Visualizing Relationships

Statistics 4868/6610 Data Visualization

Prof. Eric A. Suess

2/24/2016

Introduction

Today we will work with Regression Lines and Simpson's Paradox.

Confounding variables

Recall Simpson's Paradox

Example of Simpson's Paradox

```
### synthetic data

# Consider book price (y) by number of pages (x)

z = c("hardcover","hardcover",
      "hardcover","hardcover",
      "paperback", "paperback","paperback",
      "paperback")

x1 = c( 150, 225, 342, 185)
y1 = c( 27.43, 48.76, 50.25, 32.01 )

x2 = c( 475, 834, 1020, 790)
y2 = c( 10.00, 15.73, 20.00, 17.89 )

x = c(x1, x2)
y = c(y1, y2)
```

Example of Simpson's Paradox

Summary: Simpson's Paradox is the changing of the direction of a relationship with the introduction of another variable.

The relationship between Book Price and Number of Pages in a book changes with the introduction of the variable Type of Book (Hardcover, Paperback).

See the R Markdown document SimpsonsParadox available on RPubs.com/esuess.

Linear Regression and beyond

If you are going to work with Scatterplots to visualize relationships between quantitative variables, with qualitative variables, you should be familiar with the assumption of linear regression and linear regression models that include **qualitative variables** or **dummy variables**.

Linear Regression

$y_i = \alpha + \beta x_i + \epsilon_i$ main assumption is $\epsilon_i \sim N(0, \sigma^2)$

Assumptions:

- linear
- ϵ_i has $E[\epsilon_i] = 0$
- at least one x_i is different
- x_i uncorrelated with ϵ_i
- ϵ_i has $V(\epsilon_i) = \sigma^2$
- ϵ_i is indepdent of ϵ_j
- $\epsilon_i \sim N(0, \sigma^2)$ so $y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$

Linear Regression

Least Squares

$H_0 : \beta = 0$

R^2

$TSS = RSS + ESS$

$R^2 = \frac{RSS}{TSS}$

ANOVA

F-test

p-values

Qualitative (or Dummy) Indepdent Variables

Different Intercepts parallel lines

D = 1 or 0

Two categories, so 1 Dummy Variable included in the model.

$y_i = (\alpha_0 + \alpha_1 D_i) + \beta x_i + \epsilon_i$

D = 0 gives $\hat{y}_i = \hat{\alpha}_0 + \hat{\beta} x_i$

D = 1 gives $\hat{y}_i = (\hat{\alpha}_0 + \hat{\alpha}_1) + \hat{\beta} x_i$

Dummy Variable Trap

Suppose there are 3 categories, if we use 3 Dummy Variables in the model, then we have fallen into the **dummy variable trap**.

So when there are 3 categories we use 2 Dummy Variables and so on.

Qualitative (or Dummy) Indepdent Variables

Different Slopes ANCOVA

$y_i = \alpha + (\beta_0 + \beta_1 D_i)x_i + \epsilon_i$

D = 0 gives $\hat{y}_i = \hat{\alpha} + \hat{\beta}_0 x_i$

D = 1 gives $\hat{y}_i = \hat{\alpha} + (\hat{\beta}_0 + \hat{\beta}_1)x_i$

Qualitative (or Dummy) Indepdent Variables

Different Intercepts and Different Slopes ANCOVA

$y_i = (\alpha_0 + \alpha_1 D_i) + (\beta_0 + \beta_1 D_i)x_i + \epsilon_i$

D = 0 gives $\hat{y}_i = \hat{\alpha}_0 + \hat{\beta}_0 x_i$

D = 1 gives $\hat{y}_i = (\hat{\alpha}_0 + \hat{\alpha}_1) + (\hat{\beta}_0 + \hat{\beta}_1)x_i$

Qualitative (or Dummy) Indepdent Variables - Minitab

For the Book Price data, try to fit the last model in Minitab.

Stat > Regression > Regression > Fit Regression Model...

Add the *Categorical predictor:*

Under *Model...*

Click the **Add** next to *Cross predictors and terms in the model*

Qualitative (or Dummy) Indepdent Variables - tableau

For the Book Price data, try to fit the last model in tableau.

See

Analysis > Trend Lines > Describe Trend Model...

ggplot2

Let's make some plots using ggplot2. The basic plotting with ggplot2 is done using the *qplot* function, quick plot.

The reference for this is Hadley Wickham's book, ggplot2, Elegant Graphics for Data Analysis, Use R!, Springer 2009.

To learn more about Hadley's efforts with R, see the blog post

The Hitchhikers Guide to the Hadleyverse

An excellent reference for learning ggplot2 is the R Graphics Cookbook and Graphs.

ggplot2

The key ideas ....

data

mappings

geoms

stats

scales

coordinates

faceting

ggplot2

See the ggplot2_examples.R Handout for some examples of using ggplot2 to make Histograms, Density Plots, Scatterplots, Scatterplots with Regression lines.

Distributions

The author of our book discusses distributions and how to visualize them.

Some key points the author discusses ...

The horizontal axis is **not** time.

This needs to be pointed out to the reader of your graph.

Comparison

The author of our book discusses **comparison**.

This is mainly in time.