Regression Analysis

Multiple Linear Regression

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

Regression Parameter Estimation: Data Example

1

About This Lesson

2

Linear Regression: Example 1

Quantitative Predicting Variables:

X_1 = the amount (in hundreds of dollars) spent on advertising

X_2 = the total amount of bonuses paid

X_3 = the market share in each territory

X_4 = the largest competitor's sales

Qualitative Predicting Variable:

X_5 = a variable to indicate the region in which office is located (1 = south, 2 = west, 3 = midwest)

Response Variable:

Y = yearly sales (in thousands of dollars)

Georgia Tech

3

Example 1: Estimation & Interpretation

a. Fit a linear regression with all predictors. What are the estimated regression coefficients and the estimated regression line?

b. Interpret the coefficients. Compare the estimated coefficient for the advertisement expenditure variable under the conditional (full) model vs. the marginal (one predictor) model.

c. What change does the full regression model predict for yearly sales as the advertisement expenditure increases by an additional $1,000? Is this prediction different when compared to that from the simple linear model with the advertisement expenditure variable only?

d. What is the estimate of the error variance under the full model? Is it different from that under the simple linear regression model? Why?

Georgia Tech

4

Example 1: Estimation & Interpretation

```
meddcor = read.table("meddcor.txt", sep = "", header = FALSE)
colnames(meddcor) = c("sales", "advertising", "bonuses", "marketshare", "largestcomp", "region")
meddcor$region = as.factor(meddcor$region)
model = lm(sales ~ ., data = meddcor)
summary(model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	117.0200	192.9732	0.606	0.5518
advertising	1.4092	0.2687	5.244	5.49e-05
bonuses	1.0123	0.4641	2.181	0.0427
marketshare	3.1548	2.9802	1.059	0.3038
largestcomp	-0.2354	0.2338	-1.007	0.3275
region2	53.6285	34.7359	1.544	0.1400
region3	267.9569	47.5577	5.634	2.40e-05 ***

Residual standard error: 55.57 on 18 degrees of freedom
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9407
F-statistic: 64.42 on 6 and 18 DF, p-value: 3.466e-11

Example 1: Estimation & Interpretation

```
meddcor = read.table("meddcor.txt", sep = "", header = FALSE)
colnames(meddcor) = c("sales", "advertising", "bonuses", "marketshare", "largestcomp", "region")
meddcor$region = as.factor(meddcor$region)
model = lm(sales ~ ., data = meddcor)
summary(model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	117.0200	192.9732	0.606	0.5518
advertising	1.4092	0.2687	5.244	5.49e-05
bonuses	1.0123	0.4641	2.181	0.0427
marketshare	3.1548	2.9802	1.059	0.3038
largestcomp	-0.2354	0.2338	-1.007	0.3275
region2	53.6285	34.7359	1.544	0.1400
region3	267.9569	47.5577	5.634	2.40e-05 ***

Residual standard error: 55.57 on 18 degrees of freedom
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9407
F-statistic: 64.42 on 6 and 18 DF, p-value: 3.466e-11

a. Estimated Regression Coefficients

b. Conditional model:

$\hat{\beta}_{adv} = 1.4092$

The expected additional gain in sales in thousands for $100 additional expenditure in advertisement **while holding all other fixed.**

Marginal model:

$\hat{\beta}_{adv} = 2.772$

The expected additional gain in sales in thousands for $100 additional expenditure in advertisement **not accounting for other predicting variables.**

Example 1: Estimation & Interpretation

```
meddcor = read.table("meddcor.txt", sep = "", header = FALSE)
colnames(meddcor) = c("sales", "advertising", "bonuses", "marketshare", "largestcomp", "region")
meddcor$region = as.factor(meddcor$region)
model = lm(sales ~ ., data = meddcor)
summary(model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	117.0200	192.9732	0.606	0.5518	
advertising	1.4092	0.2687	5.244	5.49e-05	
bonuses	1.0123	0.4641	2.181	0.0427	
marketshare	3.1548	2.9802	1.059	0.3038	
largestcomp	-0.2354	0.2338	-1.007	0.3275	
region2	53.6285	34.7359	1.544	0.1400	
region3	267.9569	47.5577	5.634	2.40e-05	***

Residual standard error: 55.57 on 18 degrees of freedom
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9407
F-statistic: 64.42 on 6 and 18 DF, p-value: 3.466e-11

c. An additional **$1,000** in advertising expenditures results in **$14,092** additional sales under the full model and **$27,720** additional sales under the simple linear model.

Which is more meaningful? Because sales varies with other factors, the interpretation based on multiple regression is more meaningful.

d. Under the full model, the variance estimate is **$(55.57)^2$**. Under the simple linear model, the variance estimate was **$(101.4)^2$**.

Why? More variability in the response is explained when including multiple predicting variables versus only one.

Georgia Tech

7

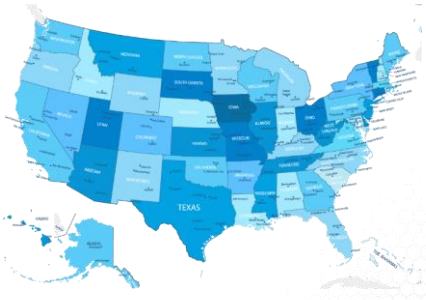
Linear Regression: Example 2

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

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

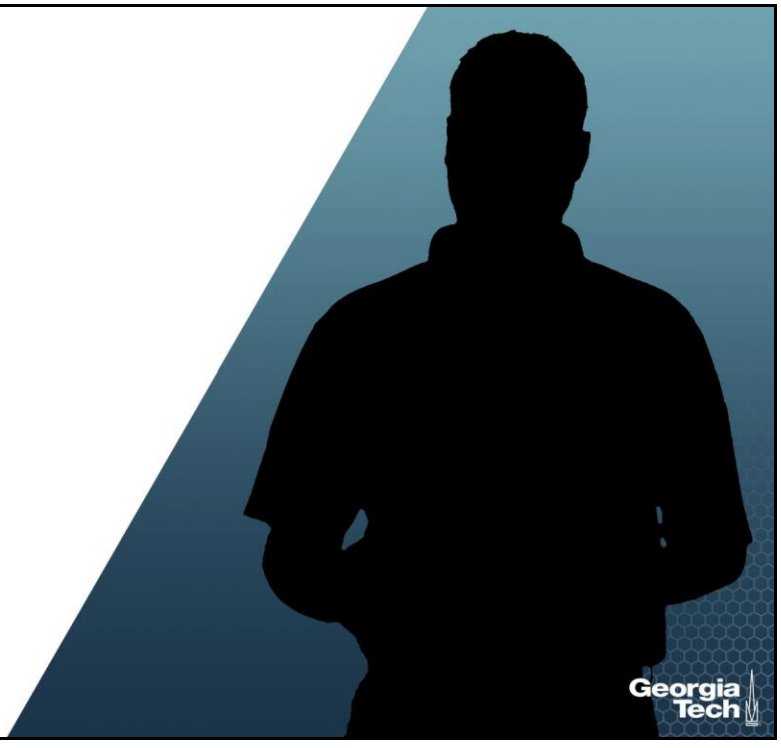


SAT Mean Score by State — Year 1982
790 (South Carolina) – 1088 (Iowa)

Georgia Tech

8

Summary



9