## Contents

5 Simple Linear Regression 5

5.1 General Regression Setting ............................... 5

5.1.1 Two variable setting ................................. 5

5.1.2 Regression versus two sample approaches .......... 8

5.1.3 Guiding principle ..................................... 8

5.2 Motivating Problem: Cholesterol and Age ............. 8

5.2.1 Definitions ........................................... 8

5.2.2 Simple Regression Model ............................ 9

5.2.3 Approximate Interpretation .......................... 12

5.2.4 Estimates and Interpretation ........................ 12

5.2.5 Uses of Regression .................................. 12

5.2.6 Linear Regression Inference ........................ 13

5.3 Simple Linear Regression ............................... 14

5.3.1 Ingredients .......................................... 14

5.3.2 Matrix Notation ...................................... 17
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4 Example Analysis of BMI (Part 1)</td>
<td>18</td>
</tr>
<tr>
<td>5.4.1 Stata code</td>
<td>18</td>
</tr>
<tr>
<td>5.4.2 R Code</td>
<td>20</td>
</tr>
<tr>
<td>5.4.3 Comments on Interpretation</td>
<td>21</td>
</tr>
<tr>
<td>5.5 Residuals</td>
<td>22</td>
</tr>
<tr>
<td>5.6 Inference with Regression</td>
<td>23</td>
</tr>
<tr>
<td>5.6.1 The Big Three Assumptions</td>
<td>23</td>
</tr>
<tr>
<td>5.6.2 Additional Assumptions for Additional Inference</td>
<td>24</td>
</tr>
<tr>
<td>5.6.3 Implications for Inference</td>
<td>25</td>
</tr>
<tr>
<td>5.7 Example Analysis Revisited: Robust Standard Errors</td>
<td>32</td>
</tr>
<tr>
<td>5.7.1 R and Stata Code</td>
<td>32</td>
</tr>
<tr>
<td>5.7.2 Stata output</td>
<td>33</td>
</tr>
<tr>
<td>5.7.3 Interpretation of the slope</td>
<td>34</td>
</tr>
<tr>
<td>5.8 Model Checking</td>
<td>35</td>
</tr>
<tr>
<td>5.9 Linear Regression and Correlation</td>
<td>37</td>
</tr>
<tr>
<td>5.9.1 Pearson’s Correlation</td>
<td>37</td>
</tr>
<tr>
<td>5.9.2 Relationship to regression</td>
<td>39</td>
</tr>
<tr>
<td>5.10 Linear Regression and t Tests</td>
<td>41</td>
</tr>
<tr>
<td>5.10.1 Example: BMI and Gender</td>
<td>42</td>
</tr>
<tr>
<td>5.11 Regression on Geometric Means: Log transformations</td>
<td>44</td>
</tr>
</tbody>
</table>
5.11.1 Interpretation of Parameters ................................. 44
5.11.2 Example: FEV with Height ................................. 45
5.11.3 Log Transformed Predictors ................................. 46
5.11.4 Looking Ahead to Multivariable Models ................. 48
Chapter 5

Simple Linear Regression

5.1 General Regression Setting

5.1.1 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
5.1.2 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

- Proportional odds regression with a binary predictor
  - Wilcoxon Rank Sum Test

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

5.1.3 Guiding principle

- Everything is regression

5.2 Motivating Problem: Cholesterol and Age

5.2.1 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

5.2.2 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 (\(\beta_0\)) and slope (\(\beta_1\))

  · 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
Cholesterol by Age w/ lowess and LS line
5.2.3 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”
  - $E[\text{Chol}|\text{Age}] = 200 + 0.33 \times \text{Age}$

5.2.4 Estimates and Interpretation

**Stata output**

```stata
. regress chol age
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 301</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1281.08911</td>
<td>1</td>
<td>1281.08911</td>
<td>F( 1, 299) = 5.44</td>
</tr>
<tr>
<td>Residual</td>
<td>70363.8865</td>
<td>299</td>
<td>235.330724</td>
<td>Prob &gt; F = 0.0203</td>
</tr>
<tr>
<td>Total</td>
<td>71644.9756</td>
<td>300</td>
<td>238.816585</td>
<td>Adj R-squared = 0.0146</td>
</tr>
</tbody>
</table>

| chol  | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|---------|-----------|------|-----|---------------------|
| age   | 0.3209091 | 0.1375408 | 2.33 | 0.020 | 0.0502384 0.5915798 |
| _cons | 203.2259 | 10.31378 | 19.70 | 0.000 | 182.9291 223.5227 |

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

5.2.5 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
    • Do we want to assume that comparisons of 65 to 66 year old subjects are the same as comparisons of 95 to 96 year old subjects?

  – 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

5.2.6 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, $\alpha$, of 0.05.

5.3 Simple Linear Regression

5.3.1 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: It 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 lie 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
CHAPTER 5. SIMPLE LINEAR REGRESSION

* 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 $\beta_0$: 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 $\beta_1$: Difference in mean $Y$ across groups differing in $X$ by 1 unit
    * Usually measures association between $Y$ and $X$
    * $E[Y|X] = \beta_0 + \beta_1 \times X$

· Derivation of interpretation
  – Simple linear regression of response $Y$ on predictor $X$
  – Mean of any arbitrary group can be derived from the $Y_i = \beta_0 + \beta_1 X_i$
  – Interpretation determined by considering possible values of $X$
  – Model: $E[Y_i|X_i] = \beta_0 + \beta_1 \times X_i$
    * When $X_i = 0$, $E[Y_i|X_i = 0] = \beta_0$
    * When $X_i = x$, $E[Y_i|X_i = x] = \beta_0 + \beta_1 x$
* When \( X_i = x + 1 \), \( E[Y_i|X_i = x + 1] = \beta_0 + \beta_1 x + \beta_1 \)

### 5.3.2 Matrix Notation

- Using scalars, the simple linear regression model can be written as
  \[
  Y_i = \beta_0 + \beta_1 \times X_i + \epsilon_i
  \]
  \[i = 1, \ldots, n\]
  * \( i \) indexes the independent sampling units (e.g. subjects)

  * \( n \) is the total number of independent sampling units

- Drop the expected value notation, add in \( \epsilon_i \)
  * \( \epsilon_i \) are the “Residuals” or “Errors”

  * \( E[\epsilon_i] = 0 \)

  * \( V[\epsilon_i] = \sigma^2 \) (constant variance assumption)

- Matrix notation: \( Y = X\beta + \epsilon \)

  \[
  Y = \begin{pmatrix}
  Y_1 \\
  Y_2 \\
  \vdots \\
  Y_n
  \end{pmatrix}_{n \times 1}
  \]

  \[
  X = \begin{pmatrix}
  1 & x_1 \\
  1 & x_2 \\
  \vdots & \vdots \\
  1 & x_n
  \end{pmatrix}_{n \times 2}
  \]

  \[
  \beta = \begin{pmatrix}
  \beta_0 \\
  \beta_1
  \end{pmatrix}_{2 \times 1}
  \]

  \[
  \epsilon = \begin{pmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_n
  \end{pmatrix}_{n \times 1}
  \]

  * \( E[\epsilon] = 0 \), where \( 0 = \begin{pmatrix}
  0 \\
  0 \\
  \vdots \\
  0
  \end{pmatrix}_{n \times 1} \)
\* \( V[\epsilon] = \sigma^2 I \), where \( I = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}_{n \times n} \)

\* \( E[Y] = X\beta \)

– I use standard notation to indicate matrices/vectors and scalars
  * Boldface indicates a vector or matrix (\( Y, X, \beta, \epsilon, 0, I \))

  * Normal typeface indicates a scalar (\( Y_i, x_i, \beta_0, \beta_1, \epsilon_i, 0, 1 \))

### 5.4 Example Analysis of BMI (Part 1)

• Example analysis conducted in class involving BMI (response) with gender (predictor 1) and age (predictor 2)

• Stata and R code to reproduce the same analysis are provided below

#### 5.4.1 Stata code

```
**************************
*** Initial data manipulation ***
**************************

* Load the inflammation dataset from the web

use http://biostat.mc.vanderbilt.edu/wiki/pub/Main/CourseBios312/inflamm.dta

*** Quick descriptive statistics
* We will focus on three variables for this analysis
describe summarize bmi age male

*** Some variable labeling
label define malelbl 0 "Female" 1 "Male"
```
label values male malelbl

*******************************
*** Part 1: BMI and gender ***
*******************************

table male, c(n bmi mean bmi sd bmi) format (%9.2f)
table male, c(median bmi p25 bmi p75 bmi) format (%9.2f)

graph box bmi, by(male)

* Compare simple linear regression results to t-test results
regress bmi male
ttest bmi, by(male)

*******************************
*** Part 2: BMI and age ***
*******************************

* Stratified descriptive statistics
table age, c(n bmi mean bmi sd bmi) format (%9.2f)
table age, c(median bmi p25 bmi p75 bmi) format (%9.2f)

* Scatter plots with various line fits
scatter bmi age
scatter bmi age || lfit bmi age
twoway (scatter chol age, jitter(3)) (line mchol age)
lowess bmi age, bwidth(.1) addplot((lfit bmi age))

sort age
by age: egen mbmi = mean(bmi)
twoway (scatter bmi age, jitter(3)) (line mbmi age)

* Simple linear regression
regress bmi age

* What about a 5-year change in age?
lincom 5*age

* Or, a 10 year change in age?
lincom 10*age
### Initial data manipulation

```r
# Load additional packages
# If you have never installed these packages, first
# install.packages("rms")
# install.packages("lattice")
library(rms)
library(lattice)

d1 <- stata.get("http://biostat.mc.vanderbilt.edu/wiki/pub/Main/CourseBios312/inflamm.dta")
```

### Quick descriptive statistics

```r
# We will focus on three variables for this analysis
describe(d1)
summary(~bmi + age + male, data=d1)
```

### Some variable labeling

```r
d1$male.factor <- factor(d1$male, levels=c(0,1), labels=c("Female", "Male"))
```

### Part 1: BMI and gender

```r
summary(male.factor~bmi, method="reverse", data=d1)
bwplot(male.factor ~ bmi, data=d1)
```

# Compare simple linear regression results to t-test results

```r
m.male <- lm(bmi ~ male, data=d1)
summary(m.male)
confint(m.male)
t.test(bmi ~ male, var.equal=TRUE, data=d1)
```

### Part 2: BMI and age

```r
# Stratified descriptive statistics
# Note you have to specify fun= to return the mean and std deviation, otherwise it
# returns the mean by default
# continuous=50 gives summaries for all ages
summary(bmi ~ age, data=d1, continuous=50, fun=function(x) c(mean(x), sd(x)))
```
# Scatter plots with various line fits
with(d1, plot(age, bmi))
with(d1, abline(lm(bmi ~ age), lwd=2))
with(d1, lines(lowess(bmi[!is.na(bmi)] ~ age[!is.na(bmi)]), lty=1, col='red', lwd=2))

# Simple linear regression
m.age <- lm(bmi~age, data=d1)
summary(m.age)
confint(m.age, "age")

# What about a 5-year change in age?
d1$age5 <- d1$age/5
m.age5 <- lm(bmi~age5, data=d1)
summary(m.age5)
confint(m.age5, "age5")

# Or, a 10 year change in age?
coef(m.age)[["age"]]*10
confint(m.age, "age")*10

5.4.3 Comments on Interpretation

- The intercept is just a mathematical construct and has no direct interpretation

- All estimates assume a common within group variance. If this is not the case, the reported variance is just the (weighted) average over all groups.

- The slope for age is expressed as a difference in group means, not the difference due to aging. We did not do a longitudinal study in which repeated measurements were taken on the same subject.

- If the group means are linear, then the slope has an exact interpretation as the change in BMI due to a one year change in (any) age
  - Otherwise, the slope estimates the first order trend of the sample data and we should not treat the estimates of group means as accurate
5.5 Residuals

- Often linear regression models are specified in terms of the response instead of the mean response
  - Include an error term in the model, $\epsilon_i$
  - Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- The regression model is divided into two parts
  - The mean, or systematic, part (the “signal”)
  - The error, or random, part (the “noise”)

- Residuals
  - $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
  - $\hat{\epsilon} = Y - X \hat{\beta}$
  - $\hat{\beta} = (X'X)^{-1} (X'Y)$

  - The mean of the residuals is 0

  - The standard deviation of the residuals is the “Root Mean Square Error”
    - In our example analysis of BMI and gender, the RMSE is exactly equal to the pooled estimate of the standard deviation from a two-sample, equal variance t-test
    - In our example analysis of BMI and age, the RMSE is the square root of the average variances across the age groups

- In many textbooks, $\epsilon_i \sim N(0, \sigma^2)$
– A common $\sigma^2$ implies constant variance across all levels of the grouping variable, “homoscedasticity”

– Normality of the residuals is a nice property, but it is not necessary (and rarely observed in practice)

– We will discuss how lack of Normality and heteroscedasticity impact statistical inference

5.6 Inference with Regression

5.6.1 The Big Three Assumptions

· Most common uses of regression
  – Prediction: Estimating what a future value of $Y$ will be based on observed $X$

  – Comparisons within groups: Describing the distribution of $Y$ across levels of the grouping variable $X$ by estimating the mean $E[Y|X]$ 

  – Comparisons across groups: Differences appear across groups if the regression parameter slope estimate $\beta_1$ is non-zero

· Valid statistical inference (CIs, p-values) about associations requires three general assumptions
  – Assumption 1: Approximately Normal distributions for the parameter estimates
    * Normal data or “large” N

    * It is often surprising how small “large” can be
      · Definition of large depends on the error distribution and relative sample sizes within each group
· With exactly Normally distributed errors, only need one observation (or two to estimate a slope)

· With very heavy tails, “large” can be very large

· See Lumley, et al., Ann Rev Pub Hlth, 2002

– Assumption 2: Independence of observations
  ∗ Classic regression: Independence of all observation (now)

  ∗ Robust standard errors: Correlated observations within identified clusters (later)

– Assumption 3: Assumption about the variance of observations within groups
  ∗ Classic regression: Homoscedasticity (equal variance across groups)

  ∗ Robust standard errors: Allow unequal variance across groups

5.6.2 Additional Assumptions for Additional Inference

· Note that some textbooks will claim there are more than three assumptions. In truth, additional assumptions are not needed to make the aforementioned statistical inference about associations. However ...

· Valid statistical inference (CIs, p-values) about means responses in specific groups requires a further assumption

  – Assumption 4: Adequacy of the linear model

    ∗ If we are trying to borrow information about the mean from neighboring groups, and we are assuming a straight line relationship, the straight line needs to be true
No longer saying there is just a linear trend in the means, but now need to believe that all the means lie on a straight line

Note that we can model transformations of the measured predictor

- For inference about individual observations (prediction intervals, P-values) in specific groups requires another assumption
  - Assumption 5: Assumptions about the distribution of the errors within each group (a very strong assumption)
    * Classically: Errors have the same Normal distribution within each grouping

Robust standard error will not help
  - Prediction intervals assume a common error distribution across groups (homoscedasticity)

Possible extension: Errors have the same distribution, but not necessarily Normal (not implemented in software)
  - Bootstrapping
  - Bayesian analysis
  - Other flexible approaches

5.6.3 Implications for Inference

The Moral: Regression based inference about associations is far more trustworthy than estimation of group means or individual predictions. Now, the story...

- We will now consider a hierarchy of null hypotheses
  - Strong null: Total independence of $Y$ and $X$
- Intermediate null: Mean of $Y$ is the same for all $X$ groups

- Weak null: No linear trend in mean of $Y$ across $X$ groups

• Under Strong Null (response and predictor are totally independent)
  - All aspects of the distribution of $Y$ would be the same in each group
    * A flat line would describe the mean response across groups (a linear model is correct, and the slope is zero)

    * Within group variance is the same in all groups

    * In large sample sizes, the regression parameters are Normally distributed

• Under Intermediate Null
  - Means for each predictor group would lie on a flat line
    * Slope is zero

    * Within group variance could vary across groups

    * In large sample sizes, the regression parameters are Normally distributed. Note that the definition of large will depend on how much the error distributions vary across groups relative to the number of samples in each group

• Under Weak Null
  - Linear trend in means across predictor groups would lie on a flat line
    * Slope of best fitting line is zero

    * Within group variance could vary across groups

    * Error distribution could vary across groups
In large sample sizes, the regression parameters are Normally distributed. Note that the definition of large will (again) depend on how much the error distributions vary across groups relative to the number of samples in each group.
Random data under intermediate/weak null
Random data under weak null
CHAPTER 5. SIMPLE LINEAR REGRESSION

• Classical Linear Regression
  – Inference about the slope tests the strong null
    * All tests make inference by assuming the strong null is true

    * If the data appear non-linear or heteroscedastic, merely evidence the strong null is not true

  – Limitations
    * We cannot be confident that there is a difference in means (valid inference about means requires homoscedasticity)

    * We cannot be confident in estimates of group means (valid estimates of group means require linearity)

• Robust Standard Errors
  – Inference about the slope tests the weak null
    * All tests make inference by assuming the weak null is true

    * Data can appear non-linear or heteroscedastic
      • Robust SE allow unequal variance

      • Nonlinearity decreases precision, but inference about first-order trends still valid

  – Only if linear relationship holds can we...
    * Test intermediate null

    * Estimate group means

• Back to the Moral: Regression based inference about associations is far more trustworthy than estimation of group means or individual predictions
– A non-zero slope suggests an association between the response and predictor

– If use robust SE, can also make inference about linear trends in means

• Interpreting “Positive” Results
  – Slope is statistically different from 0 using robust standard errors
    * Observed data is atypical of a setting with no linear trend in mean response across groups

    * Data suggests evidence of a trend toward larger (or smaller) means in groups having larger values of the predictor

    * (To the extent the data appears linear, estimates of the group means will be reliable)

• Interpreting “Negative” Results
  – Many possible reasons why the slope is not statistically different from 0 using robust standard errors
    * There may be no association between the response and predictor

    * There may be an association, but not in the parameter considered (the mean response)

    * There may be an association in the parameter considered, but the best fitting line has zero slope

    * There may be a first order trend in the parameter considered, but we lacked the precision to be confident that it truly exists (a type II error)
5.7 Example Analysis Revisited: Robust Standard Errors

5.7.1 R and Stata Code

- In Stata, for robust standard errors, just add a “robust” option to the usual regress command
  
  ```
  regress responsevar predictor, robust
  regress bmi age, robust
  ```

- In R, for robust standard error, we can use the robcov() function in the rms package...
  
  ```
  require(rms)
  m.age <- ols(bmi~age, data=d1, x=TRUE, y=TRUE)
  robcov(m.age)
  ```

- There are multiple ways to fit models with robust standard errors in R
  
  - Here, the ols() function fits the ‘ordinary least squares’ regression and robcov() gives the robust standard errors using the Huber-White “sandwich” estimator
### 5.7.2 Stata output

```
* Classical Linear Regression *

```
```
. regress bmi age

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 4987</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2161.8992</td>
<td>1</td>
<td>2161.8992</td>
<td>F( 1, 4985) = 98.30</td>
</tr>
<tr>
<td>Residual</td>
<td>109633.165</td>
<td>4985</td>
<td>21.9926108</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>111795.064</td>
<td>4986</td>
<td>22.4217938</td>
</tr>
</tbody>
</table>

| bmi     | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|---------|-------|-----------|------|------|----------------------|
| age     | -.1175937 | .0118605 | -9.91 | 0.000 | -.1408456 -.0943418 |
| _cons   | 35.23271  | .8663188 | 40.67 | 0.000 | 33.53435 36.93108   |
```
```
* Robust Standard Errors *

```
. regress bmi age, robust

```
```
Linear regression  Number of obs = 4987  F( 1, 4985) = 103.06
|        | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|        | bmi   |         |      |      |                     |
| age    | -.1175937 | .0115836 | -10.15 | 0.000 | -.1403026 -.0948848 |
| _cons  | 35.23271  | .8534722 | 41.28 | 0.000 | 33.55953 36.90589  |
```
```
5.7.3 Interpretation of the slope

- In both classical regression and robust standard errors, inference for association is based on the slope

- Classical regression
  - Strong null inference
    - \( p < 0.001 \) suggests that the distribution of BMI varies across age groups

  - Under assumption of homoscedasticity
    - Estimated trend in mean BMI by age is an average difference of \(-0.118\) per one year difference in age (BMI decreases with age, at least when you are over 65)

    - CI for trend: \([-0.141, -0.094]\)

- Robust standard errors
  - Weak null based inference
    - \( p < 0.001 \) suggests that mean BMI varies across age groups

    - Estimated trend in mean BMI by age is an average difference of \(-0.118\) per one year difference in age

    - CI for trend: \([-0.140, -0.095]\)

- So, which inference is correct?
  - Classical linear regression and linear regression with robust standard errors differ in the strength of necessary assumptions

  - As a rule, if all of the assumptions of classical linear regression hold, it will be more precise
Hence, we will have the greatest precision to detect associations if the linear model is correct.

Can we test equality of variances? Yes, but not really...
- The F-test for equal variances relies heavily on the Normality assumption.
  - The test leads you to interpret a large p-value as indicating equal variance, which we know is wrong.
  - Also multiple testing.

- The robust standard error methods are valid for detection of associations without relying on the classic assumptions.

- “All models are false, some models are useful”– George Box.
- “In statistics, as in art, never fall in love with your model”– Unknown.

5.8 Model Checking

- Much statistical literature has been devoted to methods for checking the assumptions for regression models.
  - My (modern) philosophy: Model checking is generally fraught with peril as it necessarily involves multiple comparisons.

  - We cannot reliably use the sampled data to assess whether it accurately portrays the population.
    - We are more worried about the data from the population that we might not have sampled.
    - It is not so much the abnormal points that we see, but the ones that...
are hiding in the population that will make our model perform badly

* But, do tend to worry more when we see a tendency to extreme outliers in the sample or clear departures from model assumptions

– If we over-check our model, we will inflate the type I error rate (i.e. will be more likely to claim statistical significance when it doesn’t really exist)

* Estimates are biased away from the null

* Reported standard errors are too small

– If we fish through the data, we will always find significant results

* In clinical trials, often Phase II results are not able to be replicated in Phase III trials

· Instead of extensive model checking, go back to our choices of inference when planning our analysis

– Best to plan for unusual data

– There is often little to be lost and much to be gained by using the robust standard error estimates

– By using robust errors, avoids much of the need for model checking

* Model checking is almost entirely data driven

– Robust standard errors is a more logical scientific approach

* Minimize the need to presume more knowledge than the question we are trying to answer

* E.g., if we don’t know how the means might differ, why should we presume to know how the variances or the shape of the distribution might behave?
5.9 Linear Regression and Correlation

5.9.1 Pearson’s Correlation

- **Notation**
  - \( \rho \) signifies the population value
  - \( r \) (or \( \hat{\rho} \)) is the estimated correlation from data

- **Formula**
  \[
  r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}
  \]

  - Range: \(-1 \leq r \leq 1\)

- **Interpretation**
  - Measures the linear relationship between \( X \) and \( Y \)
  - Correlation coefficient is a unitless index of strength of association between two variables (\(+ = \) positive association, \(- = \) negative, \(0 = \) no association)

- Can test for significant association by testing whether the population correlation is zero
  \[
  t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}
  \]
  which is identical to the \( t \)-test used to test whether the population \( r \) is zero; d.f. = \( n - 2 \)

- Use probability calculator for \( t \) distribution to get a 2-tailed \( P \)-value
Confidence intervals for population $r$ calculated using Fisher’s $Z$ transformation

$$Z = \frac{1}{2} \log_e \left( \frac{1 + r}{1 - r} \right)$$

- For large $n$, $Z$ follows a Normal distribution with standard error $\frac{1}{\sqrt{n-3}}$

- To calculate a confidence interval for $r$, first find the confidence interval for $Z$ then transform back to the $r$ scale

$$Z = \frac{1}{2} \log_e \left( \frac{1 + r}{1 - r} \right)$$

$$2 \cdot Z = \log_e \left( \frac{1 + r}{1 - r} \right)$$

$$\exp(2 \cdot Z) = \left( \frac{1 + r}{1 - r} \right)$$

$$\exp(2 \cdot Z) \cdot (1 - r) = 1 + r$$

$$\exp(2 \cdot Z) - r \cdot \exp(2 \cdot Z) = 1 + r$$

$$\exp(2 \cdot Z) - 1 = r \cdot \exp(2 \cdot Z) + r$$

$$\exp(2 \cdot Z) - 1 = r (\exp(2 \cdot Z) + 1)$$

$$\frac{\exp(2 \cdot Z) - 1}{\exp(2 \cdot Z) + 1} = r$$

Example (Altman 89-90): Pearson’s $r$ for a study investigating the association of basal metabolic rate with total energy expenditure was calculated to be 0.7283 in a study of 13 women. Derive a 95% confidence interval for $r$.

$$Z = \frac{1}{2} \log_e \left( \frac{1 + 0.7283}{1 - 0.7283} \right) = 0.9251$$

The lower limit of a 95% CI for $Z$ is given by

$$0.9251 - 1.96 \times \frac{1}{13 - 3} = 0.3053$$

and the upper limit is

$$0.9251 + 1.96 \times \frac{1}{13 - 3} = 1.545$$
A 95% CI for the population correlation coefficient is given by transforming these limits from the $Z$ scale back to the $r$ scale

$$\frac{\exp(2 \times 0.3053) - 1}{\exp(2 \times 0.3053) + 1} \text{ to } \frac{\exp(2 \times 1.545) - 1}{\exp(2 \times 1.545) + 1}$$

Which gives a 95% CI from 0.30 to 0.91 for the population correlation

5.9.2 Relationship to regression

- Pearson’s correlation ($\rho$) is directly related to linear regression
  - Correlation treats $Y$ and $X$ symmetrically, but we can relate $E[Y|X]$ as a function of $X$
  - $E[Y|X] = \beta_0 + \beta_1 X$
  - $\beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$
    * $E[Y|X]$: mean $Y$ within groups having equal $X$
    * $\beta_1$: difference in mean $Y$ per 1 unit difference in $X$
    * $\rho$: true correlation between $Y$ and $X$
    * $\sigma_Y$: standard deviation of $Y$
    * $\sigma_X$: standard deviation of $X$

- More interpretable formulation of $\rho$
  - $\rho \approx \frac{\beta \sqrt{\text{Var}(X)}}{\sqrt{\beta^2 \text{Var}(X) + \text{Var}(Y|X=x)}}$
    * $\beta$: slope between $Y$ and $X$
    * $\text{Var}(X)$: variance of $X$ in the sample
\( \text{Var}(Y|X = x) \): variance of \( Y \) in groups having the same value of \( X \) (the vertical spread of data)

- Correlation tends to increase in absolute value as
  - The absolute value of the slope of the line increases
  - The variance of data decreases within groups that share a common value of \( X \)
  - The variance of \( X \) increases

- Scientific uses of correlation
  - The slope between \( X \) and \( Y \) is of scientific interest
  - The variance of \( Y \) given \( X = x \) is partly of scientific interest, but can also be impacted by restricting sampling to certain values of another variable
    - The variance of height given age is less if sampling just males than if sampling both sexes
  - The variance of \( X \) is often set by study design (which is often not of scientific interest)

- Hypothesis tests for a nonzero correlation are exactly the same as a test for a nonzero slope in classical linear regression
  - The statistical significance of a given value of \( \rho \) depends only on the sample size
  - Correlation is more of a statistical measure than a scientific measure
5.10 Linear Regression and t Tests

- Linear regression with a binary predictor corresponds to familiar t tests
  - Classical linear regression: Two sample t test which presumes equal variance (exactly the same)
  - Robust standard errors estimates: Two sample t test which allows unequal variances (nearly the same)
  - Identified clusters with robust standard error estimates: Paired t test (nearly the same)
5.10.1 Example: BMI and Gender

Classic LS vs equal variance t test

```
. ttest bmi, by(male)
Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>obs</th>
<th>mean</th>
<th>std. err.</th>
<th>std. dev.</th>
<th>[95% conf. interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>2895</td>
<td>26.8592</td>
<td>.0986978</td>
<td>5.310456</td>
<td>26.66565 27.0527</td>
</tr>
<tr>
<td>combined</td>
<td>4987</td>
<td>26.6687</td>
<td>.0670526</td>
<td>4.735166</td>
<td>26.53725 26.80015</td>
</tr>
<tr>
<td>diff</td>
<td>.4540563</td>
<td>.1357396</td>
<td>.187947</td>
<td>.7201655</td>
<td></td>
</tr>
</tbody>
</table>

diff = mean(Female) - mean(Male) t = 3.3451
Ho: diff = 0 degrees of freedom = 4985

Pr(T < t) = 0.9996 Pr(|T| > |t|) = 0.0008 Pr(T > t) = 0.0004
```

```
. regress bmi male
Source | SS     df       MS         Number of obs = 4987
-------|--------|------------|-----------------|---------------------|
Model  | 250.374569 1 250.374569  Prob > F = 0.0008
Residual | 111544.689 4985 22.3760661  R-squared = 0.0022
Total  | 111795.064 4986 22.4217938  Root MSE = 4.7303

bmi | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------|-----------------|--------|--------|---------------------|
male | -.4540563 .1357396 -3.35 0.001 -.7201655 -.187947
_cons | 26.85917 .0879159 305.51 0.000 26.68682 27.03152
```

```
. lincom male + _cons
( 1)  male + _cons = 0

bmi | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------|-----------------|--------|--------|---------------------|
(1) | 26.40511 .1034215 255.32 0.000 26.20236 26.60787
```
Robust standard errors vs unequal variance t test

```
test bmi, by(male) unequal
```

Two-sample t test with unequal variances

```
<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>2895</td>
<td>26.85917</td>
<td>0.0986978</td>
<td>5.310456</td>
<td>26.66565 27.0527</td>
</tr>
<tr>
<td>Male</td>
<td>2092</td>
<td>26.40511</td>
<td>0.0827187</td>
<td>3.783421</td>
<td>26.2429 26.56733</td>
</tr>
<tr>
<td>combined</td>
<td>4987</td>
<td>26.6687</td>
<td>0.0670526</td>
<td>4.735166</td>
<td>26.53725 26.80015</td>
</tr>
<tr>
<td>diff</td>
<td>.4540563</td>
<td>.1287775</td>
<td>.2015957</td>
<td>.7065169</td>
<td></td>
</tr>
</tbody>
</table>
```

diff = mean(Female) - mean(Male)  t = 3.5259
Ho: diff = 0  Satterthwaite’s degrees of freedom = 4984.04
Ha: diff < 0  Ha: diff != 0  Ha: diff > 0
Pr(T < t) = 0.9998  Pr(|T| > |t|) = 0.0004  Pr(T > t) = 0.0002

```
regress bmi male, robust
```

Linear regression

|                      | Coef. | Std. Err. | t    | P>|t| | 95% Conf. Interval |
|----------------------|-------|-----------|------|------|------------------|
| bmi                  | -.4540563 | .1287775 | -3.53 | 0.000 | -.706517 -.2015956 |
| _cons                | 26.85917 | .0987006 | 272.13 | 0.000 | 26.66567 27.05267 |

```
lincom male + _cons
```

( 1)  male + _cons = 0

```
|                      | Coef. | Std. Err. | t    | P>|t| | 95% Conf. Interval |
|----------------------|-------|-----------|------|------|------------------|
| (1)                 | 26.40511 | .0827156 | 319.23 | 0.000 | 26.24296 26.56727 |
5.11 Regression on Geometric Means: Log transformations

- Geometric means of distributions are typically analyzed by using linear regression on the log transformed outcome.

- Used for inference under the following circumstances:
  - We are (scientifically) interested in multiplicative rather than additive models.
  - We want to down-weight the impact of outliers.
  - The standard deviation of the response in a group is proportional to the mean.
    * “Error is ±10%” instead of “Error is ±10”

5.11.1 Interpretation of Parameters

- Linear regression on the log transformed $Y$
  - Note that in statistics $\log$ almost always refers the the natural log ($\log_e$).
  - Model: $E[\log Y_i | X_i] = \beta_0 + \beta_1 \times X_i$
  - When $X_i = 0$: $E[\log Y_i | X_i = 0] = \beta_0$
  - When $X_i = x$: $E[\log Y_i | X_i = x] = \beta_0 + \beta_1 \times x$
  - When $X_i = x + 1$: $E[\log Y_i | X_i = x + 1] = \beta_0 + \beta_1 \times x + \beta_1$

- Restate the model as log link for geometric mean
  - Model: $\log \text{GM}[Y_i | X_i] = \beta_0 + \beta_1 \times X_i$
– When $X_i = 0$: $\log \text{GM}[Y_i | X_i = 0] = \beta_0$

– When $X_i = x$: $\log \text{GM}[Y_i | X_i = x] = \beta_0 + \beta_1 \times x$

– When $X_i = x + 1$: $\log \text{GM}[Y_i | X_i = x + 1] = \beta_0 + \beta_1 \times x + \beta_1$

• Interpretation of parameters by back-transforming model
  – Model: $\text{GM}[Y_i | X_i] = e^{\beta_0} \times e^{\beta_1 \times X_i}$

– When $X_i = 0$: $\text{GM}[Y_i | X_i = 0] = e^{\beta_0}$

– When $X_i = x$: $\text{GM}[Y_i | X_i = x] = e^{\beta_0} \times e^{\beta_1 \times x}$

– When $X_i = x + 1$: $\text{GM}[Y_i | X_i = x + 1] = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$

5.11.2 Example: FEV with Height

• Question: How does FEV differ across height groups?

• Scientific justification for summarizing with geometric mean
  – FEV is a volume
  
  – Height is a linear dimension
    * Each dimension of lung size is proportional to height
  
  – Standard deviation is likely proportional to height

– Science: $FEV \propto \text{Height}^3$ or $\sqrt[3]{FEV} \propto \text{Height}$

– Statistics: $\log(FEV) \propto 3 \log(\text{Height})$
• Science dictates model choice
  – Statistical preference for (any) transformations of response
    * May transform to equal variance across groups
    * Homoscedasticity allows for easier inference
  – Statistical preference for a log transformation
    * Easier interpretation: multiplicative model
    * Compare groups using ratios

5.11.3 Log Transformed Predictors

• Interpretation of log predictor and log link for response
  – Log link used to model the geometric mean of $Y$
    * Exponentiated slope estimates ratio of geometric means across groups
  – Log transformation of predictor compares groups with a $k$-fold difference in there measured predictors
    * Estimated ratio of geometric means
      * $\exp(\log(k) \times \beta_1) = k^{\beta_1}$

• Interpretation of model
  – $\log \text{GM}[FEV_i|\log(height_i)] = -11.9 + 3.12 \times \log(height_i)$
  – Estimate ratio of geometric mean FEV for two groups differing by 10% in height (a 1.1 fold difference in height)
    * To find: Exponentiate 1.1 to the slope: $1.1^{3.12} = 1.35$
∗ “Group that is 10% taller is estimated to have a geometric mean FEV that is 1.35 times higher (35% higher)"

· Why transform the predictor?
  – Typically chosen according to whether the data are likely to follow a straight line relationship

  – Linearity necessary to predict the values of the parameter in individual groups
    ∗ Linearity is not necessary to estimate existence of an association

    ∗ Linearity is not necessary to estimate a first order trend in the parameter across groups having the sampled distribution of the predictor

    ∗ (Inference about these two questions will tend to be conservative if linearity does not hold)

· Rarely do we know which transformation of the predictor provides the best linear fit
  – As always, there is significant danger in using the data to estimate the best transformation to use

  – If there is no association of any kind between the response and the predictor, a linear fit (with a zero slope) is correct

  – Trying to detect a transformation is thus an informal test for an association (multiple testing, which inflates type I error rate)

  – It is best to choose transformations of the predictor on scientific grounds
    ∗ In practice, it is often true that mean functions are well approximated by a straight line over a small range of the data
Example: In the FEV dataset, the logarithm of height is approximately linear over the range of heights sampled (look at the plot)

- We can often choose untransformed predictors when science would suggest a non-linear association
  - Can have major advantages in interpreting the results of the analysis
  - In the FEV example, it is far more natural to compare heights by difference than by ratios
    * We would rather characterize two children as differing by 4 inches in height rather than as the 44 inch child being 10% taller than the 40 inch child

5.11.4 Looking Ahead to Multivariable Models

- The relative importance of having the “true” transformation for a predictor depends on the statistical role
  - Predictor of Interest
  - Effect Modifiers
  - Confounders
  - Precision variables

- Earlier I made a comment about treating these variables different in an analysis

- Predictor of Interest (should be pre-specified)
  - In general, don’t worry about modeling the exact association until you have even established an association (binary search)
CHAPTER 5. SIMPLE LINEAR REGRESSION

* Need to have a prespecified model for the POI

* Searching for the best fit can inflate the type I error

* Make most accurate, precise inference about the presence of the association first

* Exploratory models can suggest models for future analyses

· Effect Modifiers (too difficult to do right anyway)
  – Modeling of effect modifiers is invariably just a test for the existence of an interaction
  * We rarely have a lot of precision to answer questions in subgroups of the data

  * Patterns of interaction can be so complex that it is unlikely that we will really capture the interactions across all subgroups in a single model

  * Typically, we would restrict future studies to analyses treating subgroups separately

· Confounders (matters the most)
  – It is important to have an appropriate model of the association between the confounder and the response

  – Failure to accurately model the confounder means that some residual confounding will exist

  – However, searching for the best model may inflate the type I error for inference about the predictor of interest by overstating the precision of the study
– Luckily, we rarely care about inference for the confounder, so we are free to use inefficient means of adjustment (e.g. a stratified analysis)

• Precision Variables (who cares?)
  – When modeling precision variables, it is rarely worth the effort to use the “best” transformation
  – We usually capture the largest part of the added precision using crude models
  – We generally do not care about estimating association between the response and the precision variable
  – Most often, precision variables represent known effects on the response