To Do

Read Sections 5.1 - 5.3 (Hypothesis Testing) and Sections 6.1 – 6.2.

Do 5.1 – 5.8 for Midterm Test 2.

See detailed information posted on Learn regarding material covered by Midterm Test 2 (4:40 - 6:10 on Tuesday November 15).

Today's Class

(1) Least Squares Estimates

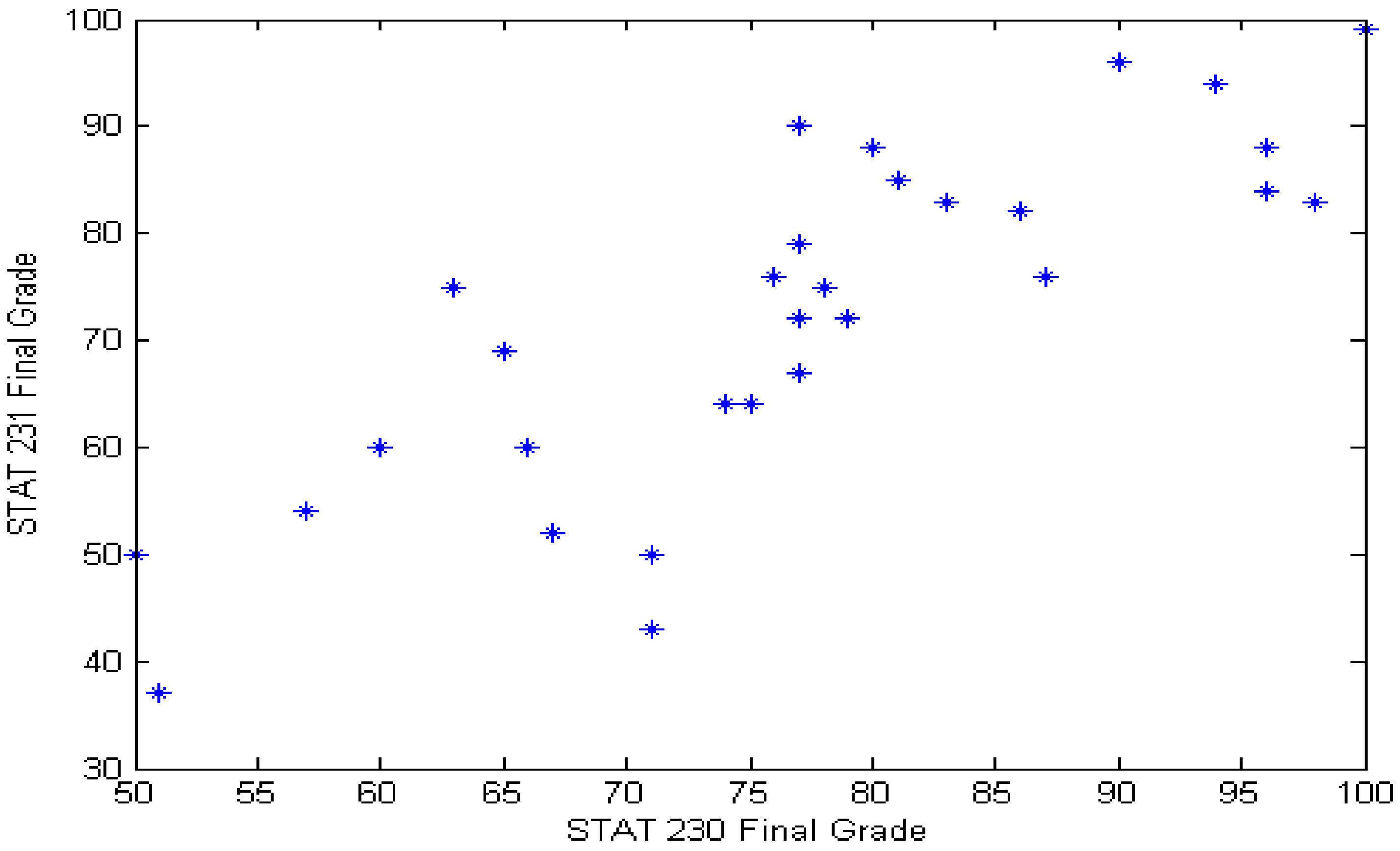
(2) Simple Linear Regression Model

(3) Maximum Likelihood Estimates

Example: STAT 230 and 231 Final Grades

No.	S230	S231		No.	S230	S231		No.	S230	S231
1	76	76		11	87	76		21	98	83
2	77	79		12	71	50		22	80	88
3	57	54		13	63	75		23	67	52
4	75	64		14	77	72		24	78	75
5	74	64		15	96	84		25	100	99
6	60	60		16	65	69		26	94	94
7	81	85		17	71	43		27	83	83
8	86	82		18	66	60		28	51	37
9	96	88		19	90	96		29	77	90
10	79	72		20	50	50		30	77	67

Scatterplot of Data



Sample Correlation for STAT 230/231 Final Grades

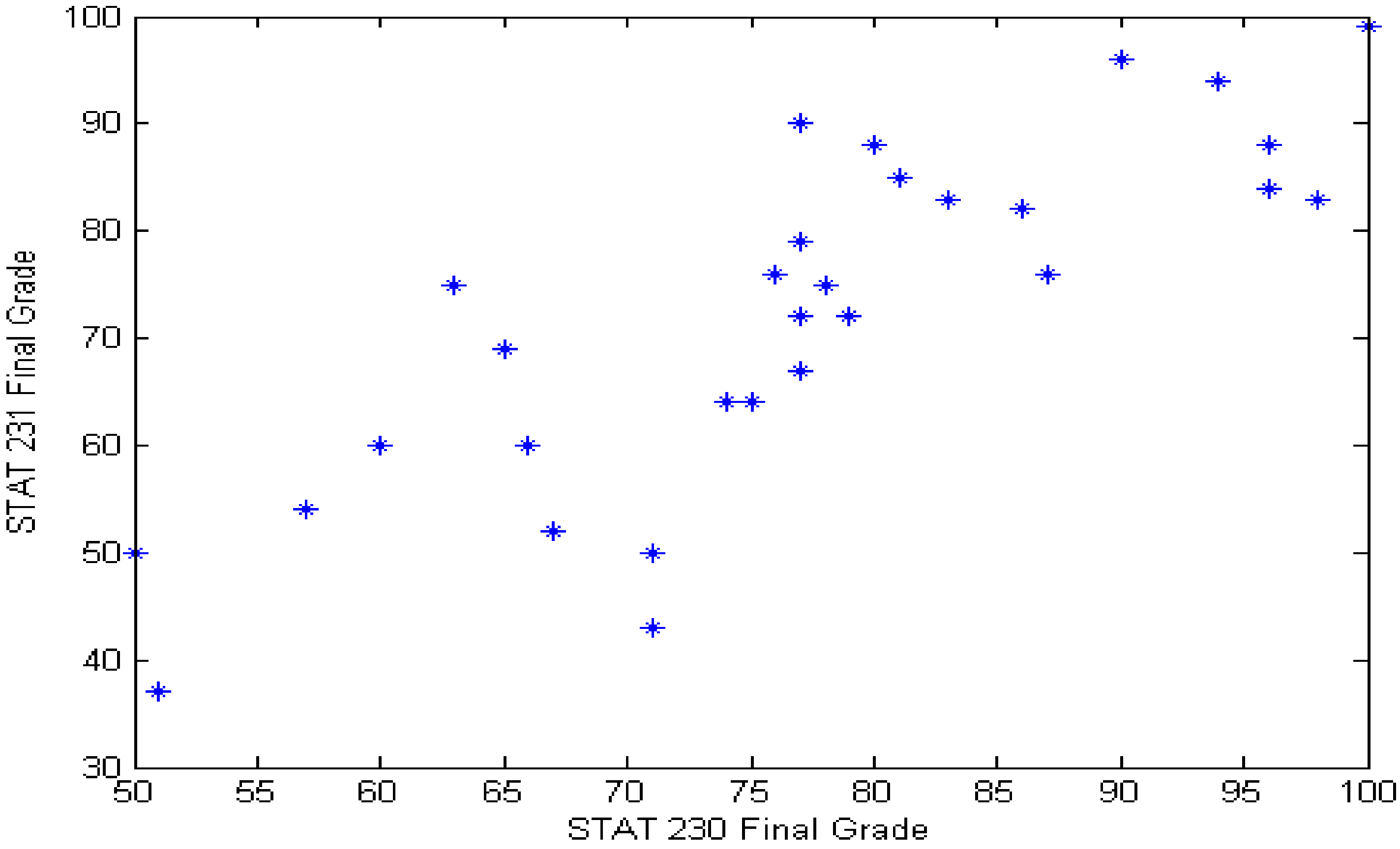
For these data

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{5106.8667}{\sqrt{(5135.8667)(7585.3667)}} = 0.82$$

Since *r* is close to 1 we would say that there is a strong positive linear relationship between STAT 230 final grades and STAT 231 final grades.

Question

How do we fit a straight line to these data?



Least Squares Estimates

The least squares estimates

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

minimize the sum of the squares of the vertical distances between the observed points (x_i,y_i), *i* = 1,2,…,*n* and the line *y* = *α*+*βx*.

That is the least squares estimates minimize the function

$$g(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)^2$$

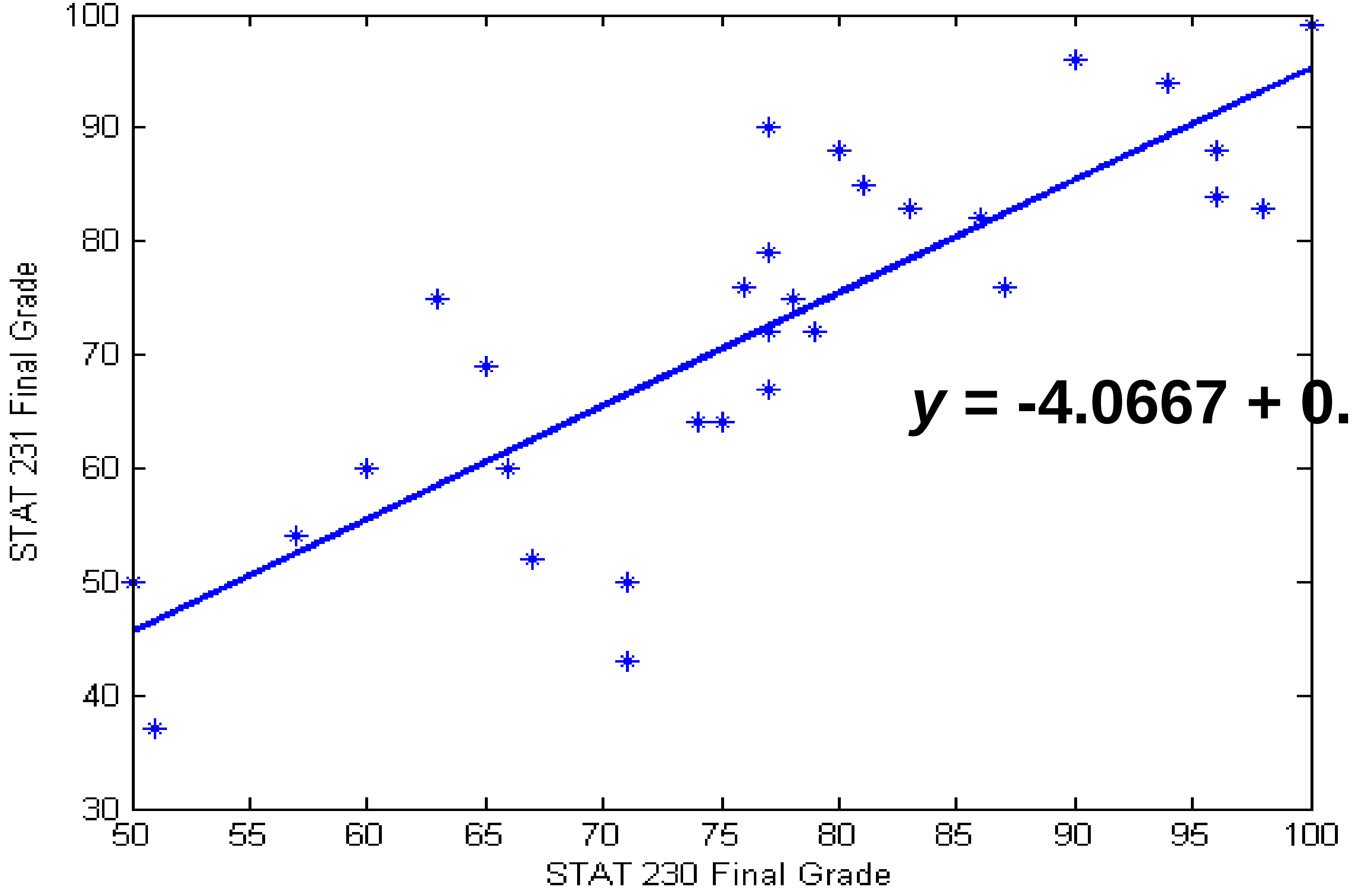
Least Squares LIne

The fitted line

$$y = \hat{\alpha} + \hat{\beta}x$$

is called the least squares line.

Scatterplot with Least Squares Line



$$y = -4.0667 + 0.9944x$$

Relationship Between Sample Correlation and Least Squares Estimate of Slope

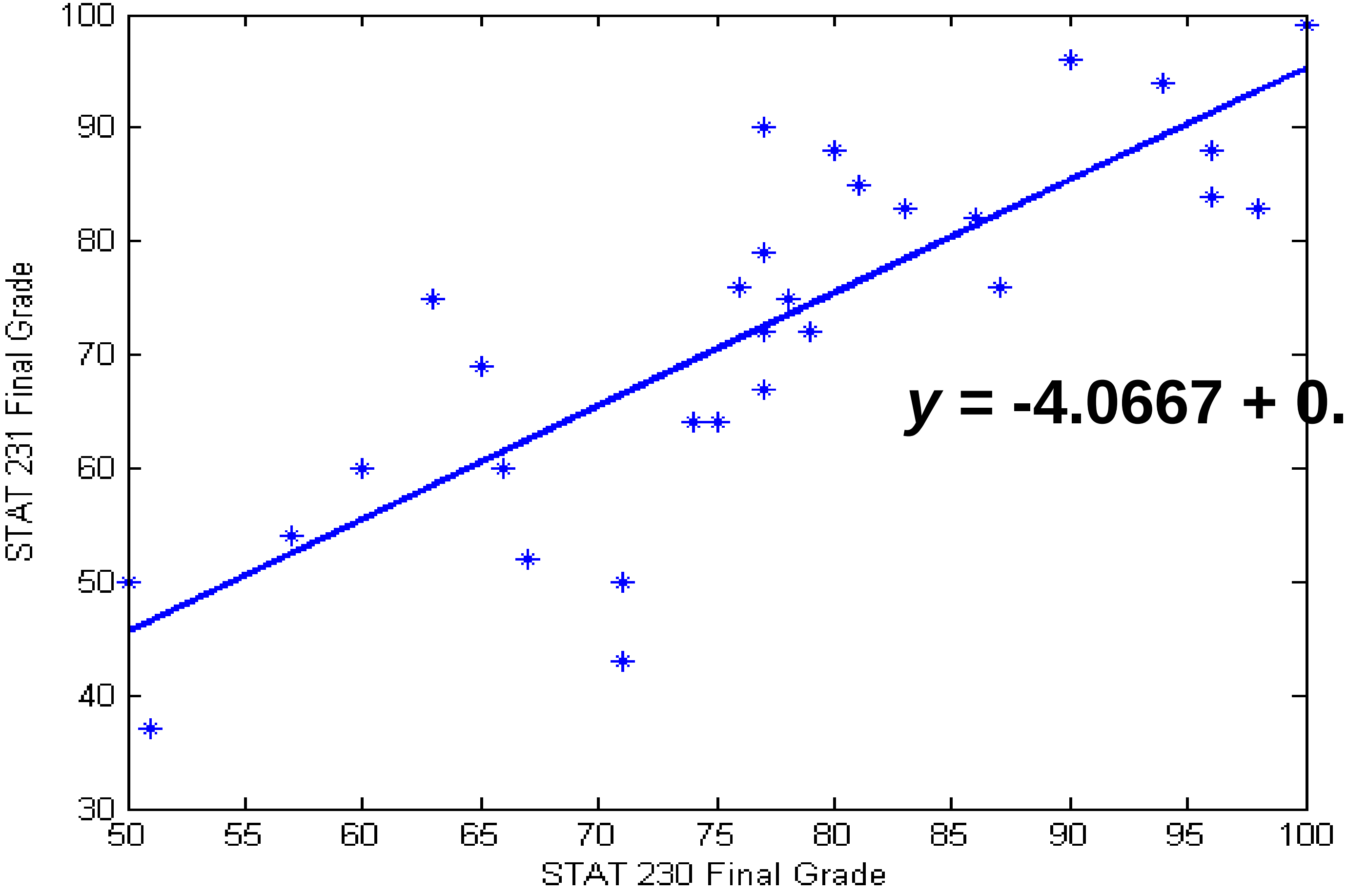
Since

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \quad \text{and} \quad \hat{\beta} = \frac{S_{XY}}{S_{XX}}$$

Therefore

$$r = \hat{\beta}\sqrt{\frac{S_{XX}}{S_{YY}}} \quad \text{and} \quad \hat{\beta} = r\sqrt{\frac{S_{YY}}{S_{XX}}}$$

STAT 231 versus STAT 230 Scatterplot with Least Squares Line



$y = -4.0667 + 0.9944x$

STAT 231 versus 230 Final Grades - Model?

If your final grade in STAT 230 was $x = 75$, then the least squares (point) estimate of your STAT 231 final grade is

$$y = -4.0667 + 0.9944(75) = 70.51$$

What can we say about the uncertainty in this estimate?

STAT 231 versus 230 Final Grades - Model?

We need a statistical model in order to obtain an interval estimate of your final grade in STAT 231.

We need a model which captures the fact that not everyone with a final grade of x = 75 in STAT 230 gets a final grade of 70.51 in STAT 231.

We need a model which models the variability in final grades for each STAT 230 final grade x.

A Model We Have Already Studied

Let's begin by considering the population of students who obtained a final grade of x = 75 in STAT 230.

Let Y = STAT 231 final grade of a student drawn at random from this population.

What distribution might we assume for Y?

A Model We Have Already Studied

Since frequency histograms for final grades often exhibit a bell shape, we might assume $Y \backsim G(\mu,\sigma)$ where μ represents the mean STAT 231 final grade for students in the study population who obtained a final grade of $x = 75$ in STAT 230.

We could then use this model along with the observed data for $x = 75$ to obtain point estimates and interval estimates for the mean μ.

Model for STAT 231 versus STAT 230 Final Grades

In our sample of 30 students we only observed one student with a final grade of 75 in STAT 230.

Does it make sense to do estimation with only one observation?

What to do?

We do have 29 other observations.

Model for STAT 231 versus STAT 230 Final Grades

Since the other 29 students had different STAT 230 final grades they are observations drawn from populations which have different means (and possibly different variances).

From the scatterplot however, the relationship between STAT 230 and STAT 231 marks looks very linear.

Model for STAT 231 versus STAT 230 Final Grades

It seems reasonable to assume a model in which the $\mu(x)$ = mean STAT 231 final grade for students in the study population who obtained a final grade of x in STAT 230 takes a linear form in x:

$$\mu(x) = \alpha + \beta x$$

Model Assumptions

For data (x_i,y_i), i = 1,2,…,n
we assume the model

$Y_i \backsim G(\alpha + \beta x_i, \sigma)$ for i =1,2,…,n

independently and where the
x_i's, i = 1,2,…,n
are assumed to be known constants.

Simple Linear Regression

This model is usually referred to as a simple linear regression model.

Note that we have assumed that the standard deviation σ does not depend on x_i.

We will look at graphical methods of assessing whether this assumption is reasonable.

Simple Linear Regression

There are three unknown parameters:

α, β and σ.

For our data, the parameter

$\mu(x) = \alpha + \beta x$

represents the mean STAT 231 final grade in the study population of students with a STAT 230 final grade equal to *x*.

Interpretation of Parameters

In the STAT 230/231 example the parameter *β* represents the change in the mean STAT 231 final grade in the study population for a one mark increase in STAT 230 final grade.

The parameter *α* represents the mean STAT 231 final grade in the study population of students with a STAT 230 final grade equal to 0.

(Note: In this example the parameter *α* is not of much interest.)

Interpretation of Parameters

In the STAT 230/231 example the parameter *σ* represents the variability in the response variate *Y* in the study population for each value of the explanatory variate *x*.

Likelihood Function for *α* and *β*

Since our model is:

$Y_i \backsim G(\alpha + \beta x_i, \sigma)$ for $i = 1,2,\ldots,n$

independently where the x_i's, $i = 1,2,\ldots,n$ are known constants then

$$L(\alpha,\beta) = \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right]$$

assuming for the moment that *σ* is known and ignoring constants with respect to *α* and *β*.

Maximum Likelihood Estimates for α and β

To maximize

$$L(\alpha,\beta)=\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2\right]$$

we would minimize

$$\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

but this is just the least squares problem!

Theorem

For the model

$$Y_i \backsim G(\alpha + \beta x_i, \sigma) \quad \text{for } i = 1,2,\ldots,n$$

independently where the x_i's, i = 1,2,…,n are known constants, the maximum likelihood estimates of α and β (often called the regression parameters) are given by

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

which are also the least squares estimates.