Lecture 10
Multiple Linear Regression

STAT 512
Spring 2011

Background Reading
KNNL: 6.1-6.5
Topic Overview

• Multiple Linear Regression Model
Data for Multiple Regression

- $Y_i$ is the response variable (as usual)
- $X_{i1}, X_{i2}, \ldots X_{i,p-1}$ are the $p - 1$ explanatory variables for cases $i = 1, 2, \ldots, n$
- Example – In HW #2, you considered predicting freshman GPA based on ACT scores. Perhaps we consider high school GPA and an intelligence test as well. Then for this problem, we would be using $p = 4$. 
Multicollinearity

- Predictor variables are often correlated to each other.
- If predictor variables are highly correlated, they will be “fighting” to explain the same part of the variation in the response variable.
- Caution: Using highly correlated predictor variables in the same model will not lead to useful parameter estimates. Want to be careful of this.
Multiple Regression Model

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \]

- \( i = 1, 2, \ldots, n \) observations
- Assumptions exactly as before:
  \[ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \]
- \( Y_i \) is the value of the response variable for the \( i^{th} \) case.
- \( X_{ik} \) is the value of the \( k^{th} \) explanatory variable for the \( i^{th} \) case.
Multiple Regression Model (2)

- $\beta_0$ is the intercept (think multidimensional).
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression (slope) coefficients for the explanatory variables.
- Parameters as usual include all of the $\beta$'s as well as $\sigma^2$. These need to be estimated from the data.
Regression Plane/Surface
Model in Matrix Form

\[ Y = X \beta + \varepsilon \]

\[ \varepsilon \sim N \left( 0, \sigma^2 I \right) \]

\[ Y \sim N \left( X\beta, \sigma^2 I \right) \]
Design matrix $X$

$$X_{n \times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$
Coefficient matrix $\beta$

$$\beta_{p \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$
Least Squares Solution

- Minimize distances between point and response surface
- Find \( b \) to minimize
  \[
  SSE = (Y - Xb)' (Y - Xb)
  \]
- Obtain normal equations as before:
  \[
  X'Xb = X'Y
  \]
- Least Squares Solution as before:
  \[
  b = (X'X)^{-1} X'Y
  \]
Fitted Values / Residuals

• Fitted (predicted) values for the mean of $Y$ are

$$\hat{Y} = Xb = X(X'X)^{-1} X'Y = HY$$

• Residuals are

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

• Note formulas are same as before, with hat matrix:

$$H = X(X'X)^{-1} X'$$
“Linear” Regression Models

- The term *linear* here refers to the **parameters**, not the predictor variables.
- We can use *linear* regression models to deal with almost any “function” of a predictor variable (e.g. $X^2$, $\log(X)$, etc.)
- We cannot use *linear* regression models to deal with nonlinear functions of the parameters (unless we can find a transformation that makes them linear).
Types of Predictors

- Continuous Predictors – we are used to these.

- Qualitative Predictors
  - Two possible outcomes (e.g. male/female) represented by 0 or 1

- Polynomial Regression
  - Use squared or higher-ordered terms in regression model.
  - Typically always include lower order terms.
  - \[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_{p-1} X_i^{p-1} + \varepsilon_i \]
Types of Predictors (2)

- Using Transformed Variables
  - Transform one or more X’s
  - Transform Y

- Interaction Effects
  - Use Product of Predictor variables as an additional variable.
  - Each variable in the product included by itself as well.

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \]

- More on some of these models later...
Analysis of Variance

Formulas for sums of squares (in matrix terms) are the same as before

\[
SSR = \sum (\hat{Y}_i - \bar{Y})^2 = b'X'Y - \left( \frac{1}{n} \right) Y'JY
\]

\[
SSE = \sum (Y_i - \hat{Y}_i)^2 = e'e = Y'Y - b'X'Y
\]

\[
SSTO = \sum (Y_i - \bar{Y})^2 = Y'Y - \left( \frac{1}{n} \right) Y'JY
\]
Analysis of Variance (2)

- Degrees of Freedom depend on the model
- Always \( n - 1 \) total degrees of freedom
- Model degrees of freedom is equal to the number of terms in the model \( (p - 1) \)
  - Each variable has at least one term
  - May be additional terms for squares, interactions, etc.
- Error degrees of freedom is difference between total and model degrees of freedom \( (n - p) \).
Analysis of Variance (3)

- Mean Squares obtained by dividing SS by DF for each source.

- The mean square error (MSE) is still, always, and forever, the estimate of $\sigma^2$. 
## ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression</strong> (Model)</td>
<td>p-1</td>
<td>$\sum (\hat{Y}_i - \bar{Y})^2$</td>
<td>$\frac{SSR}{df_R}$</td>
<td>$\frac{MSR}{MSE}$</td>
</tr>
<tr>
<td><strong>Error</strong></td>
<td>n-p</td>
<td>$\sum (Y_i - \hat{Y}_i)^2$</td>
<td>$\frac{SSE}{df_E}$</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>n-1</td>
<td>$\sum (Y_i - \bar{Y})^2$</td>
<td>$\frac{SSTO}{df_T}$</td>
<td></td>
</tr>
</tbody>
</table>
F-test for model significance

- The ratio $F = \frac{MSR}{MSE}$ is again used to test for a regression relationship.

- Difference from SLR
  - Null Hyp: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
  - Alt Hyp: $H_a : \text{at least one } \beta_k \neq 0$

- Tests model significance, not individual variables; gives no indication of which variable(s) in particular are important
F-test for model significance (2)

• Under null, has F-distribution with degrees of freedom $p - 1$ and $n - p$.
• Reject if statistic is larger than critical value for $\alpha = 0.05$; or if p-value for test (given in SAS ANOVA table) is less than 0.05
• If reject, conclude at least one of the explanatory variables is important.
• If fail to reject, and sample size large enough (power), then none of the explanatory variables are useful.
Coefficient of Multiple Determination

\[ R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \]

- Measures the percentage of variation explained by the variables in the model.
- Additional variables will make \( R^2 \) go up; so cannot really use \( R^2 \) to determine whether a variable should be added.
Adjusted $R^2$

- $R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \left( \frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$

- Recall mean squares are SS adjusted by degrees of freedom

- $R_a^2$ can increase or decrease when a new variable is introduced into the model; depending on whether the decrease in SSE is offset by the lost degree of freedom.

- $R_a^2$ can be used to decide if variables are important in a model.
Inference for INDIVIDUAL Regression Coefficients

- We already have \( b \sim N(\beta, \sigma^2(X'X)^{-1}) \), so define

\[
\begin{align*}
    s^2 \{b\} &= \text{MSE} \times (X'X)^{-1} \\
    &= \left[ s^2 \{b\} \right]_{k,k}
\end{align*}
\]

- For individual \( b_k \), the estimated variance is the \( k^{th} \) diagonal element of this matrix:

\[
    s^2 \{b_k\} = \left[ s^2 \{b\} \right]_{k,k}
\]

Note: \( k=0,1,\ldots,p-1 \).
Confidence Intervals for $\beta_k$

- CI for $\beta_k$ is $b_k \pm t_{crit} s\{b_k\}$
- Critical value comes from t-distribution with $n - p$ degrees of freedom (DF for error)
- If CI includes zero, then we cannot reject $H_0 : \beta_k = 0$ (i.e. that variable is not significant when added to the model containing all of the other variables.)
Significance Test for $\beta_k$

- Is known as a variable-added-last test; tests whether the $k^{th}$ explanatory variable is important when added to all of the other variables in the model (i.e. it is a conditional test).

- Test statistic is as before: $t^* = b_k / s \{b_k\}$

- Compare to $t$-critical value on $n - p$ degrees of freedom.

- This is the test given in the model parameters section of SAS for PROC REG.
Upcoming in Lecture 11

- Case Study: Computer Science Student Data