17 degrees of freedom
Multiple R-squared:  0.5571,  Adjusted R-squared:  0.505
F-statistic: 10.69 on 2 and 17 DF,  p-value: 0.000985
```

(a) What is the test statistic for the coefficient of *Lot*?

(b) At the $\alpha = 0.05$ level, is the overall model effective? Which test statistic is used in the test?

(c) At the $\alpha = 0.05$ level, is *Size* a significant predictor of *Price*, after accounting for *Lot*? Which p-value is used in the test?

d. What is the most likely reason the overall model is significant, but the individual t-tests have large p-values?

(a) The model needs an interaction term.

(b) There is a problem with the R code.

(c) The predictors are probably highly correlated.

(d) The standard errors are large due to measurement error.

Problem 4 (Plant Growth)**.** Suppose you are interested in whether fertilizers A and B have different effects on plant growth. You want to build a model using fertilizer type and the amount of water a plant receives as predictors of plant height. You have the following variables measured on each of 50 plants:

- *Height*: plant height in inches (after one month of fertilizer).
- *Water*: amount of water plant received each day.
- *FertA*: indicator variable (= 1 if fertilizer A).

Consider this model: $Height = \beta_0 + \beta_1 Water + \beta_2 FertA + \beta_3(Water \times FertA) + \varepsilon$

a. In terms of the βs, what is the intercept for fertilizer A?

b. In terms of the βs, what is the slope of *Water* for fertilizer A?

c. In terms of the βs, what is the slope of *Water* for fertilizer B?

d. If we don't want to allow the slopes of *Water* to differ by fertilizer type, but we do still want to let the intercepts differ, which term(s) do we need to remove from this model?

Problem 5 (Multiple Regression)**.** Suppose you have 5 quantitative predictors ($X_1, X_2, ..., X_5$) you are considering including in a model to predict Y. The most complicated model you are willing to consider is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

a. You fit the full model and find that overall it is effective at predicting Y, but you suspect you could do just as well without X_2 and X_3. You want to test whether the full model is significantly better, at the 0.05 level. State the null and alternative hypotheses (using mathematical symbols) for the appropriate test.

b. What is the reduced model?

c. What is the name of the test we should do to compare the models?

d. You fit the full model (`fit.full`) and the reduced model (`fit.reduced`), and get the following from R: `anova(fit.reduced, fit.full)`

```
Analysis of Variance Table

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     31 36.234
2     29 35.008  2    1.2259 0.5078 0.6071
```

(i) What is the SSE for the reduced model?

(ii) What is the sample size?

(iii) Do you reject or fail to reject the null hypothesis? What is your conclusion for your test?