STAT 340, MHC, F2014

Intro Activity

- Have shoebill pix up as they walk in.
- Ask for theories.
- Show subsequent pix and ask for for revisions

Theories

•

•

•

•

At 50 k ft Science

- Make observations in the natural world
  collect data
- Generate theories/models.
- Evaluate the consistency between the model and the data.
- Gather more data, revise theories
- Iterate ...

Regression: the central statistical model

- Express one variable of interest (response) as a function of some other variables (predictors)

$$Y = f(X_1, X_2, X_3 \ldots)$$

- It is a probabalistic model

$$P(Y \mid X_1, X_2 \ldots X_n)$$

Examples from NYT

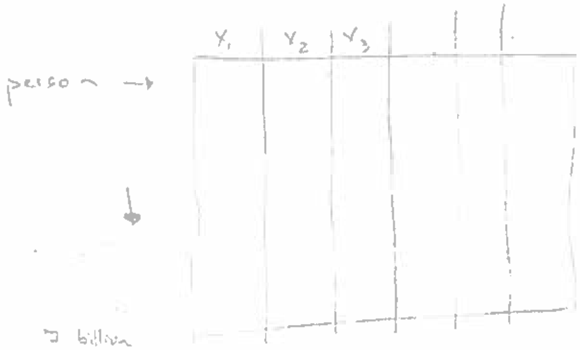
Activity

Let's build a model to predict someone's actual age.

Our only predictor for now: a guess based on a photograph

Week 2 Day 1: Describing distributions

→ Taxonomy of Data



What variables do we have data on?

What type of data is each?

Write out pseudo-code for the subsetting operator they do.

↑ mean photo.

Taxonomy of Data

- numerical
  - continuous (R: numeric)
  - discrete (R: integer)
- categorical (R: factor)
  - unordered
  - ordinal

One variable

Numerical Descriptors

center: mean, median, mode

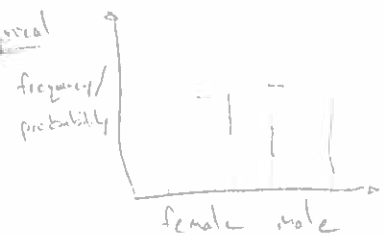
spread: standard deviation, variance, IQR

shape: unimodal/bimodal

skew:

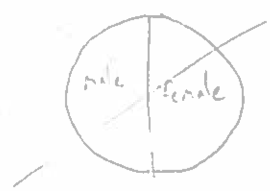
Graphical Descriptors

categorical



← barplot

barplot(table(netgrade))



Numerical

- must be aggregated by binning (hist) or smoothing (density)
- beware of binwidth, bandwidth selection!

Boxplot

- shows median, Q1, Q3, and outliers.
- good for side-by-side comparisons

Two Variables

Numerical summaries (2 num)

- shape: linear, quadratic
- direction: positive/neg, curvature
- strength: how tightly clustered? correlation coefficient r

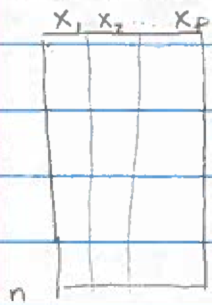
Graphical

Scatterplot

- time often on x
- beware of overplotting
- response often on y

Week 3C: From Data to Distributions

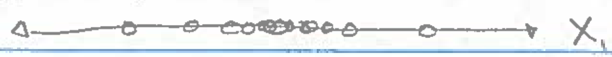
Representations of Data



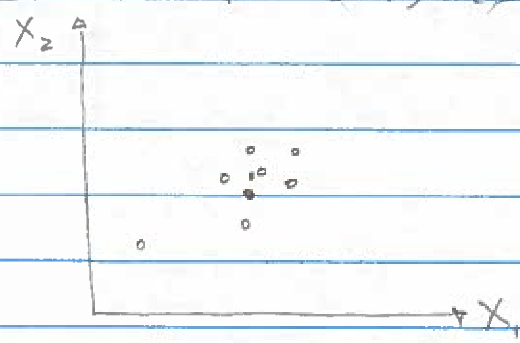
n = # rows

p = # variables

$\boxed{1D}$ Dotplot $\{X_1\}$



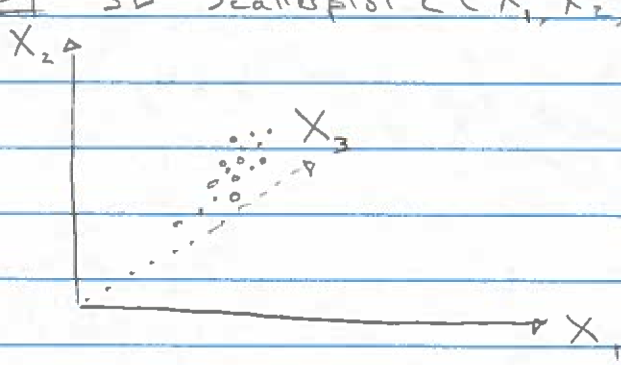
$\boxed{2D}$ Scatterplot $\{(X_1, X_2)\}$



* In all of these representations, the entire data set is recoverable. Zero information loss.

* There is value in taking the big picture view of the general, smoother structure at play

$\boxed{3D}$ 3D Scatterplot $\{(X_1, X_2, X_3)\}$



From Data to Distributions

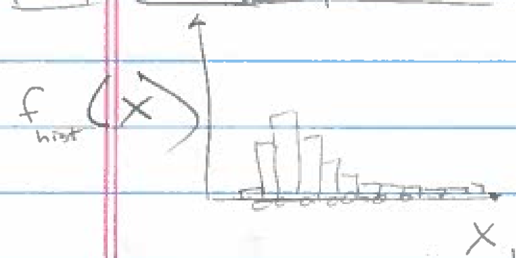
Question of interest: Where is the data dense? Where is it sparse?

To answer this, we need (A) a measure of data density.

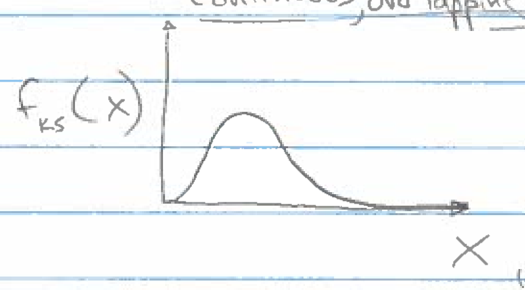
proportion of obs.
length/area/vol

(B) a functional map from X → density

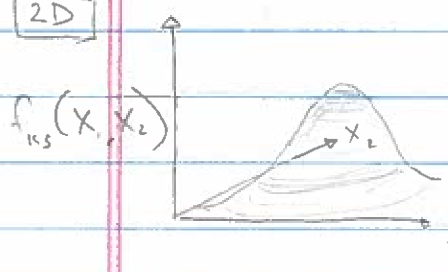
$\boxed{1D}$ Discrete, disjoint bins



Continuous, overlapping bins



$\boxed{2D}$



3 Types of distributions

Joint Distribution: $f(x_1, x_2)$; what is the dist. of X_1 and X_2

Marginal Distribution: $f(X_1)$, $f(X_2)$ what is the dist of X_1 over all X_2?

Conditional Dist: $F(X_1 \mid X_2 = x_2)$
what is the dist of X_1 given X_2 takes on this particular value?

Week 4A Simple Linear Regression

For Weds: Read 2.1 - 2.4

Notation

A random variable X takes particular values w/ particular probs.

eg1 Let X ~~be the value on a rolled fair die.~~

x	$P(X=x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

eg 2. Let X be the # of heads in 3 fair coin flips

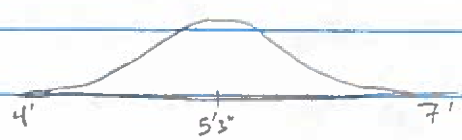
HHH
HTT
HHT
TTT
HTH
TTH
THT
THH

x	$P(X=x)$
0	1/8
1	3/8
2	3/8
3	1/8

eg. 3

Let X be the height of a randomly selected female college student. X is continuous between 4' and 7', normally distributed w/ mean 5'3" and variance 9".

$$X \sim N(\mu = 5'3", \sigma^2 = 9")$$



Expected Value:

$$E(X) = \sum_{\text{all } x} x P(X=x)$$

eg3 $$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \mu$$

eg. 2

x	$P(X=x)$	
0	1/8	
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8
		12/8 = 1.5

Variance

eg 2 $Var(X) = \sum_{all\ x} (x-\mu)^2 P(X=x) = \sigma^2$ eg 3. $Var(X) = \int (x-\mu)^2 f(x)\,dx$

Describe the relationship

Shape:

Direction:

Strength:

What you're describing is $E(Y|X=x)$ a conditional mean.

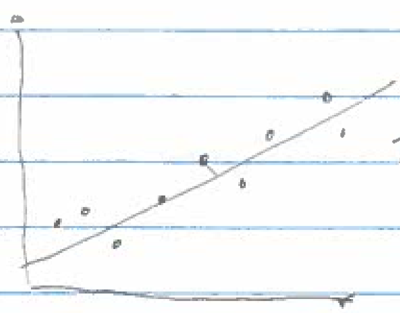
Simple Linear Regression

$$E(Y|X=x) = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} x$$ ← mean function

$$Y_i = E(Y|X=x) + e_i = \beta_0 + \beta_1 x + \underset{\text{random error}}{e_i}$$ ← data generating function

$E(e|X) = 0$
$Var(e|X) = \sigma^2$

Which line?



two options

(1) minimize distance from (x_i, y_i) to line

☆ (2) minimize distance from (y_i) to line

residual: $\hat{e}_i = y_i - \hat{y}_i$

minimize the sum of squared residuals

$\sum_{i=1}^{n} \hat{e}_i^2$ (aka RSS)

Fitting the Least Squares Line

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Since we have data for x and y, we can treat them as constant and view the RSS as a function of $\hat{\beta}_0$ and $\hat{\beta}_1$. To find the values that minimize RSS, we can take the derivative and set to zero.

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

and

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

rearrange:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

normal eqns.

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \hat{\beta}_0 - \sum \hat{\beta}_1 x_i^2 = 0$$

$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

Solve for $\hat{\beta}_0$

$$\frac{n\hat{\beta}_0}{n} = \frac{\sum y_i}{n} - \frac{\hat{\beta}_1 \sum x_i}{n} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\sum xy = (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x + \hat{\beta}_1 \sum x^2$$

$$= \bar{y} \sum x - \beta_1 \bar{x} \sum x + \beta_1 \sum x^2$$

$$\sum xy - n\bar{y}\bar{x} = \beta_1 \left(\sum x^2 - n\bar{x}^2\right)$$

rearrange

$$\hat{\beta}_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

Week 6b Outliers and leverage

We want a measure for leverage that is based on

1. the distance x_i is away from the x's.
2. the extent to which $\hat{y}_i$ depends on y_i

Recall $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\hat{\beta}_1 = \sum_{j=1}^{n} \frac{x_j - \bar{x}}{SXX} y_j$$

centered

all the data

$$\hat{y}_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$$

$$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

$$= \frac{1}{n} \sum_{j=1}^{n} y_j + \sum_{j=1}^{n} \frac{x_j - \bar{x}}{SXX} y_j (x_i - \bar{x})$$

isolate the y

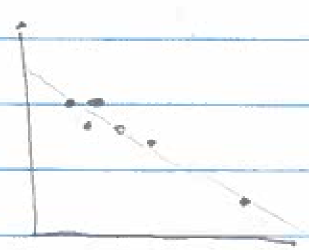
$$= \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX} \right] y_j$$

$$= \sum_{j=1}^{n} h_{ij} y_j$$

---

* predicted $\hat{y}_i$ is a weighted sum of the all the y's, with $\sum_{j=1}^{n} h_{ij} = 1$

$$\hat{y}_i = h_{i1} y_1 + h_{i2} y_2 + \dots + h_{ii} y_i + \dots h_{in} y_n$$

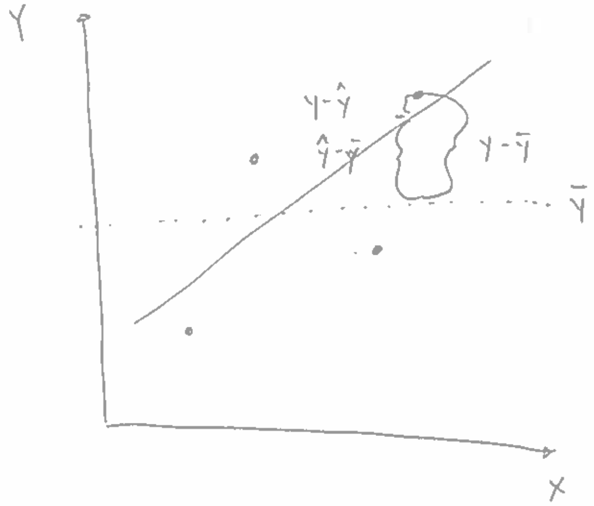


A measure of model fit: R^2 + Adjusted R^2

Recall R; Pearsons Corr. Coef: measures strength of linear relationship. $R \in [-1, 1]$

R^2: the proportion of the total variability in the y's explained by the regression model

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{RSS}{SST}$$



Total variation in the y

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$SSreg$: Total variation explained by $\hat{y}$

$$= \sum (\hat{y}_i - \bar{y})$$

RSS: Total unexplained variation

$$= \sum (y_i - \hat{y}_i)^2$$

$$SST = SSreg + RSS$$

No linear trend

RSS: big, SSreg: small $\Rightarrow$ R^2 near 0

Strong linear trend

RSS: small, SSreg: big $\Rightarrow$ R^2 near 1

Caution: Adding complexity (i.e. additional terms) to our model will always increase R^2, even if the term was bogus

Solution: Use adjusted R^2

$$R^2_{adj} = 1 - \frac{RSS(n-p-1)}{SST(n-1)}$$

n: sample size
p: # predictors

Now, new terms have to decrease RSS by more than the penalty to increase R^2_{adj}.

Model Fitting: a Metaphor

△ rice
□ quinoa
· salt →

← Data
- signal (△, □) linear, quadratic
- noise (·)

sieve (rice istopped) →

← Fitted #1
Mean fxn: captures deterministic portion (signal)
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

(quinoa + salt pass through) →

← residuals
(noise + some signal)
aka residual plot
w/ structure

Key points

- Your fitted model is a way to seperate signal from noise.
- A good model will leave behind residuals w/o any structure.
- Be wary of building a model that considers <u>everything</u> to be signal. This is <u>overfitting</u> and will damage prediction

---

Ockam's Razor aka "Laws of Parsimony"

"Plurality should not be posited w/o necessity".
complexity

aka: "Among similar explanations, the simpler is better"
→ useful heuristic in model building.

Wednesday Oct. 22

- Reading: 5.3, 5.2, Linear Algebra.
- Proposals due Wednesday.
- Survey
- Analysis of Covariance
-

m1 Questions

- Write out the eqn for the line.
- Is volume a significant predictor?
- How much of the variation in weight is explained by the model containing volume?

ANCOVA Interpretation

- Simple Linear Regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Parallel Lines $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_2$

x_1: continuous (volume)
x_2: categorical (cover) $\begin{cases} 1 & \text{if pb} \\ 0 & \text{if hb} \end{cases}$

For paperbacks: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ (is 1)

$\hat{y} = \underbrace{(\hat{\beta}_0 + \hat{\beta}_2)}_{\text{new intercept}} + \hat{\beta}_1 x_i$

For Hardbacks: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ (is 0)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$

same slope
different intercept
∴ parallel

- Two intercepts, two slopes

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

pb: $\hat{y} = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) x_1$

hb: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

October 21st

Matrix Multiplication - a review

ex 1

$$AB = \overset{A}{\begin{bmatrix} 2 & 5 & -1 \\ -4 & 3 & -3 \end{bmatrix}} \overset{B}{\begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 4 \end{bmatrix}} = \begin{bmatrix} 2\cdot 2 + 5\cdot 1 + (-1)\cdot 0 = & 2\cdot 0 + 5\cdot 1 + (-1)\cdot 4 = 1 \\ (-4)\cdot 2 + 3\cdot 1 + (-3)0 = & (-4)\cdot 0 + 3\cdot 1 + (-3)4 = 9 \end{bmatrix} = \begin{bmatrix} 9 & 1 \\ -5 & -9 \end{bmatrix}$$

2×3 3×2 ok

- You can only multiply matrices w/ the same inner dimension.
- The resulting matrix takes its dimensions from the outer dimensions.
- Each entry in the resulting matrix is a dot product of the corresponding row of the 1st matrix and the column of the 2nd.

Q40 ex 2 What is BA?

$$\begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 2 & 5 & -1 \\ -4 & 3 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 10 & -2 \\ -2 & 8 & 2 \\ -16 & 12 & 12 \end{bmatrix}$$

3×2 2×3

Def: Transpose: $A' = \begin{bmatrix} 2 & 4 \\ 5 & 3 \\ -1 & -3 \end{bmatrix}$ The transpose of a matrix is a matrix of flipped dim where the row, col indices have been flipped

ex 3 what is $B'A'$?

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 2 & -4 \\ 5 & 3 \\ 1 & -3 \end{bmatrix} = \begin{bmatrix} 9 & -5 \\ 1 & 9 \end{bmatrix}$$

* $B'A' = (AB)'$

Def Symmetric: a matrix is symmetric if $a_{ij} = a_{ji}$ (must be square)

$$S = \begin{bmatrix} 1 & 3 & 6 \\ 3 & 2 & 1 \\ 6 & 1 & 0 \end{bmatrix} = S'$$

Def Diagonal: a matrix is diagonal if $a_{ij} = 0$ for $i \neq j$.

$$D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Def. Identity matrix: a diagonal matrix with $a_{ij} = 1$ for $i=j$

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Recall from algebra

$a \cdot 1 = a$ * 1 is the identity operator

$$AI = \begin{bmatrix} 2 & 5 & -1 \\ -4 & 3 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 5 & -1 \\ -4 & 3 & -3 \end{bmatrix} = A$$

2×3 3×3

$a \cdot \frac{1}{a} = 1$

Def. Inverse the inverse of a matrix yields I when multiplied.

$AA^{-1} = I$ * A must be square!*

* Not all matrices can be inverted

$$M = \begin{bmatrix} 1 & 3 & -2 \\ 2 & 5 & -3 \\ -3 & 2 & 4 \end{bmatrix} \begin{bmatrix} 14 & -8 & -1 \\ -17 & 10 & 1 \\ 19 & 11 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrix Notation for Regression

$$Y = \underset{n\times 1}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}} \qquad X = \underset{n\times(p+1)}{\begin{array}{c} x_1 \; x_2 \; x_3 \ldots x_p \\ \begin{bmatrix} \\ \\ \\ \end{bmatrix} \end{array}} \qquad \hat{\beta} = \underset{(p+1)\times 1}{\begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

To predict the fitted value at x_i

Summation Notation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Matrix Notation

$$= \underset{1\times(p+1)}{x_{i\cdot}} \underset{(p+1)\times 1}{\hat{\beta}} = \underset{1\times 1}{\left[\hat{\beta}_0 \cdot 1 + \hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \ldots \hat{\beta}_p \cdot x_{ip}\right]} = \hat{y}_i$$

True Mean Function

$$E(Y|X=x) = X\beta$$

Fitted Mean Fxn

$$\underset{n\times 1 \text{ fit}}{\hat{Y}} = X \underset{n\times 1 \text{ coef}}{\hat{\beta}}$$

Data Gen Fxn

$$Y = X\beta + e$$

Wednesday October 29th

$\underset{n\times 1}{Y}$; $\underset{n\times(p+1)}{X}$; $\underset{(p+1)\times 1}{\beta}$; $\underset{n\times 1}{e}$; $\underset{n\times 1}{\hat{e}} = Y - X\hat{\beta}$

Estimating $\hat{\beta}$

We want to pick $\hat{\beta}$ to minimize R^2.

$$RSS(\hat{\beta}) = \hat{e}'\hat{e} = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$
$$= Y'Y + (X\hat{\beta})'X\hat{\beta} - Y'X\hat{\beta} - (X\hat{\beta})'Y$$
$$= Y'Y + \hat{\beta}'(X'X)\hat{\beta} - 2Y'X\hat{\beta}$$

* $(AB') = B'A'$
* $B'A = A'B$ when result is 1×1

Take derivative wrt $\hat{\beta}$

$$= 0 + 2(X'X)\hat{\beta} - 2X'Y$$
$$(X'X)\hat{\beta} = X'Y$$
$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y}$$

SLR

$$X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \xrightarrow[\text{page 133}]{\text{see}} \hat{\beta} = (X'X)^{-1}X'Y$$

$$= \begin{bmatrix} \bar{y} - \frac{SXY}{SXX}\bar{x} \\ \frac{SXY}{SXX} \end{bmatrix}$$

Monday 11/3/14

Properties of Least Squares Estimates

$Y = X\beta + e$, $Var(e) = \sigma^2 \underset{n\times n}{I}$ $= \begin{bmatrix} \sigma^2 & & & 0 \\ & \sigma^2 & & \\ & & \ddots & \\ 0 & & & \sigma^2 \end{bmatrix}$

Var(e_1) Var(e_2) Cov(e_1, e_2)

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

Are our estimates unbiased?

i.e., do the reach the true β in expectation?

- $E(cY) = c\,E(Y)$
- $E(c) = c$

$$E(\hat\beta|X) = E\left[(X'X)^{-1}X'Y\right]$$
$$= (X'X)^{-1}X'E(Y|X) \rightarrow E(Y|X) = E(XB) + E(e|X) = XB + 0$$
$$= (X'X)^{-1}X'X\beta$$
$$= B$$

✓ the least-squared estimates are unbiased

What is their variance?

$$Var(\hat\beta|X) = Var\left[(X'X)^{-1}X'Y\right]$$
$$= (X'X)^{-1}X'\,Var(Y|X)\,X(X'X)^{-1}$$
$$= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1} \longrightarrow \underset{p+1}{\begin{bmatrix} \cdot & \cdot & \\ & \cdot & \\ & & \end{bmatrix}}$$

* $Var(aY) = a^2 Var(Y)$

as matrix

$Var(AX) = A\,Var(X)\,A'$

symmetric

* $(AB)' = B'A'$

Var(B_1) Var(B_2) Cov(B_1, B_2) $p+1$

estimate w/ $S^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\hat e'\hat e$

Gauss-Markov Theorem

IF $Var(e) = \sigma^2 I$

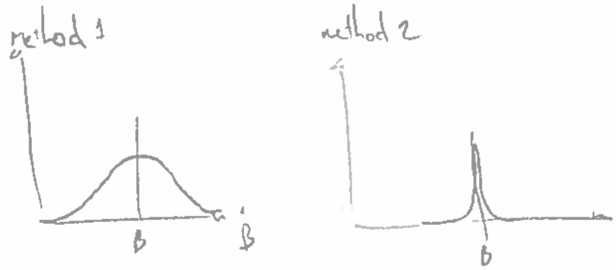
the least squares estimates are

B: est

L: inear

U: nbiased

E: stimates



Wednesday November 5th

- Special seminar + lunch
- Proposals due midnight
- BMI Paper
- Diagnostics I

Leverage

def: the extent to which $\hat{Y}_i$ is attracted by Y_i.

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X'X)^{-1}X'}_{H}Y = HY$$

$X(X'X)^{-1}X'$ is known as the "hat matrix" since it puts a hat on Y.

$$H = \begin{bmatrix} h_{11} & & \\ h_{12} & & \\ & & \end{bmatrix}_{n \times n}$$

$$\hat{Y}_i = \underset{1\times n}{H_{i.}} \cdot \underset{n\times 1}{\hat{Y}} = h_{i1}Y_1 + h_{i2}Y_2 + \dots + h_{ii}Y_i + \dots + h_{in}Y_n$$

row vector

the effect of Y_i on $\hat{Y}_i$

The leverages are the diagonal elements of the hat matrix.

* Rule of thumb: an obs has high leverage if its greater than twice the avg. lev.

$$h_{ii} > 2\frac{(p+1)}{n}$$

average leverage

Properties of Residuals

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

- $Var(CY) = C\,Var(Y)\,C'$
- H is idempotent $\cdot$ symmetric $HH' = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H$

Expected value

$$E(\hat{e}|X) = E((I-H)Y)$$
$$= (I-H)E(Y)$$
$$= (I-H)XB$$
$$= XB - X(X'X)^{-1}X'XB$$
$$= XB - XB$$
$$= 0$$

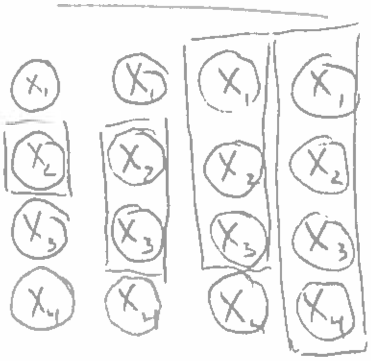
Variance

$$Var(\hat{e}|X) = Var((I-H)Y)$$
$$= (I-H)Var(Y)(I-H)'$$
$$= (I-H)\sigma^2 I(I-H)'$$
$$= \sigma^2(I-H)(I'-H')$$
$$= \sigma^2(II' - IH' - HI' - HH')$$
$$= \sigma^2(I - H - H + H)$$
$$= \sigma^2(I-H) = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \end{bmatrix}_{n \times n}$$

Wednesday Nov 19th

- Hack Ebola
- Presentation Schedule
- Exam parameters
- Searching the Model Space

Best Subsets

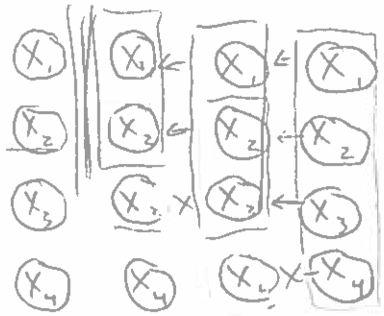


1 2 3 4

Predictors

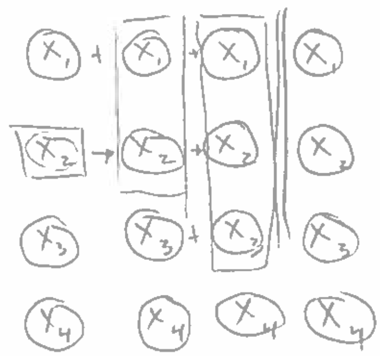
1) Find the model of each size that max's R^2_{adj}.
2) Compare those 'best' models by all criteria

Backwards Elimination



1) Fit the full model and calculate criteria of choice.
2) Remove predictor w/ largest p-val and calculate new criterion.
3) Stop iteration when the criterion is no longer improving

Forward Selection



1) Fit the simple model w/ the predictor w/ the smallest p-val. compute criterion
2) Fit new model w/ that predictor + new predictor, compute criterion.
3) Stop iteration w/ criterion is no longer improving