BIOS 312: MODERN REGRESSION ANALYSIS

James C (Chris) Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/CourseBios312

Updated January 13, 2010

Contents

Chapter 3

Simple Linear Regression

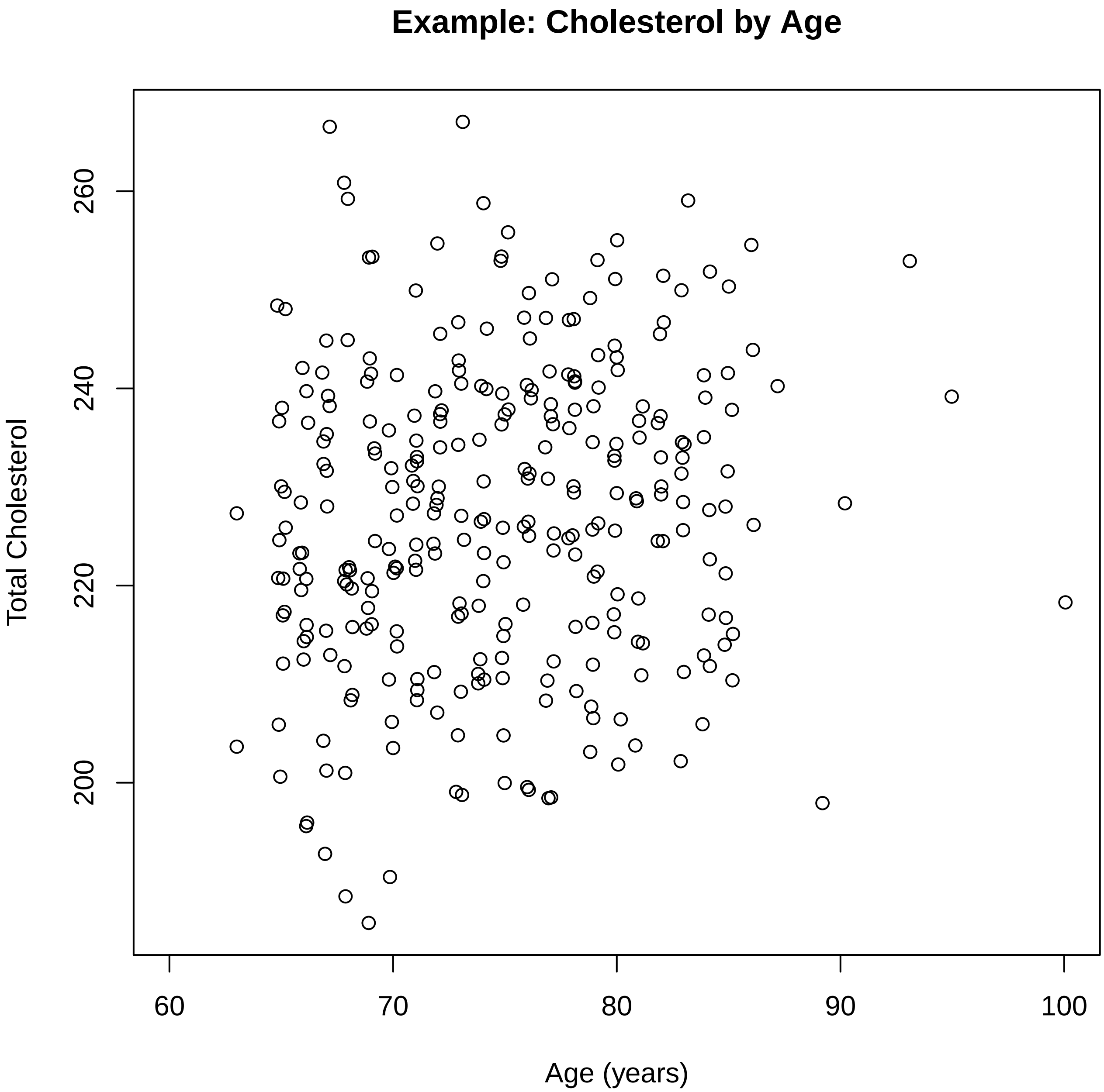
3.1 General Regression Setting

3.1.1 Two variable setting

- Many statistical problems examine the association between two variables
  - Outcome variable (response variable, dependent variable)
  - Grouping variable (covariate, predictor variable, independent variable)
- Compare distribution of the outcome variable across levels of the grouping variable
  - Groups are defined by the grouping variable
  - Within each group, the grouping variable is constant
- In intro course, statistical analysis is characterized by two factors
  - Number of groups (samples)

  - If subjects in groups are independent

- In the two variable setting, statistical analysis is more generally characterized by the grouping variable. If the grouping variable is...
  - Constant: One sample problem
  - Binary: Two sample problem
  - Categorical: k sample problem (e.g. ANOVA)
  - Continuous: Infinite sample problem (analyzed with regression)
- Regression thus *extends* the one- and two-sample problems up to infinite sample problems
  - Of course, in reality we never have *infinite* samples, but models that can handle this case are the ultimate generalization
    - Continuous predictors of interest
    - Continuous adjustment variables


Example: Cholesterol by Age
Total Cholesterol
200
220
240
260
60
70
80
90
100
Age (years)

3.1.2 Regression versus two sample approaches

- With a binary grouping variable, regression models reduce to the corresponding two variable methods

- Linear regression with a binary predictor
  - t-test, equal variance: Classic linear regression
  - t-test, unequal variance: Linear regression with robust standard errors (approximately)

- Logistic regression with a binary predictor
  - (Pearson) Chi-squared test: Score test from logistic regression

- Cox (proportional hazards) regression with a binary predictor
  - Log-rank test: Score test from Cox regression

3.1.3 Guiding principle

- Everything is regression

3.2 Motivating Problem: Cholesterol and Age

3.2.1 Definitions

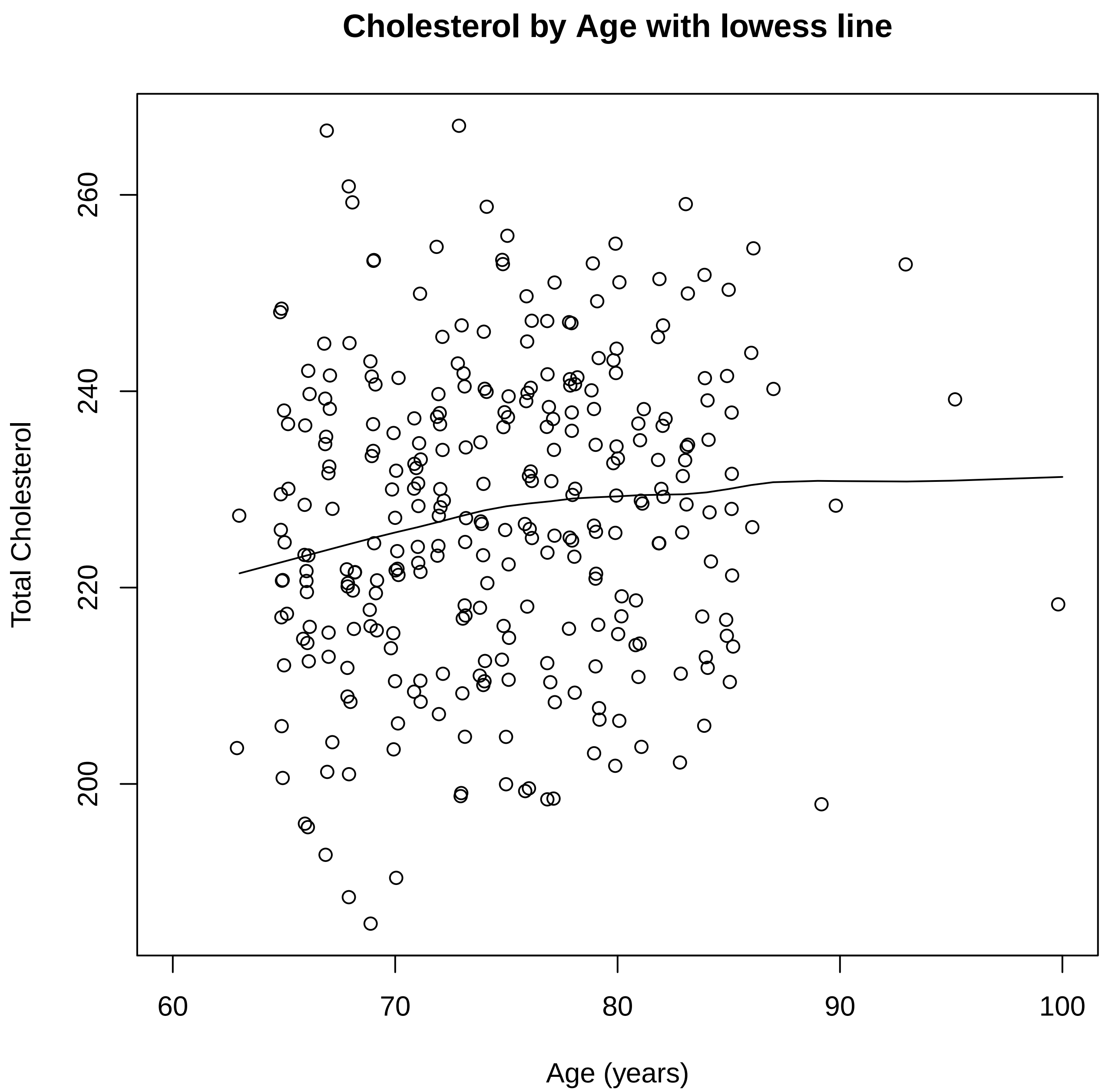
- Is there an association between cholesterol and age?
  - Scientific question: Does aging effect cholesterol?
  - Statistical question: Does the distribution of cholesterol differ across age groups?

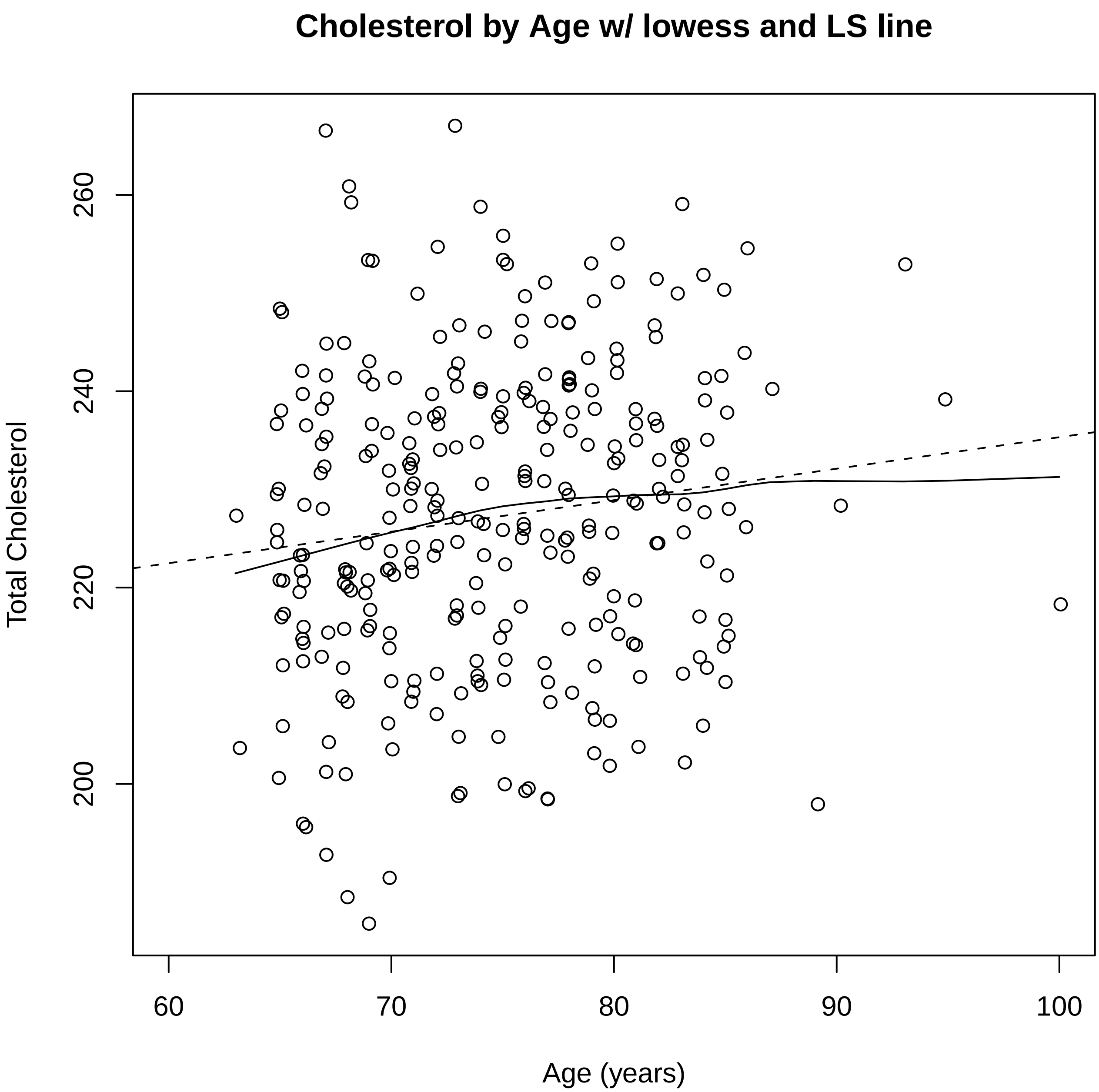
        * Acknowledges variability in the response (cholesterol)
        * Acknowledges cause-effect relationship is uncertain
            · Association does not imply causation
            · Differences could be due to calendar time of birth rather than age

- Continuous response variable: Cholesterol
- Continuous grouping variable (predictor of interest): Age
  - An infinite number of ages are possible
  - We will not sample every possible age

3.2.2 Simple Regression Model

- Attempt to answer scientific question by assessing linear trends in average cholesterol
- Estimate the best fitting line to average cholesterol within age groups
  - $E[\text{Chol}|\text{Age}] = \beta_0 + \beta_1 \times \text{Age}$
  - The expected value of cholesterol given age is modeled using an intercept (β_0) and slope (β_1)
- An association exists if the slope is nonzero
  - A non-zero slope indicates that the average cholesterol will be different across different age groups


Cholesterol by Age with lowess line
Total Cholesterol
200
220
240
260
60
70
80
90
100
Age (years)


Cholesterol by Age w/ lowess and LS line
Total Cholesterol
260
240
220
200
60
70
80
90
100
Age (years)

3.2.3 Approximate Interpretation

- The simple regression model produces an easy to remember (but approximate) rule of thumb.
  - "Normal cholesterol is 200 + one-third of your age"
  - $E[\text{Chol}|\text{Age}] = 200 + 0.33 \times \text{Age}$

3.2.4 Estimates and Interpretation

Stata output

```
. regress chol age

      Source |       SS       df       MS              Number of obs =     301
-------------+------------------------------           F(  1,   299) =    5.44
       Model |  1281.08911     1  1281.08911           Prob > F      =  0.0203
    Residual |  70363.8865   299  235.330724           R-squared     =  0.0179
-------------+------------------------------           Adj R-squared =  0.0146
       Total |  71644.9756   300  238.816585           Root MSE      =   15.34

------------------------------------------------------------------------------
        chol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3209091   .1375408     2.33   0.020     .0502384    .5915798
       _cons |   203.2259   10.31378    19.70   0.000     182.9291    223.5227
------------------------------------------------------------------------------
```

- $E[\text{Chol}|\text{Age}] = 203.2 + 0.321 \times \text{Age}$

3.2.5 Uses of Regression

- Borrowing information
  - Use other groups to make estimates in groups with sparse data
    - Intuitively, 67 and 69 year olds would provide some relevant information about 68 year olds

    * Assuming a straight line relationship tells us about other, even more distant, individuals
    * If we do not want to assume a straight line, we may only want to borrow information from nearby groups
        - Locally weighted scatterplot smooth line (lowess) added to the previous figures
        - Splines discussed in future lectures
  - Do not want to borrow too much information
    * Linear relationship is an assumption, with often low power to detect departures from linearity
    * Always avoid extrapolating beyond the range of the data (e.g. ages under 65 or over 100)

- Defining "Contrasts"
  - Define a comparison across groups to use when answering scientific questions
  - If the straight line relationship holds, the slope is the difference in mean cholesterol levels between groups differing by 1 year in age
  - If a non-linear relationship, the slope is still the average difference in mean cholesterol levels between groups differing by 1 year in age
    * Slope is a (first order or linear) test for trend

3.2.6 Linear Regression Inference

- Regression output provides
  - Estimates

    * Intercept: Estimated mean cholesterol when age is 0

    * Slope: Estimated average difference in average cholesterol for two groups differing by 1 year in age

  - Standard errors

  - Confidence intervals

  - P-values for testing ...
    * Intercept is zero (usually unimportant)

    * Slope is zero (test for linear trend in means)

- Interpretation

  - From linear regression analysis, we estimate that for each year difference in age, the difference in mean cholesterol is 0.32 mg/dL. A 95% confidence interval (CI) suggests that this observation is not unusual if the true difference in mean cholesterol per year difference in age were between 0.05 and 0.32 mg/dL. Because $p = 0.02$, we reject the null hypothesis that there is no linear trend in the average cholesterol across age groups using a significance level, α, of 0.05.

3.3 Simple Linear Regression

3.3.1 Ingredients

- Response

  - The distribution of this variable will be compared across groups
    * Linear regression models the mean of the response variable

    - Log transformation of the response corresponds to modeling the geometric mean

  - Notation: Is is extremely common to use Y to denote the response variable when discussing general methods

- Predictor
  - Group membership is measured by this variable
  - Notation
    - When not using mnemonics, will be referred to as the X variable in simple linear regression (linear regression with one predictor)
    - Later, when we discuss multiple regression, will refer to $X_1, X_2, \ldots, X_p$ when there are up to p predictors

- Regression Model
  - We typically consider a "linear predictor function" that is linear in the modeled predictors
    - Expected value (i.e. mean) of Y for a particular value of X
    - $E[Y|X] = \beta_0 + \beta_1 \times X$
  - In a deterministic world, a line is of the form $y = mx + b$
    - With no variation in the data, each value of y would like exactly on a straight line
    - Intercept b is values of y when $x = 0$
    - Slope m is the difference in y for a one unit difference in x
  - Statistics in not completely deterministic. The real world has variability

- Response with groups is variable
  - Randomness due to other variables (?)

  - Inherent randomness

- The regression line thus describes the central tendency of the data in a scatterplot of the response versus the predictor

- Interpretation of regression parameters
  - Intercept β_0: Mean Y for a group with $X = 0$
    - Often is not of scientific interest

    - May be out of the range of data, or even impossible to observe $X = 0$

  - Slope β_1: Difference in mean Y across groups differing in X by 1 unit
    - Usually measures association between Y and X

    - $E[Y|X] = \beta_0 + \beta_1 \times X$