Intro to Social Science Data Analysis

Week 12 Lecture: Multivariate Linear Regression & Presenting Regression Results

Christopher Gandrud

November 12, 2012

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

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

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

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

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

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 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?

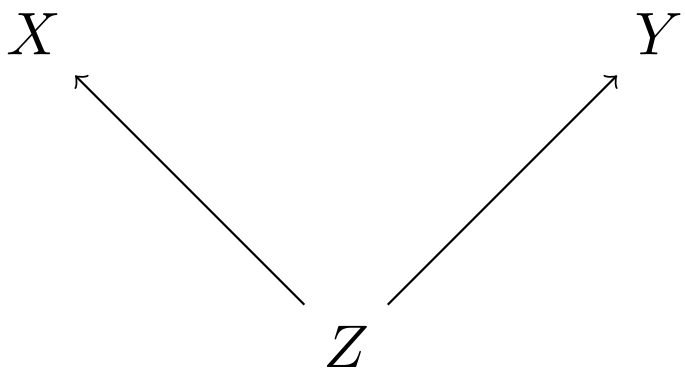
Spurious Example

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 estimated 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

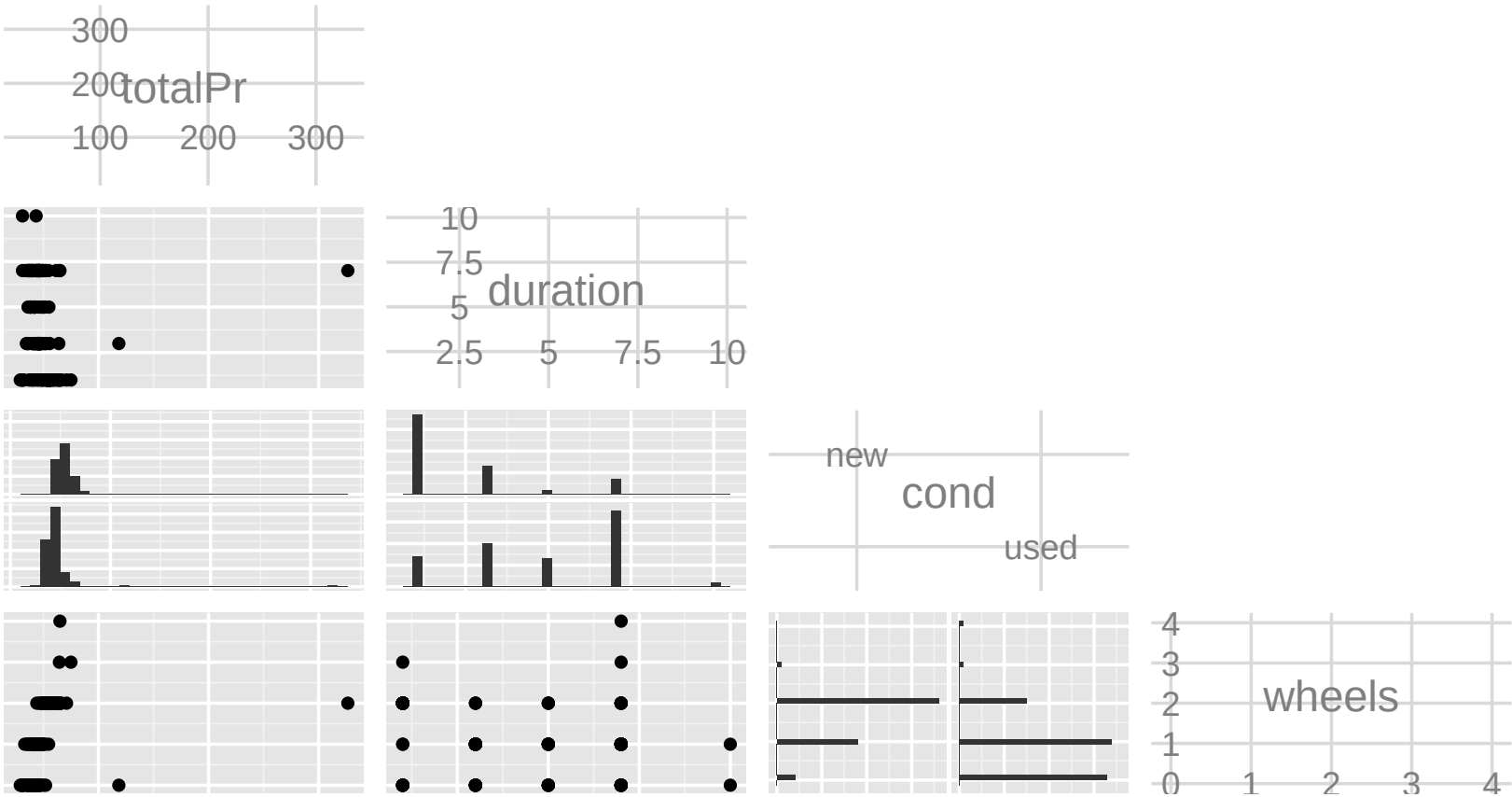
```
# Load library
library(openintro)

# 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


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

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

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 estimated linear regression equation?

Linear Regression Equation

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

Question Linear Regression Equation

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

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 Estimated Y

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

Single Variable Interpretation

We can interpret the effect of `duration` like this:

> *Controlling for the condition of the game and the number of wheels included, each day a Mario Kart EBay auction continues I expect that the total selling price will increase by 0.68.*

For the dummy variable `condused` we have the following interpretation:

> *Controlling for the duration of the auction and the number of wheels sold, I expect used Mario Kart games to sell for 0.695 less than new games in EBay auctions.*

Single Variable Interpretation

We can interpret the effect of `duration` like this:

> *Controlling for the condition of the game and the number of wheels included, each day a Mario Kart EBay auction continues I expect that the total selling price will increase by 0.68.*

For the dummy variable `condused` we have the following interpretation:

> *Controlling for the duration of the auction and the number of wheels sold, I expect used Mario Kart games to sell for 0.695 less than new games in EBay auctions.*

Remember the Units

Remember that the coefficients are in terms of the variables' units.

A large coefficient does not necessarily mean a big effect.

Multinomial Variables in Multiple Linear Regression

What if an independent variable has more than two category?

For example, the shipping method variable `shipSp` has the following categories:

```
summary(marioKart$shipSp)

## firstClass      media      other     parcel
##         22         14          3         16
##   priority   standard    ups3Day  upsGround
##         23         33          1         31
```

Linear Model with Multinomial Independent Variables

```
# Estimate model
M2 <- lm(totalPr ~ shipSp, data = marioKart)
```

The Linear Regression Equation with a Multinomial Independent Variable

$$\begin{aligned}\widehat{totalPr} = 42.35 &+ 8.5(\texttt{Media}) + 4.46(\texttt{Other}) + \\ &26.84(\texttt{Parcel}) + 1.45(\texttt{Priority}) + \\ &4.03(\texttt{Standard}) + 4.47(\texttt{ups3Day}) + \\ &10.27(\texttt{upsGround})\end{aligned}$$

Note that `firstclass` is missing.

It is the **reference category**.

Interpreting Multinomial Variable Regression Coefficients

The coefficients compare the predicted effect of the category to the reference category.

For example, the coefficient for `Parcel` post is 26.84. This means that:

> *We estimate that Mario Kart games that are shipped by parcel post will sell for 26.84 more than those sold with first class shipping.*

Question, in this example, what is the predicted selling price of a game sold with first class shipping?

Warning

There are at least two issues you should look out for with multinomial variables in linear regression:

- Is the comparison with the reference category substantively interesting? If not **change the reference category** with the `relevel` command.
- There may be too many categories with too few observations per category to estimate meaningful results. If this is the case, **combine some of the categories in a substantively meaningful way**.

We follow the same steps as before to test the null hypothesis that:

$$\beta_p = 0, \text{ when the other variables are included}$$

With the alternative hypothesis that:

$$\beta_p \neq 0, \text{ when the other variables are included}$$

First:

calculated the test statistic ($t\ test$).

Second:

find the p-value or confidence interval for the significance level we are interested in.

P-values & Summary Table

Luckily, R will quickly do this for us with the `summary` command:

```
summary(M1)
```

```
##
## Call:
## lm(formula = totalPr ~ duration + cond + wheels, data = marioKart)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17.25  -6.26  -2.73   0.51 265.80
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.735      5.889    6.07  1.2e-08
## duration       0.680      0.912    0.75  0.45724
## condused      -0.695      5.036   -0.14  0.89051
## wheels        10.455      2.697    3.88  0.00016
##
## (Intercept) ***
## duration
## condused
## wheels      ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.5 on 139 degrees of freedom
## Multiple R-squared: 0.113, Adjusted R-squared: 0.0935
## F-statistic: 5.88 on 3 and 139 DF,  p-value: 0.000825
```

Table: Linear Regression for Mario Kart Total Selling Price

	Model 1
(Intercept)	35.73 ***
	(5.89)
duration	0.68
	(0.91)
condused	-0.69
	(5.04)
wheels	10.45 ***
	(2.70)
N	143
R^2	0.11
adj. R^2	0.09
Resid. sd	24.46

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Confidence Intervals for Multiple Linear Regression Coefficients

Estimating confidence intervals for multiple linear regression coefficients is similar to what we have done before:

$$\beta_i \pm t^*_{df} SE_{\beta_i}$$

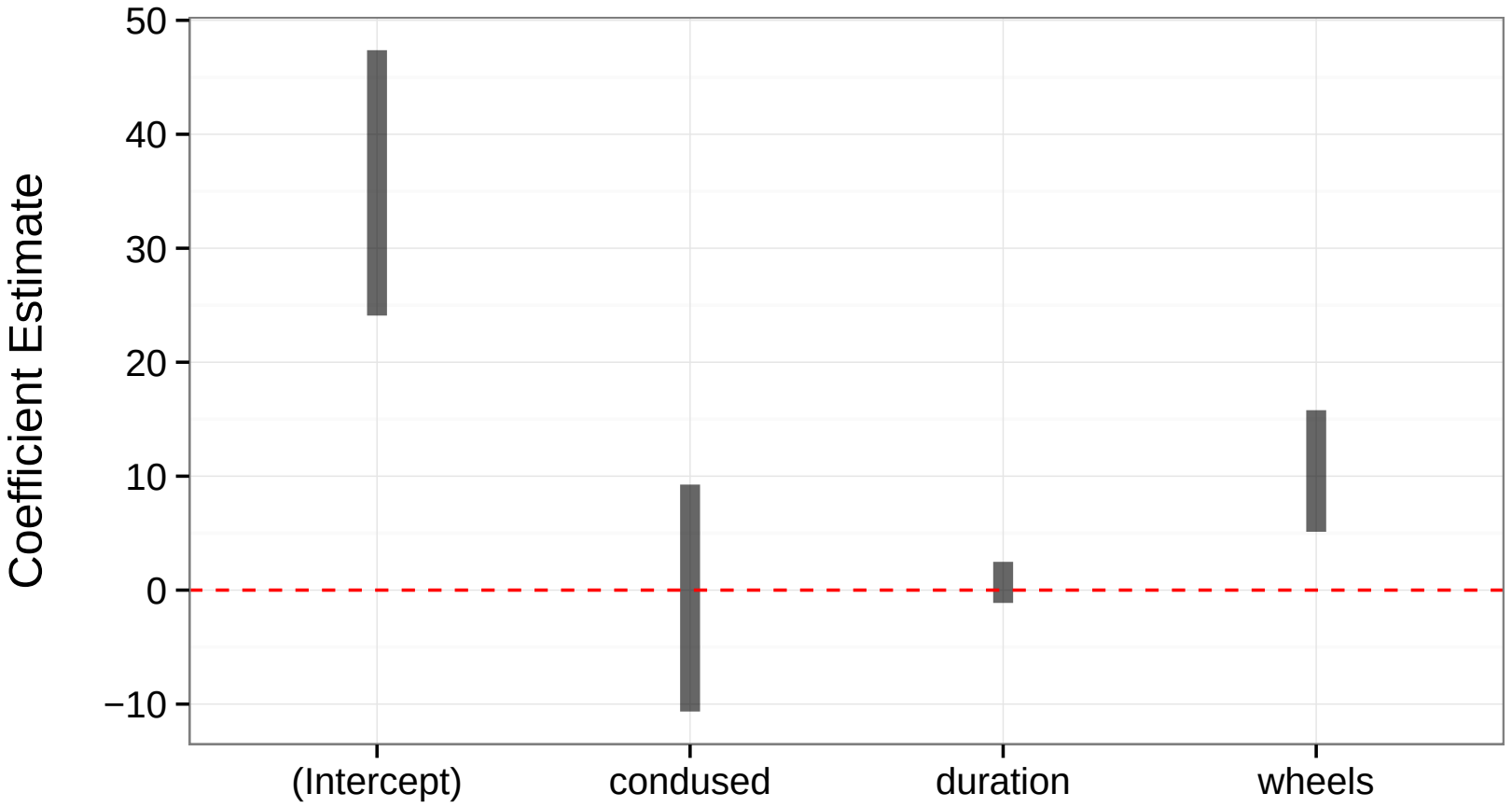
where t^*_{df} is the $t\ test$ statistic with $df = n - p - 1$.

Confidence Intervals in R for Objects of Class `lm`

To calculate the confidence intervals for regression coefficients in R use the `confint` command.

```
# Find confidence intervals
CI <- confint(M1)
CI

##                 2.5 % 97.5 %
## (Intercept)  24.091 47.379
## duration     -1.124  2.484
## condused    -10.652  9.263
## wheels         5.122 15.788
```

Graphically



Multiple Linear Regression Model Assumptions

- Nearly normally distributed residuals,
- Nearly constant residual variability,
- Residuals are independent,
- Each variable is linearly associated with the outcome

Multiple Linear Regression Model Assumptions

- Nearly normally distributed residuals,
- Nearly constant residual variability,
- Residuals are independent,
- Each variable is linearly associated with the outcome

Multiple Linear Regression Model Assumptions

- Nearly normally distributed residuals,
- Nearly constant residual variability,
- Residuals are independent,
- Each variable is linearly associated with the outcome

Multiple Linear Regression Model Assumptions

- Nearly normally distributed residuals,
- Nearly constant residual variability,
- Residuals are independent,
- Each variable is linearly associated with the outcome

Model Selection with the Adjusted R Squared

How do we decide which variables to include in our multiple linear regression model?

Omitted Variable Bias

We generally want to include all of the variables that have an important effect on the outcome.

Not including them creates **omitted variable bias**: i.e. our results are biased because we have omitted important variables.

However...

We don't want to just include every variable we can think of into one model.

- **Occam's Razor: aim for parsimony (the simplest model possible).**
- The more variables you add, the **fewer degrees of freedom** you will have to work with.
- You may include **highly correlated variables**, which can produce very biased estimates.

However...

We don't want to just include every variable we can think of into one model.

- **Occam's Razor: aim for parsimony (the simplest model possible).**
- The more variables you add, the **fewer degrees of freedom** you will have to work with.
- You may include **highly correlated variables**, which can produce very biased estimates.

However...

We don't want to just include every variable we can think of into one model.

- **Occam's Razor: aim for parsimony (the simplest model possible).**
- The more variables you add, the **fewer degrees of freedom** you will have to work with.
- You may include **highly correlated variables**, which can produce very biased estimates.

The Adjusted R^2

We can use the **adjusted** R^2 (R^2_{adj})to determine if the addition of a variable to a linear regression model **reduces the errors** we make when predicting the outcome.

$$R^2_{adj} = \frac{Var(e_i)/n - p - 1}{Var(y_i)/(n-1)}$$

It is similar to the non-adjusted R^2 we learned about last week, but **only gets bigger if the addition of a variable reduces the errors** we make predicting the outcome.

Table: Linear Regressions for Mario Kart Total Selling Price

	Model 1	Model 2	Model 3
(Intercept)	38.41 ***	35.38 ***	35.73 ***
	(3.43)	(5.27)	(5.89)
wheels	10.01 ***	10.58 ***	10.45 ***
	(2.41)	(2.53)	(2.70)
duration		0.63	0.68
		(0.83)	(0.91)
condused			-0.69
			(5.04)
N	143	143	143
R^2	0.11	0.11	0.11
adj. R^2	0.10	0.10	0.09
Resid. sd	24.34	24.37	24.46

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Model Section as Art

Model selection is something of an art, but it is useful to follow these **rules of thumb**:

- Drop highly insignificant variables, especially if they don't add to the adjusted R^2
- **Show a number of step-wise models** that could convince the reader that you have found the most **parsimonious** model.
- Include results from **theoretically important** variables,
- **Avoid including highly correlated** variables in the same model,
- Don't show estimates from models that **violate linear regression assumptions**,

Model Section as Art

Model selection is something of an art, but it is useful to follow these **rules of thumb**:

- Drop highly insignificant variables, especially if they don't add to the adjusted R^2
- **Show a number of step-wise models** that could convince the reader that you have found the most **parsimonious** model.
- Include results from **theoretically important** variables,
- **Avoid including highly correlated** variables in the same model,
- Don't show estimates from models that **violate linear regression assumptions**,

Model Section as Art

Model selection is something of an art, but it is useful to follow these **rules of thumb**:

- Drop highly insignificant variables, especially if they don't add to the adjusted R^2
- **Show a number of step-wise models** that could convince the reader that you have found the most **parsimonious** model.
- Include results from **theoretically important** variables,
- **Avoid including highly correlated** variables in the same model,
- Don't show estimates from models that **violate linear regression assumptions**,

Model Section as Art

Model selection is something of an art, but it is useful to follow these **rules of thumb**:

- Drop highly insignificant variables, especially if they don't add to the adjusted R^2
- **Show a number of step-wise models** that could convince the reader that you have found the most **parsimonious** model.
- Include results from **theoretically important** variables,
- **Avoid including highly correlated** variables in the same model,
- Don't show estimates from models that **violate linear regression assumptions**,

Model Section as Art

Model selection is something of an art, but it is useful to follow these **rules of thumb**:

- Drop highly insignificant variables, especially if they don't add to the adjusted R^2
- **Show a number of step-wise models** that could convince the reader that you have found the most **parsimonious** model.
- Include results from **theoretically important** variables,
- **Avoid including highly correlated** variables in the same model,
- Don't show estimates from models that **violate linear regression assumptions**,

Though popular, tables like these are **not an effective way to communicate your findings**.

	Model 1	Model 2	Model 3
(Intercept)	38.41 ***	35.38 ***	35.73 ***
	(3.43)	(5.27)	(5.89)
wheels	10.01 ***	10.58 ***	10.45 ***
	(2.41)	(2.53)	(2.70)
duration		0.63	0.68
		(0.83)	(0.91)
condused			-0.69
			(5.04)
N	143	143	143
R^2	0.11	0.11	0.11
adj. R^2	0.10	0.10	0.09
Resid. sd	24.34	24.37	24.46

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Simulating Expected Values

Instead, King et al (2000) argue that we should find a way to **present results**, including our **estimation uncertainty** in relatively **simple language**.

They suggest that we simulated probable outcomes from our estimated models.

They (and others) created the `Zelig` package to help us simulate and graph expected outcomes.

Zelig Simulation Steps:

1. Estimate model,
2. Set fitted values,
3. Simulate expected outcomes,
4. Graph the simulated values.

Zelig Simulation Steps:

1. Estimate model,
2. Set fitted values,
3. Simulate expected outcomes,
4. Graph the simulated values.

Zelig Simulation Steps:

1. Estimate model,
2. Set fitted values,
3. Simulate expected outcomes,
4. Graph the simulated values.

Zelig Simulation Steps:

1. Estimate model,
2. Set fitted values,
3. Simulate expected outcomes,
4. Graph the simulated values.

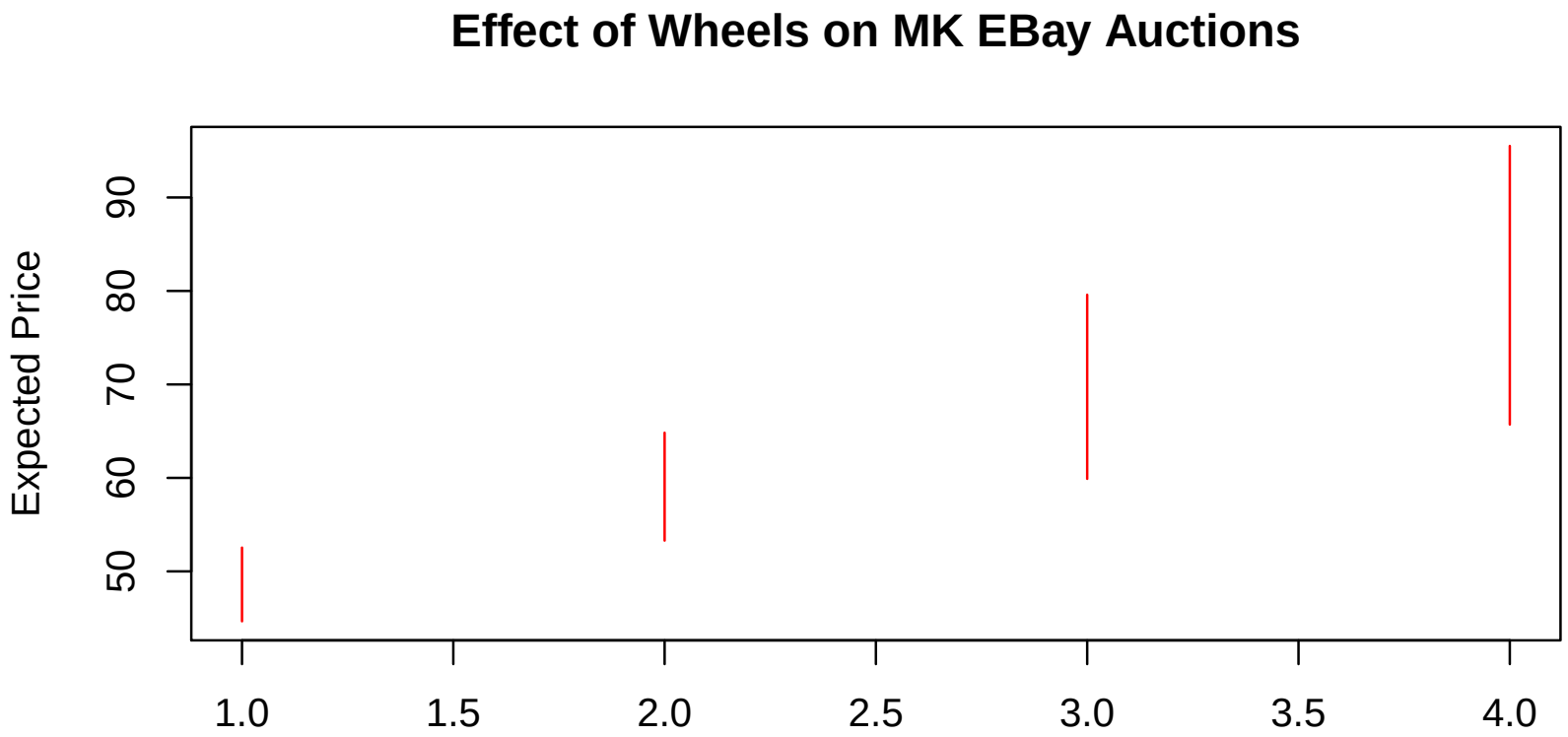
Zelig Example

Simulated expected Total Mario Kart auction price with various numbers of included steering wheels.

```
# Load Zelig
library(Zelig)

# Estimate model of total auction price with
# wheels and duration as independent variables
ZOut <- zelig(totalPr ~ wheels + duration,
              data = marioKart, model = "normal",
              cite = FALSE)
```

```r
# Find valid range of wheels values
range(marioKart$wheels)

## [1] 0 4

# Set fitted values for wheels
# Note: duration set at its mean
WValues <- 1:4
XOut <- setx(ZOut, wheels = WValues)

# Simulate expected prices
ZSim <- sim(ZOut, x = XOut)
```

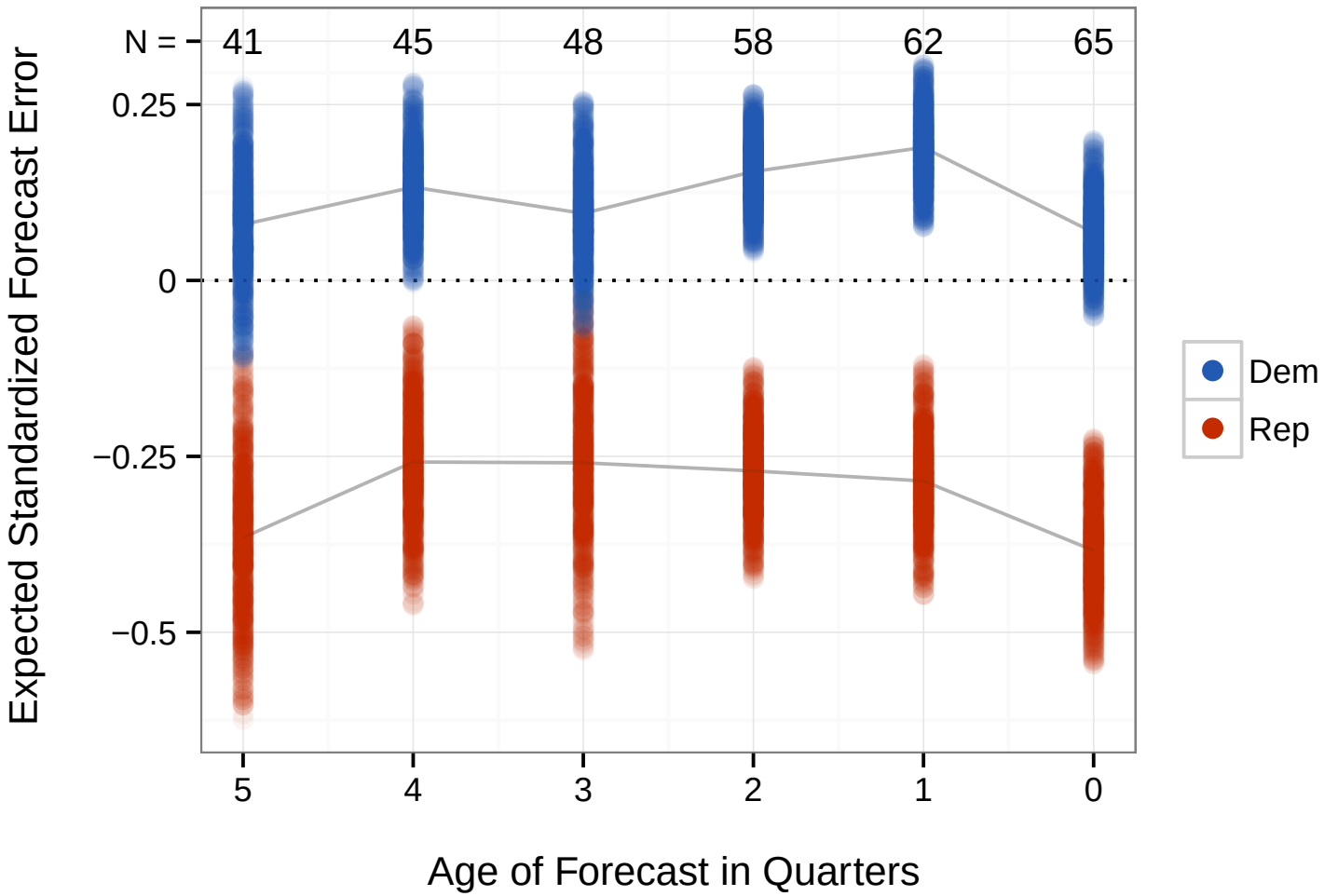
```r
# Plot middle 95% of the simulations
plot(ZSim, ylab = "Expected Price",
     main = "Effect of Wheels on MK EBay Auctions")
```



Question

Why is the predicted range of prices wider when there are 4 steering wheels?

With (a lot) more work simulation graphs can look better



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.