

ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

22-Jan-2018
Week 2

Announcement

- Assignment 1 will be released tomorrow morning
 - Due on 29 January (in-class)
 - No late submission

Week 2: Inference in Regression Analysis (Part 1)

Week 2: Inference in Regression Analysis (Part 1)

Week 2: Inference in Regression Analysis (Part 1)

- Review of Week 1
- Inferences on β_1
 - Confidence interval

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

$$\text{Model: } Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i : value of the response variable of the i^{th} observation
- β_0, β_1 : parameters
 β_1 : slope, β_0 : intercept
- ϵ_i are independent $N(0, \sigma^2)$
Thus, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

Freshmen's GPA and their college entrance test

- Study: 120 students at random from the new freshman class
- Goal: Can we predict GPA from ACT test score ?
- Y - GPA, X - ACT score; Normal error regression model

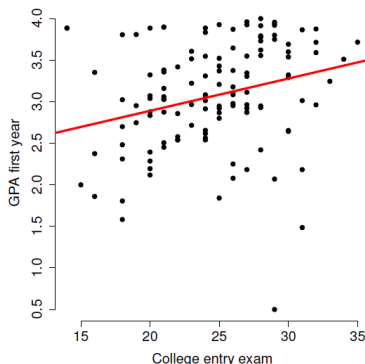
► Data:

i:	1	2	3	...	118	119	120
X_i :	21	14	28	...	28	16	28
Y_i :	3.897	3.885	3.778	...	3.914	1.860	2.948

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

Freshmen's GPA and their college entrance test



The least squares estimate of the regression line (red) is $b_0 + b_1X$,
with $b_0 = 2.11$ and $b_1 = 0.04$

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

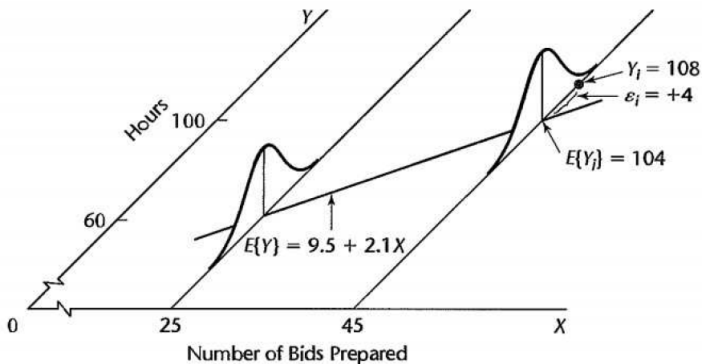
Simple linear regression vs. Single population

- Simple linear regression: $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ independently for each i
- Single population: $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ independently (special case of simple linear regression with $\beta_1 = 0$ and $\beta_0 = \mu$)

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

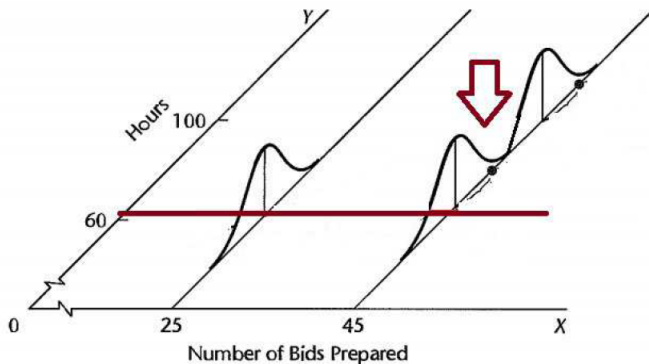
Simple linear regression



Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

Single population



Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

Least squares estimator

- Minimize $Q = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$
- LS estimator b_0, b_1 solves two **normal equations**:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

Least squares estimators of β_1, β_0

- $b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
- $b_0 = \bar{Y} - b_1 \bar{X}$

Estimator of σ^2

- $s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$

Week 2: Inference in Regression Analysis (Part 1)

Review of Week 1

Properties of estimators

- Gauss-Markov theorem ((1.11) in the text):
the LS estimators b_1 and b_0 are the Best Linear Unbiased Estimator (BLUE).
Here, “best” means giving the lowest variance among all unbiased linear estimators.
- s^2 is an unbiased estimator:
$$E[s^2] = \sigma^2$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Inference on β_1

- Often, we want to test whether there exists **linear association between X and Y** .
- The test would look like

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- Under the null, the means of Y_i 's for all i 's are equal at all levels of X_i (simple population)

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Sampling dist. of b_1

- The point estimator of b_1 :

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- **The sampling distribution of b_1** implies the distribution of the different values of b_1 that would be obtained over repeated sampling with the values of X fixed

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Sampling dist. of b_1

- For normal error regression model, the sampling distribution of $b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ is **normal**
- $E(b_1) = \beta_1$
- $Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick notes

- $b_1 = \sum_{i=1}^n k_i Y_i$
where $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$
- Note that
$$\sum k_i = 0$$
$$\sum k_i X_i = 1$$
$$\sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2}$$
- $b_1 = \sum k_i (Y_i - \bar{Y}) = \sum k_i Y_i - \bar{Y} \sum k_i = \sum k_i Y_i$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Normality of b_1

- The Y_i 's are independently, normally distributed
- A linear combination of Y_i 's is also normally distributed as follows:
For constants c_1, \dots, c_n ,
$$\sum_{i=1}^n c_i Y_i \sim N(\sum c_i E[Y_i], \sum c_i^2 \text{Var}(Y_i))$$
- b_1 is a linear combination of the Y_i 's ($b_1 = \sum k_i Y_i$)
 $\rightarrow b_1 \sim N(\sum k_i E[Y_i], \sum k_i^2 \text{Var}(Y_i))$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Mean of b_1

$$\begin{aligned} E(b_1) &= E\left(\sum k_i Y_i\right) \\ &= \sum k_i E(Y_i) \\ &= \sum k_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\ &= \beta_0 \cdot 0 + \beta_1 \cdot 1 \\ &= \beta_1 \end{aligned}$$

(Unbiased)

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Variance of b_1

$$\begin{aligned} \text{Var}(b_1) &= \text{Var}\left(\sum k_i Y_i\right) \\ &= \sum k_i^2 \text{Var}(Y_i) \\ &= \sum k_i^2 \sigma^2 \\ &= \sigma^2 \sum k_i^2 \\ &= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2} \end{aligned}$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Estimated variance of b_1

- Usually, we do not know the value of σ^2 , and thus use the estimator s^2 in place of σ^2 where $s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$
- The estimated variance of b_1 is

$$\widehat{Var}(b_1) = \frac{s^2}{\sum(X_i - \bar{X})^2}$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Recap

The sampling distribution of b_1 is

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Digression: Gauss-Markov Theorem

Theorem

In a regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$ for all i 's, and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, the LS estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimator.

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Proof of minimum variance among all unbiased linear estimator

- Let $\hat{\beta}_1$ be an unbiased linear estimator so that

$$\hat{\beta}_1 = \sum c_i Y_i \quad \text{for some constants } c_i\text{'s}$$

and $E(\hat{\beta}_1) = \beta_1$

- We show that

$$\text{Var}(\hat{\beta}_1) \geq \text{Var}(b_1)$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Proof cont.

- Being an unbiased estimator, $\hat{\beta}_1$ need to satisfy $E(\hat{\beta}_1) = \beta_1$, which is

$$\begin{aligned} E(\hat{\beta}_1) &= \sum c_i E(Y_i) \\ &= \sum c_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1 \end{aligned}$$

- Thus, $\sum c_i = 0$ and $\sum c_i X_i = 1$ should hold

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Proof cont.

- The variance of $\hat{\beta}_1$ is

$$\text{Var}(\hat{\beta}_1) = \sum c_i^2 \text{Var}(Y_i) = \sigma^2 \sum c_i^2$$

- Letting $c_i = k_i + d_i$ where $k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$,

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \sigma^2 \sum (k_i + d_i)^2 \\ &= \sigma^2 \left(\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right) \\ &= \text{Var}(b_1) + \sigma^2 \left(\sum d_i^2 + 2 \sum k_i d_i \right)\end{aligned}$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Proof cont.

- We show that $\sum k_i d_i = 0$:

$$\begin{aligned}\sum k_i d_i &= \sum k_i (c_i - k_i) \\&= \sum k_i c_i - \sum k_i^2 \\&= \sum c_i \left(\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right) - \frac{1}{\sum (X_i - \bar{X})^2} \\&= \frac{\sum c_i X_i}{\sum (X_i - \bar{X})^2} - \frac{\bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} \\&= \frac{1 \cdot 1}{\sum (X_i - \bar{X})^2} - \frac{\bar{X} \cdot 0}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2}\end{aligned}$$

since $\sum c_i = 0$ and $\sum c_i X_i = 1$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Proof cont.

- We have

$$\text{Var}(\hat{\beta}_1) = \text{Var}(b_1) + \sigma^2 \sum d_i^2 \geq \text{Var}(b_1)$$

- The minimum is attained when $\sum d_i^2 = 0$ which is equivalent to $d_i = 0$ for all i , in which case the unbiased linear estimator becomes b_1
 $\Rightarrow b_1$ has the minimum variance among all linear unbiased estimators

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Sampling dist. regarding b_1

- b_1 is normally distributed, and

$$\frac{b_1 - \beta_1}{\sqrt{\text{Var}(b_1)}} \sim N(0, 1) \quad \text{where } \text{Var}(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

- We show that

$$\frac{b_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(b_1)}} \sim t(n - 2) \quad \text{where } \widehat{\text{Var}}(b_1) = \frac{\text{MSE}}{\sum (X_i - \bar{X})^2}$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review of t-distribution

- If z and y are independent random variables such that $z \sim N(0, 1)$ and $y \sim \chi^2(\nu)$, then

$$\frac{z}{\sqrt{\frac{y}{\nu}}} \sim t(\nu)$$

(t distribution with degrees of freedom ν)

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

For the normal error regression model, we have

- $\frac{SSE}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} = \frac{\sum e_i^2}{\sigma^2} \sim \chi^2(n - 2)$
- $\frac{SSE}{\sigma^2}$ is independent of b_0 and b_1

((2.11) in the text)

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

For the normal error regression model, we have

- $\frac{SSE}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} = \frac{\sum e_i^2}{\sigma^2} \sim \chi^2(n - 2)$
- $\frac{SSE}{\sigma^2}$ is independent of b_0 and b_1

((2.11) in the text)

Quick note: why is the degrees of freedom $n - 2$, not $n - 1$ or n ?

\Rightarrow two parameters (β_1 and β_0) need to be estimated from n samples

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Now, we have

- $\frac{b_1 - \beta_1}{\sqrt{\text{Var}(b_1)}} \sim N(0, 1)$ and independent of SSE
(since SSE is independent of b_1 and b_0 from (2.11))
- $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$ (from (2.11))
- $\frac{\widehat{\text{Var}}(b_1)}{\text{Var}b_1} = \frac{MSE / \sum(X_i - \bar{X})^2}{\sigma^2 / \sum(X_i - \bar{X})^2} = \frac{SSE / (n-2)}{\sigma^2}$

$$\begin{aligned}\Rightarrow \frac{b_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(b_1)}} &= \left(\frac{b_1 - \beta_1}{\sqrt{\text{Var}(b_1)}} \right) / \left(\frac{\sqrt{\widehat{\text{Var}}(b_1)}}{\sqrt{\text{Var}(b_1)}} \right) \\ &\sim \frac{z}{\sqrt{y/n}} \quad \text{for } z \sim N(0, 1), y \sim \chi^2(n-2), x \perp y \\ &\sim t(n-2)\end{aligned}$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Remark: sampling distribution of b_1

- For fixed X_i 's, suppose we repeatedly and independently sample Y_1, \dots, Y_n

$$1^{st} \text{ sample} : Y_1^{(1)}, \dots, Y_n^{(1)}$$

$$2^{nd} \text{ sample} : Y_1^{(2)}, \dots, Y_n^{(2)}$$

$$\vdots$$

$$m^{th} \text{ sample} : Y_1^{(m)}, \dots, Y_n^{(m)}$$

$$\vdots$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Remark: sampling distribution of b_1

- For fixed X_i 's, suppose we repeatedly and independently sample Y_1, \dots, Y_n

$$\begin{aligned} 1^{st} \text{ sample} &: Y_1^{(1)}, \dots, Y_n^{(1)} \rightarrow b_1^{(1)} \\ 2^{nd} \text{ sample} &: Y_1^{(2)}, \dots, Y_n^{(2)} \rightarrow b_1^{(2)} \\ &\vdots \\ m^{th} \text{ sample} &: Y_1^{(m)}, \dots, Y_n^{(m)} \rightarrow b_1^{(m)} \\ &\vdots \end{aligned}$$

- Then, for each sample, we get different values of b_1 's

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

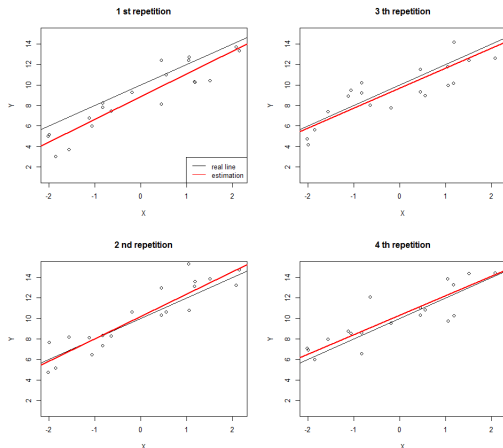
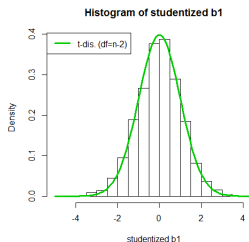
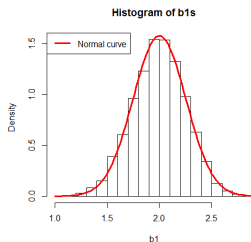


Figure: $E(Y) = 10 + 2X$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Remark: sampling distribution of b_1



- b_1 is drawn from $N(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2})$ distribution
- $\frac{(b_1 - \beta_1)}{\sqrt{MSE / \sum(X_i - \bar{X})^2}}$ is drawn from $t(n - 2)$ distribution

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Confidence interval and hypothesis test

- Now we know the sampling distribution of b_1 , and can construct confidence intervals and hypothesis test on β_1

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval

- Interval estimate of a parameter
- $(1 - \alpha) \cdot 100\%$ confidence interval implies that if we repeat sampling infinitely many times, then $(1 - \alpha) \cdot 100\%$ of those confidence interval would contain the targeted population parameter

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval Single population example with known σ^2

Suppose weekly income of senior level assembly-line workers in a company is normally distributed as $N(\mu, \sigma^2)$.

In order to investigate the population mean μ , a researcher randomly samples nine employees and check their weekly income X_1, \dots, X_9 .

- A point estimate of the mean: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
- We know $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval
Single population example with known σ^2

- To construct $(1 - \alpha)\%$ confidence interval, we want to find a and b such that

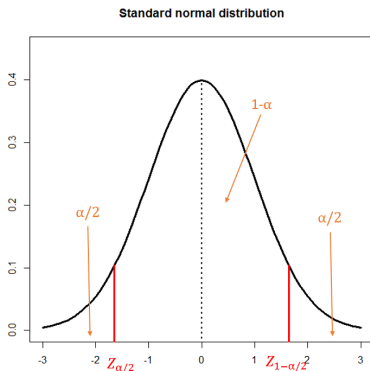
$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) = 1 - \alpha$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval Single population example with known σ^2

- We take $a = z_{\alpha/2}$ and $b = z_{1-\alpha/2}$
($z_{\alpha/2} = -z_{1-\alpha/2}$ since standard normal distribution is symmetric around 0)



Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval
Single population example with known σ^2

- We have

$$\begin{aligned}1 - \alpha &= P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) \\&= P\left(\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

- Our confidence interval becomes

$$\left(\bar{X}_{obs} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_{obs} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval True or false?

For a random sample of nine employees, a researcher obtains a 95% confidence interval of (371, 509) for the mean weekly income.

- We infer that the 95% of the employees in the population have income between \$371 and \$509
- The probability that (371, 509) contains the population mean is 95%

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval True or false?

For a random sample of nine employees, a researcher obtains a 95% confidence interval of (371, 509) for the mean weekly income.

- We infer that the 95% of the employees in the population have income between \$371 and \$509
⇒ False.
- The probability that (371, 509) contains the population mean is 95%
⇒ False.

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Quick review: confidence interval True or false?

For a random sample of nine employees, a researcher obtains a 95% confidence interval of (371, 509) for the mean weekly income.

- We infer that the 95% of the employees in the population have income between \$371 and \$509
⇒ False.
- The probability that (371, 509) contains the population mean is 95%
⇒ False.

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Confidence interval for β_1

- We have $\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n - 2)$
- Therefore,

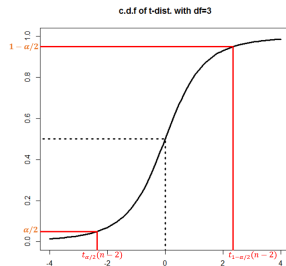
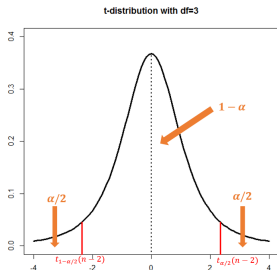
$$P\left(t_{\alpha/2}(n - 2) \leq \frac{b_1 - \beta_1}{s\{b_1\}} \leq t_{1-\alpha/2}(n - 2)\right) = 1 - \alpha$$

- $t_{\alpha/2}(n - 2)$ and $t_{1-\alpha/2}(n - 2)$ denotes $\alpha/2$ and $1 - \alpha/2$ quantiles of t-distribution with d.f. $n-2$ respectively.

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

- $t_{\alpha/2}(n-2)$ and $t_{1-\alpha/2}(n-2)$ denotes $\alpha/2$ and $1 - \alpha/2$ quantiles of t-distribution with d.f. $n-2$ respectively.



- $t_{\alpha/2}(n-2) = -t_{1-\alpha/2}(n-2)$ since t-dist. is symmetric around 0.

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Confidence interval for β_1

- Now, we have

$$\begin{aligned} 1 - \alpha &= P\left(-t_{1-\alpha/2}(n-2) \leq \frac{b_1 - \beta_1}{s\{b_1\}} \leq t_{1-\alpha/2}(n-2)\right) \\ &= P\left(b_1 - t_{1-\alpha/2}(n-2) \cdot s\{b_1\} \leq \beta_1 \leq b_1 + t_{1-\alpha/2}(n-2) \cdot s\{b_1\}\right) \end{aligned}$$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Confidence interval for β_1

- Thus, $(1 - \alpha) \cdot 100\%$ confidence interval for β_1 is

$$(b_1 - t_{1-\alpha/2}(n-2) \cdot s\{b_1\}, b_1 + t_{1-\alpha/2}(n-2) \cdot s\{b_1\})$$

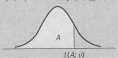
i.e., $b_1 \pm t_{1-\alpha/2}(n-2) \cdot s\{b_1\}$

- Note that this quantity can be used to calculate confidence intervals given n and α
 - Fixing α can guide the choice of sample size if a particular confidence interval is desired
 - Given a sample size n , vice versa.
- Also useful for hypothesis testing

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Entry is $t(A; \nu)$ where $P\{t(\nu) \leq t(A; \nu)\} = A$



	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Finding quantiles of t-dist.

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

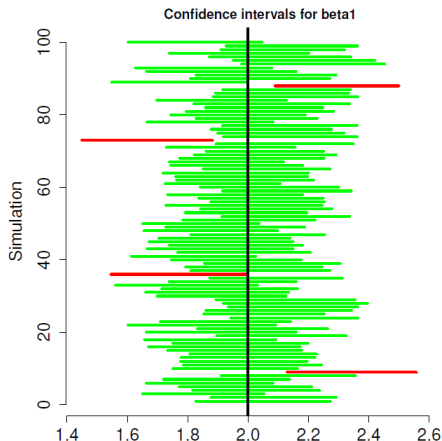
How to interpret confidence interval

- With 95% confidence interval, we can be 95% **confident** that β_1 , the true slope of the regression line, is within the 95% CI
- But what does “being confident” actually mean?
 - β_1 is an unknown but fixed parameter, so you CANNOT state that “the probability that β_1 is in its 95% CI is 0.95”!
 β_1 is either in its CI, or its not!
 - Instead: if we select n random samples (with X_i 's fixed) repeatedly, then 95% of the time the CIs for β_1 cover the true parameter β_1 in the long run

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

Simulation example to illustrate “coverage” of CIs

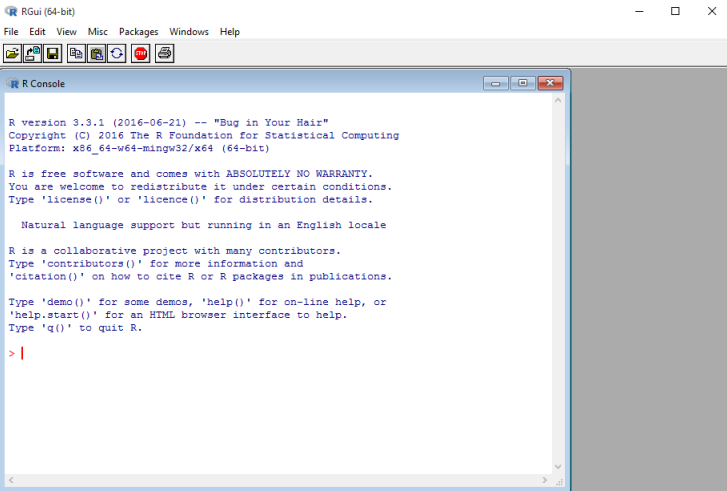


- Each horizontal line gives a 95% CI for $\beta_1 = 2$
- For 100 simulations, we see that the CI's cover β_1 96 times.

Intro to R

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1



The screenshot shows the RGui (64-bit) application window. The title bar reads "RGui (64-bit)". The menu bar includes "File", "Edit", "View", "Misc", "Packages", "Windows", and "Help". The toolbar contains icons for file operations and running code. The "R Console" window is open, displaying the R startup message:

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

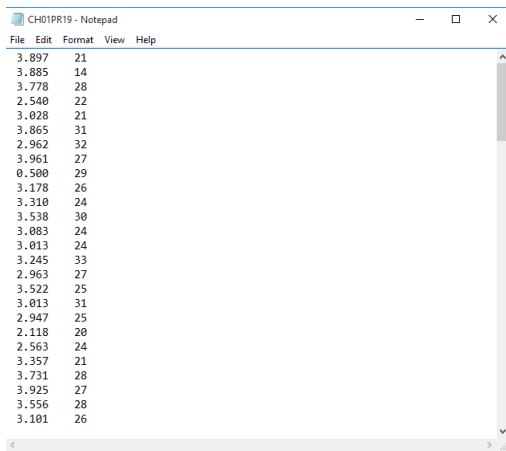
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

At the bottom right of the slide, there are navigation icons (back, forward, search, etc.) and the page number "55 / 65".

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

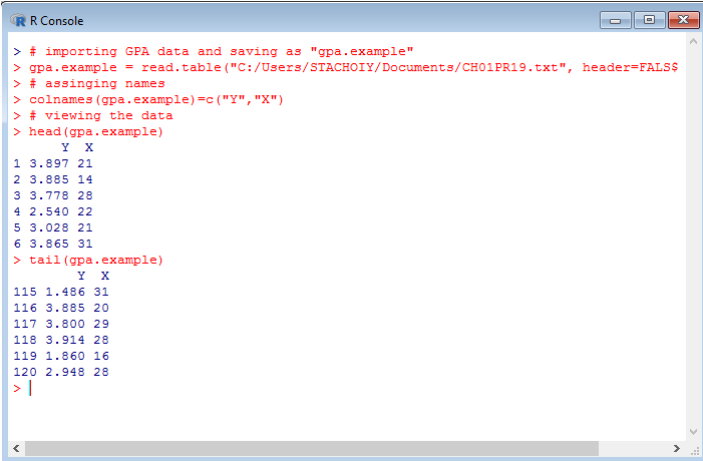


3.897	21
3.885	14
3.778	28
2.540	22
3.028	21
3.865	31
2.962	32
3.961	27
0.500	29
3.178	26
3.310	24
3.538	30
3.083	24
3.013	24
3.245	33
2.963	27
3.522	25
3.013	31
2.947	25
2.118	20
2.563	24
3.357	21
3.731	28
3.925	27
3.556	28
3.101	26

GPA vs. ACT example data in CH01PR19.txt

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

A screenshot of an R Console window. The title bar says "R Console". The console shows a series of commands and their output. The commands are: 1. A comment: "# importing GPA data and saving as 'gpa.example'". 2. A command: "gpa.example = read.table('C:/Users/STACHOIY/Documents/CH01PR19.txt', header=FALSE)". 3. A comment: "# assinging names". 4. A command: "colnames(gpa.example)=c('Y', 'X')". 5. A comment: "# viewing the data". 6. A command: "head(gpa.example)". The output for head shows a table with two columns, Y and X, and 6 rows of data. 7. A command: "tail(gpa.example)". The output for tail shows a table with two columns, Y and X, and 5 rows of data (rows 115 to 120). The console has a scrollbar on the right and a status bar at the bottom.

```
> # importing GPA data and saving as "gpa.example"
> gpa.example = read.table("C:/Users/STACHOIY/Documents/CH01PR19.txt", header=FALSE)
> # assinging names
> colnames(gpa.example)=c("Y", "X")
> # viewing the data
> head(gpa.example)
      Y X
1 3.897 21
2 3.885 14
3 3.778 28
4 2.540 22
5 3.028 21
6 3.865 31
> tail(gpa.example)
      Y X
115 1.486 31
116 3.885 20
117 3.800 29
118 3.914 28
119 1.860 16
120 2.948 28
> |
```

Importing and viewing the data

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

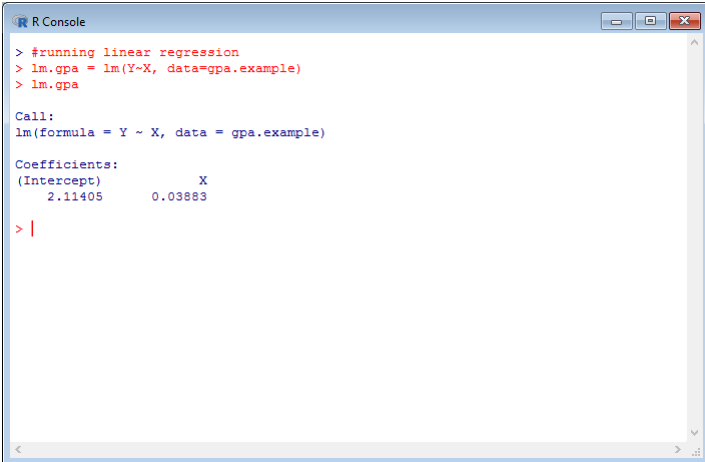
```
R Console

> # accessing the data
> ## accessing the element in the 2nd row, 1st column
> gpa.example[2,1]
[1] 3.885
> ## accessing the 100th row
> gpa.example[100,]
      Y X
100 2.311 18
> ## accessing the 2nd column
> gpa.example[,1]
 [1] 3.897 3.885 3.778 2.540 3.028 3.865 2.962 3.961 0.500 3.178 3.310 3.538 3.083
[14] 3.013 3.245 2.963 3.522 3.013 2.947 2.118 2.563 3.357 3.731 3.925 3.556 3.101
[27] 2.420 2.579 3.871 3.060 3.927 2.375 2.929 3.375 2.857 3.072 3.381 3.290 3.549
[40] 3.646 2.978 2.654 2.540 2.250 2.069 2.617 2.183 2.000 2.952 3.806 2.871 3.352
[53] 3.305 2.952 3.547 3.691 3.160 2.194 3.323 3.936 2.922 2.716 3.370 3.606 2.642
[66] 2.452 2.655 3.714 1.806 3.516 3.039 2.966 2.482 2.700 3.920 2.834 3.222 3.084
[79] 4.000 3.511 3.323 3.072 2.079 3.875 3.208 2.920 3.345 3.956 3.808 2.506 3.886
[92] 2.183 3.429 3.024 3.750 3.833 3.113 2.875 2.747 2.311 1.841 1.583 2.879 3.591
[105] 2.914 3.716 2.800 3.621 3.792 2.867 3.419 3.600 2.394 2.286 1.486 3.885 3.800
[118] 3.914 1.860 2.948
> |
```

Accessing specific data points

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

A screenshot of an R Console window. The window has a title bar that says "R Console" and standard Windows window controls (minimize, maximize, close). The console shows the following text:

```
> #running linear regression
> lm.gpa = lm(Y~X, data=gpa.example)
> lm.gpa

Call:
lm(formula = Y ~ X, data = gpa.example)

Coefficients:
(Intercept)          X
    2.11405      0.03883

> |
```

The console has a vertical scrollbar on the right and a horizontal scrollbar at the bottom.

Running simple linear regression

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

```
R Console

> summary(lm.gpa)

Call:
lm(formula = Y ~ X, data = gpa.example)

Residuals:
    Min       1Q   Median       3Q      Max
-2.74004 -0.33827  0.04062  0.44064  1.22737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
X             0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917

> |
```

Viewing the summary table

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

The screenshot shows an R Console window with the following content:

```
> summary(lm.gpa)

Call:
lm(formula = Y ~ X, data = gpa.example)

Residuals:
    Min       1Q   Median       3Q      Max
-2.74004 -0.33827  0.04062  0.44064  1.22737

Coefficients:
(Intercept)  2.11405   0.32089   6.588   1.3e-09 ***
X            0.03883   0.01277   3.040   0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262    Adjusted R-squared:  0.06476 
F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

> |
```

Annotations in the image:

- Red boxes around the coefficients table:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.11405	0.32089	6.588	1.3e-09 ***
X	0.03883	0.01277	3.040	0.00292 **
- Red arrows pointing from the following text to the boxed values:
 - b_0 and b_1 points to the Estimate column.
 - $s\{b_0\}$ and $s\{b_1\}$ points to the Std. Error column.
 - $\frac{b_0-0}{s\{b_0\}}$ and $\frac{b_1-0}{s\{b_1\}}$ points to the t value column.
 - s points to the Residual standard error value (0.6231).

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

```
> summary(lm.gpa)

Call:
lm(formula = Y ~ X, data = gpa.example)

Residuals:
    Min       1Q   Median       3Q      Max 
-2.74004 -0.33827  0.04062  0.44664  1.22737 

Coefficients:
(Intercept)  2.11405  0.32089  6.588  1.3e-09 ***
X             0.03883  0.01277  3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262    Adjusted R-squared:  0.06476 
F-statistic: 9.24 on 1 and 118 DF,  p-value: 0.002917

> |
```

Handwritten notes and arrows in the image:

- $b_0 = 0$ and $b_1 = 0$ (pointing to the Intercept and X coefficients)
- b_0 and b_1 (pointing to the Intercept and X coefficients)
- $s(b_0)$ and $s(b_1)$ (pointing to the Std. Error for Intercept and X)
- s (pointing to the Residual standard error)

```
> ##b1
> xx = sum((gpa.example$X-mean(gpa.example$X))^2)
> xy = sum((gpa.example$X-mean(gpa.example$X))*(gpa.example$Y-mean(gpa.example$Y)))
> b1=xy/xx
> b1
[1] 0.03882713
> ##b0
> b0=mean(gpa.example$Y)-b1*mean(gpa.example$X)
> b0
[1] 2.114049
> |
```

Verifying b_1 and b_0

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

```
R Console
> summary(lm.gpa)

Call:
lm(formula = Y ~ X, data = gpa.example)

Residuals:
    Min       1Q   Median       3Q      Max
-2.74004 -0.33827  0.04062  0.44664  1.22737

Coefficients:
(Intercept) 2.11405 0.32089 6.588 1.3e-09 ***
X           0.03883 0.01277 3.040 0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262    Adjusted R-squared:  0.06476
F-statistic: 9.24 on 1 and 118 DF,  p-value: 0.002917

> |
```

Annotations in the image:

- Red arrows point from $b_0 = 0$ and $b_1 = 0$ to the intercept and X coefficients respectively.
- Red arrows point from $s(b_0)$ and $s(b_1)$ to the standard errors of the intercept and X coefficients respectively.
- Red arrows point from s to the residual standard error.

```
R Console
> ##s
> dim(gpa.example)
[1] 120 2
> s2 = sum( (gpa.example$Y - (b0+b1*gpa.example$X) )^2 )/118
> s = sqrt(s2)
> s
[1] 0.623125
> ##s{b_1}
> s/sqrt(xx)
[1] 0.01277302
> |
```

Verifying s and $s\{b_1\}$

Week 2: Inference in Regression Analysis (Part 1)

Inference on β_1

```
R Console
> # constructing 95% confidence interval
> ## 0.975 quantile of t-dist. with d.f 118
> qt(1-0.05/2, df=118)
[1] 1.980272
> width = qt(1-0.05/2, df=118)*s/sqrt(xx)
> ## confidence interval
> c(b1-width, b1+width)
[1] 0.01353307 0.06412118
> ## or alternatively
> confint(lm.gpa, level=0.95)
              2.5 %      97.5 %
(Intercept) 1.47859015 2.74950842
X            0.01353307 0.06412118
> |
```

Constructing CI for β_1

Reading: entire chapter 1