Class 9

- Data, model and inference for multiple regression

Data for Multiple Regression

- Y_i is the response variable
- X_{i1}, X_{i2}, … , X_{ip-1} are *p-1* explanatory variables for cases *i = 1 to n*

Multiple Regression Model

- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1} + \xi_i$
- Y_i is the value of the response variable for the i^{th} case
- β_0 is the intercept
- β_1, β_2, … , β_{p-1} are the regression coefficients for the explanatory variables

Multiple Regression Model (2)

- X_{ik} is the value of the k^{th} explanatory variable for the i^{th} case
- ξ_i are independent normally distributed random errors with mean 0 and variance σ^2

Many interesting special cases

- $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_{p-1} X_i^{p-1} + \xi_i$
- Xs can be *indicator* or *dummy* variables with 0 and 1 (or any other two distinct numbers) as possible values
- Interactions
- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \xi_i$

Multiple Regression Parameters

- β_0 the intercept
- β_1, β_2, … , β_{p-1} the regression coefficients for the explanatory variables
- σ^2 the variance of the error term

Model in Matrix Form

$$\underset{nx1}{\mathbf{Y}} = \underset{nxp}{\mathbf{X}} \underset{px1}{\beta} + \underset{nx1}{\xi}$$

$$\xi \sim \mathrm{N}(0, \sigma^2 \mathbf{I})$$

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

Least Squares

$$\mathbf{Y} = \mathbf{X}\beta + \xi$$

$$\min(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$$

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

Least Squares Solution

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

Fitted (predicted) values

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

$$= \mathbf{HY}$$

Residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$= \mathbf{Y} - \mathbf{HY}$$

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

$\mathbf{I} - \mathbf{H}$ is symetric and idempotent i.e.

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$$

Covariance Matrix of residuals

- **Cov(e)=σ^2(I-H)(I-H)'= σ^2(I-H)**
- **So,**
- **Var(e_i)= σ^2(1-h_{ii})**
- **h_{ii}= $X'_i(X'X)^{-1}X_i$**
- **X'_i =(1,X_{i1},…,$X_{i(p-1)}$)**
- **Residuals are usually correlated**
- **Cov(e_i,e_j)= - $\sigma^2 h_{ij}$**

Estimation of σ

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-p}$$

$$= \frac{(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})}{n-p}$$

$$= \frac{SSE}{df\mathbf{e}} = MSE$$

$$s = \sqrt{s^2} = Root\ \ MSE$$

Distribution of b

- b= $(X'X)^{-1}X'Y$
- $Y \sim N(X\beta, \sigma^2 I)$
- $E(b)=((X'X)^{-1}X')X\beta=\beta$
- $Cov(b)=\sigma^2 ((X'X)^{-1}X') ((X'X)^{-1}X')'$
  $= \sigma^2 (X'X)^{-1}$

Estimation of variance of b

- $b \sim N(\beta, \sigma^2 (X'X)^{-1})$
- $\sigma^2 (X'X)^{-1}$
- Is estimated by
- $s^2 (X'X)^{-1}$

ANOVA Table

- To organize arithmetic
- Sources of variation are
  - Model
  - Error or Residual
  - Total
- SS and df add
  - SSM + SSE =SST
  - dfM + dfE = dfT

SS

$$\text{SSM} = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

df

$$df\ \text{M} = \text{p - 1}$$

$$df\ \text{E} = \text{n - p}$$

$$df\ \text{T} = \text{n - 1}$$

Mean Squares

$$\text{MSM} = \text{SSM/dfM}$$

$$\text{MSE} = \text{SSE/dfE}$$

$$\text{MST} = \text{SST/dfT}$$

Mean Squares (2)

$$\text{MSM} = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 /(p-1)$$

$$\text{MSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 /(n-p)$$

$$\text{MST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 /(n-1)$$

ANOVA Table

Source	SS	df	MS	F
Model	SSM	dfM	MSM	MSM/MSE
Error	SSE	dfE	MSE	
Total	SST	dfT	(MST)	

ANOVA F test

- H_0: $\beta_1 = \beta_2 = \ldots \beta_{p-1} = 0$
- H_a: $\beta_k \neq 0$, for at least one *k=1, … , p-1*
- Under H_0, F ~ F(p-1,n-p)
- Reject H_0 if F is large, use P value

Study of CS students

- Study of computer science majors at Purdue
- Large drop out rate
- Can we find predictors of success ?
- Predictors must be available at time of entry into program

Data available

- GPA after three semesters
- High school math grades
- High school science grades
- High school English grades
- SAT Math
- SAT Verbal
- Gender (of interest for other reasons)

Example

```
Data a1;
   infile‘…/csdata.dat';
   input id gpa
         hsm hss hse
         satm satv genderm1;
proc reg data=a1;
   model gpa=hsm hss hse;
run;
```

CS ANOVA Table

Source	DF	Sum of Squares	Mean Square	F
Model	3	27.71	9.23	18.86
Error	220	107.75	0.48	
Total	223	135.46		

Hypothesis Tested by F

- H_0: $\beta_1 = \beta_2 = \ldots \beta_{p-1} = 0$
- F = MSM/MSE
- Reject H_0 if the P value is $\leq$ .05

What do we conclude?

Sum of Source	Mean Square	F	P
Model	9.23744	18.86	<.0001
Error	0.48977		

R^2

- The squared multiple regression correlation (R^2) gives the proportion of variation in the response variable explained by the explanatory variables included in the model
- It is usually expressed as a percent
- It is sometimes called the coefficient of multiple determination

R^2 (2)

- R^2 = SSM/SST, the proportion of variation explained
- $R^2 = 1 -$ (SSE/SST), 1 – the proportion of variation not explained
- $F = [(R^2)/(p-1)] / [(1- R^2)/(n-p)]$
- For the CS data R^2= 27.71/135.46=0.205

- The P-value for the F significance test tells us one of the following:
  - there is no evidence to conclude that *any* of our explanatory variables can help us to model the response variable using this kind of model (P $\geq$ .05)
  - one or more of the explanatory variables in our model *is* potentially useful for predicting the response variable in a linear model (P $\leq$ .05)

Inference for individual regression coefficients

- $b \sim N(\beta, \sigma^2 (X'X)^{-1})$
- $S^2_b = s^2 (X'X)^{-1}$
- $s^2(b_i) = S^2_b(i,i)$
- CI: $b_i \pm t_c s(b_i)$, where $t_c = t(.975, n-p)$
- Significance test for $H_{0i}: \beta_i, = 0$ uses the test statistic $t = b_i/s(b_i)$, df=dfE=n-p, and the P-value computed from the t(n-p) distribution

CS Data results

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.58988	0.29424	2.00	0.0462
hsm	1	0.16857	0.03549	4.75	<.0001
hss	1	0.03432	0.03756	0.91	0.3619
hse	1	0.04510	0.03870	1.17	0.2451

Example

- Dwaine Studios operates portrait studios in 21 cities
- Y is sales
- X_1 is number of persons aged 16 and under
- X_2 is per capita disposable income
- n = 21 cities

Check the data

```
data a1;
infile `…/ch06fi05.txt';
   input young income sales;
 proc print data=a1;
run;
```

Print

```
Obs     young     income     sales

  1      68.5      16.7      174.4
  2      45.2      16.8      164.4
  3      91.3      18.2      244.2
  4      47.8      16.3      154.6
  5      46.9      17.3      181.6
```

Proc Reg

```
proc reg data=a1;
   model sales=young income;
run;
```

Output

```
        Sum of  Mean
Source DF    Sq    Sq  F  Pr > F

Model  2  24015 12008 99. <.0001
Error 18   2180   121
C Tot 20  26196
```

Output (2)

```
Root MSE 11.00  R-Sq 0.916
Dep Mean 181.9  Adj R-Sq 0.907
Coeff Var  6.0
```

Output (3)

```
           Par   St
Var    DF  Est  Err     t      P

Int     1 -68.8 60.0 -1.15 0.2663
young   1 1.454 0.21  6.87 <.0001
income  1 9.365 4.06  2.30 0.0333
```

CLB option

```
proc reg data=a1;
 model sales=young income/clb;
run;
```

Output (4)

```
95% Confidence Limits

Int     -194.94   57.23
young      1.00    1.89
income     0.82   17.90
```

Estimation of $E(Y_h)$

- **X_h is now a vector**
- **$(1, X_{h1}, X_{h2}, \ldots, X_{h1})'$**
- **We want a point estimate and a confidence interval for the subpopulation mean corresponding to X_h**

Theory for $E(Y_h)$

$$E(Y_h) = \mu_h = X'_h \beta$$

$$\hat{\mu}_h = X'_h b$$

$$\sigma^2(\hat{\mu}_h) = X'_h \Sigma_b X_h = \sigma^2 X'_h (X'X)^{-1} X_h$$

$$s^2(\hat{\mu}_h) = s^2 X'_h (X'X)^{-1} X_h$$

$$CI : \hat{\mu}_h \pm s\ (\hat{\mu}_h) t_{(0.975,n-p)}$$

Estimation of $E(Y_h)$ (CLM)

```
proc reg data=a1;
 model sales=young income/clm;
 id young income;
run;
```

$E(Y_h)$ CLM Output

Obs	young	inco	DV sal	Pre Val	SE Pre
1	68.5	16.7	174	187	3.8
2	45.2	16.8	164	154	3.5
3	91.3	18.2	244	234	4.5
4	47.8	16.3	154	153	3.2

$E(Y_h)$ CLM Output (2)

Obs	95% CL	Mean
1	179.1146	195.2536
2	146.7591	161.6998
3	224.7569	244.0358
4	146.5361	160.1210

Prediction of Y_h

- X_h is now a vector
- $(1, X_{h1}, X_{h2}, \ldots, X_{h1})'$
- We want a prediction for Y_h with an interval that expresses the uncertainty in our prediction

Theory for Y_h

$$Y_h = X'_h \beta + \xi$$

$$\hat{Y}_h = \hat{\mu}_h = X'_h b$$

$$\sigma^2(pred) = Var(\hat{Y}_h - Y_h) = Var\ \hat{Y}_h + \sigma^2$$
$$= \sigma^2 (1 + X'_h (X'X)^{-1} X_h)$$

$$s^2(pred) = s^2 (1 + X'_h (X'X)^{-1} X_h)$$

$$CI : \hat{\mu}_h \pm s\ (pred) t_{(0.975,n-p)}$$

Prediction of Y_h (CLI)

```
proc reg data=a1;
 model sales=young income/cli;
 id young income;
run;
```

Prediction Intervals Output

Obs	95% CL Predict	
1	162.6910	211.6772
2	129.9271	178.5317
3	209.3421	259.4506
4	129.2260	177.4311