Lecture 6

- Data, model and inference for multiple regression

Data for Multiple Regression

- Y_i is the response variable
- $X_{i1}, X_{i2}, \ldots, X_{ip-1}$ 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
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ 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
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ 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

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

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
  - 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

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

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

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

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: $\beta_1 = \beta_2 = \dots \beta_{p-1} = 0$
- 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

```
cs<-read.table('csdata.dat',
col.names=c("id", "gpa", "hsm",
"hss", "hse", "satm", "satv",
"gen"));
reg1<-lm(gpa~hsm+hss+hse, cs);
Anova(reg1);
summary(reg1);
```

CS ANOVA Table

```
    Df Sum   Mean   F   Pr(>F)
hsm 1 25.81   25.8 52.7 6.6e-12
hss 1  1.24   1.23  2.5 0.1134
hse 1  0.67   0.67  1.4 0.2451
Res 220 107.7 0.49

F-stat: 18.86 on 3 and 220 DF
p-value: 6.359e-11
```

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 ?

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
- $F = [\ (R^2)/(p-1)\ ]\ /\ [\ (1- R^2)/(n-p)\ ]$

- 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 ($P \leq .05$)

Stat 512 Class 14

- Review multiple linear regression
  - data
  - Model

  Inference for multiple regression (continued)

  Diagnostics and remedies

Data for Multiple Regression

- Y_i is the response variable
- $X_{i1}, X_{i2}, \ldots, X_{ip-1}$ are *p-1* explanatory variables for cases *i = 1* to *n*
- $Y_i, X_{i1}, X_{i2}, \ldots, X_{ip-1}$ is the data for case i, where *i = 1* to *n*
- *Y | X is the data*

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
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ 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

Model in Matrix Form

$$\underset{n\times1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times1}{\beta} + \underset{n\times1}{\xi}$$

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

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

Least Squares Solution

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

Estimation of σ^2

s^2 = MSE

s = Root MSE

ANOVA F test

- H_0: $\beta_1 = \beta_2 = \dots \beta_{p-1} = 0$
- 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, using P value we reject if P leq 0.05

R^2

- R^2 = SSM/SST, the proportion of variation explained by the explanatory variables

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: $b_i \pm t^* s(b_i)$, where $t^* = t(.975, n-p)$
- Significance test for H_{0i}: β_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

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

R code

```
dwst<-read.table('ch06fi05.txt',
col.names=c("young", "income",
"sales"));
reg<-lm(sales~young+income,
dwst);
summary(reg)
```

```
        Est     Std     t      p-val
Int    -68.86 60.02   -1.15 0.2663
young    1.45  0.21    6.87 2e-06
income   9.37  4.06    2.31 0.0333

Residual standard error: 11.01
on 18 degrees of freedom
Multiple R-squared: 0.9167,
Adjusted R-squared: 0.9075
F-statistic:  99.1 on 2 and 18
DF,  p-value: 1.921e-10
```

```
confint(reg)

              2.5 %     97.5 %
Int   -194.9480130 57.233867
young    1.0096226  1.899497
income   0.8274411 17.903560
```

Estimation of $E(Y_h)$

- X_h is now a vector
- $(1, X_{h1}, X_{h2}, \ldots, X_{h1})'$
- We want an 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)

```
predict.lm(reg,
interval='confidence');
```

$E(Y_h)$ CI Output

```
      fit      lwr      upr
1  187.1841 179.1146 195.2536
2  154.2294 146.7591 161.6998
3  234.3963 224.7569 244.0358
4  153.3285 146.5361 160.1210
5  161.3849 152.0778 170.6921
```

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 (PI)

```
predict.lm(reg,
interval='prediction');
```

Prediction Intervals Output

```
        fit      lwr      upr
1  187.1841 162.6910 211.6772
2  154.2294 129.9271 178.5317
3  234.3963 209.3421 259.4506
4  153.3285 129.2260 177.4311
5  161.3849 136.4566 186.3132
```

Diagnostics

- Look at the distribution of each variable
- Look at the relationship between pairs of variables
- Plot the residuals versus
  - Each explanatory variable
  - Time

Diagnostics (2)

- Are the residuals approximately normal?
  - Look at a histogram
  - Normal quantile plot
- Is the variance constant?
  - Plot the squared residuals vs anything that might be related to the variance (e.g. residuals vs predicted)

Remedial measures

- Transformations such as Box-Cox
- Analyze without outliers

Scatter Plot Matrix

pairs(~gpa+satm+satv,cs)