Stat 111 Spring 2025 Week 12: Multiple Linear Regression

1. **The Multiple Regression Model**

   a) Write out the Normal multiple regression model in scalar form and explain the least squares criterion. Describe the difference in the interpretation of the coefficient for an explanatory variable x_j in a multiple regression and the coefficient for x_j in a simple regression. Consider this contrived example: A cauldron contains nickels, dimes and pennies. Each subject randomly chooses a small handful of coins and reports x_1 = number of pennies, x_2 = total number of coins, and Y = total value of the coins in cents. Compare the coefficient for x_1 (number of pennies) in a simple regression of Y on x_1 and in a multiple regression of Y on x_1 and x_2. Summarize by noting how you can't refer to the "effect" of x on y without considering what other variables are in the model.

   b) Show how to compute the fitted multiple regression coefficients from simple regression output. Fit a simple regression of Y on x_2 and save the residuals. Then fit a regression of x_1 on x_2 and save the residuals. Finally, fit a simple regression of the first set of residuals on the second set of residuals. The fitted slope is the coefficient for x_1 in the multiple regression. Note how this estimates how the part of x_1 not explained by x_2 explains the part of y not explained by x_2. An analogous procedure gives the coefficient for x_2. The intercept assures that the fitted value is $\bar{y}$ when all x's equal their averages.

   c) Explain how part a is an example of Simpson's paradox. As a less contrived example, consider data on winning times in the Olympic 1500 meter run since 1948. The times for men and for women are both decreasing as year increases (people are running faster), but the least squares for time vs. year is positive. Make a graph to show how this is possible.

   d) A simple ANCOVA model has one numeric explanatory variable x_1 and one indicator variable x_2. Use the 1500m data as an example. Let x_1 be the year and let $x_2 = 1$ if the time was for a men's race ($x_2 = 0$ for women's races). For the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$, give an equation for the fitted time for men and for women as a function of year. Compare the interpretation of β_1 in this model to β_1 in a simple regression of time on year.

   e) Show how we can allow for non-parallel lines by adding an *interaction variable* $x_{12} = x_1 x_2$, the product of x_1 and the indicator variable. Carry out a test to see if an interaction is needed.

2. **Matrix Representation**
   The linear model becomes much easier to manage when represented in vector and matrix form. Let $\mathbf{Y}$ be the $n \times 1$ vector of $Y_1, \ldots, Y_n$ and let $\mathbf{X}$ be an $n \times p$ matrix with a column of 1's and $p-1$ columns for the explanatory variables $x_1, \ldots, x_{p-1}$. Define $\boldsymbol{\beta}$ to be the $p \times 1$ vector of $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

   a) Show that the regression model can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, for $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

   b) Explain why we **can't solve** $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ to get $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{Y}$, but we can solve $\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$ to get $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

   c) Show the generalization of the sum of squares expansion, and explain how it shows that $\hat{\boldsymbol{\beta}}$ is the MLE.

   $$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \;=\; (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \;+\; (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

   d) Recall that, for an $n \times 1$ vector random variable $\mathbf{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$, if $\mathbf{M}$ is an $m \times n$ matrix of constants, then $\mathbf{MY}$ has mean $\mathbf{M}\boldsymbol{\mu}$ and covariance matrix $\mathbf{MVM}^{\mathrm{T}}$. Show that $\hat{\boldsymbol{\beta}}$ has mean vector $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Verify the variance and covariance results we found for simple regression ($p = 2$).

3. **Fitted and Residual Vectors**

   a) A matrix $\mathbf{M}$ is a projection matrix if and only if $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$ and $\mathbf{MM} = \mathbf{M}$ ($\mathbf{M}$ is symmetric and idempotent). The vector of fitted mean values is $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{HY}$, and the vector of fitted residuals is $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, for $\mathbf{I}$ the $n \times n$ identity matrix. Show that $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{I} - \mathbf{H}$ are orthogonal projection matrices. Note that $\mathbf{Y} = \mathbf{HY} + (\mathbf{I} - \mathbf{H})\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\epsilon}}$.

   b) Identify the distributions of $\hat{\mathbf{Y}}$ and of $\hat{\boldsymbol{\epsilon}}$. Show that $\hat{\mathbf{Y}}$ and $\hat{\boldsymbol{\epsilon}}$ are uncorrelated by showing Var($\mathbf{Y}$) = Var($\hat{\mathbf{Y}}$)+Var($\hat{\boldsymbol{\epsilon}}$). Show directly that the cross-covariance of $\hat{\mathbf{Y}}$ and $\hat{\boldsymbol{\epsilon}}$ is 0. See Theorem A of Rice 14.4.4 on p. 577.

   c) Explain why the diagonal elements h_{ii} of the matrix $\mathbf{H}$ are referred to as "leverages." Show that the h_{ii} are all between 0 and 1, and sum to p, the dimension of $\mathbf{X}$. Hint: recall that the trace of a square matrix is the sum of its diagonal elements, and that for an $n \times m$ matrix $\mathbf{A}$ and an $m \times n$ matrix $\mathbf{B}$, trace($\mathbf{AB}$) = trace($\mathbf{BA}$). Note the complementary nature of $Var(\hat{Y}_i)$ and $Var(\hat{\boldsymbol{\epsilon}}_i)$. Demonstrate visually for the $p = 2$ case using the Correlation and Regression app:

      `http://digitalfirst.bfwpub.com/stats_applet/asset/applet_index.html`

      Vary a y_i value in a range to represent $\pm 2\sigma$. Show how the variability in $\hat{y}_i$ (the value on the regression line) and $\hat{\boldsymbol{\epsilon}}_i$ (the distance between the point and the line) differ for a point with x_i close to $\bar{x}$ and for a point with x_i far from $\bar{x}$. Explain how, with more than one explanatory variable, a point can have high leverage even if none of the corresponding x values is particularly extreme.

   d) Show how to use the mean square error and the leverages to find the standard error for a confidence interval for a mean response and for a prediction interval. For example, predict the mean BAC and an individual BAC for someone who drinks 9 beers and weighs 110 pounds (one of the observed combinations). Show how the intervals change as a function of weight.

4. **ANOVA Table for Regression**

   The Analysis of Variance table for regression breaks the total sum of squares (SST) into a sum of squares model (SSM) and a sum of squares error (SSE). These are used to compute R^2 and the F statistic.

   a) Write out the ANOVA table for multiple regression, with columns for Source, Degrees of Freedom, Sums of Squares and Mean Square and rows for Model, Error and Total. Define R^2 and Adjusted R^2. Explain how Adjusted R^2 is more useful for assessing the impact of adding another covariate to the model, but that it is not bounded by 0 and 1.

   b) Suppose $\mathbf{H}$ is an $n \times n$ projection matrix with tr($\mathbf{H}$) = m and let $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I})$. Show that $\mathbf{Z}'\mathbf{HZ} \sim \chi^2_{(m)}$. Use the symmetric matrix decomposition $\mathbf{H} = \boldsymbol{\Gamma\Lambda\Gamma}'$, where $\boldsymbol{\Gamma\Gamma}' = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \mathbf{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{H}$.

   c) Use the projection matrix theorem from part b to identify the sampling distributions of SSE, and the null sampling distributions for SSM and SST. State hypotheses for the whole-model F test and show that the generalized likelihood ratio test rejects for large values of the F ratio. Find the expected sums of squares under the null and alternative hypotheses to show that the F statistic tends to be larger under H_a. Also identify the null sampling distribution of R^2.

   d) Contrast the interpretations of the F test and the t tests for individual β_j's. As an example, a professor gives 4 quizzes and a final exam. They run a multiple regression of the final scores on the five quiz scores and find the F test is very significant, but none of the four t tests are significant. Explain how this is possible (and perhaps likely).

5. **Extra Sum of Squares and Lack of Fit Tests**
The Extra Sum of Squares F test allows tests of hypotheses that involve multiple β's, such as testing for a set of indicators defining multiple groups, or interaction effects. The lack of fit test is a special case of extra sum of squares, comparing the fitted model to a saturated model.

a) For testing whether a subset of the β_j's in a model are 0, we fit regression models with and without these β_j's and compare the sums of squares. To get an F test, we need for the extra sum of squares to be independent of the SSE from the full model. Let $\mathbf{H}_\mathrm{f}$ be the projection matrix for the full model and let $\mathbf{H}_\mathrm{r}$ be the projection matrix for the reduced model. The residuals for the full model are given by $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H}_\mathrm{f})\mathbf{Y}$ and the residuals for the reduced model by $\hat{\boldsymbol{\epsilon}}_\mathrm{r} = (\mathrm{I} - \mathbf{H}_\mathrm{r})\mathbf{Y}$. Show that the "extra sum of squares" is the sum of squares for the difference $\hat{\boldsymbol{\epsilon}}_\mathrm{r} - \hat{\boldsymbol{\epsilon}}_\mathrm{f}$, and that this difference in residual vectors is uncorrelated with the residuals for the full model. One fact you will need is that the composition $\mathbf{H}_\mathrm{f}\mathbf{H}_\mathrm{r} = \mathbf{H}_\mathrm{r}\mathbf{H}_\mathrm{f} = \mathbf{H}_\mathrm{r}$. This is true because $\mathbf{H}_\mathrm{r}$ projects into a subspace of R^n that is contained in the image space of $\mathbf{H}_\mathrm{f}$. Show how we get a statistic that follows an F distribution under the null hypothesis that the β_j's for the extra predictors are all 0.

b) Demonstrate the Extra Sum of Squares using the soccer data with weight predicted from height and position and possible interactions. Show graphs of weight by height for different positions, with and without interactions.

c) Explain how the Lack of Fit test is a special case of the Extra Sum of Squares test, where the full model is the null model, with the alternative is a *saturated* model with a different mean for every distinct combination of x values (e.g. a different mean weight for each distinct height, or for each position by height combination in the soccer data).

d) Suppose for a simple regression there are n individuals with m distinct x values $(1 < m < n)$. Define $\bar{Y}_j$ to be the average Y value for the n_j individuals with covariate x_j, $j = 1, \ldots, m$. Define $\mathbf{H}$ to be the usual regression projection matrix that projects the $n \times 1$ vector Y to the p dimensional subspace spanned by the columns of the $\mathbf{X}$ matrix. Define $\mathbf{H}_s$ to be the projection matrix for the saturated model, projecting $\mathbf{Y}$ to the vector of averages $\bar{Y}_j$, with each $\bar{Y}_j$ is replicated n_j times, for $j = 1, \ldots, m$. Show that

$$\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}/\sigma^2 \;=\; \mathbf{Y}'(\mathbf{I} - \mathbf{H}_s)\mathbf{Y}/\sigma^2 \;+\; \mathbf{Y}'(\mathbf{H}_s - \mathbf{H})\mathbf{Y}/\sigma^2.$$

Show that, under the null hypothesis that the Normal linear model is correct, this decomposition yields independent χ^2 random variables with $n - m$ and $m - p$ degrees of freedom. Use this fact to define the Lack of Fit F test. Demonstrate using the soccer player data.