4 - Intro to R

Models and Testing

Load data

- `library(ggplot2)`
- Automatically loads tips data
- `head(tips)`

Basic Tests

- `binom.test()`
- `chisq.test()`
- `t.test() #talk about this later`

Clean the data

- Data is not usually ready to work with.
- Examine data first.
- Rename variables?

Fitting Models

- Model grammar, Formula
- Linear Models
- Diagnostics

lm

- `lm(formula, data, weight, subset, na.action)`
- Formula
  - `y ~ x1 + x2 + x3*x4`

Simple linear regression

- Plot beforehand?
- Formula: `y ~ x`
- Let's use:
  - `tips.reg <- lm(tip ~ bill, data = tips)`

Output

- Coefficients, fitted values, residuals
- Anova table
- Can get predictions for future values

Diagnostics

- Residual sums of squares (deviance)
- AIC (Akaike Information Criterion)
- Residual plots, normal quantile plots

Your Turn

- Do a simple linear regression using psize as your predictor.
- Does it provide a good fit?
- Check the diagnostics
- Compare this fit to the previous model.

Multiple Regression

- Pretty much the same but we're adding more continuous variables.
- Formula:
  - `y ~ x1 + x2 (additive)`
  - `y ~ x1 + x2 + x1:x2 (with interaction)`

Your Turn

- Should we keep the interaction term in the model?
- Based on the model you choose what is the predicted value for a bill of 40 for a party size 3.

Anova

- Want to test means for multiple categories
- Formula is same: `y ~ x`
- Only difference is that x is categorical.
- R fits a treatment effects model by default

Constraints

- To make the model full rank R uses a "set first to 0" identifiability constraint. (`contr.treatment`)
- SAS uses "set last to 0" (`contr.SAS`)
- Another option: "sum to 0" (`contr.sum`)
- use `getOption("contrasts")` or `options()$contrasts` to check which one(s) you're using

Example

Friday	Saturday	Sunday	Thursday
μ	μ + Sat	μ + Sun	μ + Thur

Ancova and other models

- Can have categorical and continuous variables in a model.
- When adding terms
  - `a:b` is the interaction between a and b
  - `a*b` gives linear terms and interaction
  - `a*b` is the same as `a + b + a:b`

Your Turn

- Calculate basic summary stats for the variables.
- Create a few plots to show the relationships between the variables and tip.
- Are there any unusual points?
- Explore the data. Find a good model to predict tips.

Your Turn

- Create a new variable called tiprate
  - `tips$tiprate <- tips$tip/tips$bill`
  - Explore the relationship between the variables and tiprate.
  - Find a good model to explain tiprate.