Intermediate statistics: introduction

Frank Edwards

1/25/2019

School of Criminal Justice, Rutgers - Newark

Basics

Contact: `frank.edwards@rutgers.edu`

Office hours: email for appointments

Course webpage and syllabus:

`https://f-edwards.github.io/intermediate_stats/`

Slack: `https://ru-intermed-stats.slack.com/messages`

Before we begin

Remember: All models are wrong, some are useful.

Linear regression is a hammer. This course provides a variety of other tools to add to your toolbox. None of them are direct reperesentations of the real phenomena we investigate in the social sciences.

However, they can be incredibly useful ways to abstractly represent complex processes.

What we will cover

- How to program in R
- How to explore, visualize, and model diverse kinds of data
- How to design and write quantitative social science

Quick assessment of where we’re at with programming

1. Explain what this code does and expected output

```r
k<-2
for(i in 1:10){
  k<-i*k
}
```

2. Explain what this code does and expected output

```r
a<-c(1, 2, 3)
b<-c(2, 3, 4)
a*b
```

3. Explain what this code does and expected output

```
whatsitdo<-function(x){
  a<-min(x)
  return(1/a)
}
z<-c(4,5,6)
whatsitdo(z)
```

4. Explain what this code does and expected output

```r
library(dplyr)
dat<-data.frame("var1" = c(1,2,3),
                "var2" = c(4, 5, 6))
dat%>%
  summarise(total = sum(var1 + var2))
```

5. Explain what z is and what m1 is

```r
y<-c(1,2,3,4,5)
x<-c(3,4,5,6,7)
z<-solve(t(x)%*%x)%*%t(x)%*%y
m1<-lm(y~x)
```

Discussion on how we could proceed with programming in the course

Self assessment:

Were these problems easy? Hard? Completely foreign? Which parts were most unfamiliar?

Discussion on how we could proceed with programming in the course

Self assessment:

Were these problems easy? Hard? Completely foreign? Which parts were most unfamiliar?

Question for the class:

Would it be helpful to cover basic programming concepts (i.e. functions, loops)?

Discussion on how we could proceed with programming in the course

Self assessment:

Were these problems easy? Hard? Completely foreign? Which parts were most unfamiliar?

Question for the class:

Would it be helpful to cover basic programming concepts (i.e. functions, loops)? Using the tidyverse packages?

Discussion on how we could proceed with programming in the course

Self assessment:

Were these problems easy? Hard? Completely foreign? Which parts were most unfamiliar?

Question for the class:

Would it be helpful to cover basic programming concepts (i.e. functions, loops)? Using the tidyverse packages? Using RMarkdown?

Discussion on how we could proceed with programming in the course

Self assessment:

Were these problems easy? Hard? Completely foreign? Which parts were most unfamiliar?

Question for the class:

Would it be helpful to cover basic programming concepts (i.e. functions, loops)? Using the tidyverse packages? Using RMarkdown?

Review the syllabus

https://f-edwards.github.io/intermediate_stats/

How I will run seminars

- Basic statistical theory

How I will run seminars

- Basic statistical theory
- Applied data analysis and modeling in R

Expectations

- Bring a laptop: we will be writing code in class

Expectations

- Bring a laptop: we will be writing code in class
- Make space for everyone: respect varying levels of comfort with statistics and programming

Expectations

- Bring a laptop: we will be writing code in class
- Make space for everyone: respect varying levels of comfort with statistics and programming
- Come prepared and complete assignments on time

My general approach to data analysis

1. Explore and visualize data

My general approach to data analysis

1. Explore and visualize data
2. Fit models

My general approach to data analysis

1. Explore and visualize data
2. Fit models
3. Assess model fit

My general approach to data analysis

1. Explore and visualize data
2. Fit models
3. Assess model fit
4. Interpret and describe results through simulation

The Generalized Linear Model

The linear model

We know we can model data as:

$$y = \beta_0 + \beta_1 x_1 \cdots \beta_n x_n + \varepsilon$$

The linear model

We know we can model data as:

$$y = \beta_0 + \beta_1 x_1 \cdots \beta_n x_n + \varepsilon$$

Or, more succinctly:

$$y = \mathsf{X}\beta + \varepsilon$$

The linear model

We know we can model data as:

$$y = \beta_0 + \beta_1 x_1 \cdots \beta_n x_n + \varepsilon$$

Or, more succinctly:

$$y = \mathsf{X}\beta + \varepsilon$$

Where the likelihood for the outcome conditional on the data takes the form:

$$Y|X \sim Normal(\mu, \sigma^2)$$

Generalzing the linear model

The linear model:

$$Y|X \sim Normal(\mu, \sigma^2)$$

Can be written as a more general formulation for a likelihood function f

$$Y|X \sim f(\mu, \sigma^2)$$

Now we can extend the (very) useful linear model to data with discrete outcomes

A linear predictor η:

$$\eta = \mathsf{X}\beta$$

A link function g

$$g\big(E(Y|X)\big) = \eta$$

A mean expectation $E\big(Y|X\big) = \mu$

$$\mu = g^{-1}(\eta)$$

From OLS to GLM

OLS:

$$Y|X \sim Normal(\mu, \sigma^2)$$

GLM:

$$Y|X \sim f(\mu, \sigma^2)$$

Diverse likelihood functions

- Binary data: linear probability and logistic models

Diverse likelihood functions

- Binary data: linear probability and logistic models
- Categorical data: Multinomial model

Diverse likelihood functions

- Binary data: linear probability and logistic models
- Categorical data: Multinomial model
- Count data: Poisson and negative binomial models

Diverse likelihood functions

- Binary data: linear probability and logistic models
- Categorical data: Multinomial model
- Count data: Poisson and negative binomial models
- Positive continuous data: Gamma model

Getting started: software

Required installations

All software we are using is free and open source.

Install R:

`https://cran.r-project.org/`

Install RStudio:

`https://www.rstudio.com/products/rstudio/download/`

Recommended software: Git and GitHub

Git and GitHub are powerful tools for backing up and sharing your research.

All course materials, source code, and most of my research are hosted on GitHub (`https://github.com/f-edwards`).

Install Git:

`https://git-scm.com/`

Set up a GitHub account:

`https://github.com/`

Using GitHub for social science:

`https://happygitwithr.com/`

Recommended software: LaTeX

LaTeX is a powerful typesetting tool that works well with RMarkdown. It makes very attractive academic papers and slides.

Install it here: *Install TexLive*:
`https://tug.org/texlive/acquire-netinstall.html`

Questions so far?

Break

Returning to the linear model

What do we know about the linear regression model?

$$y = \mathsf{X}\beta + \varepsilon$$

$$\varepsilon \sim Normal(0, \sigma^2)$$

Review

1. What forms can y take?
2. What assumptions does the linear regression model require?
3. What are some contexts where the linear regression model can be misleading?

Let's build some models to review

I'm showing you the code and R output using Markdown - you'll learn how to do this as we keep going.

```
data(USArrests)
```

From the help file (access help on anything in R with ?, e.g. ?USArrests, ?data, etc.):

Violent Crime Rates by US State

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Evaluate the structure of the data

```r
str(USArrests)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.
##  $ Assault : int  236 263 294 190 276 204 110 238 335 21
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8
```

R relies heavily on data frames

```r
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Call individual variables (columns) in a data frame with $, like \texttt{USArrests$Murder}

```r
summary(USArrests$Murder)
```

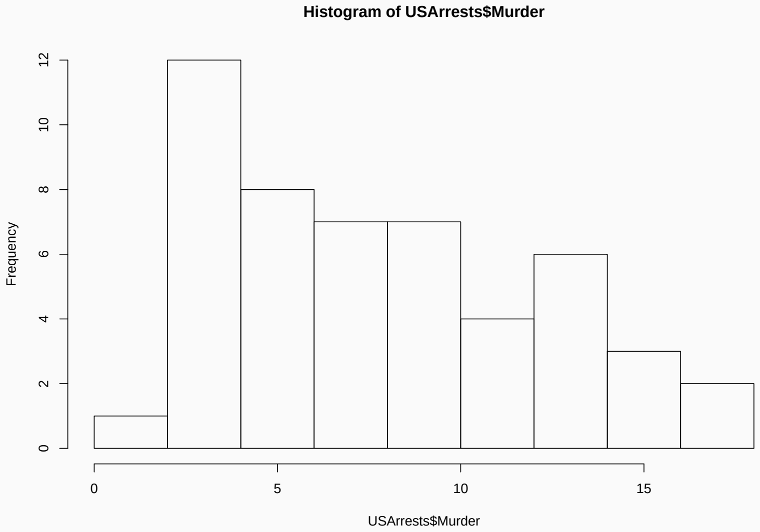
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

```r
sd(USArrests$Murder)
```

```
## [1] 4.35551
```

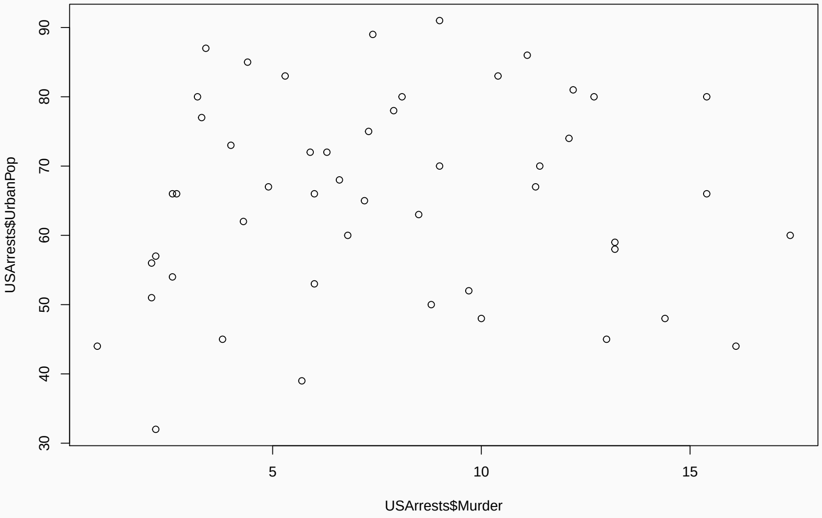
R has powerful tools for plotting data

```r
hist(USArrests$Murder)
```



R has powerful tools for plotting data

```
plot(USArrests$Murder, USArrests$UrbanPop)
```



Fitting a linear model

```
model_1<-lm(Murder ~
              UrbanPop,
            data = USArrests)
```

Display the model fit

```r
summary(model_1)
```

```
##
## Call:
## lm(formula = Murder ~ UrbanPop, data = USArrests)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.537 -3.736 -0.779  3.332  9.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.41594    2.90669   2.207   0.0321 *
## UrbanPop     0.02093    0.04333   0.483   0.6312
## ---
```

Visualize the model fit

```
library(ggplot2)
ggplot(USArrests,
       aes(x=UrbanPop, y=Murder))+
  geom_smooth(method = "lm",
              formula = y~x)
```