Introduction to R for Epidemiologists

Jenna Krall, PhD

Thursday, Feburary 26, 2015

Outline

1. Single linear regression
2. Multiple linear regression
3. One-way ANOVA

Simple linear regression

Definitions:

- Regression is used to estimate associations between a predictor (or set of predictors) and an outcome
- Linear regression for a continuous outcome

When would we want to apply linear regression?

- Is age associated with blood pressure?
- Is air pollution associated with birthweight?
- Is population associated with murder rate?

Simple linear regression

Simple linear regression:

$E(y|x) = \beta_0 + \beta_1 x$

where

- $E(y|x)$ is the expectation or mean of y
- β_0 is the intercept, $E(y|x)$ when $x = 0$
- β_1 is the slope, the change in $E(y|x)$ for a one unit change in x

Assumptions

1. Linearity
2. Independence
3. Normality
4. Equal variances

Simple linear regression

Violent crime rates in Georgia

- Uniform Crime Reporting
- FBI collected statistics from law enforcement agencies across the US
- `http://www.ucrdatatool.gov/Search/Crime/State/OneYearofData.cfm`

Multiple linear regression

Allows us to estimate the association between a predictor and outcome while controlling for other variables

- Is age associated with blood pressure, after controlling for family history?
- Is air pollution associated with birthweight, adjusting for temperature?
- Is log population associated with log murder rate, controlling for robbery rate?

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_P x_P$$

where

- $E(y|x)$ is the expectation (mean) of y
- β_0 is the intercept or the $E(y|x)$ when $x_1 = x_2 = ... = x_P = 0$
- β_1 is the slope, or the change in $E(y|x)$ for a one unit change in x_1 when all other covariates $(x_2, ..., x_P)$ are held constant.

Multiple linear regression

- After controlling for robbery rate, we found that the association between log population and log murder rate was -0.61.
- For two agencies that differ by 1 unit in log population that have the same robbery rate, we expect a difference in log murder rates of -0.61.

Robbery rate is statistically significantly associated with murder rate, controlling for population

- For an increase of one robbery per 100,000 people, log murder rate increased by 0.002, holding population constant.
- Since the p-value < 0.05, we reject the null hypothesis that robbery rate is not associated with murder rate, controlling for population

Multiple linear regression

We can also use linear regression with categorial covariates

- Categorical covariates are factor variables that have multiple levels
- Can have two levels (income $<$ 80,000/year, income $\geq$ 80,000/year)
- Can have multiple levels (race: white, black or african american, asian, etc.)
- Can be nominal or ordinal
  - nominal, no levels (e.g. race)
  - ordinal, ordered (e.g. education: $<$ high school, high school, some college, college)

Multiple linear regression

Linear regression with categorical covariates

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_P x_P$$

where

- x_1, x_2, ..., x_P are indicator variables for whether an observation belongs in that group
  - Indicator variable is 1 if condition is met, 0 otherwise
  - When $x_1 = x_2 = ... = x_P = 0$, indicates reference group
  - education: reference group is less than high school, $x_1 = 1$ if high school education, $x_2 = 1$ if some college, etc.
- β_0 is the intercept, or the $E(y|x)$ when $x_1 = x_2 = ... = x_P = 0$. In other words, the mean y for the reference group
- β_1 is the slope for x_1, or the difference in $E(y|x)$ for comparing group 1 to the reference group
- $\beta_0 + \beta_1$ is the mean y for group 1

Multiple linear regression

We also might want to test to see if any cities differ in murder rates across categories of population, where population is categories as small, medium, or large agencies. In other words, we want to test:

- Null hypothesis H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu$
- Null hypothesis H_0 : $\beta_0 = \beta_0 + \beta_1 = \beta_0 + \beta_2$
- Null hypothesis H_0 : $\beta_1 = \beta_2 = 0$
- Alternative hypothesis H_1 : at least one group not equal

Analysis of Variance (ANOVA)

Recall that ANOVA tests to see whether a continuous variable differs across levels of another variable

- Does number of emails that Emory students receive differ by academic program?
- Is the average air pollution the same across 20 US cities?
- Does the murder rate differ across population levels in Georgia agencies?

This is the same as applying multiple linear regression using one categorical covariate!

Analysis of Variance (ANOVA)

ANOVA tests the "global" null hypothesis for P different groups:

- Null hypothesis H_0 : $\mu_1 = \mu_2 = \mu_3 = ... = \mu_P = \mu$
- Alternative hypothesis H_1 : at least one group not equal

Assumptions:

- Normality
- Independence
- Equal variance: Variance is the same in each of the groups