390R
Modeling notes
10/27/2020

This part uses some math and stats but we will work around that.
When you look at data you should choose what model to use.
In this course we will only use linear models
Use simple models to summarize data
For a linear model we want the best line to fit the data, which is defined as the sum of residuals

function	package	fits
lm()	stats	linear models
glm()	stats	generalized linear models
gam()	mgcv	generalized additive models
glmnet()	glmnet	penalized linear models
rlm()	MASS	robust linear models
rpart()	rpart	trees
randomForest()	randomForest	random forests
xgboost()	xgboost	gradient boosting machines

Use Wages data
Import it using read_excel()in readxl package

```
wages <- read_excel("/Users/Zhichao/Dropbox/courses/UMass/R\course/lectures/modeling/wages.xlsx",
na="NA")
```

First question: How does education affect income level

Functions

1. lm() for linear models lm(formula,data)
   Ex. lm(log(income)~education,data = wages)
   This creates a linear model of income vs education
   Intercept 8.5577 Education/slope = 0.1418

Formula:
Y = alpha + beta(x) + e
y~x
y=response, x = predictors
Alpha = intercept, beta =slope of linear model ,e = random error

2. Broom package
    a. Includes tidy() returns model coefficients, characterizes the uncertainty
    b. Includes glance() returns model diagnostics: how “good” is the model?
    c. Includes augment() returns predictions, residuals, and other raw values

Tidy function
mod_e%>%tidy() will return the estimates, standard error, test statistics, and p value.
Estimates are the same as the intercept and slope of linear regression
Glance function
mod_e%>%glance() returns r.squared, adj.r.squared, sigma, p value, test statistic, deviance,...
Augment function
mod_e%>%augment() gives residuals and predictions for each observation

Modelr package
Add_predictions function can predict response based on the value of the predictor
Ex. with wages
x<-tibble(education=12)
add_predictions(x,mod_e) will estimate regression coefficients with education = 12

With a linear regression we can have more than one predictors, the result will have the intercept and one slope for each predictor
Ex. lm(log(income)~education + height,data = wages) will output an intercept and 2 slopes for height and education

QUESTION: Model log(income) against education and height and sex. Can you interpret the coefficients?
lm(log(income)~education + height + sex,data = wages) only output sexmale, because the sex variable is categorical.

These are the only models we will be taught in this class, if you want to learn more about other models there are other courses one can take, or even just console ?glm to find out more.

How to visualize a model
Geom_smooth will vizuale a regression model and fit a smooth line on your figure
wages%>%ggplot(aes(height,log(income)))+geom_point(alpha=0.1)+geom_smooth(method=”lm”)

For this course we will have a final project of data exploratory analysis where you will find a data set, and use the techniques you've learned to write a report. Within the next two weeks I'll post instructions on the final exam, where to get data sets, and the rubric.