Course Materials May Not Be Distributed or Posted Electronically

These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service.

PSTAT 126
Regression Analysis

Dr. Todd Gross

Department of Statistics and Applied Probability

UCSB

Lecture 3

Inference in Simple Regression

Text Readings for this Lecture

In Introduction to Statistical Learning (ISL)

- Ch 3 Section 3.1.3 – Assessing the Accuracy of the Model

Lecture Outline

- The Logic of Hypothesis Testing and Confidence Intervals
- Testing the Regression Parameters (β_1 and β_0)

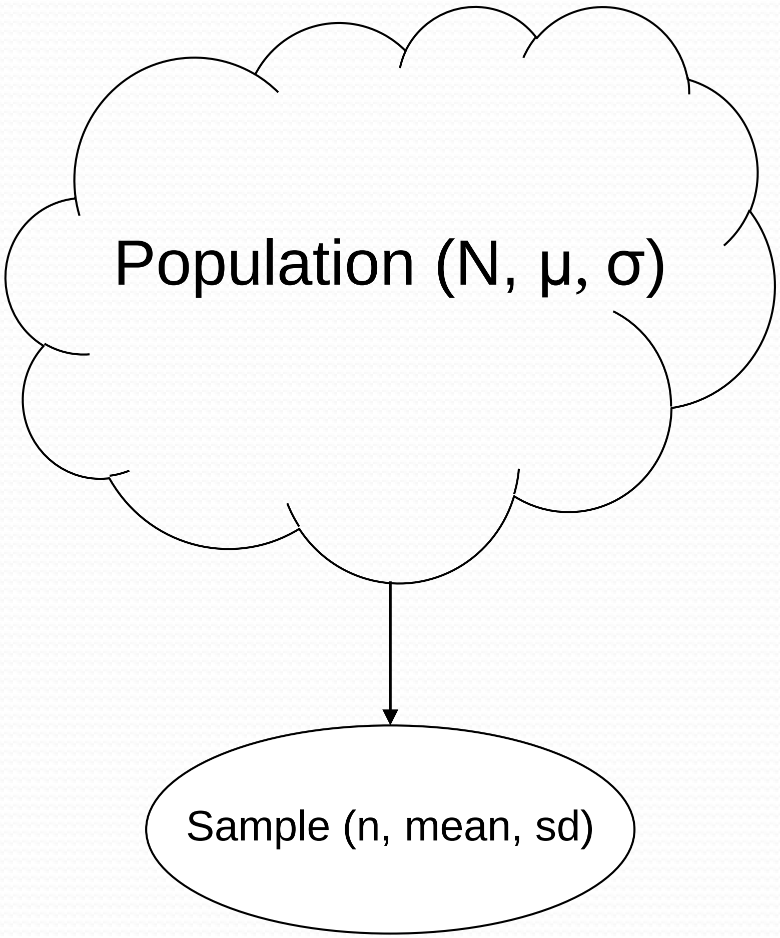
The Logic of Hypothesis Testing

Inferential Statistics

- Inferential statistics allow us to draw conclusions about a population based on information in a sample.
  - Contrast with Descriptive Statistics, which allows us to describe information in a sample
- The question we need to answer is:

"Is there sufficient evidence in the sample to conclude that there is a relationship between X and Y in the population?"

Uncertainty - Sampling Error



The Logic of Hypothesis Testing

- Research Question
  - What you want to prove, stated as a question
- Statistical Hypothesis
  - Define the Null (H_0) and Alternative (H_1) hypotheses
  - H_0 is the opposite of what you want to prove
- Conduct Study
  - Collect data to test the hypothesis
- Statistical Analysis
  - Calculate the probability of observing your results GIVEN that H_0 is true
- Statistical Conclusion
  - If the probability is low enough, then Reject H0
  - If not, Retain H_0 (can't accept it yet)

Exercise: Is My Coin Fair?

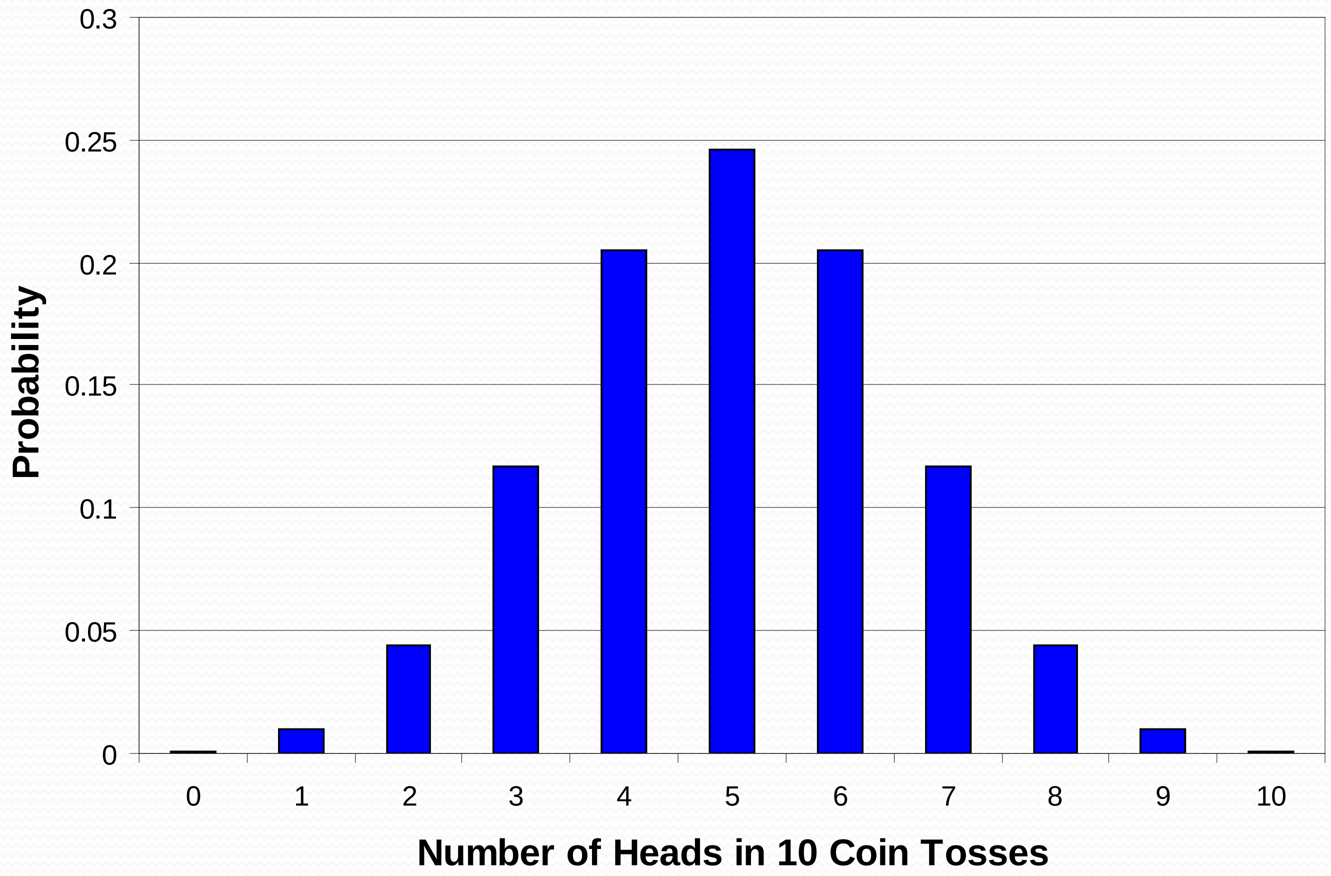
Research Question: Is my coin "fair" (i.e., are heads and tails are equally likely outcomes)

- How many heads would lead you to conclude that coin is biased?

Exercise:

- Toss your coin 10 times
- Record the number of Heads
  - Example: 7 heads out of 10 tosses
- Do the results of your "experiment" suggest that your coin is <u>not</u> fair?
  - Assuming a fair coin, what is the probability of obtaining your results?
  - If the results are very unlikely, reject the hypothesis that coin is fair

Sampling Distribution for Coin Tosses



Hypothesis Test for Coin Toss

- Research Question: Is my coin fair (i.e., equally likely to produce a heads or a tails)?
- Statistical Hypotheses
  - H_o: $\pi = 0.5$
  - H_1: $\pi \neq 0.5$
- Data: 9 heads out of 10 tosses
- What is the probability of 9 heads out of 10 tosses, <u>assuming that the coin is fair</u>?
- From the binomial distribution:
  - P(9 out of 10| $\pi = 0.5$) = 0.00977
  - P(10 out of 10| $\pi = 0.5$) = 0.0001
- The probability of 9 or more heads out of 10 tosses, <u>given a fair coin</u>, is about 1%, <u>therefore</u> conclude that coin is not fair!

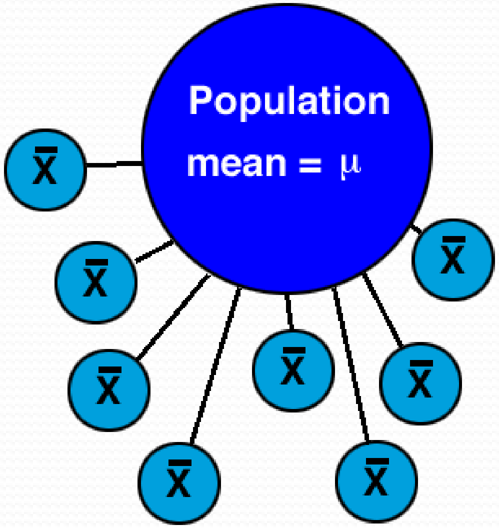
Sampling Distribution of Sample Mean (Review)

- I believe students sleep an average of 6 hours per night
- I select 25 students at random and find:
  - Mean = 7.2

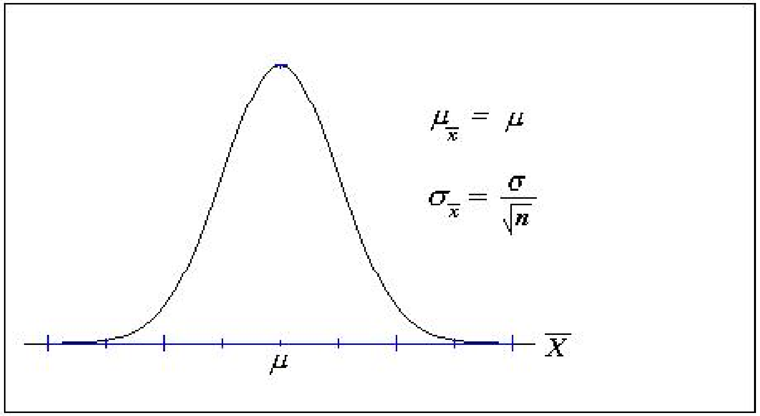
- Did my sample come from a population with a mean of 6?
  - P($\bar{X}$ = 7.2 | μ = 6) = ?
  - We need to know the sampling distribution of the sample mean

Sampling Dist of Mean (con't)

- There are an infinite number of samples of size 25 that I could have drawn, each with its own mean



- Under the Central Limit Theorem the distribution of sample means is approximately Normal, with mean = Mu



t-test for sample mean

- We can construct a statistical test using a t-test

$$t = \frac{\boxed{\bar{X} - \mu_0}}{s/\sqrt{n}}$$

Increases when null hypothesis is false

- In the numerator, we are taking the difference between the observed sample mean and the expected mean under the null hypothesis.
  - **Increases** when the null hypothesis is false
- The denominator is the standard error of the sample mean, which is an estimate of the variability of the sample mean given the sample size.
  - **Decreases** when variance is small, or sample size is large

t-test for sample mean

$$t = \frac{\boxed{\bar{X} - \mu_0}}{s/\sqrt{n}}$$

Increases when null hypothesis is false

Smaller variance decreases standard error

Larger sample size decreases standard error

Hypothesis Test for Sample Mean

- Research Question:
  - Do students sleep longer than six hours?
- Null and Alternative Hypotheses:
  - H_0: μ = 6 and H_1: μ ≠6
- Data:
  - n = 25, $\bar{X}$ = 7.2, s = 1.9
  - Standard error of the mean, $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1.9}{\sqrt{25}} = 0.38$
- Calculate probability of data given H_0
  - $t_{observed} = \frac{\bar{X}-\mu_0}{s_{\bar{x}}}$ = (7.2 – 6)/0.38 = $\boxed{3.16}$
  - Alpha = 0.05, alpha/2 = 0.25, df = n-1 = 24
  - $t_{critical}(0.025, 24) = \boxed{2.064}$
- If $|t_{observed}| > t_{critical}$ then Reject H_0
- Conclude the Research Question is confirmed.

Calculations in R

```
> tobs = (7.2-6)/(1.9/sqrt(25))
> tobs
[1] 3.157895
> tcrit = qt(.025,24)
> tcrit
[1] -2.063899
> abs(tobs)> abs(tcrit)
[1] TRUE
```

Confidence Intervals

- Interval estimation is an alternative to point estimation
- Example: the point estimate for μ is $\bar{X}$
- We can create an interval that contains most likely values of μ
- Lower Limit = $\bar{X} - t_{crit}(s_{\bar{x}})$ = 7.2 – 2.064 * 0.38 = 6.42
- Upper Limit = $\bar{X} + t_{crit}(s_{\bar{x}})$ = 7.2 + 2.064 * 0.38 = 7.98
- We are 95% confidence that our sample mean of 7.2 comes from a population with a mean between 6.42 and 7.98

Inference in Regression

Research Questions in Regression

- Slope: Is $\beta_1 \neq 0$?
- Intercept: Is $\beta_0 \neq 0$? (often not meaningful)
- Full Model: Does the regression equation fit better than chance alone?
- Confidence Intervals
  - Slope
  - Predicted value of Y
  - Regression Line

Testing the Slope (B_1)

- We will test the following Null and Alternative Hypotheses
  - H0: $\beta_1 = 0$
  - H1: $\beta_1 \neq 0$

- If the slope is zero, then there is <u>no</u> linear relationship between X and Y (no predictive value)

- We will need to know:
  - sampling distribution of b_1
  - variance and standard error of b_1
  - appropriate test statistic (t)

Gauss and the Normal Curve

Normality Assumption

- The regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ assumes that:
  - Y is a random variable
  - X is known and measured without error
  - The residual errors are unbiased, $E[\varepsilon_i] = 0$
  - The values of Y are independent of each other, $Cov(\varepsilon_i, \varepsilon_j) = 0$
  - It does NOT assume that Y is normally distributed.
- In order to do **hypothesis testing**, we need to add the assumption of a normal distribution.
- This gives us the **Normal Regression Model**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ where } \varepsilon_i \sim iid\, N(0, \sigma^2)$$

Sampling Distribution of b_1

- We estimate the population parameter β_1 using the sample statistic b_1
- Under the normal regression model, we assume that Y is normally distributed.
- We can show that b_1 is a linear combination of Y
- Therefore, b_1 is normally distributed because:
  a linear combinations of normally distributed random variables is also normally distributed

$$b_1 \sim N(\beta_1, Var(b_1))$$

T-test for the slope

- We can now define a t-test for H_0: $\beta_1 = 0$

$$t = \frac{b1}{SE(b_1)}$$

- The standard error of b_1 is:

$$SE(b_1) = \sqrt{\frac{\sum(Y-\hat{Y})^2/(n-2)}{\sum(X-\bar{X})^2}} = \sqrt{\frac{SSRes/dfres}{SSX}}$$

Hours of Study and Exam Score

Students Study for an Exam

For each student we measure:

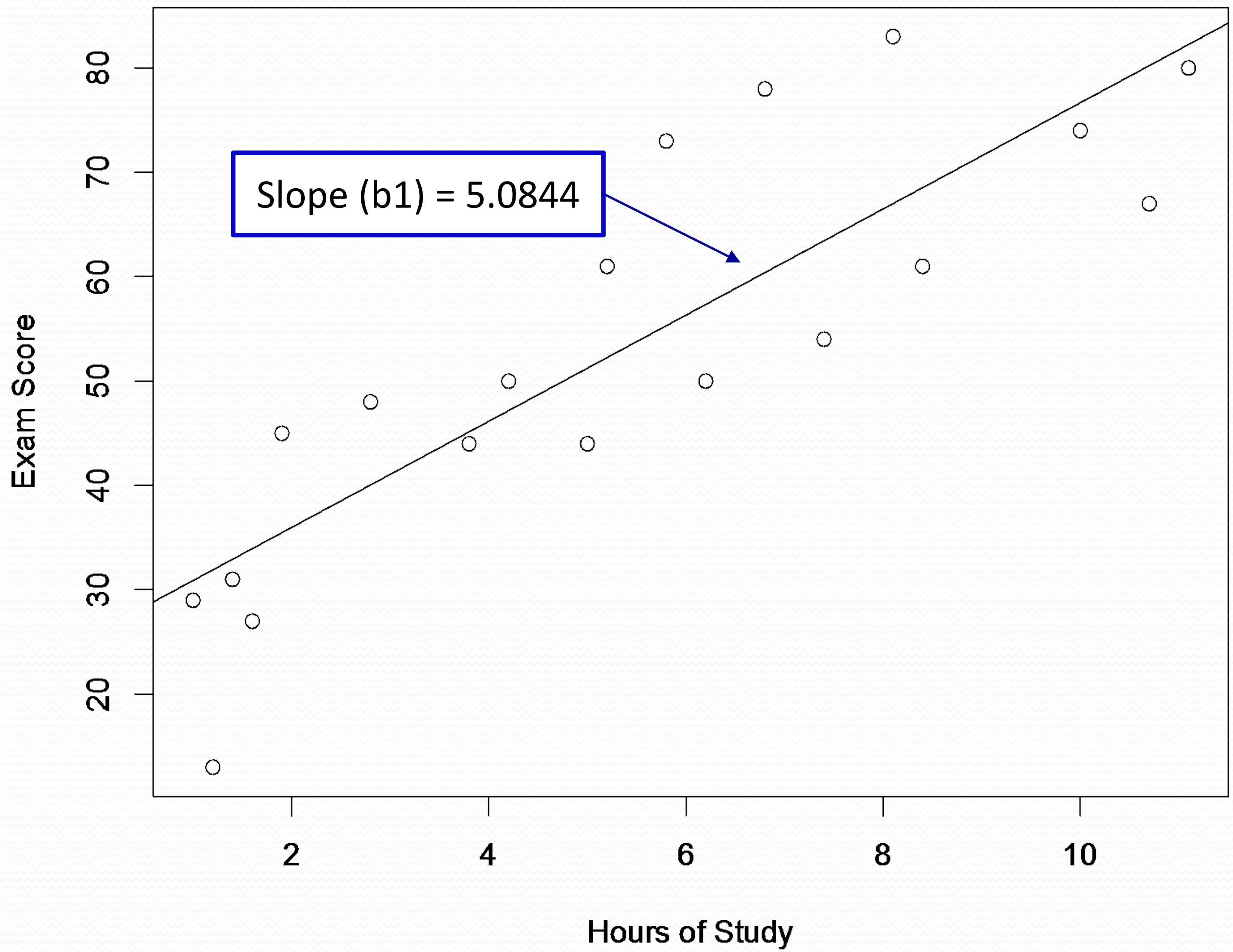
- the number of hours of study (X)
- the score on the exam (Y)

We want to know:

1. Is there a relationship between hours of study and score?
2. Is the relationship linear?
3. Describe the relationship

Student	Hours of Study	Exam Score
1	1.0	29
2	1.2	13
3	1.4	31
4	1.6	27
5	1.9	45
6	2.8	48
7	3.8	44
8	4.2	50
9	5.0	44
10	5.2	61
11	5.8	73
12	6.2	50
13	6.8	78
14	7.4	54
15	8.1	83
16	8.4	61
17	10.0	74
18	10.7	67
19	11.1	80

R Scatterplot with Regression Line



Using R to Perform Regression

```
x<-c(1, 1.2, 1.4, 1.6, 1.9, 2.8, 3.8, 4.2, 5, 5.2, 5.8, 6.2, 6.8, 7.4, 8.1, 8.4, 10, 10.7,
11.1)
> y<-c(29, 13, 31, 27, 45, 48, 44, 50, 44, 61, 73, 50, 78, 54, 83, 61, 74, 67, 80)
> fit1<-lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-18.909  -7.280  -1.925   8.355  17.703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.8073     4.8378   5.335 5.48e-05 ***
x             5.0844     0.7703   6.601 4.50e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.77 on 17 degrees of freedom
Multiple R-squared:  0.7193,	Adjusted R-squared:  0.7028
F-statistic: 43.57 on 1 and 17 DF,  p-value: 4.497e-06

> plot(x,y,xlab="Hours of Study",ylab="Exam Score")
> abline(fit1)
```

Regression parameters

T-test for b_1 in R (manual calc)

```
> SSX = sum((x-mean(x))^2)
> SSX
[1] 195.44
> SSres = sum((y-predict(fit1))^2)
> SSres
[1] 1971.291
> dfres=length(y)-2
> dfres
[1] 17
> seb1 = sqrt((SSres/dfres)/SSX)
> seb1
[1] 0.7702722
> b1=5.084425
> tobs=b1/seb1
> tobs
[1] 6.600816
> tcrit = qt(.975,24)
> tcrit
[1] 2.063899
```

Hypothesis test for slope

- H0: $\beta_1 = 0$ vs. H1: $\beta_1 \neq 0$
- Data:
  - $b_1 = 5.08$
  - SSX = 195.44, SSres = 1971.29, dfres = 17
  - se(b1) = sqrt((1971.29/17)/195.44) = 0.77
  - $t_{obs} = 5.08/0.77 = 6.60$
  - $t_{crit}(0.025, df=17) = 2.063$
  - Note: the probability of observing a slope of 5.08 in this sample, <u>assuming that β1 = 0</u> is less than 0.05 ($p < 0.05$)
- Conclusion:
  - $|t_{obs}| > t_{crit}$, <u>therefore</u> Reject H_0
  - There is sufficient evidence to conclude that there is a statistically significant linear relationship between Hours of Study and Exam Score

T-test for the slope in R (summary)

```
x<-c(1, 1.2, 1.4, 1.6, 1.9, 2.8, 3.8, 4.2, 5, 5.2, 5.8, 6.2, 6.8, 7.4, 8.1, 8.4, 10, 10.7,
11.1)
> y<-c(29, 13, 31, 27, 45, 48, 44, 50, 44, 61, 73, 50, 78, 54, 83, 61, 74, 67, 80)
> fit1<-lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.909  -7.280  -1.925   8.355  17.703 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  25.8073     4.8378   5.335 5.48e-05 ***
x             5.0844     0.7703   6.601 4.50e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.77 on 17 degrees of freedom
Multiple R-squared:  0.7193,     Adjusted R-squared:  0.7028 
F-statistic: 43.57 on 1 and 17 DF,  p-value: 4.497e-06

> plot(x,y,xlab="Hours of Study",ylab="Exam Score")
> abline(fit1)
```

t-test for the slope

Confidence Interval for B_1

- What values of β_1 could have resulted in our sample slope (b_1)?

- Just like the CI for Mu, we need:
  - the point estimate: b_1
  - Standard error of slope: $se(b_1)$
  - The sampling distribution: $t(\alpha/2, dfres)$

- Lower Limit = $b_1 - t_{crit} * se(b_1)$
- Upper Limit = $b_1 + t_{crit} * se(b_1)$

Confidence Interval for slope

- For Hours of Study and Exam Score
  - LL = 5.08 – 2.063(0.77) = 3.49
  - UL = 5.08 + 2.063(0.77) = 6.67

- "We are 95% confident that the population slope (β_1) is between 3.49 and 6.67"

Confidence Intervals for Regression Coefficients in R

95% CI for slope and intercept

```
> confint(fit1)
                2.5 %    97.5 %
(Intercept) 15.600409 36.014118
x            3.459293  6.709557
```

90% CI for slope and intercept

```
> confint(fit1,level=0.90)
                  5 %      95 %
(Intercept) 17.391403 34.223124
x            3.744454  6.424396
```