36-463/663: Multilevel & Hierarchical Models

Regression Basics

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Reading and HW

- Reading:
  - G&H Ch's 3-4 for today and next Tues
  - G&H Ch's 5-6 starting next Thur
    - I will not cover everything in the chapters
    - You will need to read & try some things on your own!
- HW02
  - Due next Tue Sept 13, on Dropbox.

Outline

- Interpretation of Coefficients
  - Interpretation of regression coefficient
  - Interpretation of intercept
  - Causal vs predictive interpretations
- Interpretation of the fitted model
- Multiple predictors, interactions
- Simple Diagnostics
- ***NOTE***: There is more in the R code online today
  - Check http://www.stat.cmu.edu/~brian/463-663

Interpretation of Coefficients

- The simple linear model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

- It can be fitted in R like this

```
fit.lm <- lm(y ~ x)
```

- And examined like this

```
summary(fit.lm)
plot(y~x)
abline(fit.lm)
```

Interpretation of Coefficients

- The basic linear model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

- We can interpret β_1 as representing the expected change in y for a 1-unit change in x
  - ***Predictive interpretation***: β_1 is the expected difference in y between two groups who only differ in x by one unit
  - ***Counterfactual interpretation***: If we could change x by one unit, β_1 is the change we would expect to see in y.

Interpretation of Coefficients

$$y = \beta_0 + \beta_1 x + \epsilon$$

- If x is a binary predictor (0 = "mom didn't finish high school"; 1 = "mom did finish high school") then
  - β_0 is the mean of group 0
  - $\beta_0 + \beta_1$ is the mean of group 1

  This simple interpretation is why people like working with binary predictors.
- If x is a continuous predictor (mom's iq) then
  - β_0 is the mean of y when mom's iq = 0 (??)
  - β_1 is the change in y per unit change in x.

Interpretation of Coefficients

- For the intercept to be meaningful, it can be helpful to standardize, or at least, center the data.

```
x.c <- x - mean(x)
lm.fit <- lm(y ~ x.c)
```

- …fits a model like this

$$y = \beta_0 + \beta_1(x - mean(x)) + \epsilon$$

  - The intercept is now the average test score of kids whose mothers have average IQ score.

Interpretation of the Fitted Model

- The result of

```
fit.lm <- lm(y ~ mom.iq)
```

might be a fitted function

$$f(x) = 26 + 0.6x$$

- Two interpretations:
  - f(60) = average y among kids whose mom's IQ's are 60
  - f(60) = prediction of kid's y for a mom whose IQ is 60.

The different interpretations lead to different measures of uncertainty…

Multiple Predictors, Interactions

- `lm(y ~ mom.hs + mom.iq)`

  fits the model

  $$y = \beta_0 + \beta_1 mom.hs + \beta_2 mom.iq + \epsilon$$

  - Among kids whose moms didn't go to high school (mom.hs = 0):

    $$y = \beta_0 + \beta_2 mom.iq + \epsilon$$

  - Among kids whose moms did go to high school (mom.hs = 1):

    $$y = (\beta_0 + \beta_1) + \beta_2 mom.iq + \epsilon$$

Multiple Predictors, Interactions

- `lm(y ~ mom.hs + mom.iq + mom.hs:mom.iq)`

  fits the model

  $$y = \beta_0 + \beta_1 mom.hs + \beta_2 mom.iq + \beta_3 mom.hs \cdot mom.iq + \epsilon$$

  - Among kids whose moms didn't go to high school (mom.hs = 0):

    $$y = \beta_0 + \beta_2 mom.iq + \epsilon$$

  - Among kids whose moms did go to high school (mom.hs = 1):

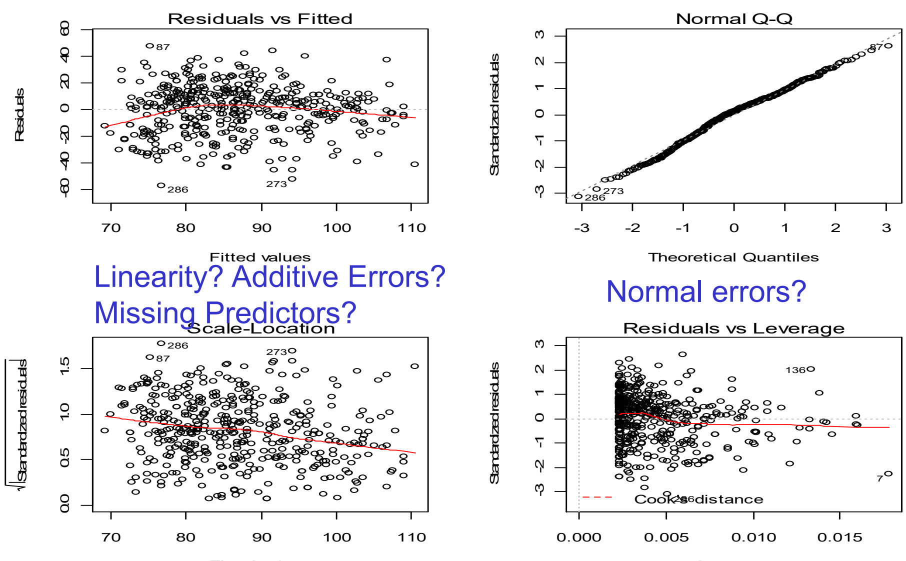
    $$y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) mom.iq + \epsilon$$

Simple Diagnostics

- After fitting the model, we want to check several things:
  - ***Validity***: does the general linear model setup match up with the scientific question? Do the variables you have bear on the answers you want?
  - ***Linearity and additive errors***: Does it make sense to add contributions of x1, x2, x1:x2, etc. to build up a model? Is the error additive rather than multiplicative?
  - ***Independent errors***: are the errors statistically independent, or does the error in one case depend somehow on the errors in other cases?
  - ***Equal-variance errors***: do the errors look like they all came from the same distribution?
  - ***Normal errors***: do the errors look like they came from a symmetric unimodal distribution without too many outliers?

Simple Diagnostics

```
fit.lm <- lm(y ~ x)
par(mfrow=c(2,2))
plot(fit.lm)
```



Linearity? Additive Errors?
Missing Predictors?

Normal errors?

Equal Variance of Errors?

Influential points?
Missing Predictors?

Summary

- Interpretation of Coefficients
  - Interpretation of regression coefficient
  - Interpretation of intercept
  - Causal vs predictive interpretations
- Interpretation of the fitted model
- Multiple predictors, interactions
- Simple Diagnostics
- Reading:
  - G&H Ch’s 3-4 today and next Tues
  - G&H Ch’s 5-6 starting next Thur
    - I will not cover everything in the chapters
    - You will need to read & try some things on your own!
- HW02
  - Due next Tue Sept 13, on Dropbox.