Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Statistical Inference:
Data Example

1

About This Lesson

2

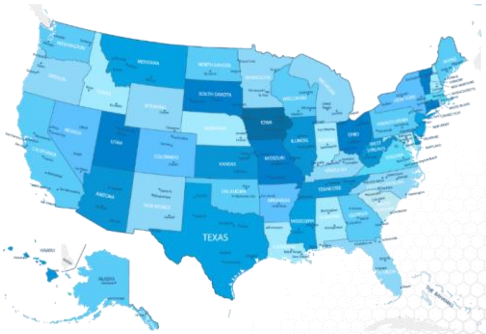
Linear Regression: Example 2

Controlling factors:

- X_1 = % of total eligible students in the state who took the exam
- X_6 = median percentile of ranking of test takers within their secondary school classes

Explanatory Factors:

- X_2 = median income of families of test takers, in hundreds of dollars
- X_3 = average number of years that test takers had in social sciences, natural sciences, and humanities
- X_4 = % of test takers who attended public schools
- X_5 = state expenditure on secondary schools, in hundreds of dollars per student



Example 2: Inference on Coefficients

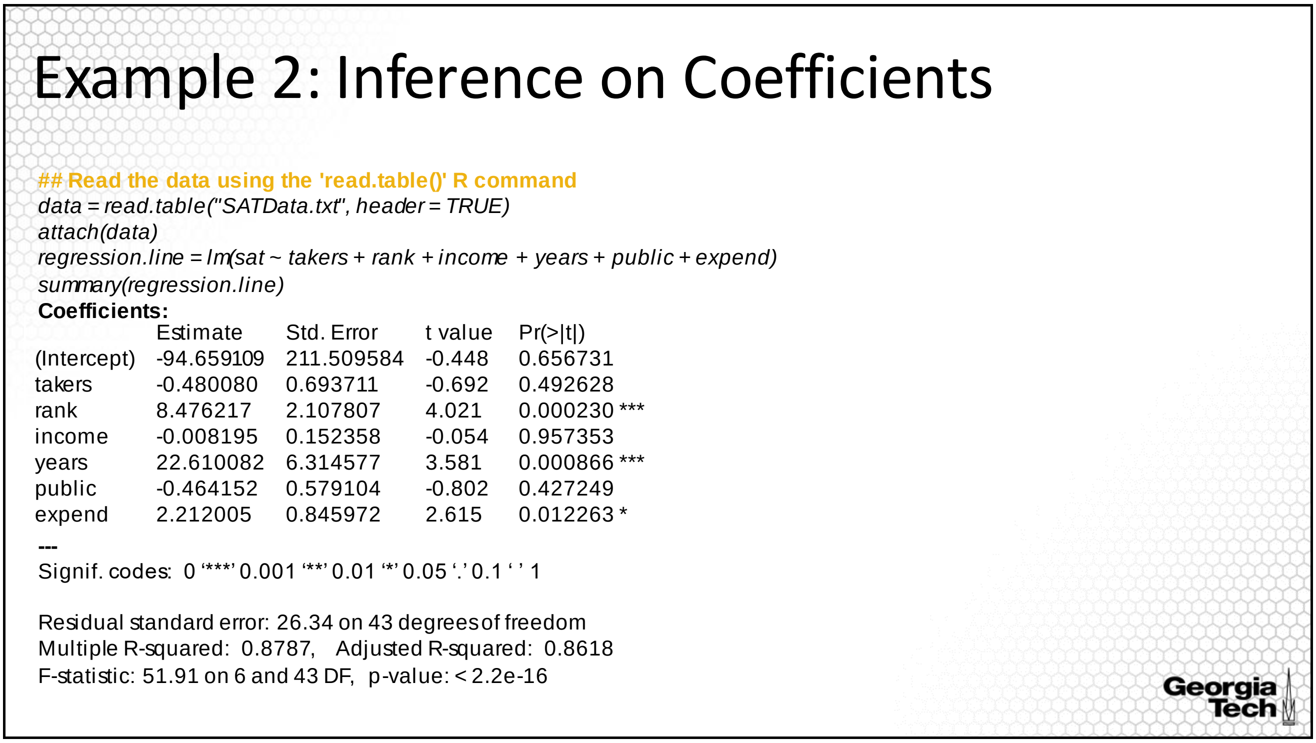
a. What is the estimate of the coefficient β_1 and its variance? Interpret. What is its sampling distribution?

b. Is the coefficient β_1 statistically significant? What is the p-value of the test. Interpret.

c. What is the F-statistic for overall regression? Do we reject the null hypothesis that all regression coefficients are zero?

d. Obtain the 99% confidence interval for β_1.

e. Given the controlling factors, test the null hypothesis that the coefficients of the other variables are zero. Clearly state the hypothesis test. Show how you perform the test. Interpret the results.



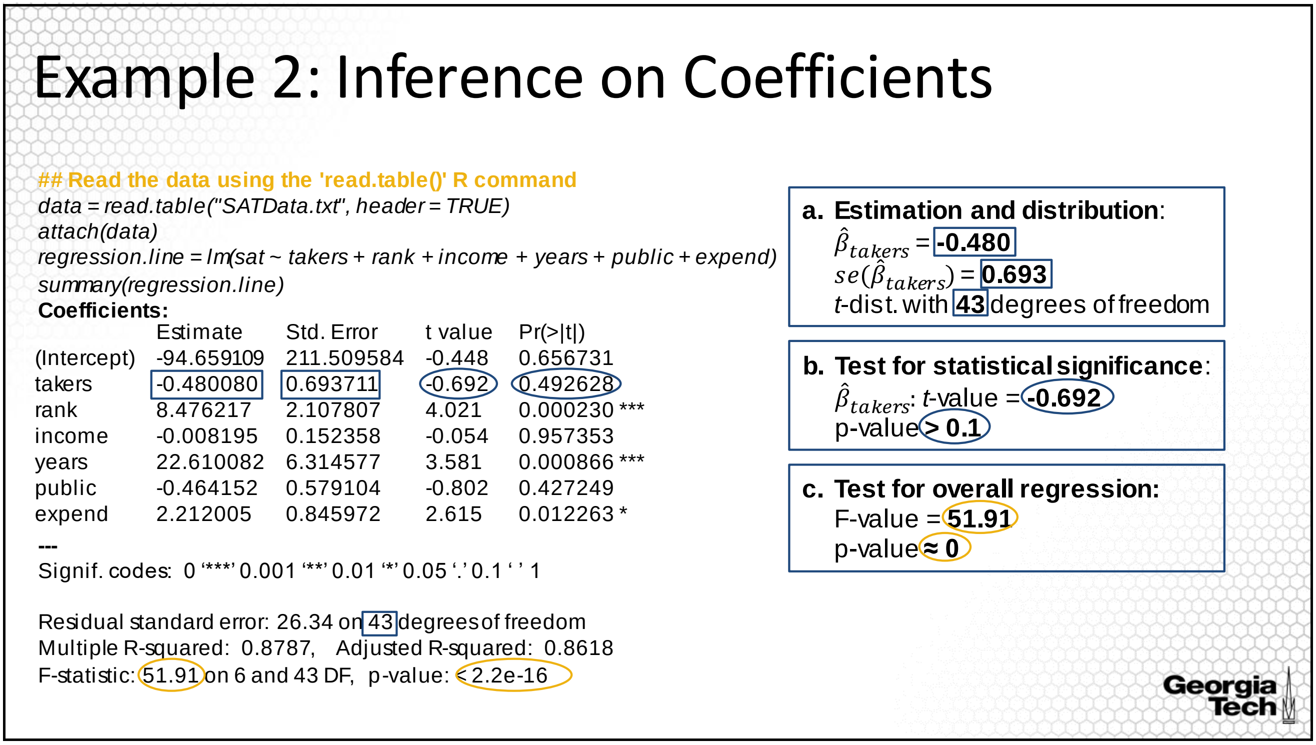
Example 2: Inference on Coefficients

## Read the data using the 'read.table()' R command

```
data = read.table("SATData.txt", header = TRUE)
attach(data)
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
Coefficients:
            Estimate    Std. Error  t value  Pr(>|t|)
(Intercept) -94.659109  211.509584  -0.448   0.656731
takers      -0.480080   0.693711    -0.692   0.492628
rank        8.476217    2.107807    4.021    0.000230 ***
income      -0.008195   0.152358    -0.054   0.957353
years       22.610082   6.314577    3.581    0.000866 ***
public      -0.464152   0.579104    -0.802   0.427249
expend      2.212005    0.845972    2.615    0.012263 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.34 on 43 degrees of freedom
Multiple R-squared:  0.8787,    Adjusted R-squared:  0.8618
F-statistic: 51.91 on 6 and 43 DF,  p-value: < 2.2e-16
```

Georgia Tech

5



Example 2: Inference on Coefficients

## Read the data using the 'read.table()' R command

```
data = read.table("SATData.txt", header = TRUE)
attach(data)
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
Coefficients:
            Estimate    Std. Error  t value  Pr(>|t|)
(Intercept) -94.659109  211.509584  -0.448   0.656731
takers      -0.480080   0.693711    -0.692   0.492628
rank        8.476217    2.107807    4.021    0.000230 ***
income      -0.008195   0.152358    -0.054   0.957353
years       22.610082   6.314577    3.581    0.000866 ***
public      -0.464152   0.579104    -0.802   0.427249
expend      2.212005    0.845972    2.615    0.012263 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.34 on 43 degrees of freedom
Multiple R-squared:  0.8787,    Adjusted R-squared:  0.8618
F-statistic: 51.91 on 6 and 43 DF,  p-value: < 2.2e-16
```

a. Estimation and distribution:

$\hat{\beta}_{takers} = \boxed{\mathbf{-0.480}}$

$se(\hat{\beta}_{takers}) = \boxed{\mathbf{0.693}}$

t-dist. with $\boxed{\mathbf{43}}$ degrees of freedom

b. Test for statistical significance:

$\hat{\beta}_{takers}$: *t*-value = **-0.692**

p-value **> 0.1**

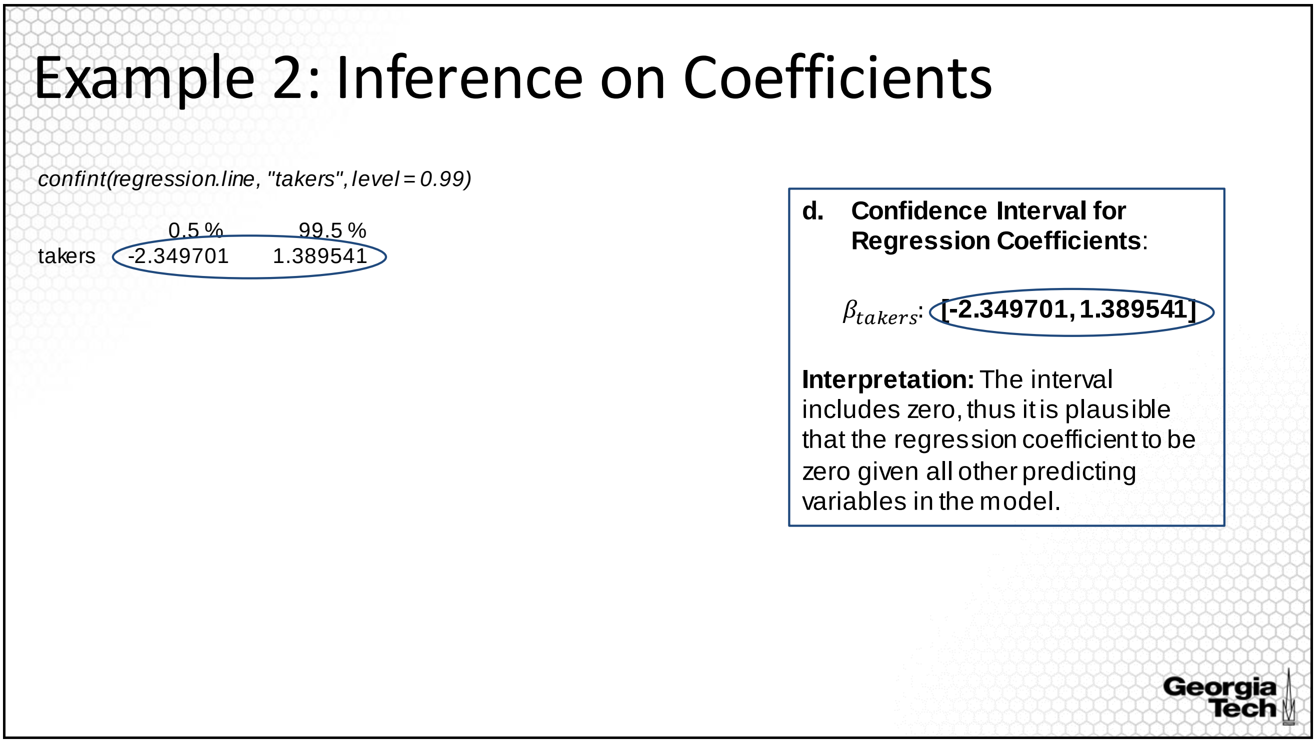
c. Test for overall regression:

F-value = **51.91**

p-value **≈ 0**

Georgia Tech

6



Example 2: Inference on Coefficients

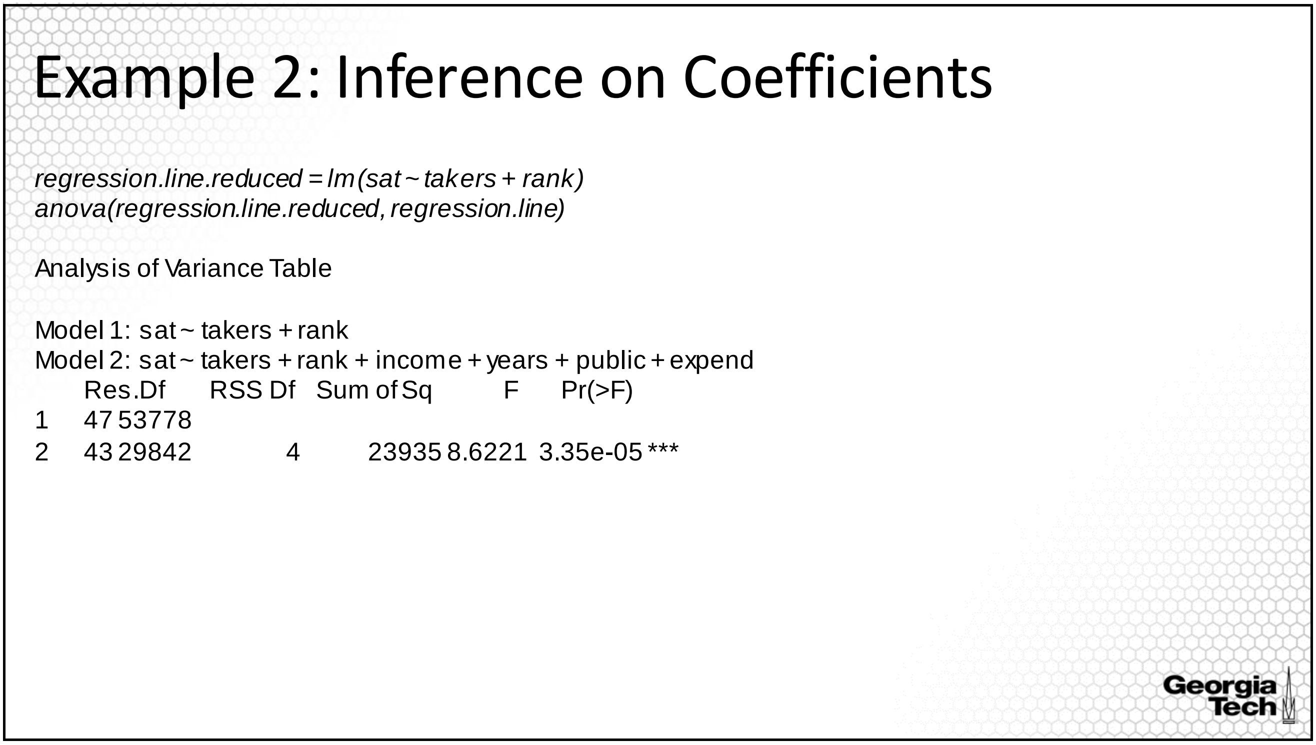
```
confint(regression.line, "takers", level = 0.99)

             0.5 %        99.5 %
takers  -2.349701      1.389541
```

d. Confidence Interval for Regression Coefficients:

β_{takers}: **[-2.349701, 1.389541]**

Interpretation: The interval includes zero, thus it is plausible that the regression coefficient to be zero given all other predicting variables in the model.

7



Example 2: Inference on Coefficients

```
regression.line.reduced = lm(sat ~ takers + rank)
anova(regression.line.reduced, regression.line)

Analysis of Variance Table

Model 1: sat ~ takers + rank
Model 2: sat ~ takers + rank + income + years + public + expend
    Res.Df     RSS Df  Sum of Sq       F     Pr(>F)
1       47 53778
2       43 29842           4       23935 8.6221  3.35e-05 ***
```

8

Example 2: Inference on Coefficients

```
regression.line.reduced = lm(sat ~ takers + rank)
anova(regression.line.reduced, regression.line)

Analysis of Variance Table

Model 1: sat ~ takers + rank
Model 2: sat ~ takers + rank + income + years + public + expend
     Res.Df     RSS Df  Sum of Sq        F     Pr(>F)
1       47  53778
2       43  29842        4       23935 8.6221 3.35e-05 ***
```

e. Testing for a subset of regression coefficients:

H_0: Reduced Model (*takers* and *rank* only) vs. H_A: Full Model

Partial F Test:

F-value = **8.6221**

P-value **≈ 0**

Example 2: Inference on Coefficients

e. Testing for a subset of regression coefficients *(continued)*:

Test H_0: $\beta_{income} = \beta_{years} = \beta_{public} = \beta_{expend} = 0$

How was the F-statistic computed?

$$\text{F-statistic} = \frac{\text{SSReg}(income, years, public, expend \mid takers, rank)/4}{\text{SSE}/(50 - 6 - 1)}$$

The p-value is computed as

$$\text{Prob}(\text{F}_{4,43} > \text{F} - \text{statistic}) = 1 - \text{Prob}(\text{F}_{4,43} < \text{F} - \text{statistic})$$

Interpretation: The p-value is approximately 0, so reject the null hypothesis. We conclude that at least one predictor among *income*, *years*, *public* and *expend* will be significantly associated with states' average SAT scores.

Summary

11