**HW 9** **Due: See Canvas**

1. Consider a multiple linear regression model with *n* observations and *m* continuous predictor variables. Recall that multiple $R^2$ = *SSModel* / (*SSModel* + *SSError*) and the overall F-ratio is *F* = *MSModel*/*MSError*. Use this to express *F* in terms of $R^2$ (and *m* and *n*).

2. Health plans use many tools to try to control the cost of prescription medicines. For older drugs, generic substitutes that are equivalent to name-brand drugs are sometimes available at a lower cost. Another tool that may lower costs is restricting the drugs that physicians may prescribe. For example, if three similar drugs are available for treating the same symptoms, a health plan may require physicians to prescribe only one of them. Since the usage of the chosen drug will be higher, the health plan may be able to negotiate a lower price for that drug.
   The data in the file *drugcost.txt*, can be used to explore the effectiveness of these two strategies in controlling drug costs. The response variable is COST, the average cost of drugs per prescription per day, and predictors include GS (the extent to which the plan uses generic substitution, a number between zero, no substitution, and 100, always use a generic substitute if available) and RI (a measure of the restrictiveness of the plan, from zero, no restrictions on the physician, to 100, the maximum possible restrictiveness). Other variables that might impact cost were also collected, and are described in Table 1. The data are from the mid-1990s, and are for 29 plans throughout the United States with pharmacies administered by a national insurance company.

   Table 1: The Drug Cost Data (I have removed ID as it is not required).

| Variable | Description |
|---|---|
| *COST* | Average cost to plan for one prescription for one day, dollars |
| *RXPM* | Average number of prescriptions per member per year |
| *GS* | Percent generic substitution used by the plan |
| *RI* | Restrictiveness index (0=none, 100=total) |
| *COPAY* | Average member copayment for prescriptions |
| *AGE* | Average member age |
| *F* | Percent female members |
| *MM* | Member months, a measure of the size of the plan |
| *ID* | An identifier for the name of the plan |

   a. Fit a multiple linear regression model of COST on the other predictor variables. Report the overall F-ratio, p-value and the multiple R-squared.
   b. Summarize your results with regard to the importance of GS and RI. In particular, can we infer that more use of GS and RI will reduce drug costs?
   c. What are the other important variables and how do they affect the cost?
   d. Find a 95% confidence interval for the coefficient of GS and interpret it in the given context.
   e. Run model diagnostics and comment.

3. The data in *longnose_dace.txt* gives the data on the abundance of longnose dace in streams in Maryland. The columns are:
   i. stream : Name of the stream [**ignore** this column for fitting model]
   ii. longnosedace : number of longnose in a 75m section of the stream.
   iii. acreage : area (in acres) drained by the stream

iv. do2 : dissolved oxygen (in mg/litre)
v. maxdepth : maximum depth (in cm) of the 75-meter segment of stream
vi. no3 : nitrate concentration (mg/liter)
vii. so4 : sulfate concentration (mg/liter)
viii. temp : water temperature on the sampling date (in $^{o}$C).

a. Use a a multiple linear model with the number of longnose dace as the response variable. Run model diagnostics and check for model adequacy.
b. Use an appropriate transformation [I mean log] on the response, fit another MLR and run model diagnostics. Do things look better than (a)?

   Irrespective of your answer, use the model in (b) to answer the following questions.

c. Check for predictive power of each variable after accounting for the rest [That is, test each regression coefficient]. Report the table of estimates, s.e.'s, T-ratios and p-values. Write a short conclusion [1-2 sentences, include direction of an effect if present]
d. Find 95% confidence intervals for the coefficients of important predictor variables [i.e. the ones you found "significant" in c.]
e. What are the units in the study? Is this an experimental study or observational?
f. Use your model to predict the median abundance of longnose dace and a 95% interval for the abundance in a Maryland river where the conditions are as follows:
   i) acreage: 6298
   ii) do2: 9.7 mg/l
   iii) maxdepth = 65cm.
   iv) NO3 = 7.5mg/l
   v) SO4 = 44mg/l
   vi) temperature = 20$^{o}$C

4. [Sleuth 9.12, Mammal Brain Weights] The data is given in *brainsize.txt*. Ignore book's questions and answer the following.
   Use a MLR of log(Brain) on log(Body), log(Gestation) and log(Litter) to answer the following questions.
   a. Report the overall F-test and the p-value associated with it, multiple R-squared and the estimate of error variance (i.e. \sigma\hat}
   b. Report the table of estimates for the coefficients. Which predictors seem to be important in predicting brain size?
   c. Suppose we want to construct a 95% interval for the average brain size of red kangaroos. Should it be a confidence interval or a prediction interval?
   d. Obtain a 95% interval for (c). Use the following information on red kangaroos
      i) Average body weight = 63 lb. [warning: watch the unit]
      ii) Average gestation period = 34 days
      iii) Average Litter size = 2
      [Note: You'll not be double penalized if you get (c) wrong and if your in part (d) is consistent with (c). ]
   e. Obtain the plot of residual-vs-fitted values. Point out a limitation of your model. [Hint: Check the signs of the residuals associated with large body weights.]