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)
- $X_{i1}, X_{i2}, \cdots X_{i,p-1}$ 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:

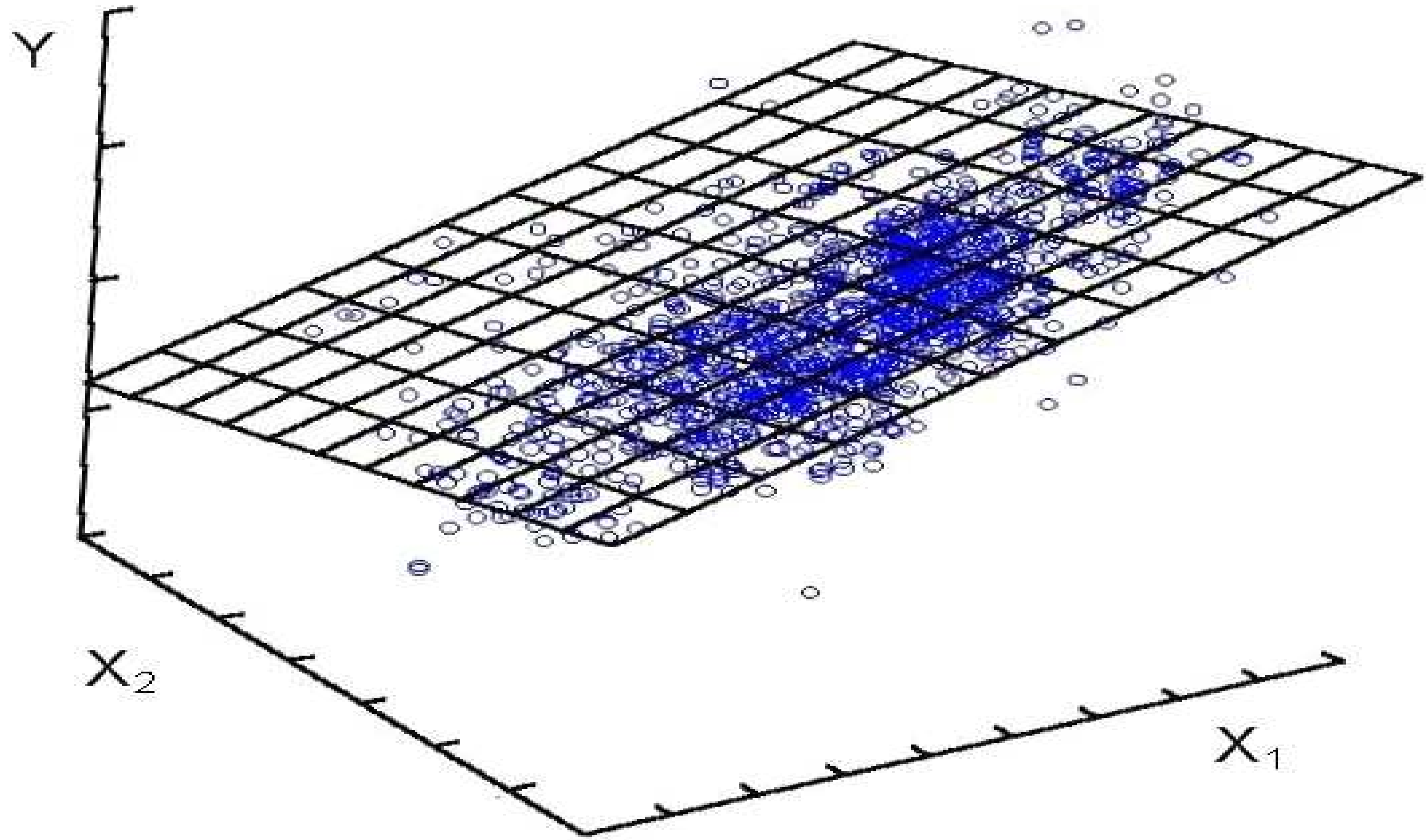
$$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

- 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)

- β_0 is the intercept (think multidimensional).
- $\beta_1, \beta_2, \cdots, \beta_{p-1}$ are the regression (slope) coefficients for the explanatory variables.
- Parameters as usual include all of the β's as well as σ^2. These need to be estimated from the data.

Regression Plane/Surface



Model in Matrix Form

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \underset{n\times n}{\mathbf{I}}\right)$$

$$\mathbf{Y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$$

Design matrix X

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

Coefficient matrix β

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Least Squares Solution

- Minimize distances between point and response surface
- Find b to minimize

$$SSE = (\mathbf{Y} - \mathbf{Xb})' (\mathbf{Y} - \mathbf{Xb})$$

- Obtain normal equations as before:

$$\mathbf{X'Xb} = \mathbf{X'Y}$$

- Least Squares Solution as before:

$$\mathbf{b} = \left(\mathbf{X'X}\right)^{-1} \mathbf{X'Y}$$

Fitted Values / Residuals

- Fitted (predicted) values for the mean of Y are

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

- Residuals are

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

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

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

“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

$$SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

$$SSE = \sum\left(Y_i - \hat{Y}_i\right)^2 = \mathbf{e}'\mathbf{e} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

$$SSTO = \sum\left(Y_i - \bar{Y}\right)^2 = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

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 σ^2.

ANOVA Table

Source	df	SS	MS	F
Regression (Model)	p-1	$\sum\left(\hat{Y}_i - \bar{Y}\right)^2$	$\dfrac{SSR}{df_R}$	$\dfrac{MSR}{MSE}$
Error	n-p	$\sum\left(Y_i - \hat{Y}_i\right)^2$	$\dfrac{SSE}{df_E}$	
Total	n-1	$\sum\left(Y_i - \bar{Y}\right)^2$	$\dfrac{SSTO}{df_T}$	

F-test for model significance

- The ratio F = MSR / MSE is again used to test for a regression relationship.
- Difference from SLR
  - Null Hyp: $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$
  - Alt Hyp: $H_a :$ 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

- $R^2 = \dfrac{SSR}{SSTO} = 1 - \dfrac{SSE}{SSTO}$
- 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 $\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, so define

$$\underset{p \times p}{\mathbf{s^2}\{\mathbf{b}\}} = MSE \times \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

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

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

Note: k=0,1,…,p-1.

Confidence Intervals for β_k

- CI for β_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 <u>when added to the model containing all of the other variables</u>.)

Significance Test for β_k

- Is known as a ***variable-added-last*** test; tests whether the k^{th} explanatory variable is important <u>when added to all of the other variables in the model</u> (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