Introduction to Multiple Linear Regression

In multiple linear regression, we assume that

1. The mean of a response variable Y is related to multiple explanatory variables X_1, $X_2, \ldots, X_m$ through the equation
$$\mu\{Y|X_1, X_2, \ldots, X_m\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m,$$
where $\beta_0, \beta_1, \ldots, \beta_m$ are unknown parameters. This equation defines the population regression function.
2. For any particular values of $X_1, \ldots, X_m$ the distribution of possible Y values is normal.
3. The normal distribution of Y values corresponding to any particular values of $X_1, \ldots, X_m$ has standard deviation σ. The standard deviation σ is unknown but assumed to be the same for any values of $X_1, \ldots, X_m$.
4. All Y values are independent of one another.

The β parameters are partial regression coefficients. The parameter β_j can help us understand the association that may exist between Y and X_j, after accounting for the relationship between Y and the other explanatory variables. The parameter β_j represents the average change in Y per unit increase in X_j when all other explanatory variables are held constant.

As an aid to understanding some basic issues in multiple linear regression we will examine some fictitious data on reading level of grade school children. The data set contains information on three variables – *readlev*, *tvtime*, and *grade* – for each of 15 grade school children. The variable *readlev* contains the students' scores on a standardized reading test. The variable *tvtime* is the number of hours spent watching television per week. The variable *grade* takes the value 1, 2, 3, 4, or 5 depending on whether a child is in 1st, 2nd, 3rd, 4th, or 5th grade.

1. There is a scatterplot of *readlev* vs. *tvtime* on page 1 of the output accompanying this worksheet. Based on your visual inspection of this scatterplot, would you say the linear correlation between *readlev* and *tvtime* is negative, zero, or positive?
2. Examine the regression output on page 2. What percent of the variation in *readlev* was explained by its linear relationship with *tvtime*.
3. Do you think the percentage of variation in *readlev* that is explained by *tvtime* would be lower or higher with real data?
4. Test $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. Give a test statistic, p-value, and conclusion.
5. Estimate the equation of the least squares regression line for predicting *readlev* from *tvtime*.
6. Predict *readlev* for a student who watches 9 hours of TV per week.
7. Estimate the mean change in *readlev* for each additional hour spent watching television each week.
8. Now examine the scatterplot of *readlev* vs. *tvtime* on page 3 of the output accompanying this worksheet. Numbers corresponding to the grades of the students have been used as points in the scatterplot. Consider for a moment only the points for 1st graders. Based on your visual inspection of this data, would you say the linear correlation between *readlev* and *tvtime* is negative, zero, or positive? What can you say about the correlation between *readlev* and *tvtime* for children in each of the other grades?

9. Page 4 shows SAS commands and output corresponding to the multiple regression of $Y =$ *readlev* on $X_1 =$ *tvtime* and $X_2 =$ *grade*. Just as in simple linear regression the total sum of squares can be written as the regression sum of squares plus the error sum of squares as follows.

$$\begin{array}{ccccc} SSTO & = & SSREG & + & SSE \\ \sum_{i=1}^{n}(Y_i - \bar{Y})^2 & = & \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 & + & \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \end{array}$$

In general $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X2 + \cdots + \hat{\beta}_m X_m$ where $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m$ are least-squares estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$. Find $SSTO$, $SSREG$, SSE, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ in the SAS output.

10. The ANOVA table is just like the simple linear regression ANOVA table except for a change in the degrees of freedom. Give the general formulas for the regression and error degrees of freedom.

11. Find an estimate of σ^2 in the regression output.

12. The F test associated with the multiple-regression ANOVA table tests

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0 \text{ against } H_A : \beta_j \neq 0 \text{ for some } j = 1, \ldots, m.$$

The null hypothesis says that there is no linear relationship between the mean of Y and any subset of the explanatory variables $X_1, X_2, \ldots, X_m$. Test this hypothesis for the regression of *readlev* on *tvtime* and *grade*. Give the test statistic, its degrees of freedom, a p-value, and a conclusion.

13. For the jth explanatory variable X_j, we may conduct a t-test of $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$. The null hypothesis says that the mean of Y does not change linearly with changes in X_j when all the other explanatory variables are held constant. Is there evidence that the mean of *readlev* changes linearly with *tvtime* when *grade* is held constant? Find a test statistic and p-value in the SAS output that will help you answer this question.

14. Is there evidence that the mean of *readlev* changes linearly with *grade* when *tvtime* is held constant? Find a test statistic and p-value in the SAS output that will help you answer this question.

15. Write down the equation of the least-squares regression function for predicting *readlev* from *tvtime* and *grade*.

16. Predict the *readlev* score for a 3rd grader who watches 9 hours of television per week.

17. For 3rd graders, estimate the mean change in *readlev* for each additional hour spent watching television each week.

18. The regression function estimated in problem 15 can be simplified for students in each grade. Write down a simplified regression function for students in 1st grade. Do the same for each group of students up through 5th graders.

19. In the problem above, you should have produced five equations relating *readlev* to *tvtime* – one for students in each grade. Sketch the lines corresponding to these equations on the scatterplot on page 3 of the output. Also sketch the original least squares regression line that relates *readlev* to *tvtime* without considering *grade*.