

Lecture 4 Stat 102 Summer 2008
Inference in Simple Regressions

- Read Chapter 3.3
- Tests and CI^s for β_1, β_0 .
- Example: 2003 house prices in ZIP 30062

Regression Model (review)

$$Y = \beta_0 + \beta_1 x + e$$

$$\mu_{Y|x} = \beta_0 + \beta_1 x,$$

– a straight line

– the model for the mean of Y at given values of x

Y = dependent variable

x = independent variable

β_0 = y-intercept

β_1 = slope of line

e = error – normally distributed with mean = 0 and

$$\text{Var}(e) = \sigma_e^2$$

The errors are assumed to be independent of each other.

Sampling Distribution of b_0, b_1

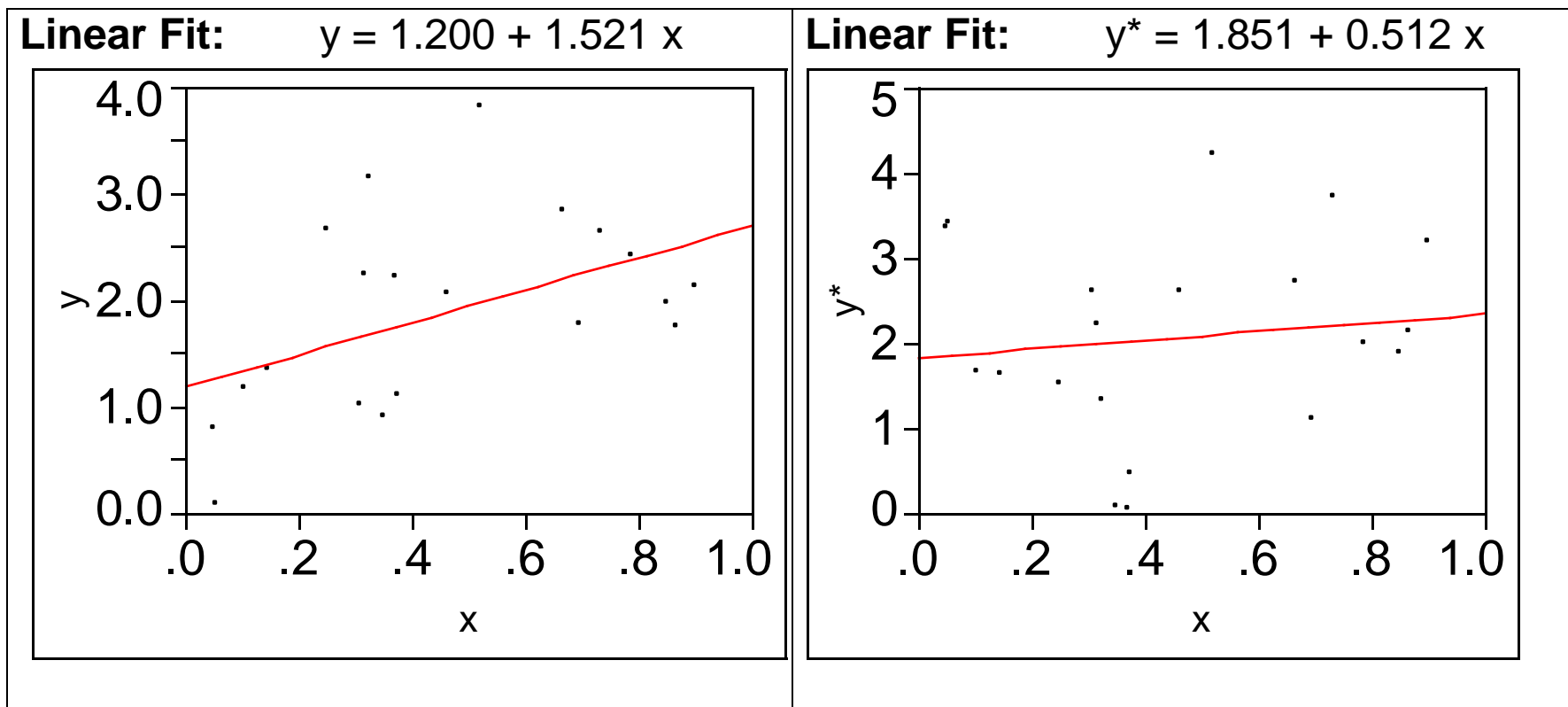
- b_0, b_1 are the least squares estimates of β_0, β_1
- The “sampling distribution” of b_0, b_1 is the probability distribution of the estimates over repeated samples y_1, \dots, y_n from the ideal linear regression model with fixed values of β_0, β_1 and σ_e^2 and x_1, \dots, x_n .

Simulation of the Sampling Distribution

- A “simulation” is a computer produced version of the sampling distribution, or of repeated, independent versions of the sampling distribution.
- The file “Standardregression.jmp” contains a simulation of pairs $(x_1, y_1), \dots, (x_{20}, y_{20})$ from a simple linear regression model with $\beta_0 = 1, \beta_1 = 2, \sigma_e^2 = 1$. AND
- It contains another simulation labeled $(x_1, y_1^*), \dots, (x_{20}, y_{20}^*)$ from the same model.

(The values of x_1, \dots, x_{20} are the same in both sets of data.)

- Notice the difference in the estimated coefficients calculated from the y 's and from the y^* 's.



Two outcomes from “standardregression.jmp”

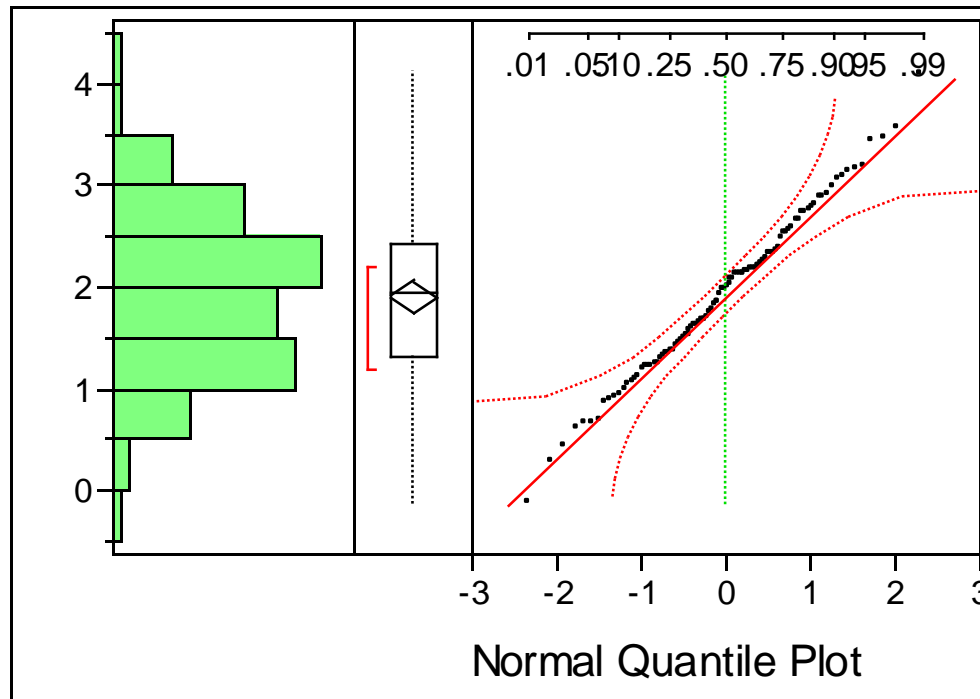
Each data set comes from the model with $\beta_0 = 1, \beta_1 = 2, \sigma_e^2 = 1$

The values of x_1, \dots, x_{20} are the same in both data sets.

New Simulation: Shows results of 100 samples from same setup

Note: Population value is $\beta_1 = 2$. Each sample size is $n = 20$

Distribution of b_1



Moments

| | |
|---------|-------|
| Mean | 1.910 |
| Std Dev | 0.792 |
| N | 100 |

Sampling Distribution (Details)

- b_0 and b_1 have normal distributions, with nice formulas for their expectation and variance
- Sampling distribution of b_0 is normal with

$$E(b_0) = \beta_0$$

$$Var(b_0) = \sigma_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right) \text{ where } s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

- Sampling distribution of b_1 is normal with

$$E(b_1) = \beta_1$$

$$Var(b_1) = \frac{\sigma_e^2}{(n-1)s_x^2}$$

Hence:

- **If** we knew σ_e^2 we could produce a confidence interval for β_1 as

$$b_1 \pm z_{1-\alpha/2} \sqrt{\sigma_e^2 / (n-1) s_x^2},$$

where $z_{1-\alpha/2}$ comes from the normal tables. Here, the term $\sqrt{\sigma_e^2 / (n-1) s_x^2}$ is the standard deviation of b_1 .

- **If** we knew σ_e^2 we could similarly test $H_0 : \beta_1 = 0$, *etc.*
- **But** we *hardly ever* know σ_e^2 .
- **So** we must first estimate σ_e^2 .

Estimating σ_e^2

- σ_e^2 is the variance of the error terms e_i .
- In other words, it is the variance of $y_i - [\beta_0 + \beta_1 x_i]$
- It can be estimated by

$$s_e^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (y_i - [b_0 + b_1 x_i])^2 .$$

{SSE stands for “**Sum of Squares of Error**”.

- Sample quantities involving s_e^2 – like t-statistics and F-statistics to be described later – have $n-2$ **Degrees of Freedom**.
 - The **$n-2$** occurs because the estimate s_e^2 involves the estimation of **2** unknown parameters – by b_0 and b_1 , resp.

Confidence Interval for the Slope, β_1

- The estimate of β_1 is b_1
- *Estimate* of standard deviation of b_1 :

$$SE(b_1) \square s_e \sqrt{\frac{1}{(n-1)s_x^2}} \quad \text{where } s_e^2 = \frac{SSE}{n-2}.$$

– This is also called the “**Standard Error**” of b_1 , since it is an estimated standard deviation of the *statistic*, b_1 .

- $100(1-\alpha)\%$ Confidence Interval

$$b_1 \pm t_{n-2;1-\alpha/2} SE(b_1),$$

where $t_{n-2;1-\alpha/2}$ comes from the t-table with **$n-2$** DF.

- Typically, one uses a 95% CI.

Hypothesis Tests for the Slope

- Two sided tests:
 - Null Hypothesis $H_0 : \beta_1 = \beta_1^*$ (for a pre-chosen value β_1^*)
 - The most common choice here is $\beta_1^* = 0$. This is what is implemented in JMP.
 - Alternative $H_1 : \beta_1 \neq \beta_1^*$.

- Test statistic:

$$t = \frac{b_1 - \beta_1^*}{SE(b_1)}$$

- Reject if $|t| > t_{n-2; 1-\alpha/2}$.
- One sided tests are also possible – see p. 82-83 & Fig 3.15.
- **P-values** are possible for all tests; see p 83-84
- Why a t-statistic, and not a normal statistic? Because we need to compensate for the fact that $SE(b_1)$ is only an *estimate* of $SD(b_1)$; it is not the exact value.

Inference for the Intercept

- Tests and CI^s are also available for the intercept β_0 .
- They resemble tests and CI^s for β_1 ;
 - but they use b_0 in place of b_1 , and
 - they use $SE(b_0)$ in place of $SE(b_1)$ where

$$SE(b_0) = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} .$$

- See p. 81, 83 and Fig 3.15

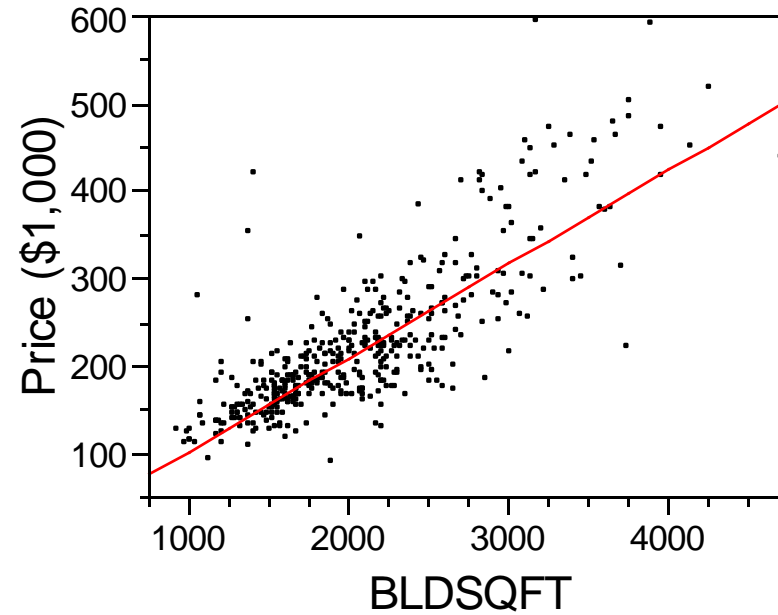
Example: 2003 House Prices in ZIP 30062

- The data lists the sale price for single family homes in the Zip-code 30062 (Northern Atlanta) sold during 2003.
- We will analyze the effect of
 - $x =$ Building Size (Sq. Feet) on
 - $Y =$ House Price (\$1,000)
- There were $n = 439$ complete data pairs in the data-set.
- Here is the Scatterplot with least-squares line and the Parameter Estimates table.

Inference in JMP

- Use “Fit Line” inside “Analyze → Fit Y by X”
- See the “Parameter Estimates” table for values of
 - $b_0 =$ “Intercept”
 - $b_1 =$ [*name of Y variable*]
 - Standard Errors of both b_0 and b_1 .
 - t-ratios and P-values for the 2-sided tests of $H_0 : \beta_{\bullet} = 0$.
 - You can get 95% CI^s for either β as follows:
 - Right click on the area inside the table. Then click “Columns” and select “Lower 95%” and “Upper 95%”
 - For other values of α you’ll need to find critical values of t from a table or from inside JMP, and then use the basic formulas.

2003 House Prices in ZIP 30062 (cont)



Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob> t | Lower 95% | Upper 95% |
|-----------|----------|-----------|---------|---------|-----------|-----------|
| Intercept | -2.52 | 7.87 | -0.32 | 0.750 | -17.98 | 12.95 |
| BLDSQFT | 0.107 | 0.0036 | 29.84 | <.0001 | 0.100 | 0.114 |

House Prices: Questions of Interest

1. What is the least squares estimate of the population regression line?
2. According to this estimate, what should be the average cost of a 3000 sq foot house?
3. For two typical, representative houses, if one is 200 sq feet larger than the other, how much more, or less, will the larger house cost?
4. Find the 95% CI for the value of the slope. Find the 99% interval.
5. What is the P-value of the 2-sided test that the slope is 0? What is the P-value of the 2-sided test that the slope is 0.10? Do these tests reject at level $\alpha=0.05$, at level $\alpha=0.01$?
6. A naïve statistician suggests after examining this analysis that a 95% confidence interval for the average price of an empty lot says that such an average lot should cost less than \$12,950.
 - a. What is the basis for this assertion?
 - b. Does the data provide reasonable support for such an assertion?

Answers in lecture