Intro to Social Science Data Analysis

Week 12 Lecture: Multivariate Linear Regression & Presenting Regression Results

Christopher Gandrud

November 12, 2012

1. Assignment 4

2. Recap

3. Multiple Linear Regression

Assignment 4

Due: Friday 30 November

Research Design

With your partner plan your research by answering the following questions:

1. What difference or anomaly do you want to explain?
2. What is your best guess explanation? Draw your best guess in a diagram.
3. Can you test your hypothesis using data? If so, what data do you need to collect and what tests could you use?
4. What rival explanations are their?
5. How could you use data to test whether your best guess or the rival explanations are better? Write this as an **equation** if possible.
6. What other factors may influence the relationship you observe?

Questionnaire from: modified from Cheryl Schonhardt-Bailey

Assignment 4

Due: Friday 30 November

Research Design

With your partner plan your research by answering the following questions:

1. What difference or anomaly do you want to explain?
2. What is your best guess explanation? Draw your best guess in a diagram.
3. Can you test your hypothesis using data? If so, what data do you need to collect and what tests could you use?
4. What rival explanations are their?
5. How could you use data to test whether your best guess or the rival explanations are better? Write this as an **equation** if possible.
6. What other factors may influence the relationship you observe?

Questionnaire from: modified from Cheryl Schonhardt-Bailey

Assignment 4

Due: Friday 30 November

Research Design

With your partner plan your research by answering the following questions:

1. What difference or anomaly do you want to explain?
2. What is your best guess explanation? Draw your best guess in a diagram.
3. Can you test your hypothesis using data? If so, what data do you need to collect and what tests could you use?
4. What rival explanations are their?
5. How could you use data to test whether your best guess or the rival explanations are better? Write this as an **equation** if possible.
6. What other factors may influence the relationship you observe?

Questionnaire from: modified from Cheryl Schonhardt-Bailey

Assignment 4

Due: Friday 30 November

Research Design

With your partner plan your research by answering the following questions:

1. What difference or anomaly do you want to explain?
2. What is your best guess explanation? Draw your best guess in a diagram.
3. Can you test your hypothesis using data? If so, what data do you need to collect and what tests could you use?
4. What rival explanations are their?
5. How could you use data to test whether your best guess or the rival explanations are better? Write this as an **equation** if possible.
6. What other factors may influence the relationship you observe?

Questionnaire from: modified from Cheryl Schonhardt-Bailey

Assignment 4

Due: Friday 30 November

Research Design

With your partner plan your research by answering the following questions:

1. What difference or anomaly do you want to explain?
2. What is your best guess explanation? Draw your best guess in a diagram.
3. Can you test your hypothesis using data? If so, what data do you need to collect and what tests could you use?
4. What rival explanations are their?
5. How could you use data to test whether your best guess or the rival explanations are better? Write this as an **equation** if possible.
6. What other factors may influence the relationship you observe?

Questionnaire from: modified from Cheryl Schonhardt-Bailey

Assignment 4

Due: Friday 30 November

Research Design

With your partner plan your research by answering the following questions:

1. What difference or anomaly do you want to explain?
2. What is your best guess explanation? Draw your best guess in a diagram.
3. Can you test your hypothesis using data? If so, what data do you need to collect and what tests could you use?
4. What rival explanations are their?
5. How could you use data to test whether your best guess or the rival explanations are better? Write this as an **equation** if possible.
6. What other factors may influence the relationship you observe?

Questionnaire from: modified from Cheryl Schonhardt-Bailey

Quick Quiz 1

Interpret the following correlations (R):

- $R = 0.91$
- $R = 0.02$
- $R = -0.3$

Quick Quiz 1

Interpret the following correlations (R):

- $R = 0.91$
- $R = 0.02$
- $R = -0.3$

Quick Quiz 1

Interpret the following correlations (R):

- $R = 0.91$
- $R = 0.02$
- $R = -0.3$

Quick Quiz 2

What is a residual?

Discuss at least **two** things that residuals are used for in simple linear regression?

Quick Quiz 3

What assumptions does the linear regression model make?

Quick Quiz 4

Create a hypothesis test for a linear regression coefficient ($\hat{\beta}$)

Quick Quiz 5

How do you interpret a linear regression coefficient for a dummy variable?

Intro to Multiple Linear Regression

Last class we learned how to use the tools of simple linear regression to examine the **bivariate** relationship between a dependent variable and **one independent** variable.

What if we want to examine **multivariate relationships**, i.e. the relationship between a dependent variable and **multiple** independent variables at the same time?

In these cases we can use **multiple linear regression**.

Why?

Why would we want to examine multiple independent variables at the same time?

Minimal criteria for making a causal argument

To make a **probabilistic causal argument**, i.e. "X caused Y" we need to meet *at least* three criteria:

- X is **statistically associated** with Y,
- X happens before Y (i.e. **time order**),
- all **alternative explanations** for the association are ruled out.

Minimal criteria for making a causal argument

To make a **probabilistic causal argument**, i.e. "X caused Y" we need to meet *at least* three criteria:

- X is **statistically associated** with Y,
- X happens before Y (i.e. **time order**),
- all **alternative explanations** for the association are ruled out.

Minimal criteria for making a causal argument

To make a **probabilistic causal argument**, i.e. "X caused Y" we need to meet *at least* three criteria:

- X is **statistically associated** with Y,
- X happens before Y (i.e. **time order**),
- all **alternative explanations** for the association are ruled out.

Time Order & Causality

Linear regression per se can't help us establish time order.

We need to understand our data to do that.

We may also need to take measurements at multiple points in time and use more advanced statistical tools than the ones covered in this course.

Simple Linear Regression & Causality

Simple Linear Regression is a tool we can use to establish the statistical association between X and Y.

How can we determine how much if at all, the value of Y is actually explained by the value of X and not some other alternative factor(s)?

Spurious Example

A silly example:

You are interested in what causes expensive fire damage.

You observe that the most expensive fires have the most fire trucks on the scene.

Did having more fire trucks cause more damage?

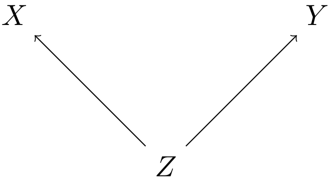
Spurious Example

Clearly, the *size of the fire* caused both the amount of fire damage and the number of fire trucks that responded.

There is a **spurious relationship** between number of fire trucks and fire damage.

Spurious Diagram

Figure: Spurious Relationship



Controlling for Z

If we were able to run an experiment where we **randomized** the units who are given ‘treatment’ X and those that are not (the ‘control’ group).

On average the units will have the same values of Z.

We can say that we are **controlling for Z.**

If, after randomization, the association between X and Y still exists, then we have found evidence to rule out alternative explanations.

Observational Data

However, in many social science situations we cannot run an experiment with randomized control and treatment groups.

For example, we cannot randomly assign people to live in dictatorships and democracies.

In these cases we need to use **statistical control** like **multiple linear regression.**

Note:

In many cases social scientists actually do conduct randomized experiments.

For example, the Obama campaign randomized the email messages it sent to people asking for donations.

Also, there are more advanced statistical techniques that can be combined with multiple linear regression to enhance statistical control. For example, matching.

Multi-causality

Also, in the social sciences something *rarely has one cause*.

Instead, phenomenon usually have multiple causes; each making a contribution to the probable value of an outcome.

Multiple linear regression is a statistical tool that we can use to help identify the **individual** contribution of some factor to an outcome, **controlling for** other factors.

The Multiple Linear Regression Model

An estiamted multiple linear regression model for predictors $x_1 \ldots x_p$:

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Today's Data

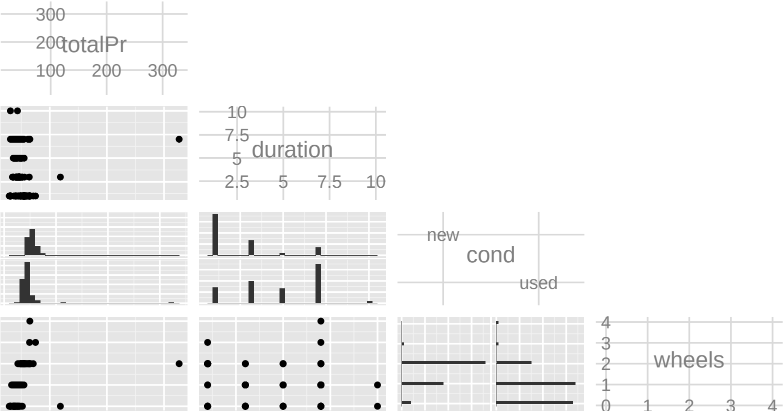
```
# Load library
library(openintro)

# Load data
data(marioKart)

# Show Variables
names(marioKart)

##  [1] "ID"         "duration"   "nBids"
##  [4] "cond"       "startPr"    "shipPr"
##  [7] "totalPr"    "shipSp"     "sellerRate"
## [10] "stockPhoto" "wheels"     "title"
```

Example Multiple Linear Regression Model

Imagine we are interested in what explains the EBay selling price of the game Mario Kart (`totalPr`)?

We want to see if the duration of the auction in days (`duration`), whether the game was in used condition (`condused`), and the number of wheels included in the auction (`wheels`) impacted the selling price.

Scatter



300
200
totalPr
100
200
300
10
7.5
duration
5
2.5
5
7.5
10
new
cond
used
4
3
2
wheels
1
0
1
2
3
4

Sum of Squared Residuals

Estimating the coefficients and making inferences about them is similar to simple linear regression.

For example, to find the sum of the squared residals:

$$SSR = \sum_{i=1}^{n} = e_i^2 = \sum_{i=1}^{n} = (y_i + \hat{y}_i)^2$$

Estimation

Estimating the βs by hand in multiple linear regression is very difficult, but it is relatively easy if you let R do the hard work with the `lm` command:

```
# Estimated multivariate linear regression model
M1 <- lm(totalPr ~ duration + cond +
           wheels, data = marioKart)
```

Showing the Coefficient Estimates

```
# Show coefficient point estimates
M1

##
## Call:
## lm(formula = totalPr ~ duration + cond + wheels, data =
##
## Coefficients:
## (Intercept)     duration     condused
##      35.735        0.680       -0.695
##      wheels
##      10.455
```

What is the estiatmed linear regression equation?

Linear Regression Equation

$$\widehat{\texttt{totalPr}} = 35.74 + 0.68\texttt{duration} + -\texttt{0.695}\texttt{condused} + \texttt{10.46}\texttt{wheels}$$

Question Linear Regression Equation

$$\begin{aligned}\widehat{\texttt{totalPr}} = 35.74 + 0.68(\texttt{duration}) \\ + - 0.695(\texttt{condused}) + \texttt{10.46}(\texttt{wheels}))\end{aligned}$$

What do you estimated will be the total selling price for a Mario Kart auction that was 5 days long, was in new condition, and included 2 wheels?

Linear Regression Equation

$$60.06 = 35.74 + 0.68(5) + -0.695(0) + 10.46(2)$$

References I

Crawley, Michael J. 2005. Statistics: An Introduction Using R. Chichester: John Wiley Sons. Ltd.

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2011. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. American Journal of Political Science 44(2): 347361.