"Fit the analysis to the data,

NOT the data to the analysis."

-Statistical Maxim

COHEN CHAP 10. LINEAR REGRESSION

For EDUC/PSY 6600

MOTIVATING EXAMPLE

Dr. Ramsey conducts a ***non-experimental*** *study to evaluate what she refers to as the 'strength-injury hypothesis'. It states that overall body strength in elderly women determines the number and severity of accidents that cause bodily injury. If the results of her prediction study support her hypothesis, she plans to conduct an experimental study to assess whether weight training reduces injuries in elderly women.*

Data from 100 women who range in age from 60 to 70 years old are collected. The women initially undergo a series of measures that assess upper and lower body strength, and these measures are summarized into an overall index of body strength.

Over the next 5 years, the women record each time they have an accident that results in a bodily injury and describe fully the extent of the injury. On the basis of these data, Dr. Ramsey calculates an overall injury index for each woman.

A simple regression analysis is conducted with the overall index of body strength as the ***predictor*** *(independent) variable and the overall injury index as the* ***outcome*** *(dependent) variable.*

CORRELATION VS. REGRESSION

Correlation

- Does a relationship exist between 2 variables?
  - No IV / DV distinction
- How strong is relationship?
- In what direction is relationship?

Regression

- What is the form of relationship?
- Can we predict values of one variable from knowledge of one or more, highly correlated variable(s)?
  - Perfect correlation = Perfect prediction
  - Imperfect correlation = Imperfect prediction (with error)
- Simple linear regression (SLR): 1 DV, 1 IV
- Multiple linear regression (MLR): 1 DV, >1 IV

REGRESSION BASICS

Y usually predicted variable

- Dependent, criterion, outcome, response variable
- Predicting Y from X = 'Regressing Y on X'
- Predicting X from Y = 'Regressing X on Y'

X usually variable used to predict Y

- Independent, predictor, explanatory variable

Different results when X & Y switched

- Lecture assumes DV is Y, IV is X

Regression analysis is procedure for obtaining THE line that best fits data

(Assuming relationship is best described as linear)

Equation for straight line:

$$Yhat_i = b_0 + b_1 X_i$$

$Yhat_i$ = predicted (unobserved) value of Y for a given case i

b_0 = y-intercept
- Constant
- *Yhat* when $X = 0$
- only interpreted if $X = 0$ is meaningful
- Alternative notation: 'a' or 'a_{XY}'

b_1 = slope of regression line for 1st IV
- Constant
- Rate of change in Y for every 1-unit change in X
- Alternative notation: 'b_{XY}'

X_i = value of predictor for a given case i

ACCURACY OF PREDICTION

Correlation ≠ Causation

All points do not fall on regression line

- Prediction works for most, but not all in sample

W/out knowledge of X, best prediction of Y is mean of Y ($\bar{Y}$)

- s_Y: best measure of prediction error

With knowledge of X, best prediction of Y is from the equation ($\hat{Y}$)

- Standard error of estimate (SE_E or $s_{Y.X}$): best measure of prediction error
  - Estimated *SD* of residuals in population

ACCURACY OF PREDICTION

$s_{y.x}$ = standard error of estimate

$$s_{Y \cdot X} = \sqrt{\frac{\Sigma(Y_i - Y'_i)^2}{N-2}} = \sqrt{\frac{SS_{\text{residual}}}{df}}$$

$s^2_{y.x}$ = residual or error variance or mean square error

$$s^2_{Y \cdot X} = \frac{\Sigma(Y_i - Y'_i)^2}{N-2} = \frac{SS_{\text{residual}}}{df}$$

$df = N - 2$

- 2 df lost in estimating regression coefficients

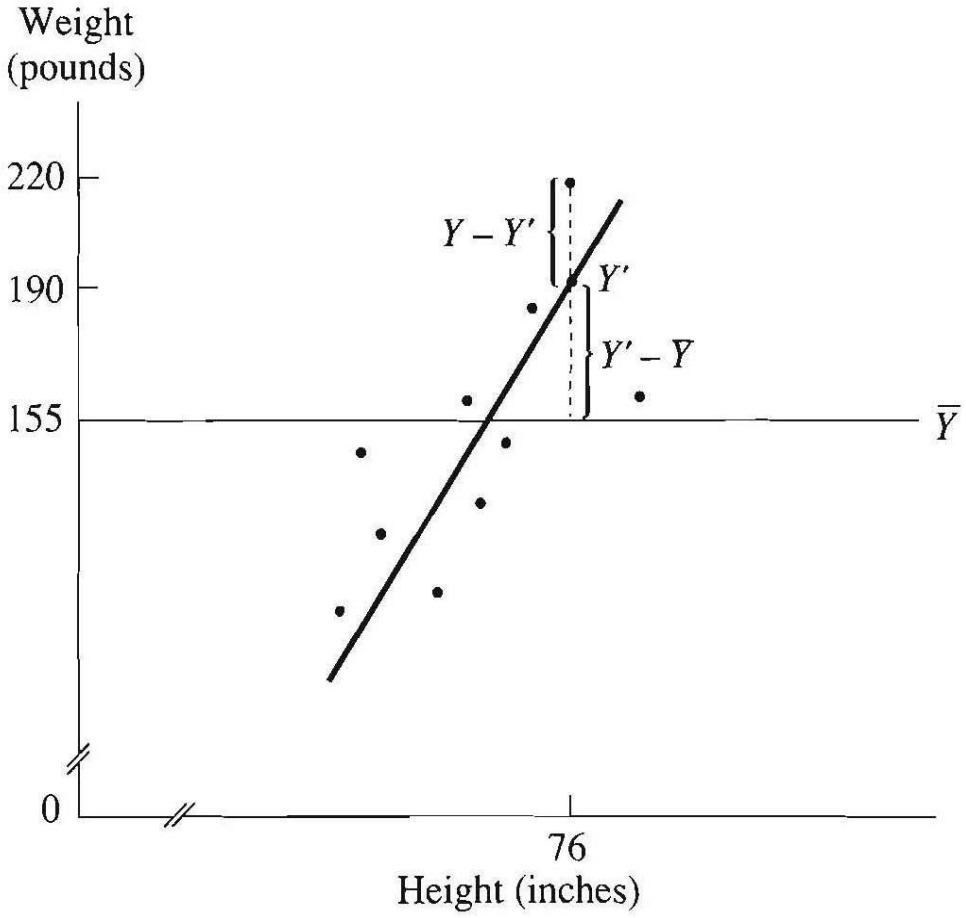
Seeking smallest $s_{Y.X}$ as it is a measure of variation of observations around regression line

"LINE OF BEST FIT"

As prediction is not usually perfect, regression coefficients (b_0, b_1) computed to minimize error as much as possible

- **Error or residuals:** difference between observed Y and predicted Y'
  - $e_i = (Y_i - Y'_i)$
- **Technique:** ordinary least squares (OLS) regression
  - Goal: minimize *SS* error, i.e. *SS* residuals

$$SS_{residuals} = \sum_{i=1}^{n} (Y_i - Y'_i)^2$$



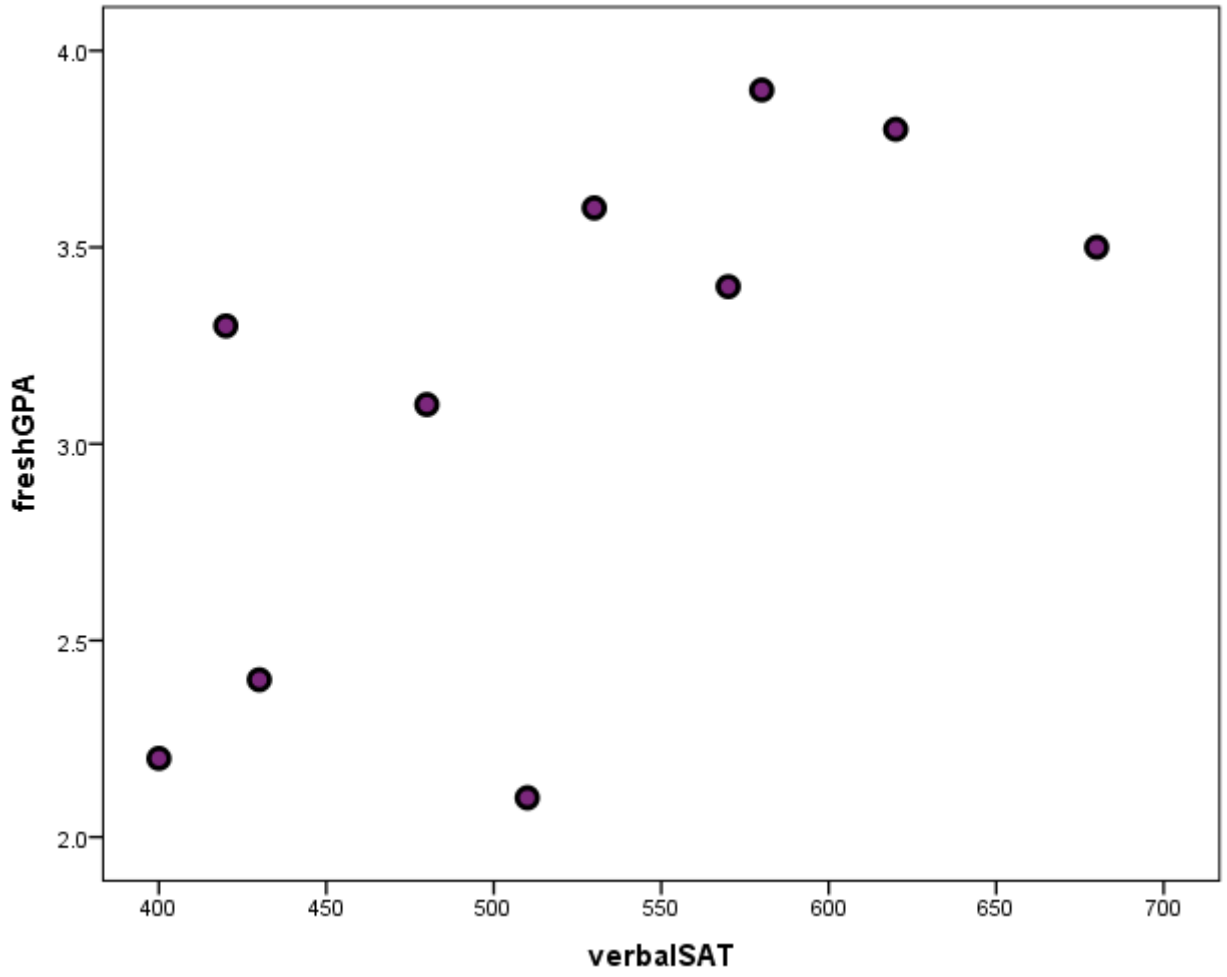
SLOPE: $b_1 = r\frac{s_y}{s_x} = .661\frac{.66}{91.87} = \mathbf{0.0047}$

INTERCEPT: $b_0 = \bar{Y} - b_1\bar{X} = 3.13 - .0047 * 522 = \mathbf{0.68}$

EQUATION: $\hat{Y} = 0.68 + 0.0047 * X$

Verbal SAT	GPA
510	2.1
620	3.8
400	2.2
480	3.1
580	3.9
430	2.4
530	3.6
680	3.5
420	3.3
570	3.4

Pearson Correlation	.661*
Sig. (2-tailed)	.038



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
verbalSAT	10	400	680	522.00	91.869
freshGPA	10	2.1	3.9	3.130	.6634
Valid N (listwise)	10				

$SS_{total} = SS_{explained} + SS_{unexplained}$

(mean to the point) (mean to the line) (line to the point)

"Residuals"

EXPLAINING VARIANCE

$$SS_{Total\ (Y)} = SS_{Regression\ (Y')} + SS_{Residuals(e)}$$

Coefficient of determination (r^2)

- computed to determine how well regression equation predicts Y from X
- r^2 = Explained variation / Total variation -OR- $r^2 = SS_{Regression\ (Y')} / SS_{Total\ (Y)}$
- Ranges from 0 to +1

If each SS was divided by its df:

- Explained variance: $SS_{Reg} / (1)$
  "Mean Square Regression" or $MS_{Regression}$
- Unexplained variance: $SS_{Res} / (N - 2)$
  "Mean Square Residuals" or MS_{Error} or $s^2_{Y.X}$

Interpret r^2 as a percent of variance in outcome variable…

"accounted for"
"attributable to"
"predictable from"
"associated with"
"explained by"

…knowledge of predictor variable

STANDARDIZED REGRESSION COEFFICIENTS

Aka: "Beta weights" or *β*

1 *SD*-unit change in X represents a *β* *SD* change in Y

Intercept = 0 and is not reported when using *β*

For simple regression ONLY

- $r = \beta$ and $r^2 = \beta^2$
- When raw scores are transformed into z-score units $r = b = \beta$

Useful for variables w/ abstract unit of measurement

EXAMPLE:

```
REGRESSION
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /DEPENDENT freshGPA
  /METHOD=ENTER verbalSAT.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.661[a]	.437	.366	.5282

a. Predictors: (Constant), verbalSAT

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.729	1	1.729	6.197	.038[b]
	Residual	2.232	8	.279		
	Total	3.961	9			

a. Dependent Variable: freshGPA

b. Predictors: (Constant), verbalSAT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.640	1.014		.631	.546	-1.699	2.978
	verbalSAT	.005	.002	.661	2.489	.038	.000	.009

a. Dependent Variable: freshGPA

ASSUMPTIONS

Independence of observations

Y normally distributed

- Does NOT apply to predictor variable(s) X
  - Can be categorical or continuous

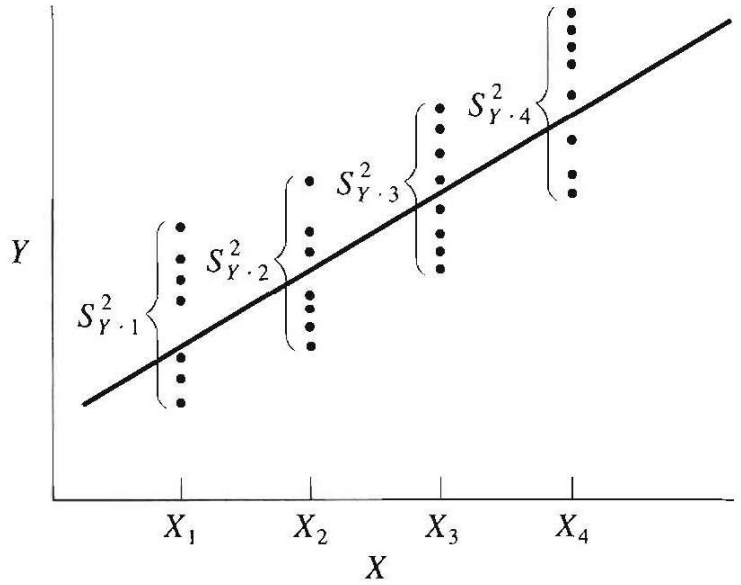
Sampling distribution of the slope (b_1) assumed normally distributed

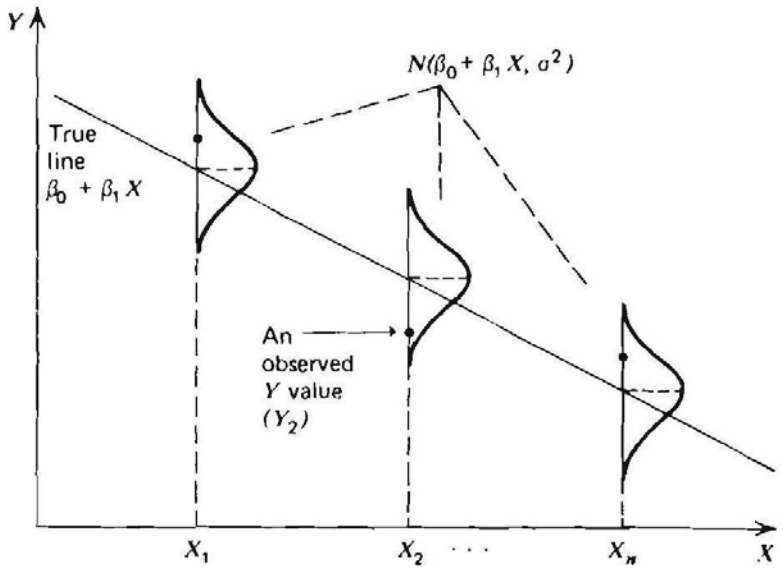
Straight line best fits data

ASSUMPTIONS

Homogeneity of variance of Y for all values of X

- Computed error variance ($s^2_{Y.X}$ or *MSE*) is representative of all individual conditional error variances (for each value of X)
- 'Homoscedasticity'
  - Violation = 'Heteroscedasticity'





SPSS - BASIC

```
* simplest syntax: must tell SPSS what is Y and what
   (leave everything we can to default settings).

REGRESSION
  /DEPENDENT WEIGHIN
  /METHOD=ENTER AGE.
```

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	AGE Patient's Incoming Age[b]	.	Enter

a. Dependent Variable: WEIGHIN Patient's Incoming Weight in pounds

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.288[a]	.083	.043	31.284

a. Predictors: (Constant), AGE Patient's Incoming Age

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2029.604	1	2029.604	2.074	.163[b]
	Residual	22510.336	23	978.710		
	Total	24539.940	24			

a. Dependent Variable: WEIGHIN Patient's Incoming Weight in pounds

b. Predictors: (Constant), AGE Patient's Incoming Age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	220.690	30.108		7.330	.000
	AGE Patient's Incoming Age	-.711	.494	-.288	-1.440	.163

a. Dependent Variable: WEIGHIN Patient's Incoming Weight in pounds

SPSS

```
* most of the optional commands used here (from point-click).

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /DEPENDENT WEIGHIN
  /METHOD=ENTER AGE
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```