SISG
2014

SISG Module 5:
Regression and Analysis of Variance

19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Summer Institute in Statistical Genetics
Module 5: Regression and Analysis of Variance
July 9-11, 2014

Instructors:

Rebecca Hubbard, PhD
Associate Investigator, Group Health Research Institute

Lurdes Inoue, PhD
Associate Professor, Dept. of Biostatistics, University of Washington

Schedule:

1. **Wed 1:30-3:00 pm**
   Simple Linear Regression
2. **Wed 3:30-5:00 pm**
   Lab: Introduction to R and Simple Linear Regression
3. **Thurs 8:30-10:00 am**
   Prediction and Model Checking
4. **Thurs 10:30 am-12:00 pm**
   Multiple Linear Regression
5. **Thurs 1:30-3:00 pm**
   Lab: Model Checking and Multiple Linear Regression
6. **Thurs 3:30-5:00 pm**
   One-Way ANOVA
7. **Fri 8:30-10:00 am**
   Two-Way ANOVA
8. **Fri 10:30 am-12:00 pm**
   Lab: One-Way and Two-Way ANOVA
9. **Fri 1:30-3:00 pm**
   ANCOVA; Experimental Design [if time permits]
10. **Fri 3:30-5:00 pm**
    Lab: ANCOVA

Email: hubbard.r@ghc.org/ linoue@uw.edu

UW School of Public Health
Department of Biostatistics

REGRESSION AND ANALYSIS OF VARIANCE

Motivation

- Objective: Investigate associations between two or more variables
- What tools do you already have?
  - T-test
    - Comparison of means in two populations
- What will we cover in this module?
  - Linear Regression
    - Association of a continuous outcome with one or more predictors (categorical or continuous)
  - Analysis of Variance
    - Comparison of a continuous outcome over a fixed number of groups

UW School of Public Health
Department of Biostatistics

REGRESSION MODELS

SIMPLE LINEAR REGRESSION

Outline: Simple Linear Regression

- Motivation
- The equation of a straight line
- Least Squares Estimation
- Inference
  - About regression coefficients
  - About predictions
- Model Checking
  - Residual analysis
  - Outliers versus Influential observations

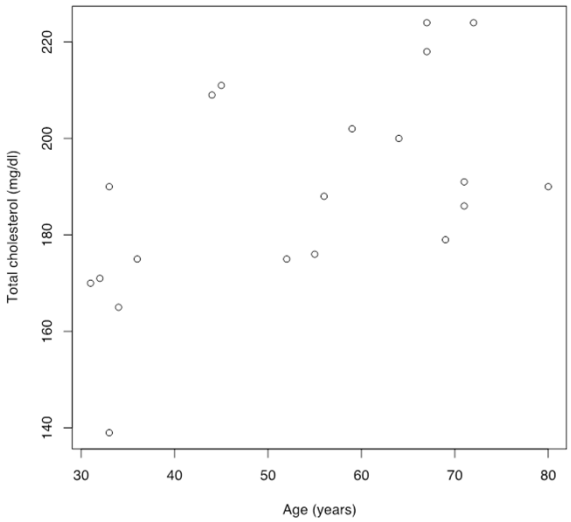
Motivation: Cholesterol Example

- Data: Factors affecting serum total cholesterol

	sex	**age**	**chol**	BMI	TG	apoE	rs174548	rs4775401
1	1	74	215	26.2	367	4	1	2
2	1	51	204	24.7	150	4	2	1
3	0	64	205	24.2	213	4	0	1
4	0	34	182	23.8	111	1	1	1
5	1	52	175	34.1	328	1	0	0
6	1	39	176	22.7	53	4	0	2

- Our goal:
  - Investigate the relationship between cholesterol (mg/dl) and age in adults

Motivation: Cholesterol Example

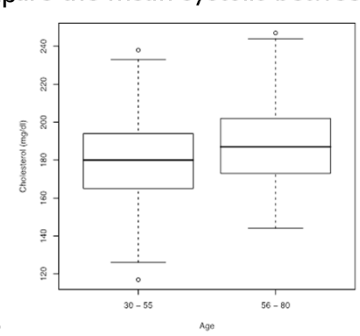


Motivation: Cholesterol Example

- Is serum cholesterol associated with age?
  - You could dichotomize age and compare the mean cholesterol between two groups: t-test

Motivation: Cholesterol Example

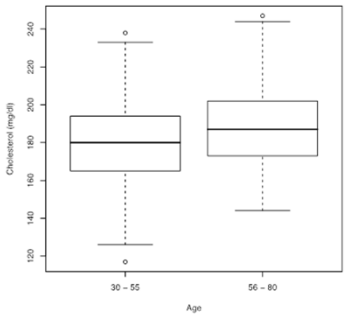
- Is cholesterol associated with age?
  - You could dichotomize age and compare the mean systolic between two groups: t-test




```
> group = 1*(age > 55)
> t.test(chol ~ group)

         Welch Two Sample t-test

data:  chol by group
t = -3.637, df = 393.477, p-value = 0.0003125
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.200209  -3.638487
sample estimates:
mean in group 0 mean in group 1
       179.9751        187.8945
```

Motivation: Cholesterol Example

- Question: What does this plot and t-test tell us about the relationship between age and cholesterol?



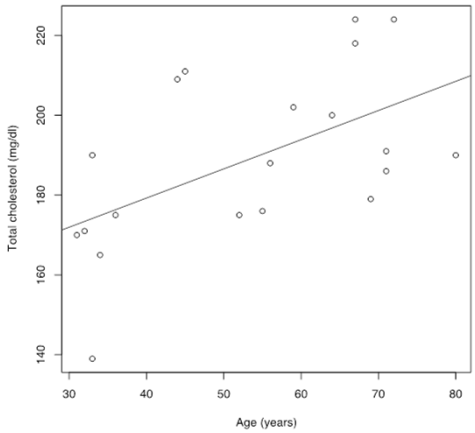

```
> group = 1*(age > 55)
> t.test(chol ~ group)

        Welch Two Sample t-test

data:  chol by group
t = -3.637, df = 393.477, p-value = 0.0003125
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.200209  -3.638487
sample estimates:
mean in group 0 mean in group 1
       179.9751        187.8945
```


Motivation: Cholesterol Example

- Using t-test:
  - There is a statistical association between cholesterol and age
  - There appears to be a positive association between cholesterol and age
    - Is there any way we could estimate the magnitude of this association without breaking the "continuous" measure of age into subgroups?

Motivation: Cholesterol Example

- Can we find the equation for a straight line



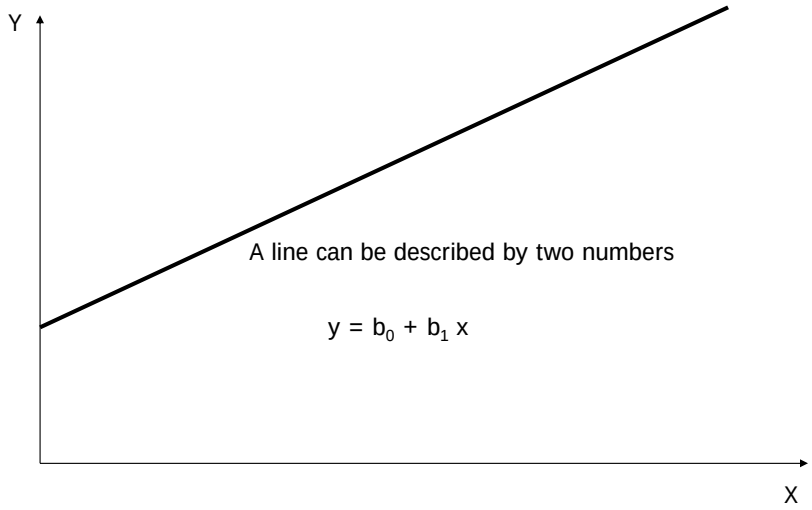
that best fits these data?

Linear Regression

- Statistical method for modeling the relationship between a continuous variable [response/outcome/dependent] and other variables [predictors/exposure/independent]
  - Most commonly used statistical model
  - Flexible
  - Well-developed and understood properties
  - Easy interpretation
  - Building block for more general models
- Goals of analysis:
  - Study the association between response and predictors

  or,
  - Predict response values given the values of the predictors.
- We will start our discussion studying the relationship between a response and <u>a single predictor</u>
  - Simple linear regression model

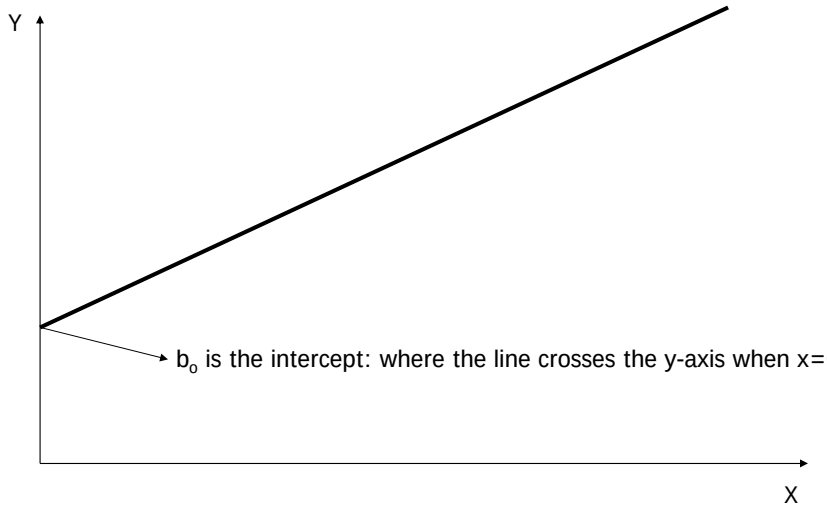
The straight line equation



A line can be described by two numbers

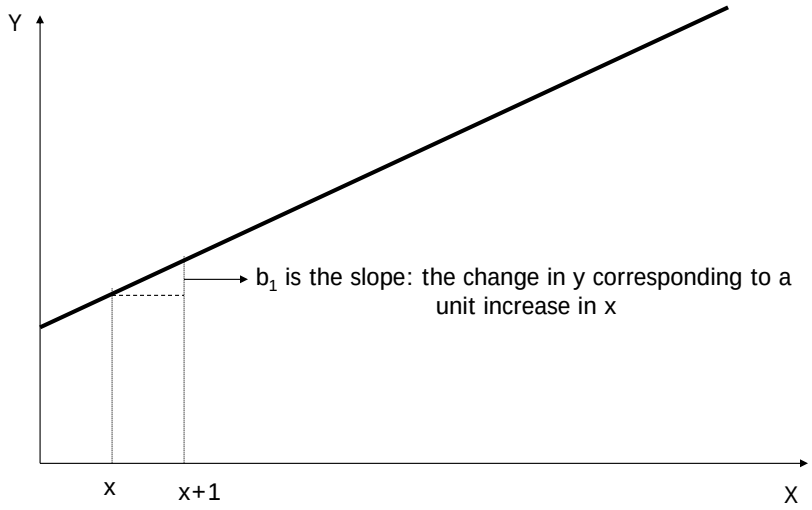
$y = b_0 + b_1 x$

The straight line equation

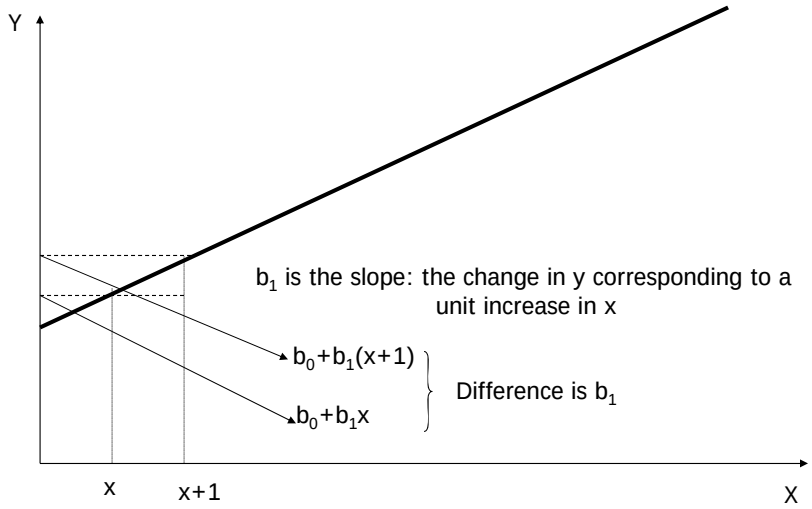


b_o is the intercept: where the line crosses the y-axis when x=0

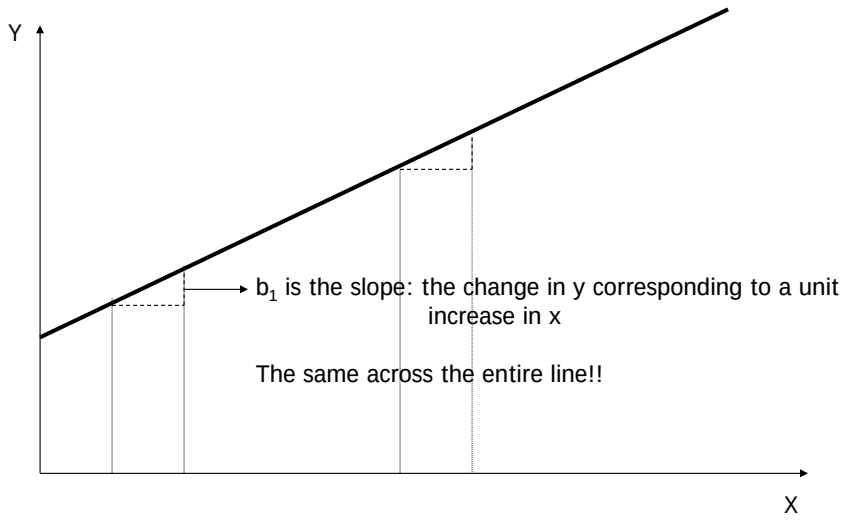
The straight line equation



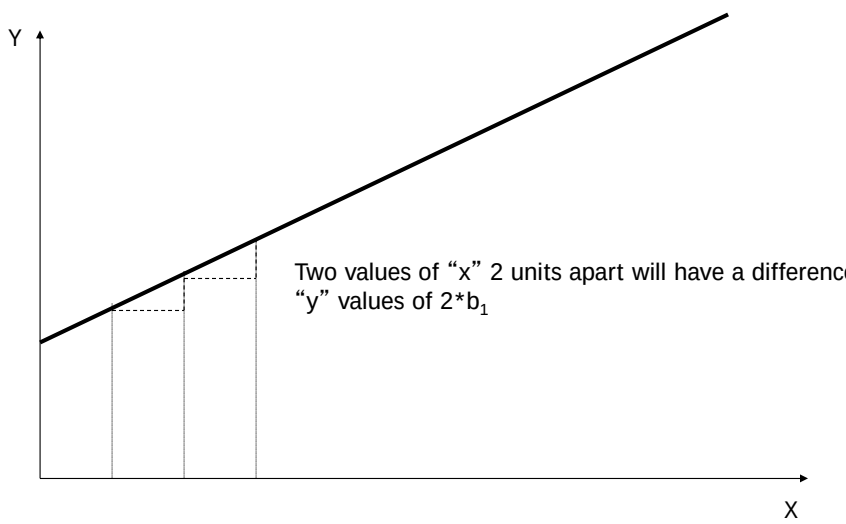
The straight line equation



The straight line equation



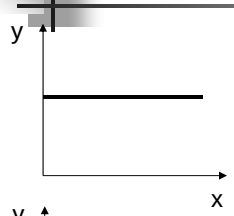
The straight line equation



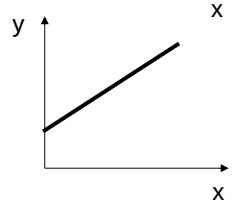
The straight line equation

- Slope b_1 is the change in y corresponding to a unit increase in x
- Slope gives information about magnitude and direction of the association between x and y

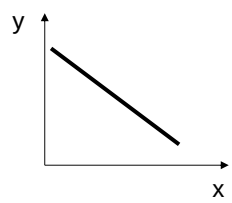
The straight line equation



(b_1=0) No association between x and y
(values of y are the same regardless of x)



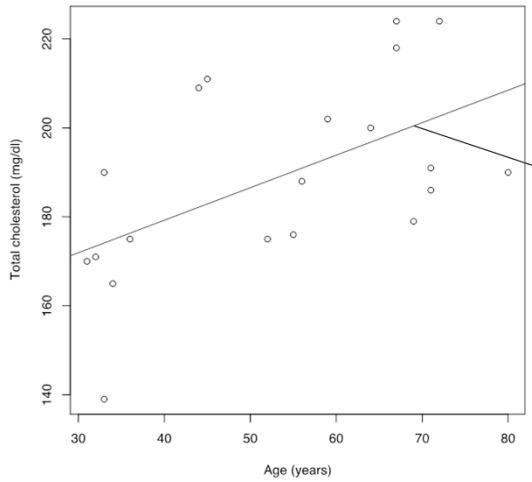
(b_1 > 0) Positive association between x and y
(values of y increase as values of x increase)



(b_1 < 0) Negative association between x and y
(values of y decrease as values of x increase)

Simple Linear Regression

- Dealing with situations where points don't fit exactly to the straight line



We estimate a straight line describing trends in the **mean** of an outcome Y as a function of predictor X

Simple Linear Regression

- In **regression:**
  - *X* is used to predict or explain outcome *Y*.
- **Response** or **dependent** variable (Y):
  - variable we want to predict or explain
- **Explanatory** or **independent** variable (X):
  - attempts to explain the response
- **Simple Linear Regression Model**:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

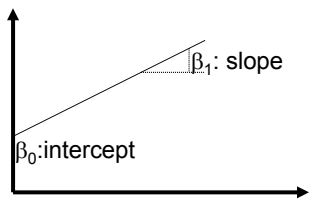
Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Model consists of two components:

•Systematic component: $E[Y \mid X = x] = \beta_0 + \beta_1 x$

Mean population value of Y at X=x



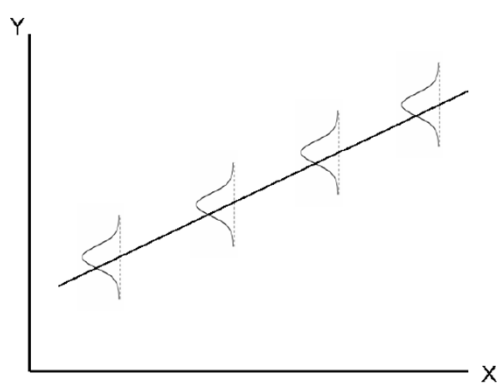
•Random component: $Var[Y \mid X = x] = \sigma^2$

Variance does not depend on x

Simple Linear Regression: Assumptions

MODEL: $E[Y \mid X = x] = \beta_0 + \beta_1 x$ $\quad Var[Y \mid X = x] = \sigma^2$

Distribution of Y at different x values:



Simple Linear Regression: Interpreting model coefficients

- Model: $E[Y|x] = \beta_0+\beta_1 x \quad Var[Y|x] = \sigma^2$

- Question: How do you interpret β_0?
- Answer:
  - $\beta_0 = E[Y|x=0]$, that is, the mean response when x=0

Your turn: interpret β_1!

Simple Linear Regression: Interpreting model coefficients

- Model: $E[Y|x] = \beta_0+\beta_1 x \quad Var[Y|x] = \sigma^2$

- Question: How do you interpret β_1?
- Answer:

$E[Y|x] = \beta_0 + \beta_1 x$

$E[Y|x+1] = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1$

$E[Y|x+1] - E[Y|x] = \beta_1$ independent of x (linearity)

i.e. β_1 is the difference in the mean response associated with a one unit positive difference in x

Example: Cholesterol and age

- Recall: Our motivating example was to determine if there is an association between age (a continuous predictor) and cholesterol (a continuous outcome)
- Suppose: We believe they are associated via the linear relationship $E[Y|x] = \beta_0+\beta_1 x$
- Question: How would you interpret β_1?
- Answer:

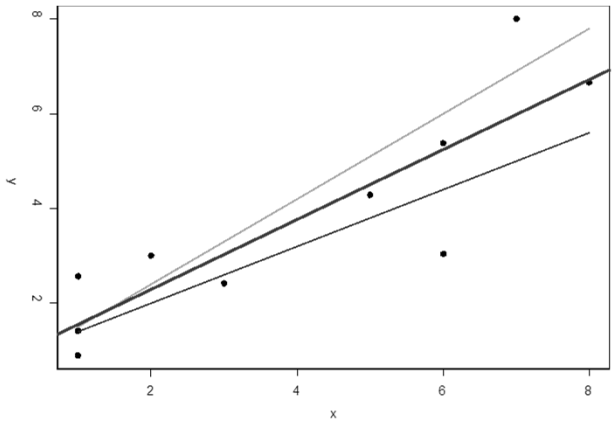
Example: Cholesterol and age

- Recall: Our motivating example was to determine if there is an association between age (a continuous predictor) and cholesterol (a continuous outcome)
- Suppose: We believe they are associated via the linear relationship $E[Y|x] = \beta_0+\beta_1 x$
- Question: How do you interpret β_1?
- Answer:

  β_1 is the difference in mean serum cholesterol associated with a one year increase in age

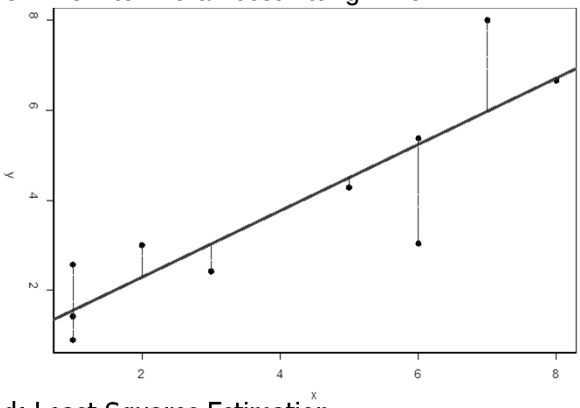
Least Squares Estimation

- Question: How to find a “best-fitting” line?



Least Squares Estimation

- Question: How to find a “best-fitting” line?



- Method: Least Squares Estimation
  - Idea: minimizes the sum of squares of the vertical distances from the observed points to the least squares regression line.

Least Squares Estimation

- The least squares regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- So the (squared) distance between the data (y) and the least squares regression line is

$$D = \sum_i (y_i - \hat{y}_i)^2$$

- We estimate β_0 and β_1 by finding the values that minimize D

Least Squares Estimation

- These values are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

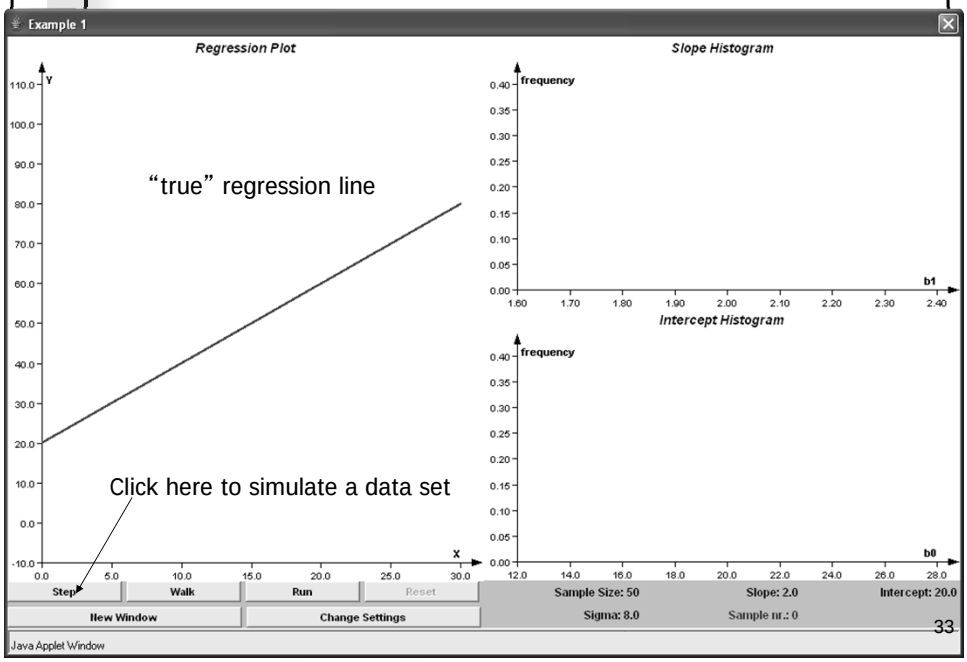
- We estimate the variance as

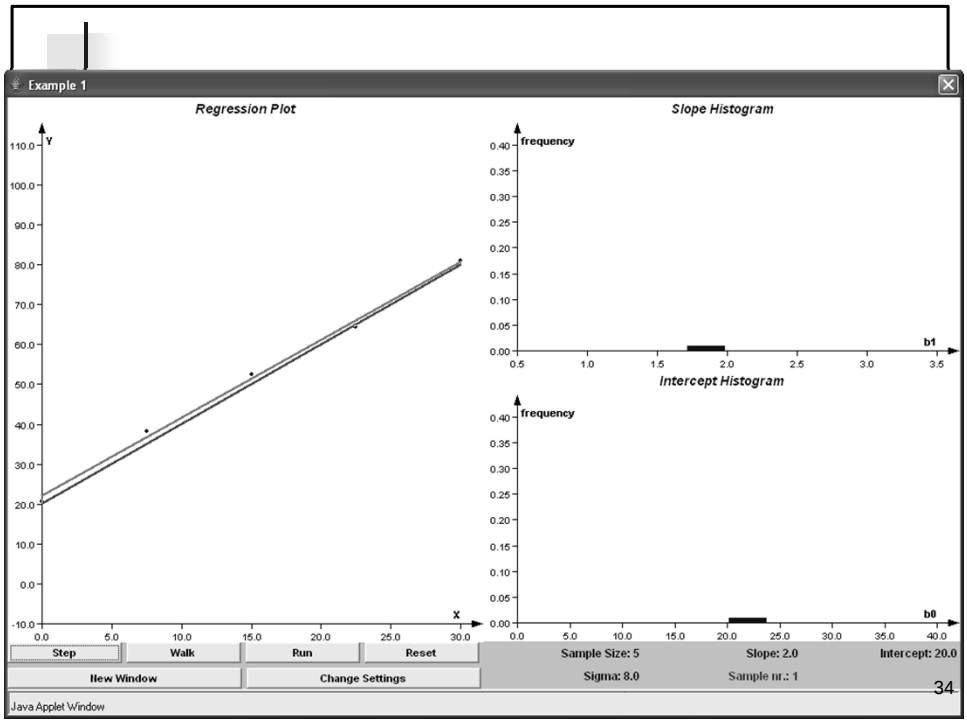
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

Estimated Standard Errors

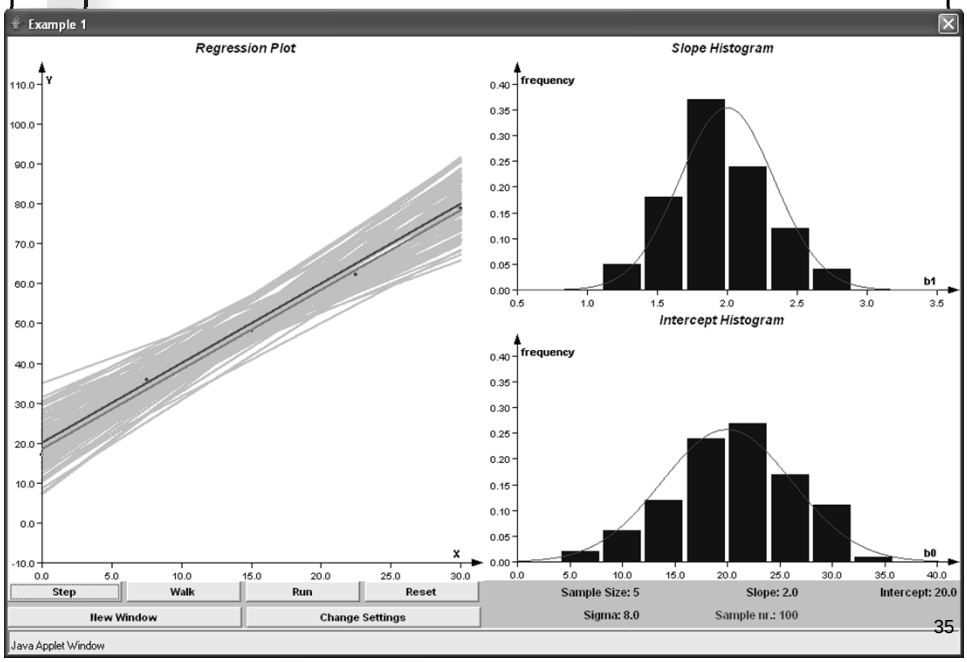
- Recall that when estimating parameters, sampling variability exists in our estimates
- Same is true for regression parameter estimates
- Looking at the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$, we can see that these are just complicated means
- In repeated sampling we would get different estimates
- Knowledge of sampling distribution of parameter estimates can help us make inference about the line

http://lstat.kuleuven.be/java/version2.0/Applet003.html

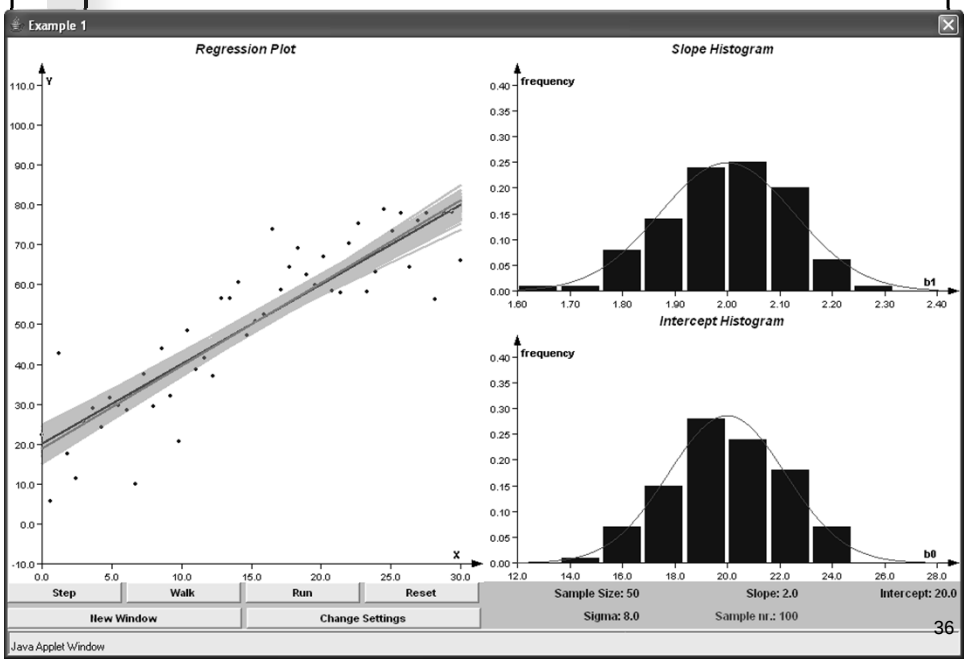


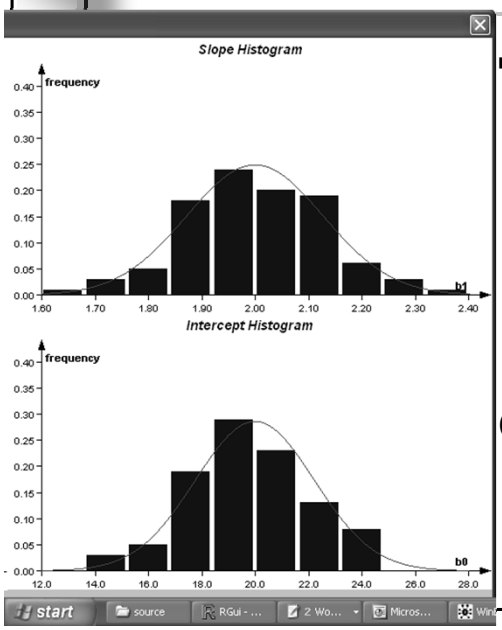


Sampling Distribution



Sampling Distribution





- From statistical theory

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2\left(\frac{1}{(n-1)s_x^2}\right)\right)$$

(in the figure, the theoretical distribution is shown in pink)

Estimated Standard Errors

- Estimate the variability of $\hat{\beta}_0$, $\hat{\beta}_1$ in repeated sampling

$$SE(\hat{\beta}_0)=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x^2}}$$

$$SE(\hat{\beta}_1)=\hat{\sigma}\sqrt{\frac{1}{(n-1)s_x^2}}$$

Inference

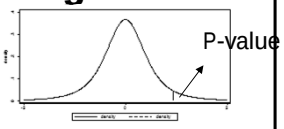
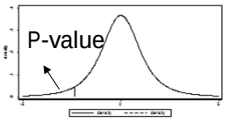
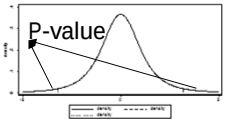
- About regression model parameters
  - Hypothesis testing: H_0: $\beta_j=0$
    - Test Statistic:
      - Large Samples: $\frac{\hat{\beta}_j-(null\ hyp)}{se(\hat{\beta}_j)}\sim N(0,1)$
      - Small Samples: $\frac{\hat{\beta}_j-(null\ hyp)}{se(\hat{\beta}_j)}\sim T_{n-2}$
  - Confidence Intervals:

$$\hat{\beta}_j \pm (critical\ value)\times se(\hat{\beta}_j)$$

[Don't worry about these formulae: we will use R to fit the model!]

Inference: Hypothesis Testing

Null Hypothesis: $\beta_j = 0$

Alternative	**P-Value**	**Figure**
$\beta_j > 0$	$P(T_{n-2} > T)$	
$\beta_j < 0$	$P(T_{n-2} < T)$	
$\beta_j \neq 0$	$2P(T_{n-2} > \lvert T \rvert)$	

Inference: Confidence Intervals

100 (1-α)% Confidence Interval for β_j (j=0,1)

$$\hat{\beta}_j \pm t_{n-2,\alpha/2} SE(\hat{\beta}_j)$$

Gives intervals that (1- α)100% of the time will cover the true parameter value (β_0 or β_1).

We say we are "(1- α)100% confident" the interval covers β_j.

Example:
Scientific Question: Is cholesterol associated with age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       Max
-60.45306 -14.64250  -0.02191  14.65925  58.99527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.90168    4.26488  39.134  < 2e-16 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.69 on 398 degrees of freedom
Multiple R-squared: 0.04099,      Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF,  p-value: 4.522e-05
```

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age       0.1624211   0.4582481
```


Example:
Scientific Question: Is cholesterol associated with age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       Max
-60.45306 -14.64250  -0.02191  14.65925  58.99527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.90168    4.26488  39.134  < 2e-16 ***
age          0.31033      0.07524   4.125 4.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.69 on 398 degrees of freedom
Multiple R-squared: 0.04099,      Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF,  p-value: 4.522e-05
```

Estimates of the model parameters and standard errors

$\hat{\beta}_0 = 166.90$; $se(\hat{\beta}_0) = 4.26$

$\hat{\beta}_1 = 0.31$; $se(\hat{\beta}_1) = 0.08$

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age         0.1624211   0.4582481
```

Example:
Scientific Question: Is cholesterol associated with age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       Max
-60.45306 -14.64250  -0.02191  14.65925  58.99527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.90168    4.26488  39.134  < 2e-16 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.69 on 398 degrees of freedom
Multiple R-squared: 0.04099,      Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF,  p-value: 4.522e-05
```

95% Confidence intervals

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age           0.1624211   0.4582481
```


Example:
Scientific Question: Is cholesterol associated with age?

- What do these models results mean in terms of our scientific question?
  - Parameter estimates and confidence intervals:

    $\hat{\beta}_0 = 166.90$ 95% CI: (158.5, 175.3)

    $\hat{\beta}_1 = 0.31$ 95% CI: (0.16, 0.46)
  - Answer: $\hat{\beta}_0$: The estimated average serum cholesterol for someone of age = 0 is 166.9
  - Your turn: What about $\hat{\beta}_1$?

Example:

Scientific Question: Is cholesterol associated with age?

- What do these models results mean in terms of our scientific question?
  - Parameter estimates and confidence intervals:

$\hat{\beta}_0 = 166.90$ 95% CI: (158.5, 175.3)

$\hat{\beta}_1 = 0.31$ 95% CI: (0.16, 0.46)

  - Answer: $\hat{\beta}_1$: mean cholesterol is estimated to differ by 0.31 mg/dl for each one year difference in age.
  - Question: What about the confidence intervals?

Example:

Scientific Question: Is cholesterol associated with age?

- What do these models results mean in terms of our scientific question?
  - Parameter estimates and confidence intervals:

$\hat{\beta}_0 = 166.90$ 95% CI: (158.5, 175.3)

$\hat{\beta}_1 = 0.31$ 95% CI: (0.16, 0.46)

  - Answer: 95% CIs give us a range of values that will cover the true intercept and slope 95% of the time
    - For instance, we can be 95% confident that the true difference in mean cholesterol associated with a one year difference in age lies between 0.16 and 0.46 mg/dl

Example:

Scientific Question: Is cholesterol associated with age?

- Presentation of the results?
  - The mean serum total cholesterol is significantly higher in older individuals ($p < 0.001$). For each additional year of age, we estimate that the mean total cholesterol differs by approximately 0.31 mg/dl (95% CI: 0.16, 0.46).
    - Note:
      - Emphasis on slope parameter (sign and magnitude)
      - Confidence interval
      - Units for predictor and response

Some basics of R syntax

- In the labs we will be using R to analyze data using the concepts we have been discussing
- The file rcommands-analysis.R on your thumb drive provides some basic R commands
- In R we use variables or "objects" to store data, model results, functions...
- We manipulate objects using functions
  - Functions always consist of a name followed by parentheses
  - "Arguments" inside parentheses specify what we want the function to do; arguments are separated by a comma
  - For instance, in `lm(chol ~ age, data = cholesterol)`,
    - `lm()` is a function
    - Its first argument specifies a regression formula
    - Its second argument specifies the data set to use to fit the model
  - To get more information about a function or its arguments use `help(function_name)` or `?function_name`

Some basics of R syntax

- Some basic operators
  - # : comment, lines beginning with # will be ignored by R
  - = or <- : "assignment operators," tells R to store the information on the right hand side in the object on the left hand side
    - For instance `x <- 5` or `x = 5` tells R to store the number 5 in the object x
    - Or `fit = lm(chol ~ age)` tells R to store the results of a linear model in the object fit
  - $: Used to select a subset of an object by name
    - For instance `fit$coef` refers to the coefficients of the linear model stored in `fit`
    - You can also select subsets of certain types of objects using square brackets; for instance `fit$coef[1]` refers to the first element in the vector of coefficients stored in `fit$coef`

Inference for predictions

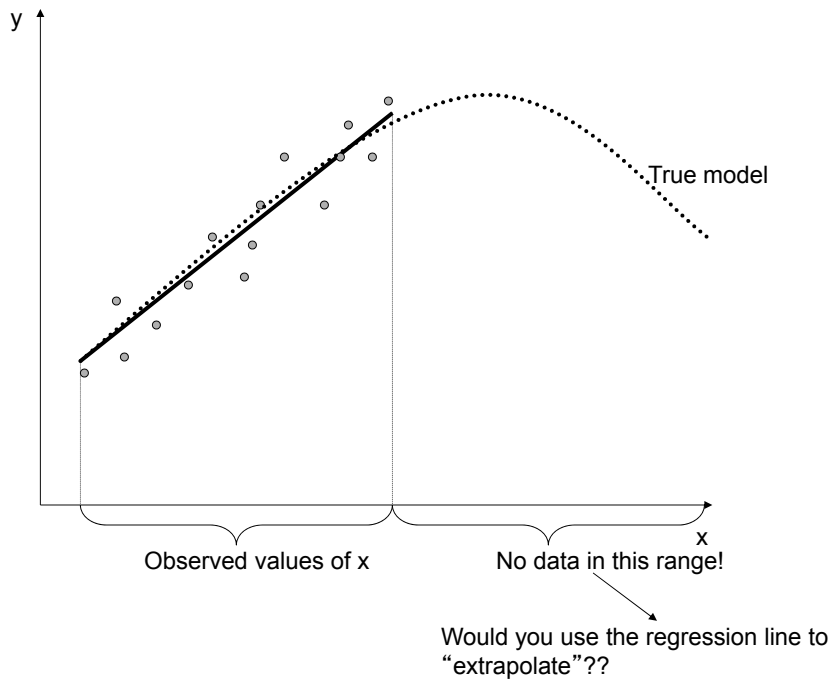
- Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$ we can find the **predicted value**, $\hat{y}_i$ for any value of x_i as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Interpretation of $\hat{y}_i$:
  - Estimated mean value of Y at $X = x_i$.

<u>Be Cautious: It assumes the model is true</u>.

- May be a reasonable assumption within the range of your data.
- It may not be true outside the range of your data!!



Prediction

- Prediction of the mean E[Y|X=x]:
  - Point Estimate: $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$
  - Standard Error: $se(\hat{y})=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

    Note that as x diverges from $\bar{x}$, variance increases!
  - 100 (1-α)% confidence interval for E[Y|X=x]:

    $$\hat{y} \pm t_{n-2,1-\alpha/2}\, se(\hat{y})$$

Prediction

- Prediction of a <u>new future observation</u>, y*, at X=x:
  - Point Estimate: $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Standard Error: $se(\hat{y}^*) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
  - 100 (1-α)% prediction interval for a new future observation: $\hat{y}^* \pm t_{n-2,1-\alpha/2}\, se(\hat{y}^*)$

    Standard error for the prediction of a future observation is bigger:

    It depends not only on the precision of the estimated mean, but also on the amount of variability in Y around the line.

Cholesterol Example: Prediction

Prediction of the mean

```
> predict.lm(fit, newdata=data.frame(age=c(46,47,48)), interval="confidence")
       fit      lwr      upr
1 181.1771 178.6776 183.6765
2 181.4874 179.0619 183.9129
3 181.7977 179.4392 184.1563

> predict.lm(fit, newdata=data.frame(age=c(46,47,48)), interval="prediction")
       fit      lwr      upr
1 181.1771 138.4687 223.8854
2 181.4874 138.7833 224.1915
3 181.7977 139.0974 224.4981
```

Prediction of a new observation

Example:

Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For x = 46

    $\hat{y} = 181.2$ 95% CI: (178.7, 183.7)

    $\hat{y}^* = 181.2$ 95% CI: (138.5, 223.9)
  - Question: How do our interpretations for $\hat{y}$ and $\hat{y}^*$ differ?

Example:

Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For x = 46

    $\hat{y} = 181.2$ 95% CI: (178.7, 183.7)

    $\hat{y}^* = 181.2$ 95% CI: (138.5, 223.9)
  - Question: How do our interpretations for $\hat{y}$ and $\hat{y}^*$ differ?
  - Answer: The point estimates represent our predictions for the mean serum cholesterol for individuals age 46 ($\hat{y}$) and for a single new individual of age 46 ($\hat{y}^*$)

Example:

Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For x = 46

    $\hat{y} = 181.2$ 95% CI: (178.7, 183.7)

    $\hat{y}^* = 181.2$ 95% CI: (138.5, 223.9)
  - Question: Why are the confidence intervals for $\hat{y}$ and $\hat{y}^*$ of differing widths?

Example:

Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For x = 46

    $\hat{y} = 181.2$ 95% CI: (178.7, 183.7)

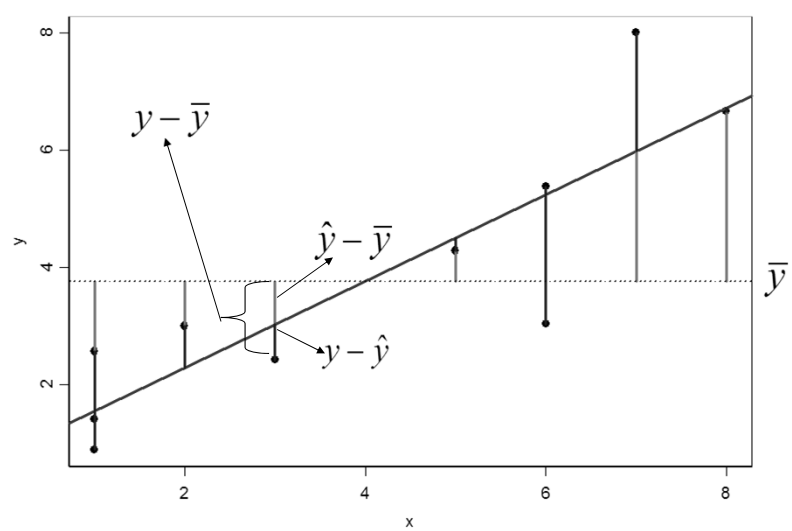
    $\hat{y}^* = 181.2$ 95% CI: (138.5, 223.9)
  - Question: Why are the confidence intervals for $\hat{y}$ and $\hat{y}^*$ of differing widths?
  - Answer: The interval is broader when we make a prediction for a single individual because it must incorporate random variability around the mean.

Simple Linear Regression: R^2

- Given no linear association:
  - We could simply use the sample mean to predict E(Y). The variability using this simple prediction is given by SST.
- Given a linear association:
  - The use of X permits a potentially better prediction of Y by using E(Y|X).
  - **Question:** What did we gain by using X?

Let's examine this question with the following figure

Decomposition of sum of squares



Decomposition of sum of squares

It is always true that: $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

It can be shown that:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SST \quad = \quad SSE \quad + \quad SSR$$

SST: describes the total variation of the Y_i.

SSE: describes the variation of the Y_i around the regression line.

SSR: describes the structural variation; how much of the variation is due to the regression relationship.

This decomposition allows a characterization of the usefulness of the covariate *X* in predicting the response variable *Y*.

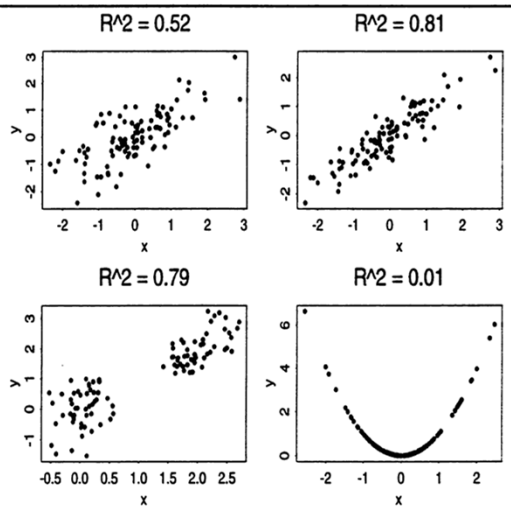
Simple Linear Regression: R^2

- Given no linear association:
  - We could simply use the sample mean to predict E(Y). The variability between the data and this simple prediction is given as SST.
- Given a linear association:
  - The use of *X* permits a potentially better prediction of Y by using E(*Y*| *X*).
  - **Question:** What did we gain by using *X*?
  - **Answer:** We can answer this by computing the proportion of the total variation that can be explained by the regression on *X*

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- This R^2 is, in fact, the correlation coefficient squared.

Examples of R^2



Low values of R^2 indicate that the model is not adequate. However, high values of R^2 do not mean that the model is adequate!!

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       Max
-60.45306 -14.64250  -0.02191  14.65925  58.99527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.90168    4.26488  39.134  < 2e-16 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.69 on 398 degrees of freedom
Multiple R-squared: 0.04099,       Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF,  p-value: 4.522e-05
```

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age           0.1624211   0.4582481
```

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- R^2=0.04
- What does R^2 tell us about our model for cholesterol?

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- R^2=0.04
- What does R^2 tell us about our model for cholesterol?
- Answer: 4% of the variability in cholesterol is explained by age. Although mean cholesterol increases with age, there is much more variability in cholesterol than age alone can explain

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- Decomposition of Sum of Squares and the F-statistic

Degrees of freedom

Decomposition of the Sum of Squares

Mean Squares: SS/df

F-statistic: MSR/MSE

```
> anova(fit)
Analysis of Variance Table

Response: chol
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1    8002  8001.7  17.013 4.522e-05 ***
Residuals 398 187187   470.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSR=age

SSE=Residuals

In simple linear regression:

F-statistic = (t-statistic for slope)^2

Hypothesis being tested: H_0: $\beta_1=0$, H_1: $\beta_1 \neq 0$.

Simple Linear Regression: Assumptions

1. E[Y|x] is related linearly to x
2. Y’s are independent of each other
3. Distribution of [Y|x] is normal
4. Var[Y|x] does not depend on x

Linearity
Independence
Normality
Equal variance

Can we assess if these assumptions are valid?

Model Checking: Residuals

- **(Raw or unstandardized) Residual**: difference (r_i) between the observed response and the predicted response, that is,

$$r_i = y_i - \hat{y}_i$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

The residual captures the component of the measurement y_i that cannot be "explained" by x_i.

Model Checking: Residuals

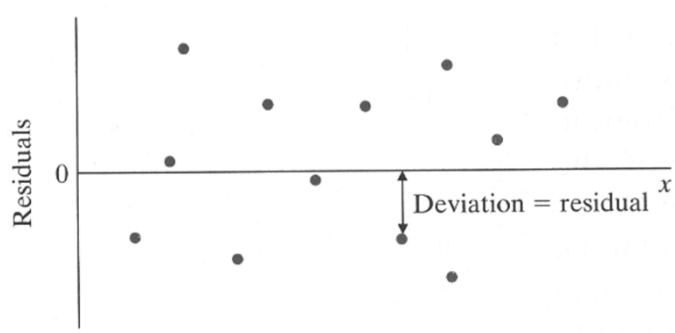
- Residuals can be used to
  - Identify poorly fit data points
  - Identify unequal variance (heteroscedasticity)
  - Identify nonlinear relationships
  - Identify additional variables
  - Examine normality assumption

Model Checking: Residuals

Linearity	Residual vs X or vs $\hat{Y}$ Q: Is there any trend?
Independence	Q: Any scientific concerns?
Normality	Residual histogram or qq-plot Q: Symmetric? Normal?
Equal variance	Residual vs X Q: Is there any pattern?

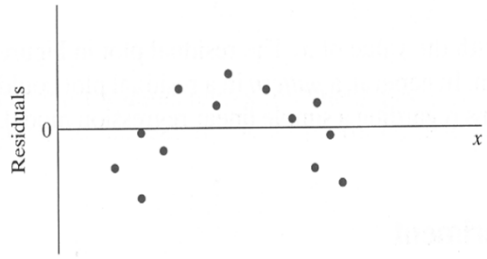
Model Checking: Residuals

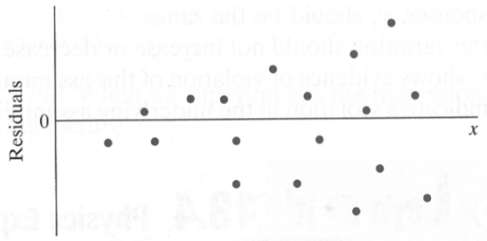
- If the linear model is appropriate we should see an **unstructured horizontal band of points centered at zero** as seen in the figure below



Model Checking: Residuals

The model does not provide a good fit in these cases!





Violations of the model assumptions? How?

Simple Linear Regression: Residual Analysis: Non-normality of errors

- QQ-plot
  - Graphical technique that allows us to assess whether or not a data set follows a given distribution (such as the normal distribution)
  - The data are plotted against a given theoretical distribution
    - Points should approximately fall in a straight line
    - Departures from the straight line indicate departures from the specified distribution.

Simple Linear Regression: Residual Analysis: Non-normality of errors

- Construction of QQ-Plot: (an example)

residuals
0.30
-0.25
-0.91
0.56
-0.79
-1.45
-0.42
-0.80
-0.39
-1.09
0.37
-0.56
1.15
-1.14
0.06
0.60
0.11
0.51
-0.02
-0.24

Sort →

sorted	index(i)
-1.45	1
-1.14	2
-1.09	3
-0.91	4
-0.80	5
-0.79	6
-0.56	7
-0.42	8
-0.39	9
-0.25	10
-0.24	11
-0.02	12
0.06	13
0.11	14
0.30	15
0.37	16
0.51	17
0.56	18
0.60	19
1.15	20

Empirical Probab. →

pr1 (i/n)	pr2 ((i-0.5)/n)
0.05	0.025
0.10	0.075
0.15	0.125
0.20	0.175
0.25	0.225
0.30	0.275
0.35	0.325
0.40	0.375
0.45	0.425
0.50	0.475
0.55	0.525
0.60	0.575
0.65	0.625
0.70	0.675
0.75	0.725
0.80	0.775
0.85	0.825
0.90	0.875
0.95	0.925
1.00	0.975

Get z-values (quantiles from Normal distr.) →

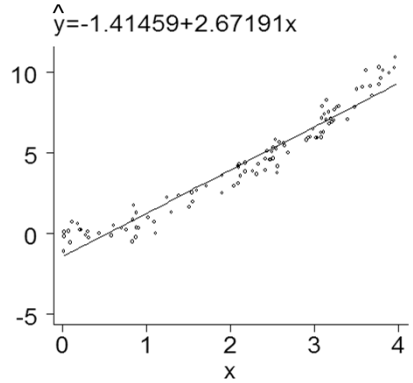
Z-quantile
-1.96
-1.44
-1.15
-0.93
-0.76
-0.60
-0.45
-0.32
-0.19
-0.06
0.06
0.19
0.32
0.45
0.60
0.76
0.93
1.15
1.44
1.96

Plot of sorted residuals (sample quantiles) versus z-quantile (theoretical quantiles) = QQ-plot

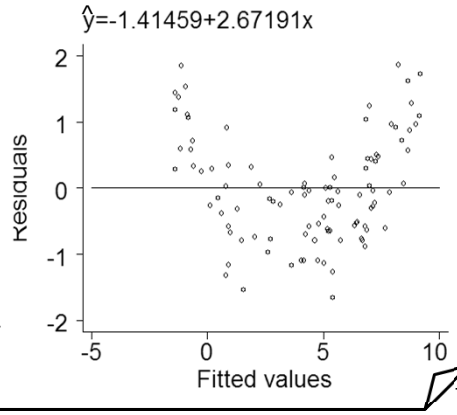
Simple Linear Regression: Residual Analysis: Nonlinear Association

True model: $y = x$^1.7

Plot of Fitted Model:



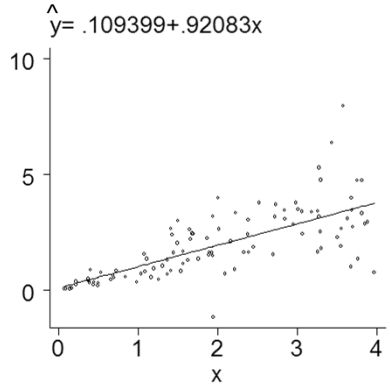
Plot fitted (prediction) vs. residual:



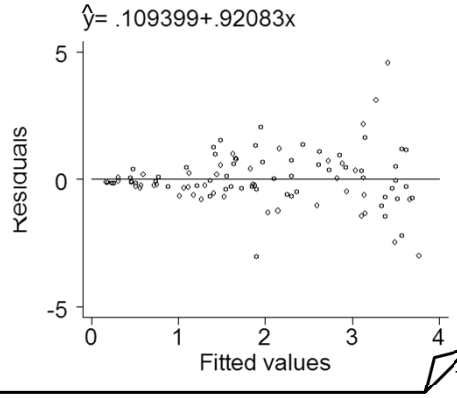
Simple Linear Regression: Residual Analysis: Non Constant Variance

True model: $y = x$ + errors increasing with x

Plot of Fitted Model:



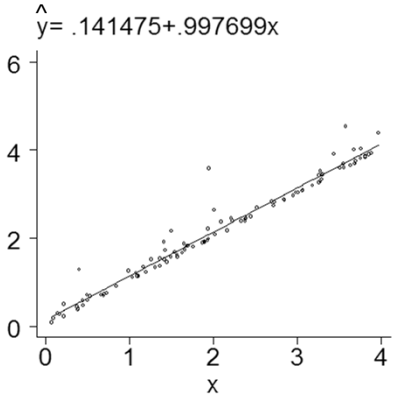
Plot fitted (prediction) vs. residual:



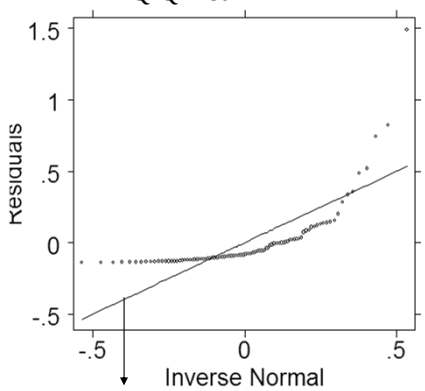
Simple Linear Regression: Residual Analysis: Non-normality of errors

True model: $y = x$ + chi-squared errors

Plot of Fitted Model:



Q-Q Plot

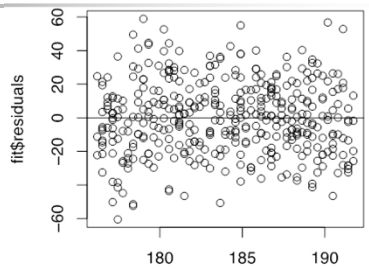


Under normality, residuals should fall on the straight line!

Cholesterol Example: Residuals

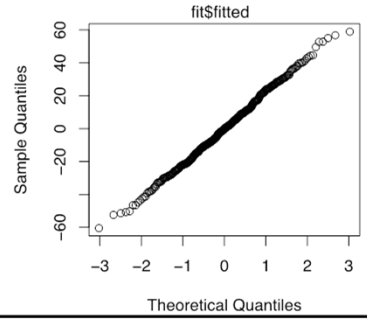
Plot of residuals versus fitted values
  Curvature?
  Heteroscedasticity?

R COMMANDS:
plot(fit$fitted, fit$residuals)



Plot of residuals versus quantiles of a normal distribution(for n > 30)
  Normality?

R COMMANDS:
qqnorm(fit$residuals)



Non-constant variance

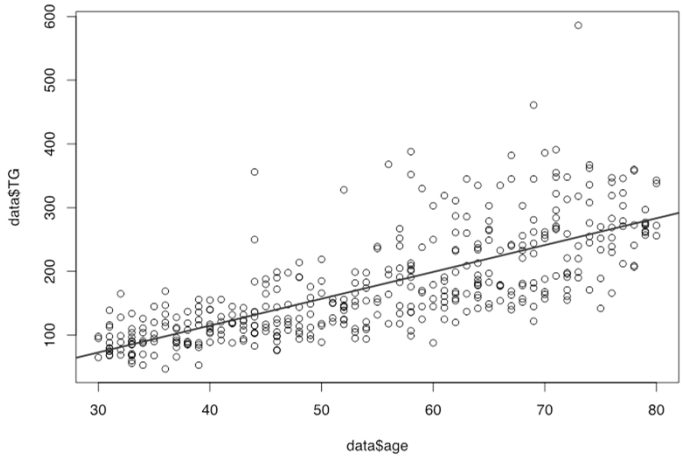
- Sometimes variance of y is not constant across the range of x (heteroscedasticity)
- Little effect on point estimates but variance estimates will be incorrect
- This affects confidence intervals and p-values
- To account for heteroscedasticity we can
  - Use robust standard errors
  - Transform the data
  - Fit a model that does not assume constant variance (GLM)

Robust standard errors

- Robust standard errors correctly estimate variability of parameter estimates even under non-constant variance
- Regression point estimates will be unchanged
- Robust or empirical standard errors will give correct confidence intervals and p-values

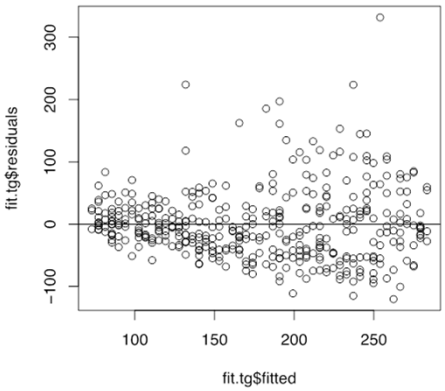
Cholesterol example: Robust standard errors

- Linear regression for association between age and triglycerides



Cholesterol example: Robust standard errors

- Residuals analysis suggests mean-variance relationship
- Use robust standard errors to get correct variance estimates



Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
age           4.2090     0.1964  21.429  < 2e-16 ***
```

Point estimates are unchanged

- Results incorporating robust SEs:

```
> summary(fit.tg.ese)

Call:
gee(formula = TG ~ age, id = seq(1, length(age)), data = data)

Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -53.305930 11.1339178 -4.787706   8.7387366 -6.099958
age           4.208964  0.1964165 21.428771   0.1813358 23.210880
```

Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
age           4.2090     0.1964  21.429  < 2e-16 ***
```

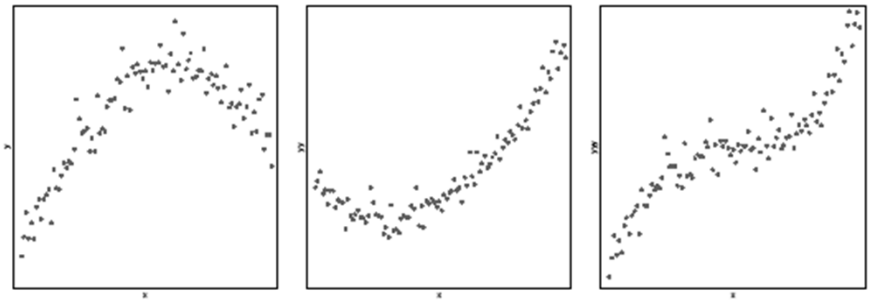
Standard errors are corrected

- Results incorporating robust SEs:

```
> summary(fit.tg.ese)

Call:
gee(formula = TG ~ age, id = seq(1, length(age)), data = data)

Coefficients:
               Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -53.305930 11.1339178 -4.787706   8.7387366 -6.099958
age           4.208964  0.1964165 21.428771   0.1813358 23.210880
```


Transformations

- Sometimes the relationship between Y and X is not linear



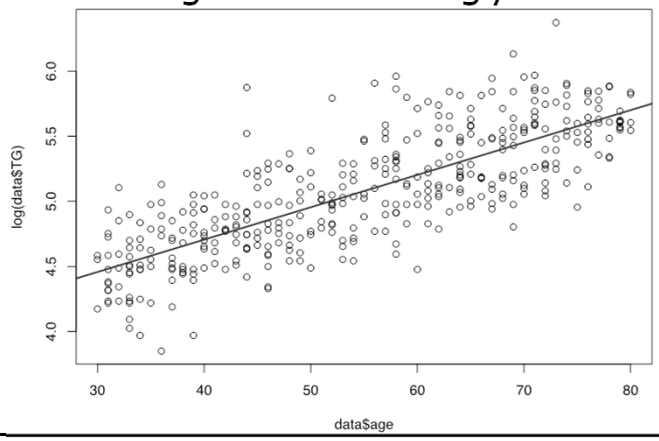
To model "curvilinear relationships" one can look at transformations in X or Y [or both]

Transformations

- Some reasons for using data transformations
  - Original data suggest nonlinearity
  - Equal variance assumption violated
  - Normality assumption violated
- Transformations may be applied to the response, predictor or both
  - Be careful with the interpretation of the results

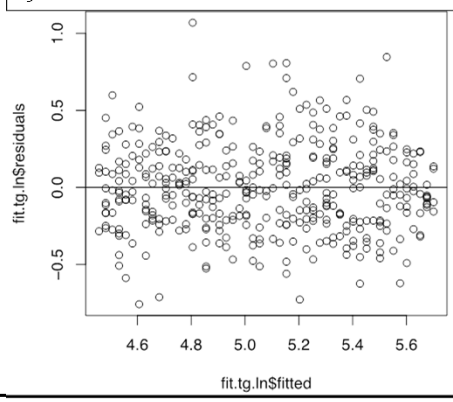
Cholesterol example: Transformations

- We have seen that triglycerides are associated with age but display non-constant variance
- What about log transformed triglycerides?



Cholesterol example: Transformations

```
> summary(fit.tg.ln)

Call:
lm(formula = log(TG) ~ age)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.7115803  0.0559237   66.37   <2e-16 ***
age         0.0248646  0.0009866   25.20   <2e-16 ***
```



- Heteroscedasticity is corrected
- But interpretation of model is more complicated

Transformations

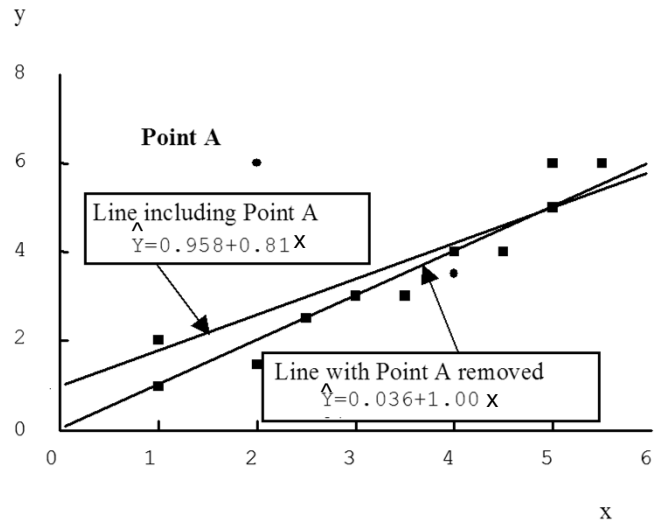
- Rarely do we know which transformation of the predictor provides best "linear" fit
  - As always, there is a danger in using the data to estimate the best transformation to use
    - If there is no association of any kind between the response and the predictor, a "linear" fit (with a zero slope) is the correct one
    - Trying to detect a transformation is thus an informal test for an association
      - Multiple testing procedures inflate the type I error
- It is best to choose the transformation of the predictor on scientific grounds
  - However, sometimes it doesn't matter – it is often the case that many functions are well approximated by a straight line over a small range of the data

Model Checking: Outlier vs Influential observations

- **Outlier:** an observation with a residual that is unusually large (positive or negative) as compared to the other residuals.

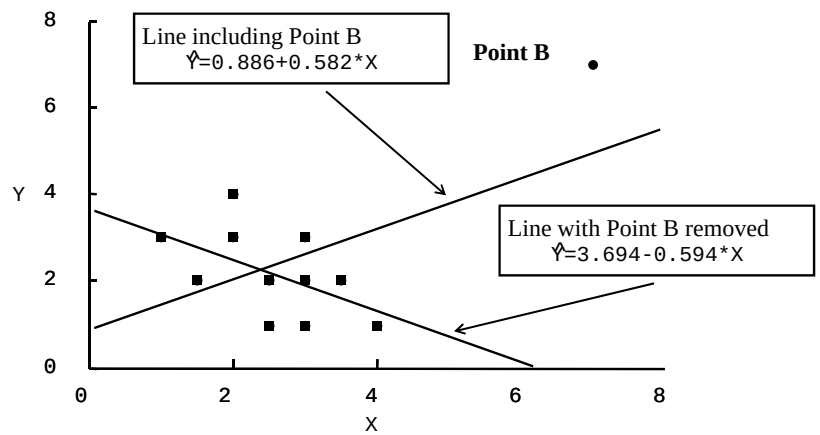
- **Influential point:** an observation that has a great deal of influence in determining the regression equation.
  - Removing such a point would markedly change the position of the regression line.
  - Observations that are somewhat extreme for the value of x are often influential.

Outlier vs Influential observations



Point A is an *outlier, but is not influential*.

Outlier vs Influential observations



Point B *is influential, but not an outlier*.

Model Checking: Deletion diagnostics

$$\Delta\beta_{(i)} = \hat{\beta} - \hat{\beta}_{(-i)}$$:delta-beta

$$\frac{\Delta\beta_{(i)}}{se(\hat{\beta})}$$:Standardized delta-beta

Delta-beta : tells how much the regression coefficient changed by including the ith observation

Standardized delta-beta : approximates how much the t-statistic for a coefficient changed by adding the ith observation

Cholesterol Example: Deletion diagnostics

```
> dfb = dfbeta(fit)
> dfb[order(abs(dfb[,2]), decreasing = T)[1:15],]
    (Intercept)          age
114  -0.9893663  0.015268514
166  -0.6827966  0.014888475
255  -0.6190643  0.013902713
186  -0.8544144  0.013279531
113   0.5376293 -0.011943495
325  -0.7517511  0.011308451
365   0.7676508 -0.011297278
257  -0.7374003  0.011092575
290  -0.7024787  0.010757541
144   0.7120264 -0.010710881
197  -0.6784150  0.010469720
296  -0.6499386  0.010101515
231  -0.6293174  0.009712016
7     0.4403297 -0.009524470
252  -0.5981020  0.009412761
```

No evidence of influential points. The largest (in absolute value) delta beta is 0.015 compared to 0.31 for the regression coefficient.

Model Checking: Deletion diagnostics

- What to do if you find an influential observation:
  - Check it for accuracy
  - Decide (based on scientific judgment) whether it is best to keep it or omit it
    - If you think it is representative, and likely would have appeared in a larger sample, keep it
    - If you think it is very unusual and unlikely to occur again in a larger sample, omit it
    - Report its existence [whether or not it is omitted].

Simple Linear Regression: Impact of Violations to Model Assumptions

	Non Linearity	Non Normality	Unequal Variances	Dependence
Estimates	Rubbish	Minimal for most departures. Outliers can be a disaster.	Minimal impact.	Often the estimates are unbiased.
Tests/CIs	Rubbish	Minimal for most departures. CIs for correlation are sensitive.	Variance estimates are wrong, but the effect is usually not dramatic.	Variance estimates are wrong (overestimate the precision and inflate test)
Correction	Transform or Choose a nonlinear model.	Delete outliers (if warranted) or Use robust regression	Transform or Use robust standard error.	Regression for dependent data.

UW School of Public Health and Community Medicine
Department of Biostatistics

REGRESSION MODELS

MULTIPLE LINEAR REGRESSION

Outline: Multiple Linear Regression

- Motivation
- Model and Interpretation
- Estimation and Inference
- Interaction

Motivation

- The response or dependent variable, Y, may depend on several predictors not just one!
- Multiple regression is an attempt to consider the <u>simultaneous</u> influence of several variables on the response
- It may reveal relationships that are completely hidden in univariate regression models

Motivation

- Why not fit multiple separate simple linear regressions?
  - A confounder can make the observed association between the predictor of interest and the response variable look
    - stronger than the true association,
    - weaker than the true association, or
    - even the reverse of the true association
- What could we do?
  - We can adjust for the effects of the confounder by adding a corresponding term to our linear regression! (more details later)

Motivation: Cholesterol Example

- Data

	sex	**age**	**chol**	**BMI**	TG	apoE	rs174548	rs4775401
1	1	74	215	26.2	367	4	1	2
2	1	51	204	24.7	150	4	2	1
3	0	64	205	24.2	213	4	0	1
4	0	34	182	23.8	111	1	1	1
5	1	52	175	34.1	328	1	0	0
6	1	39	176	22.7	53	4	0	2

- Our goal:
  - Investigate the relationship between age (years), BMI (kg/m^2) and serum total cholesterol (mg/dl)

Motivation

In general, the multiple regression equation can be written as follows:

$$E[Y \mid x_1, x_2, \ldots, x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Prediction: we use multiple variables if we think more than one variable will be useful in predicting future outcomes accurately
- Association: we use multiple variables when:
  - The variable is categorical with more than two groups
  - We need polynomials, splines or other functions to model the shape of the relationship(s) accurately
  - We want to adjust for confounding by other variables
  - We want to allow the association to differ for different values of other variables (interaction)

Model and Interpretation

- Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$

  where we assume $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

  Extension of simple linear regression!

  - Systematic component:

    $$E[Y \mid x_1, \ldots, x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

  - Random component:

    $$Var[Y \mid x_1, \ldots, x_p] = \sigma^2$$

Model and Interpretation

- For example, let us assume that there are two predictors in the model and so

$$E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Consider two observations with the same value for x_2, but one observation has x_1 one unit higher, that is,

Obs 1: $E[Y|x_1=k+1, x_2=c] = \beta_0 + \beta_1 (k+1) + \beta_2 c$
Obs 2: $E[Y|x_1=k, x_2=c] = \beta_0 + \beta_1 (k) + \beta_2 c$

Thus, $E[Y|x_1=k+1, x_2=c] - E[Y|x_1=k, x_2=c] = \beta_1$

That is, β_1 is the expected mean change in y per unit change in x_1 if x_2 is held constant (adjusted/controlling for x_2)!

Similar interpretation applies to β_2!

Model and Interpretation

- To facilitate our discussion let's assume we have two predictors with binary values

- Model:

$$E[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

mean	$X_2=0$	$X_2=1$
$X_1=0$	β_0	$\beta_0+\beta_2$
$X_1=1$	$\beta_0+\beta_1$	$\beta_0+\beta_1+\beta_2$

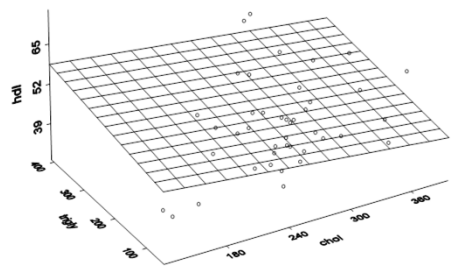
$E[Y|x_1=1, x_2=0]-E[Y|x_1=0,x_2=0] = \beta_1$
$E[Y|x_1=1, x_2=1]-E[Y|x_1=0,x_2=1] = \beta_1$
$E[Y|x_1=0, x_2=1]-E[Y|x_1=0,x_2=0] = \beta_2$
$E[Y|x_1=1, x_2=1]-E[Y|x_1=1,x_2=0] = \beta_2$

Estimation

- Least Squares Estimation:
  - minimizes the residual sum of squares

$$\sum_i (y_i - \hat{y}_i)^2$$

  - Computation more difficult, but statistical software (R) will do that for you!



Estimation and Inference

- Inference
  - About regression model parameters
    - **Hypothesis Testing** H_0: $\beta_j=0$

Interpretation: Is there a statistically significant relationship between the response y and x_j after adjusting for all other factors (predictors) in the model?

Test Statistic: $$\frac{\hat{\beta}_j - (null\ hyp)}{se(\hat{\beta}_j)} \sim T_{n-p-1}$$

Note: The square of the t-statistic gives the F-statistic and the test is known as the **partial F-Test**

- **Confidence Intervals**

$$\hat{\beta}_j \pm (critical\ value) \times se(\hat{\beta}_j)$$

Estimation and Inference

- About the full model
  - Hypotheses

    H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$ vs. H_1: At least one β_j is not null

  - Analysis of variance table

Source	df	SS	MS	F
Regression	p	SSR= $\sum(\hat{y}_i - \bar{y})^2$	MSR= SSR/p	MSR/ MSE
Residual	n-p-1	SSE= $\sum(y_i - \hat{y}_i)^2$	MSE= SSE/n-p-1	
Total	n-1	SST= $\sum(y_i - \bar{y})^2$		

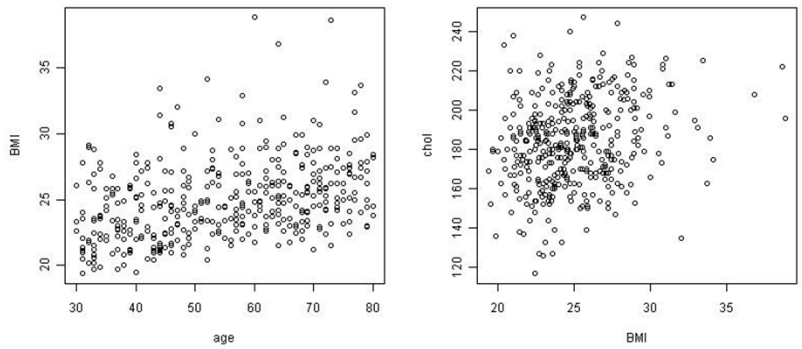
Estimation and Inference

- The F-value is tested against a F-distribution with p, n-p-1 degrees of freedom
  - If we reject the null hypothesis, then the predictors do aid in predicting Y [in this analysis we do not know which ones are important!]
  - Failing to reject the null-hypothesis does not mean that none of the covariates are important, since the effect of one or more covariates may be "masked" by others. The hard part is choosing which covariates to include or exclude.
- This is known as the **global (multiple) F-test**

Scientific example: Modeling cholesterol using age and BMI

- We have seen that there is a significant relationship between age and cholesterol
- Can we better understand variability in cholesterol by incorporating additional covariates?

Scientific example: Modeling cholesterol using age and BMI



Scientific example: Modeling cholesterol using age and BMI

- It appears that BMI increases with age
- And cholesterol increases with BMI
- What if we want to estimate the association between age and cholesterol while holding BMI constant?
- Multiple regression!

Scientific example: Modeling cholesterol using age and BMI

```
Call:
lm(formula = chol ~ age + BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-58.994 -15.793   0.571  14.159  62.992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.1612     9.0061  15.230  < 2e-16 ***
age           0.2023     0.0795   2.544 0.011327 *
BMI           1.4266     0.3822   3.732 0.000217 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.34 on 397 degrees of freedom
Multiple R-squared: 0.07351,  Adjusted R-squared: 0.06884
F-statistic: 15.75 on 2 and 397 DF,  p-value: 2.62e-07
```

Scientific example: Modeling cholesterol using age and BMI

- Our estimated regression equation is

$$\hat{y} = 137.16 + 0.20Age + 1.43BMI$$

- Question: How do we interpret the age coefficient?

Scientific example: Modeling cholesterol using age and BMI

- Our estimated regression equation is

$$\hat{y} = 137.16 + 0.20Age + 1.43BMI$$

- Question: How do we interpret the age coefficient?
- Answer: This is the estimated average difference in cholesterol associated with a one year difference in age for two subjects with the same BMI.

Scientific example: Modeling cholesterol using age and BMI

- Our estimated regression equation is

$$\hat{y} = 137.16 + 0.20Age + 1.43BMI$$

- The age coefficient from our simple linear regression model was 0.31.
- Question: Why do the estimates from the two models differ?

Scientific example: Modeling cholesterol using age and BMI

- Our estimated regression equation is

$$\hat{y} = 137.16 + 0.20Age + 1.43BMI$$

- The age coefficient from our simple linear regression model was 0.31.
- Question: Why do the estimates from the two models differ?
- Answer: We are now conditioning on or controlling for BMI so our estimate of the age association is among subjects with the same BMI.

Cholesterol Example:

- Did adding BMI improve our model?

```
> anova(fit,fit2)
Analysis of Variance Table

Model 1: chol ~ age
Model 2: chol ~ age + BMI
 Res.Df  RSS      Df       Sum of Sq        F        Pr(>F)
1 398   187187
2 397 1  80842   1        6345.8           13.931   0.0002174 ***
--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How does this model compare to a model that contains only the mean?

```
> anova(fit0,fit2)
Analysis of Variance Table

Model 1: chol ~ 1
Model 2: chol ~ age + BMI
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    399 195189
2    397 180842  2     14347 15.748 2.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Interaction and Linear Regression

- Statistical interaction (aka effect modification) occurs when the relationship between an outcome variable and one predictor is different depending on the levels of a second predictor
- Interactions are usually investigated because of *a priori* assumptions/hypotheses on the part of the researchers
- Linear regression models allow for the inclusion of interactions with cross-product terms

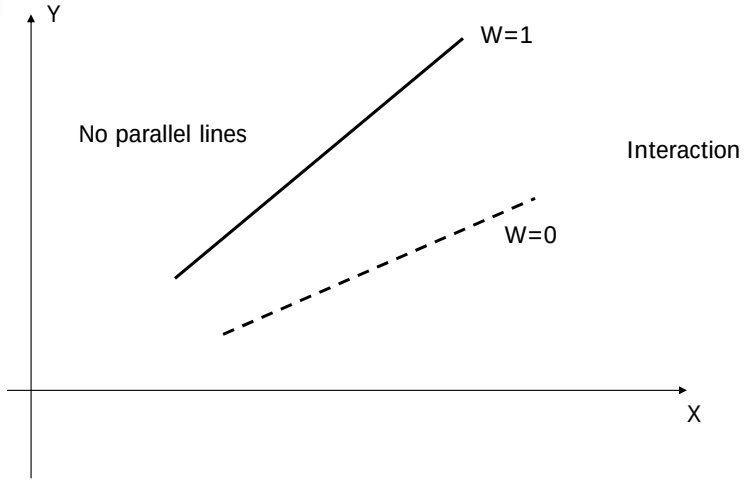
Discriminating between different classifications

- It is often very difficult to decide whether a new variable should be treated as a confounding or effect modification variable
- Data and scientific assessments help discriminate between confounding and effect modifying variables:
  - Confounder: Associated with predictor and response; Association between response and predictor constant across strata of the new variable
  - Effect modifier/interaction: Association between response and the predictor vary across strata of the new variable

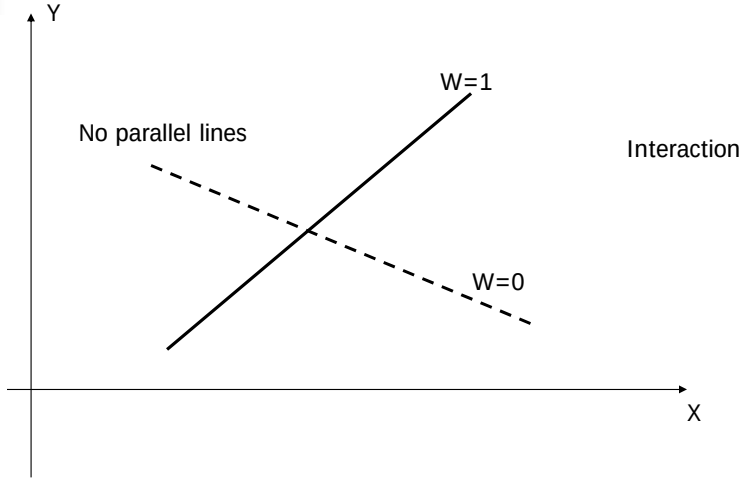
Confounding vs. Interaction/Effect Modification

- Estimates of association from unadjusted analysis are markedly different from estimates of association from adjusted analysis
  - Association within each stratum is similar, but different from the association in the combined data (ignoring the strata)
  - In linear regression, these symptoms are diagnostic of confounding
- Effect modification would show differences between adjusted analysis and unadjusted analysis, but would also show different associations in the different strata

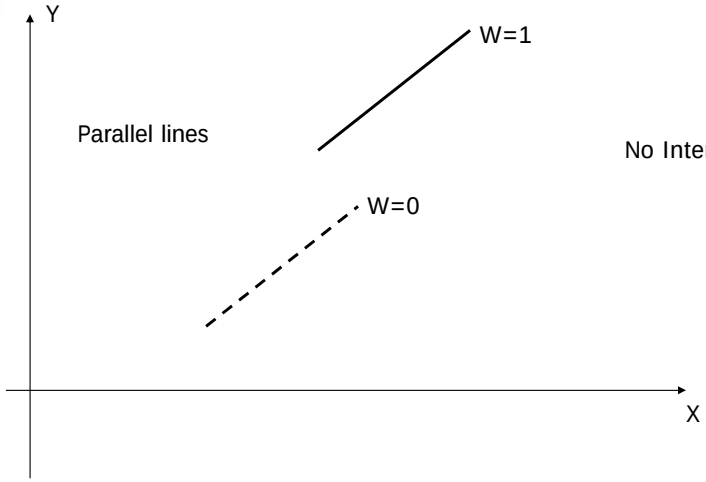
Graphical Representation



Graphical Representation

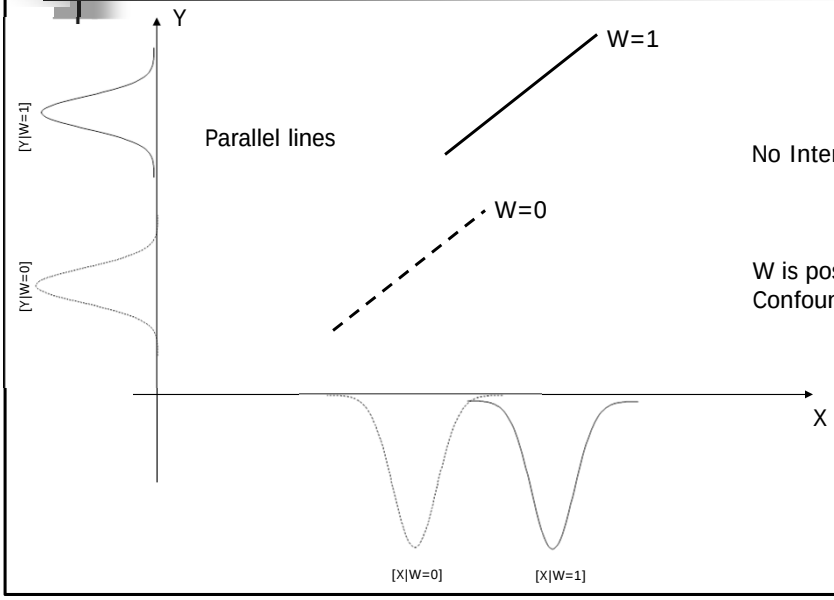


Graphical Representation



No Interaction

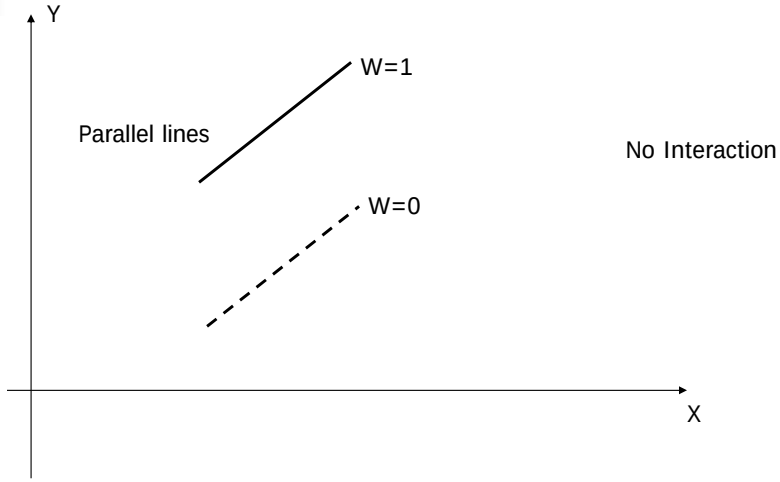
Graphical Representation



No Interaction

W is possibly a Confounder

Graphical Representation



Model and Interpretation: interaction

- Assume that there are two predictors in the model

$$E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2}$$

Consider two observations with the same value for x_2, but one observation has x_1 one unit higher

Obs 1: $E[Y|x_1=k+1, x_2=c] = \beta_0 + \beta_1 (k+1) + \beta_2 c + \beta_3 (k+1)c$

Obs 2: $E[Y|x_1=k, x_2=c] = \beta_0 + \beta_1 (k) + \beta_2 c + \beta_3 kc$

Thus, $E[Y|x_1=k+1, x_2=c] - E[Y|x_1=k, x_2=c] = \beta_1 + \underline{\beta_3 c}$

That is, the difference in means depends now on the value of x_2!

Model and Interpretation: interaction

- Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2}$

- Difference in Means:

  $E[Y|x_1=k+1, x_2=c] - E[Y|x_1=k, x_2=c] = \beta_1 + \underline{\beta_3 c}$

  The difference in means depends now on the value of x_2!
  - The difference in means is β_1 if c=0.
  - The difference in means is $\beta_1 + \beta_3$ if c=1
  - The difference in means changes by β_3 for each unit difference in c (that is, in x_2) [that is, β_3 is the difference of differences!]

Model and Interpretation: interaction

- Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2}$

- Another way to look at this

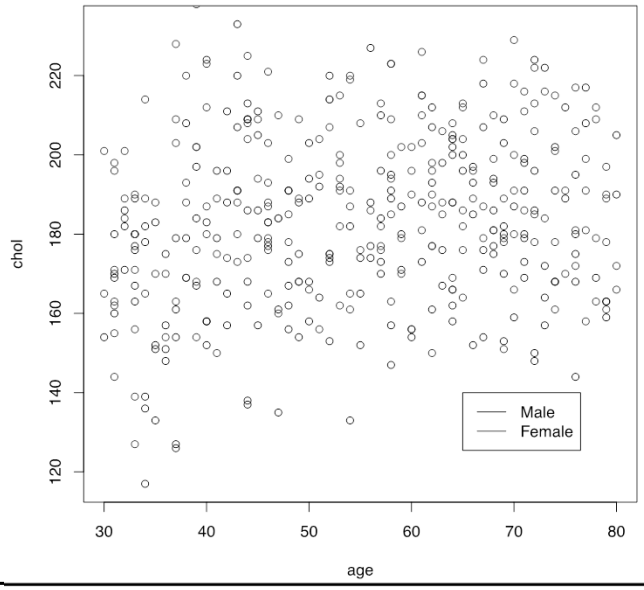
- Factor terms involving x_1:

  $E[Y|x_1, x_2] = \beta_0 + \underline{(\beta_1 + \beta_3 x_2)} x_1 + \beta_2 x_2$

  Slope of x_1 changes with x_2 =

  Difference in means for each unit difference in x_1 changes with x_2 (for each one unit difference in x_2, the difference in means changes by β_3)

Cholesterol Example: Does gender affect the age – cholesterol relationship?



Cholesterol Example: Does gender affect the age – cholesterol relationship?

We first fit the model with age and sex terms only

```
> fit2 = lm(chol ~ age+sex)
> summary(fit2)

Call:
lm(formula = chol ~ age + sex)

Residuals:
    Min      1Q  Median      3Q     Max
-55.662 -14.482  -1.411  14.682  57.876

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.35445    4.24184  38.275  < 2e-16 ***
age           0.29697    0.07313   4.061 5.89e-05 ***
sex          10.50728    2.10794   4.985 9.29e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.06 on 397 degrees of freedom
Multiple R-squared: 0.09748,    Adjusted R-squared: 0.09293
F-statistic: 21.44 on 2 and 397 DF,  p-value: 1.440e-09
```

Cholesterol Example: Does gender affect the age – cholesterol relationship?

- This model indicates that, after controlling for the effect of sex, the average cholesterol differs by 0.30 for each additional year of age
- The age effect in this model is very similar to the effect from our simple linear regression (0.31)
- However, this does not mean that the age/cholesterol relationship is the same in males and females
- To answer this question we must add the interaction term

Cholesterol Example: Does gender affect the age – cholesterol relationship?

Model with age and sex main effects, plus interaction effect

```
Call:
lm(formula = chol ~ age * sex)

Residuals:
    Min      1Q  Median      3Q     Max
-56.474 -14.377  -1.215  14.764  58.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.31151    5.86268  27.344  < 2e-16 ***
age           0.33460    0.10442   3.204  0.00146 **
sex          14.56271    8.29802   1.755  0.08004 .
age:sex      -0.07399    0.14642  -0.505  0.61361
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.08 on 396 degrees of freedom
Multiple R-squared: 0.09806,      Adjusted R-squared: 0.09123
F-statistic: 14.35 on 3 and 396 DF,  p-value: 6.795e-09
```

Cholesterol Example: Does gender affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * sex)

Residuals:
    Min      1Q  Median      3Q     Max
-56.474 -14.377  -1.215  14.764  58.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.31151    5.86268  27.344  < 2e-16 ***
age           0.33460    0.10442   3.204  0.00146 **
sex          14.56271    8.29802   1.755  0.08004 .
age:sex      -0.07399    0.14642  -0.505  0.61361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.08 on 396 degrees of freedom
Multiple R-squared: 0.09806,  Adjusted R-squared: 0.09123
F-statistic: 14.35 on 3 and 396 DF,  p-value: 6.795e-09
```

Mean cholesterol for males at age 0

Cholesterol Example: Does gender affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * sex)

Residuals:
    Min      1Q  Median      3Q     Max
-56.474 -14.377  -1.215  14.764  58.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.31151    5.86268  27.344  < 2e-16 ***
age           0.33460    0.10442   3.204  0.00146 **
sex          14.56271    8.29802   1.755  0.08004 .
age:sex      -0.07399    0.14642  -0.505  0.61361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.08 on 396 degrees of freedom
Multiple R-squared: 0.09806,  Adjusted R-squared: 0.09123
F-statistic: 14.35 on 3 and 396 DF,  p-value: 6.795e-09
```

Difference in mean cholesterol between males and females at age 0

Cholesterol Example: Does gender affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * sex)

Residuals:
    Min      1Q  Median      3Q     Max
-56.474 -14.377  -1.215  14.764  58.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.31151    5.86268  27.344  < 2e-16 ***
age           0.33460    0.10442   3.204  0.00146 **
sex          14.56271    8.29802   1.755  0.08004 .
age:sex      -0.07399    0.14642  -0.505  0.61361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.08 on 396 degrees of freedom
Multiple R-squared: 0.09806,  Adjusted R-squared: 0.09123
F-statistic: 14.35 on 3 and 396 DF,  p-value: 6.795e-09
```

Difference in mean cholesterol associated with each one year change in age for males

Cholesterol Example: Does gender affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * sex)

Residuals:
    Min      1Q  Median      3Q     Max
-56.474 -14.377  -1.215  14.764  58.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.31151    5.86268  27.344  < 2e-16 ***
age           0.33460    0.10442   3.204  0.00146 **
sex          14.56271    8.29802   1.755  0.08004 .
age:sex      -0.07399    0.14642  -0.505  0.61361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.08 on 396 degrees of freedom
Multiple R-squared: 0.09806,  Adjusted R-squared: 0.09123
F-statistic: 14.35 on 3 and 396 DF,  p-value: 6.795e-09
```

Difference in change in mean cholesterol associated with each one year change in age for females compared to males

Cholesterol Example: Does gender affect the age – cholesterol relationship?

- Interpretation?
  - Estimated model:

160.3 + 0.33 Age + 14.56 Sex- 0.07 Age x Sex

Subject 1: Age = a+1, sex = b

Subject 2: Age = a, sex = b

Difference in the estimated cholesterol:

[160.3 + 0.33(a+1) + 14.56(b) – 0.07 (a+1)(b)] –

[160.3 + 0.33(a) + 14.56 (b) – 0.07 (a)(b)] = 0.33-0.07b

  - Sex exerts a small (not statistically significant) effect on the age/cholesterol relationship

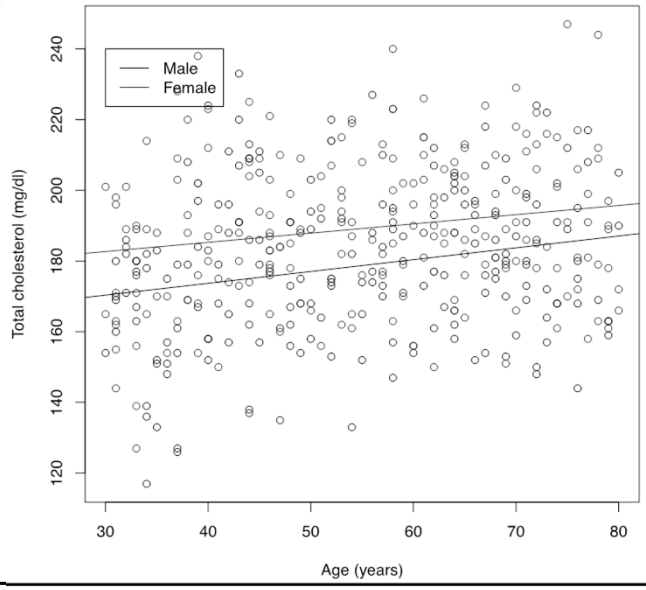
Cholesterol Example: Does gender affect the age – cholesterol relationship?

- We can also test the significance of interaction terms using an F-test

```
> anova(fit2,fit3)
Analysis of Variance Table

Model 1: chol ~ age + sex
Model 2: chol ~ age * sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    397 176162
2    396 176049  1    113.52 0.2554 0.6136
```

- Adding the interaction term did not significantly improve model fit

Cholesterol Example: Does gender affect the age – cholesterol relationship?

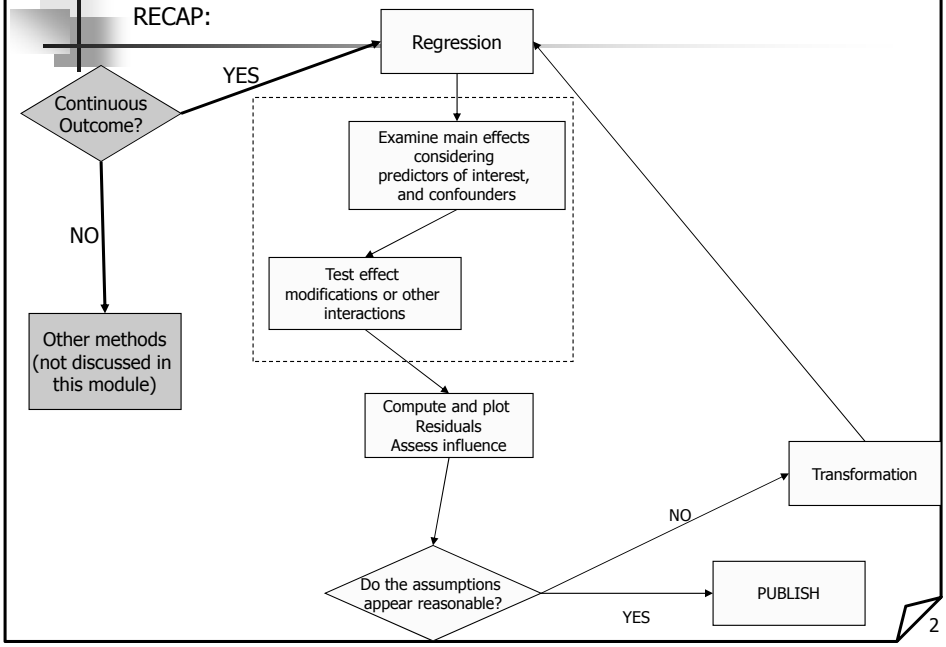


UW School of Public Health and Community Medicine
Department of Biostatistics

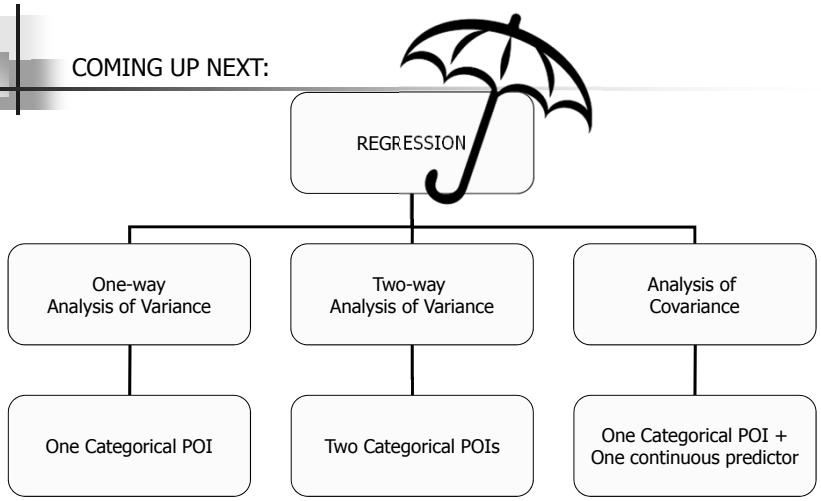
REGRESSION MODELS

ANOVA MODELS

RECAP:



COMING UP NEXT:



Uses dummy variables to represent categorical variables!

Outline

- Motivation
- ANOVA as a regression model
  - Dummy variables
- One-way ANOVA models
  - Contrasts
  - Multiple comparisons
- Two-way ANOVA models
  - Interactions
- ANCOVA models
- Experimental Designs and ANOVA models

ANOVA

Motivation

Motivation

- Let's investigate if genetic factors are associated with cholesterol levels.
  - Ideally, you would have a confirmatory analysis of scientific hypotheses formulated prior to data collection
  - [Alternatively, you could consider an exploratory analysis – hypotheses generation for future studies]

ANOVA/ANCOVA: Motivation

- Scientific hypotheses of interest:
  - Assess the effect of rs174548 on cholesterol levels.
  - Assess the effect of rs174548 and gender on cholesterol levels
    - Does the effect of rs174548 on cholesterol differ between males and females?
  - Assess the effect of rs174548 and age on cholesterol levels
    - Does the effect of rs174548 on cholesterol differ depending on subject's age?

ANOVA: One-Way Model Motivation:

- Scientific question:
  - Assess the effect of rs174548 on cholesterol levels.

Motivation: Example

Here are some descriptive summaries:

```
> tapply(chol, as.factor(rs174548), mean)
       0        1        2
181.0617 187.8639 186.5000


> tapply(chol, as.factor(rs174548), sd)
       0        1        2
21.13998 23.74541 17.38333
```

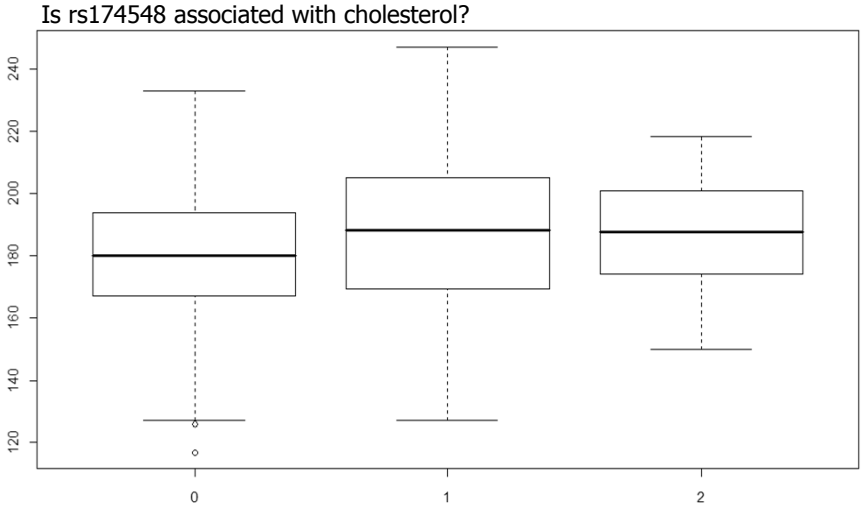

Motivation: Example

Another way of getting the same results:

```
> by(chol, as.factor(rs174548), mean)
as.factor(rs174548): 0
[1] 181.0617
------------------------------------------------------------
as.factor(rs174548): 1
[1] 187.8639
------------------------------------------------------------
as.factor(rs174548): 2
[1] 186.5

> by(chol, as.factor(rs174548), sd)
as.factor(rs174548): 0
[1] 21.13998
------------------------------------------------------------
as.factor(rs174548): 1
[1] 23.74541
------------------------------------------------------------
as.factor(rs174548): 2
[1] 17.38333
```

Motivation: Example

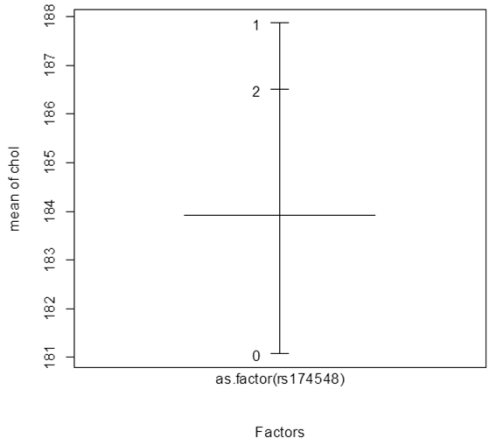
Is rs174548 associated with cholesterol?



R command: boxplot(chol ~ as.factor(rs174548))

Motivation: Example

Another graphical display:



R commands:
plot.design(chol ~ as.factor(rs174548))

Motivation: Example

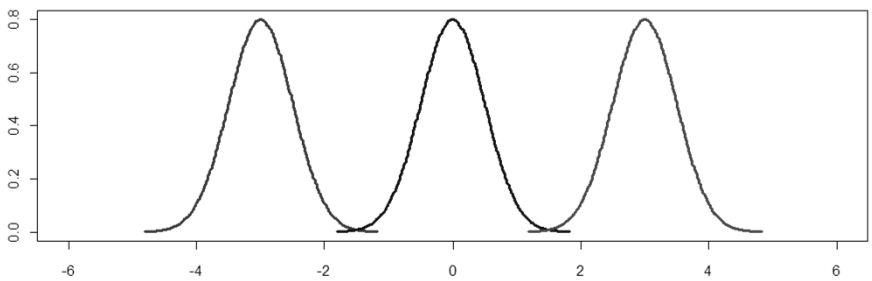
- Feature:
  - How do the mean responses compare across different groups?
    - Categorical/qualitative predictor

ANOVA

As a regression model

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations

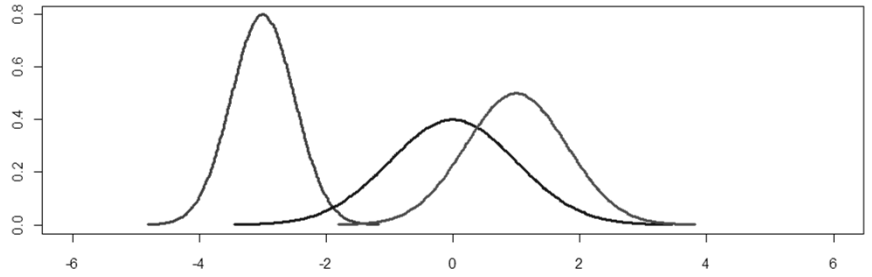


Assumptions for Classical ANOVA Framework:

Independence
Normality
Equal variances

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations



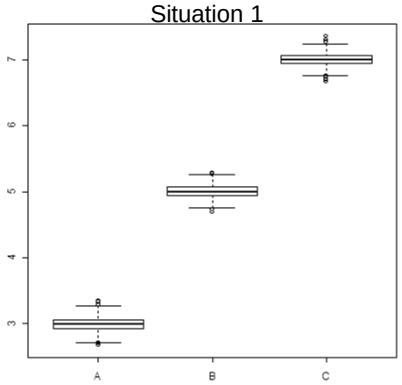
ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations
  - Counter-intuitive name!

ANalysis Of VAriance Models (ANOVA)

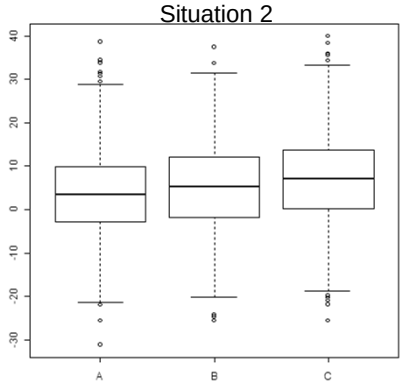
In both data sets, the true population means are: 3 (A), 5 (B), 7(C)

Situation 1



Low variance within groups

Situation 2



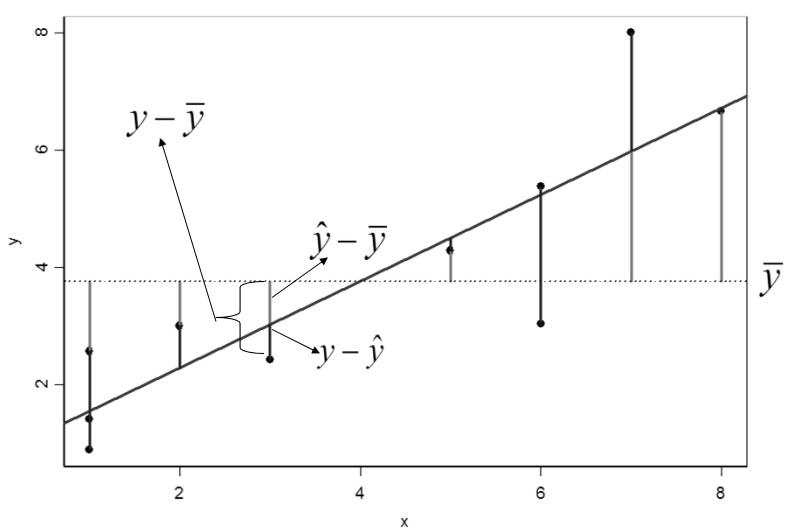
High variance within groups

Where do you expect to detect difference between population means?

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations
  - Counter-intuitive name!
    - Underlying concept:
      - To assess whether the population means are equal, compares:
        - Variation between the sample means (MSR) to
        - Natural variation of the observations within the samples (MSE).
      - The larger the MSR compared to MSE the more support that there is a difference in the <u>population means</u>!
      - The ratio MSR/MSE is the F-statistic.

Decomposition of sum of squares



ANalysis Of VAriance Models (ANOVA)

- Equivalent to regression with categorical predictors.
  - Predictors represented with “dummy” variables

ANOVA as a multiple regression model

- Dummy Variables:
  - Suppose you have a categorical variable C with k categories. To represent that variable we can construct k-1 dummy variables of the form

$$x_1 = \begin{cases} 1, & \text{if subject is in category 2} \\ 0, & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{if subject is in category 3} \\ 0, & \text{otherwise} \end{cases}$$

…

$$x_{k-1} = \begin{cases} 1, & \text{if subject is in category k} \\ 0, & \text{otherwise} \end{cases}$$

The omitted category (here category 1) is the **reference group**.

ANOVA as a multiple regression model

- Dummy Variables:
  - Back to our motivating example:
    - Predictor: rs174548 (coded 0=C/C, 1=C/G, 2=G/G)
    - Outcome (Y): cholesterol

    Let's take C/C as the reference group.

$$x_1 = \begin{cases} 1, & \text{if code 1 (C/G)} \\ 0, & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{if code 2 (G/G)} \\ 0, & \text{otherwise} \end{cases}$$

ANOVA as a multiple regression model

rs174548	X_1	X_2
C/C	0	0
C/G	1	0
G/G	0	1

ANOVA as a multiple regression model

- Regression with Dummy Variables:
  - Example:

    Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- Interpretation of model parameters?

ANOVA as a multiple regression model

- Regression with Dummy Variables:
  - Example:

    Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- Interpretation of model parameters?
  - β_0: mean cholesterol when rs174548 is C/C
  - $\beta_0+\beta_1$: mean cholesterol when rs174548 is C/G
  - $\beta_0+\beta_2$: mean cholesterol when rs174548 is G/G

ANOVA as a multiple regression model

- Regression with Dummy Variables:
  - Example:

    Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- Interpretation of model parameters?
  - β_0: mean cholesterol when rs174548 is C/C
  - $\beta_0+\beta_1$: mean cholesterol when rs174548 is C/G
  - $\beta_0+\beta_2$: mean cholesterol when rs174548 is G/G
  - Alternatively
    - β_1: difference in mean cholesterol levels between groups with rs174548 equal to C/G and C/C.
    - β_2: difference in mean cholesterol levels between groups with rs174548 equal to G/G and C/C.

ANOVA as a multiple regression model

- Alternative parameterization
  - Each group with its own mean!
- Let's re-write the model:

  Model: $E[Y_{ij}] = \mu_i$

  (i: genotype index, j: subject index)

ANOVA as a multiple regression model

- Regression Model:
  Model 1: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

- ANOVA Model:
  Model 2: $E[Y_{ij}] = \mu_i$

ANOVA as a multiple regression model

Mean	Regression Model
μ_1	β_0
μ_2	$\beta_0 + \beta_1$
μ_3	$\beta_0 + \beta_2$

ANOVA as a multiple regression model

- Regression Model:
  Model 1: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

- ANOVA Model:
  Model 2: $E[Y_{ij}] = \mu_i$

Key Message:
ANOVA is a special case of a regression model!

ANOVA as a multiple regression model

- The same idea applies to problems with several categorical predictors [aka: factors]
  - One-way ANOVA: one factor
  - Two-way ANOVA: two factors
  - …

- Model assumptions
  - Equal variances
  - Normality
  - Independence

ANOVA

One-way ANOVA models

ANOVA: One-Way Model

- Goal:
  - Compare the means of K independent groups (defined by a categorical predictor)
    - Statistical Hypotheses:
      - (Global) Null Hypothesis:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_K.$$

      - Alternative Hypothesis:

$$H_1: \text{not all means are equal}$$

  - If the means of the groups are not all equal (i.e. you rejected the above H_0), determine which ones are different (multiple comparisons)

Estimation and Inference

- Global Hypotheses

  H_0: $\mu_1 = \mu_2 = \ldots = \mu_K$ vs. H_1: not all means are equal

- Analysis of variance table

Source	df	SS	MS	F
Regression	K-1	SSR= $\sum_i (\bar{y}_i - \bar{y})^2$	MSR= SSR/(K-1)	MSR/ MSE
Residual	n-K	SSE= $\sum_{i,j} (y_{ij} - \bar{y}_i)^2$	MSE= SSE/n-K	
Total	n-1	SST= $\sum_{i,j} (y_{ij} - \bar{y})^2$		

ANOVA as a multiple regression model

Back to example:

Mean	Regression Model
μ_1	β_0
μ_2	$\beta_0 + \beta_1$
μ_3	$\beta_0 + \beta_2$

Estimation and Inference

- Global Hypotheses

  H_0: $\beta_1 = \ldots = \beta_{K-1} = 0$ vs. H_1: not all coeffs are zero

- Analysis of variance table

Source	df	SS	MS	F
Regression	K-1	SSR= $\sum_i (\bar{y}_i - \bar{y})^2$	MSR= SSR/(K-1)	MSR/ MSE
Residual	n-K	SSE= $\sum_{i,j} (y_{ij} - \bar{y}_i)^2$	MSE= SSE/n-K	
Total	n-1	SST= $\sum_{i,j} (y_{ij} - \bar{y})^2$		

ANOVA: One-Way Model

- How to fit a one-way model as a regression problem?
  - Need to use "dummy" variables
    - Create on your own (can be tedious!)
    - Most software packages will do this for you
      - R creates dummy variables in the background <u>as long as</u> you state you have a categorical variable (may need to use: as.factor)

ANOVA: One-Way Model

By hand:
Creating "dummy" variables:

```
> dummy1 = 1*(rs174548==1)
> dummy2 = 1*(rs174548==2)
```

Fitting the ANOVA model:

```
> fit0 = lm(chol ~ dummy1 + dummy2)
> summary(fit0)
Call:
lm(formula = chol ~ dummy1 + dummy2)

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.062      1.455 124.411  < 2e-16 ***
dummy1         6.802      2.321   2.930  0.00358 **
dummy2         5.438      4.540   1.198  0.23167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,     Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit0)
Analysis of Variance Table

Response: chol
           Df Sum Sq Mean Sq F value   Pr(>F)
dummy1      1   3624    3624  7.5381 0.006315 **
dummy2      1    690     690  1.4350 0.231665
Residuals 397 190875     481
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


ANOVA: One-Way Model

Better:
Let R do it for you!

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16 ***
as.factor(rs174548)1    6.802      2.321   2.930  0.00358 **
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,     Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA: One-Way Model

- Your turn!
  - Compare model fit results (fit0 & fit1.1)
    What do you conclude?

ANOVA: One-Way Model

```
> fit0 = lm(chol ~ dummy1 + dummy2)
> summary(fit0)
Call:
lm(formula = chol ~ dummy1 + dummy2)

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.062      1.455 124.411  < 2e-16 ***
dummy1         6.802      2.321   2.930  0.00358 **
dummy2         5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221, Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184
```

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16 ***
as.factor(rs174548)1    6.802      2.321   2.930  0.00358 **
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,  Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184
```

```
> anova(fit0)
Analysis of Variance Table

Response: chol
           Df Sum Sq Mean Sq F value   Pr(>F)
dummy1      1   3624    3624  7.5381 0.006315 **
dummy2      1    690     690  1.4350 0.231665
Residuals 397 190875     481
---
```

```
> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
```

ANOVA: One-Way Model

```
> fit0 = lm(chol ~ dummy1 + dummy2)
> summary(fit0)
Call:
lm(formula = chol ~ dummy1 + dummy2)

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.062      1.455 124.411  < 2e-16 ***
dummy1         6.802      2.321   2.930  0.00358 **
dummy2         5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221, Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184
```

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16 ***
as.factor(rs174548)1    6.802      2.321   2.930  0.00358 **
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,  Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184
```

```
> anova(fit0)
Analysis of Variance Table

Response: chol
           Df Sum Sq Mean Sq F value   Pr(>F)
dummy1      1   3624    3624  7.5381 0.006315 **
dummy2      1    690     690  1.4350 0.231665
Residuals 397 190875     481
---
```

```
> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
```

```
> 1-pf(4.4865,2,397)
[1] 0.01183671
```


ANOVA: One-Way Model

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16
as.factor(rs174548)1    6.802      2.321   2.930  0.00358
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,  Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
```

- Let's interpret the regression model results!
  - What is the interpretation of the regression model coefficients?

ANOVA: One-Way Model

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16
as.factor(rs174548)1    6.802      2.321   2.930  0.00358
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,  Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
```

- Interpretation:
  - Estimated mean cholesterol for C/C group: 181.062 mg/dl
  - Estimated difference in mean cholesterol levels between C/G and C/C groups: 6.802 mg/dl
  - Estimated difference in mean cholesterol levels between G/G and C/C groups: 5.438 mg/dl

ANOVA: One-Way Model

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16
as.factor(rs174548)1    6.802      2.321   2.930  0.00358
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,  Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
```

- Overall F-test shows a significant p-value. We reject the null hypothesis that the mean cholesterol levels are the same across groups defined by rs174548 (p=0.01184).
  - This does not tell us which groups are different!

    (Need to perform multiple comparisons! More soon…)

ANOVA: One-Way Model

Alternative form:
(better if you will perform multiple comparisons)

```
> fit1.2 = lm(chol ~ -1 + as.factor(rs174548))
> summary(fit1.2)
Call:
lm(formula = chol ~ -1 + as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
as.factor(rs174548)0  181.062      1.455  124.41   <2e-16 ***
as.factor(rs174548)1  187.864      1.809  103.88   <2e-16 ***
as.factor(rs174548)2  186.500      4.300   43.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.9861,     Adjusted R-squared: 0.986
F-statistic:  9383 on 3 and 397 DF,  p-value: < 2.2e-16

> anova(fit1.2)
Analysis of Variance Table
Response: chol
                     Df   Sum Sq Mean Sq F value    Pr(>F)
as.factor(rs174548)   3 13534205 4511402  9383.2 < 2.2e-16 ***
Residuals           397   190875     481
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


ANOVA: One-Way Model

Alternative form:
- Different command!

```
> fit1.3 = aov(chol ~ as.factor(rs174548))
> summary(fit1.3)
                      Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314 2157.10  4.4865 0.01184 *
Residuals           397 190875  480.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(fit1.3)
Analysis of Variance Table

Response: chol
                      Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314 2157.10  4.4865 0.01184 *
Residuals           397 190875  480.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> fit1.3$coeff
         (Intercept) as.factor(rs174548)1 as.factor(rs174548)2
          181.061674             6.802272             5.438326
```

ANOVA: One-Way Model

How about this one?
How is rs174548 being treated now?

Compare model fit results from (fit1.1 & fit2).

```
> fit2 = lm(chol ~ rs174548)
> summary(fit2)

Call:
lm(formula = chol ~ rs174548)

Residuals:
    Min      1Q  Median      3Q     Max
-64.575 -16.278  -0.575  15.120  60.722

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.575      1.411 128.723  < 2e-16 ***
rs174548       4.703      1.781   2.641  0.00858 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.95 on 398 degrees of freedom
Multiple R-squared: 0.01723,    Adjusted R-squared: 0.01476
F-statistic: 6.977 on 1 and 398 DF,  p-value: 0.008583

> anova(fit2)
Analysis of Variance Table

Response: chol
           Df Sum Sq Mean Sq F value   Pr(>F)
rs174548    1   3363    3363  6.9766 0.008583 **
Residuals 398 191827     482
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


ANOVA: One-Way Model


```
> fit2 = lm(chol ~ rs174548)
> summary(fit2)

Call:
lm(formula = chol ~ rs174548)

Residuals:
    Min      1Q  Median      3Q     Max
-64.575 -16.278  -0.575  15.120  60.722

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.575      1.411 128.723  < 2e-16 ***
rs174548       4.703      1.781   2.641  0.00858 **

Residual standard error: 21.95 on 398 degrees of freedom
Multiple R-squared: 0.01723, Adjusted R-squared: 0.01476
F-statistic: 6.977 on 1 and 398 DF,  p-value: 0.008583

> anova(fit2)
Analysis of Variance Table

Response: chol
           Df Sum Sq Mean Sq F value   Pr(>F)
rs174548    1   3363    3363  6.9766 0.008583 **
Residuals 398 191827     482
```


- Model: $E[Y|x] = \beta_0 + \beta_1 x$

where Y: cholesterol, x: rs174548

- Interpretation of model parameters?
  - β_0: mean cholesterol in the C/C group [estimate: 181.575 mg/dl]
  - β_1: mean cholesterol difference between C/G and C/C – or – between G/G and C/G groups [estimate: 4.703 mg/dl]
- This model presumes differences between “consecutive” groups are the same (in this example, linear dose effect of allele) – more restrictive than the ANOVA model!

Back to the ANOVA model…

ANOVA: One-Way Model

```
> fit1.1 = lm(chol ~ as.factor(rs174548))
> summary(fit1.1)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16
as.factor(rs174548)1    6.802      2.321   2.930  0.00358
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,  Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit1.1)
Analysis of Variance Table

Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
```

- We rejected the null hypothesis that the mean cholesterol levels are the same across groups defined by rs174548 (p=0.01184).
  - What are the groups with differences in means?

MULTIPLE COMPARISONS

ANOVA

MULTIPLE COMPARISONS

ANOVA: One-Way Model

- What are the groups with differences in means?

MULTIPLE COMPARISONS:

$\mu_0 = \mu_1$?

$\mu_0 = \mu_2$?

$\mu_1 = \mu_2$?

} Pairwise comparisons

$(\mu_1 + \mu_2)/2 = \mu_0$? ⟶ Non-pairwise comparison

Multiple Comparisons: Family-wise error rates

- Illustrating the multiple comparison problem
  - Truth: null hypotheses
  - Tests: pairwise comparisons - each at the 5% level.

What is the probability of rejecting at least one?

#groups = K	2	3	4	5	6	7	8	9	10
#pairwise comparisons = K(K-1)/2	1	3	6	10	15	21	28	36	45
P(at least one sig) $=1-(1-0.05)^c$	0.05	0.143	0.265	0.401	0.537	0.659	0.762	0.842	0.901

That is, if you have three groups and make pairwise comparisons, each at the 5% level, your family-wise error rate (probability of making at least one false rejection) is over 14%!

Need to address this issue!

Several methods!!!

Multiple Comparisons

- Several methods:
  - None (no adjustment)
  - Bonferroni
  - Holm
  - Hochberg
  - Hommel
  - BH
  - BY
  - FDR
  - …

Available in R

Multiple Comparisons

- **Bonferroni** adjustment: for k tests performed, use level α/k (or multiply *P*-values by k).
  - Simple
  - Conservative
  - Must decide on number of tests beforehand
  - Widely applicable
  - Can be done without software!

Multiple Comparisons

This option considers all pairwise comparisons

```
> ## call library for multiple comparisons
> library(multcomp)
>
> ## fit model
> fit1 = lm(chol ~ -1 + as.factor(rs174548))
>
> ## all pairwise comparisons
> ## -- first, define matrix of contrasts
> M = contrMat(table(rs174548), type="Tukey")
> M

         Multiple Comparisons of Means: Tukey Contrasts

       0  1 2
1 - 0 -1  1 0
2 - 0 -1  0 1
2 - 1  0 -1 1
>
> ## -- second, obtain estimates for multiple comparisons
> mc = glht(fit1, linfct =M)
```

Stands for general linear hypothesis testing

Multiple Comparisons

```
> ## -- third, adjust the p-values (or not) for multiple comparisons
> summary(mc, test=adjusted("none"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(rs174548))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
1 - 0 == 0    6.802      2.321   2.930  0.00358 **
2 - 0 == 0    5.438      4.540   1.198  0.23167
2 - 1 == 0   -1.364      4.665  -0.292  0.77015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- none method)
```

Multiple Comparisons

```
> summary(mc, test=adjusted("bonferroni"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(rs174548))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
1 - 0 == 0    6.802      2.321   2.930   0.0107 *
2 - 0 == 0    5.438      4.540   1.198   0.6950
2 - 1 == 0   -1.364      4.665  -0.292   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- bonferroni method)
```


Multiple Comparisons

- What if nonpairwise comparison?
  - Suppose you want to compare the mean cholesterol among those with genotype C/C with the mean cholesterol for the combined group with genotypes C/G and G/G.

$$\mu_0 = (\mu_1 + \mu_2)/2$$

Or equivalently,

$$2\mu_0 = (\mu_1 + \mu_2)$$

Or equivalently,

$$2\mu_0 - \mu_1 - \mu_2 = 0$$

Multiple Comparisons

- What if nonpairwise comparison?
  - Your turn: Suppose you want to compare the mean cholesterol among those with genotype C/G with the mean cholesterol for the combined group with genotypes C/C and G/G.

Multiple Comparisons

- What if nonpairwise comparison?
  - Your turn: Suppose you want to compare the mean cholesterol among those with genotype C/G with the mean cholesterol for the combined group with genotypes C/C and G/G.

$$(\mu_0 + \mu_2)/2 = \mu_1$$

Or equivalently,

$$\mu_0 + \mu_2 = 2\mu_1$$

Or equivalently,

$$\mu_0 - 2\mu_1 + \mu_2 = 0$$

Multiple Comparisons

Using R for multiple comparisons with "user-defined" contrasts:

```
> contr = rbind("mean(C/G+G/G) - mean(C/C)" = c(-2, 1, 1))
> mc2 = glht(fit1, linfct =contr)
> summary(mc2, test=adjusted("none"))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = chol ~ -1 + as.factor(rs174548))

Linear Hypotheses:
                              Estimate Std. Error t value Pr(>|t|)
mean(C/G+G/G) - mean(C/C) == 0   12.241      5.499   2.226   0.0266 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- none method)
```


Multiple Comparisons

```
> ## more than one contrast (again user-defined)
> contr2 = rbind("mean(C/G+G/G) - mean(C/C)" = c(-2, 1, 1),
+                "mean(C/C+G/G) - mean(C/G)" = c(1, -2, 1))
> mc3 = glht(fit1, linfct =contr2)
> summary(mc3, test=adjusted("none"))

         Simultaneous Tests for General Linear Hypotheses
Fit: lm(formula = chol ~ -1 + as.factor(rs174548))
Linear Hypotheses:
                              Estimate Std. Error t value Pr(>|t|)
mean(C/G+G/G) - mean(C/C) == 0   12.241      5.499   2.226   0.0266 *
mean(C/C+G/G) - mean(C/G) == 0   -8.166      5.805  -1.407   0.1603
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- none method)

> summary(mc3, test=adjusted("bonferroni"))

         Simultaneous Tests for General Linear Hypotheses
Fit: lm(formula = chol ~ -1 + as.factor(rs174548))
Linear Hypotheses:
                              Estimate Std. Error t value Pr(>|t|)
mean(C/G+G/G) - mean(C/C) == 0   12.241      5.499   2.226   0.0531 .
mean(C/C+G/G) - mean(C/G) == 0   -8.166      5.805  -1.407   0.3205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- bonferroni method)
```

Multiple Comparisons

- What about using other adjustment methods?
  - For example, we used:

    > summary(mc, test=adjusted("bonferroni"))

    (all pairwise comparisons, with Bonferroni adjustment)

  - Other options, in place of "bonferroni", are:
    - summary(mc, test=adjusted("holm"))
    - summary(mc, test=adjusted("hochberg"))
    - summary(mc, test=adjusted("hommel"))
    - summary(mc, test=adjusted("BH"))
    - summary(mc, test=adjusted("BY"))
    - summary(mc, test=adjusted("fdr"))

Results, in this particular example, are basically the same, but they don't need to be! Different criteria could lead to different results!

Multiple Comparisons

```
> summary(mc, test=adjusted("fdr"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(rs174548))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
1 - 0 == 0    6.802      2.321   2.930   0.0107 *
2 - 0 == 0    5.438      4.540   1.198   0.3475
2 - 1 == 0   -1.364      4.665  -0.292   0.7702
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- fdr method)
```

Multiple Comparisons

- FDR (False Discovery Rate)
  - Less conservative procedure for multiple comparisons
  - Among rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (that is, type I errors).

UW School of Public Health and Community Medicine
Department of Biostatistics

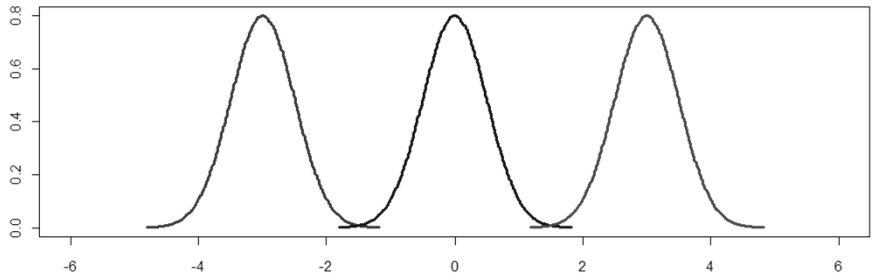
ANOVA

MODEL CHECKING

ANOVA Assumptions

- Recall the assumptions for classical ANOVA are:

Independence
Normality
Equal variance



Bartlett's test

- We assume that variances are the same across populations
- Bartlett's test allows you to test the hypothesis that the population variances are the same (versus they are not all equal).

```
> bartlett.test(chol ~ as.factor(rs174548))

        Bartlett test of homogeneity of variances

data:  chol by as.factor(rs174548)
Bartlett's K-squared = 4.8291, df = 2, p-value = 0.0894
```

Bartlett's test?

- No real need to test variances!
  - You can perform one-way ANOVA allowing for unequal variances!
  - You can perform one-way ANOVA – using the regression framework with robust standard errors!

One-Way ANOVA allowing for unequal variances

```
> oneway.test(chol ~ as.factor(rs174548))

        One-way analysis of means (not assuming equal variances)

data:  chol and as.factor(rs174548)
F = 4.3258, num df = 2.000, denom df = 73.284, p-value = 0.01676
```

One-Way ANOVA with robust standard errors

```
> summary(gee(chol ~ as.factor(rs174548), id=seq(1,length(chol))))
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
         (Intercept) as.factor(rs174548)1 as.factor(rs174548)2
          181.061674             6.802272             5.438326

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent

Call:
gee(formula = chol ~ as.factor(rs174548), id = seq(1, length(chol)))

Summary of Residuals:
         Min           1Q       Median           3Q          Max
-64.06167401 -15.91337769  -0.06167401  14.93832599  59.13605442

Coefficients:
                       Estimate Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept)          181.061674   1.455346 124.411431    1.400016 129.328297
as.factor(rs174548)1   6.802272   2.321365   2.930290    2.402005   2.831914
as.factor(rs174548)2   5.438326   4.539833   1.197913    3.624271   1.500530

Estimated Scale Parameter:  480.7932
Number of Iterations:  1
```


Kruskal-Wallis Test

- Non-parametric analogue to the one-way ANOVA
  - Based on ranks
- In our example:

```
> kruskal.test(chol ~ as.factor(rs174548))

        Kruskal-Wallis rank sum test

data:  chol by as.factor(rs174548)
Kruskal-Wallis chi-squared = 7.4719, df = 2, p-value = 0.02385
```

  - Conclusion:
    - Evidence that the cholesterol distribution is not the same across all groups.
    - With the global null rejected, you can also perform pairwise comparisons [Wilcoxon rank sum], but adjust for multiplicities!

Multiple Comparisons (following Kruskal-Wallis Test)

```
> wilcox.test(chol[rs174548!=0] ~rs174548[rs174548!=0])

        Wilcoxon rank sum test with continuity correction

data:  chol[rs174548 != 0] by rs174548[rs174548 != 0]
W = 1974.5, p-value = 0.789
alternative hypothesis: true location shift is not equal to 0


> wilcox.test(chol[rs174548!=1] ~rs174548[rs174548!=1])

        Wilcoxon rank sum test with continuity correction

data:  chol[rs174548 != 1] by rs174548[rs174548 != 1]
W = 2482, p-value = 0.1849
alternative hypothesis: true location shift is not equal to 0


> wilcox.test(chol[rs174548!=2] ~rs174548[rs174548!=2])

        Wilcoxon rank sum test with continuity correction

data:  chol[rs174548 != 2] by rs174548[rs174548 != 2]
W = 14025.5, p-value = 0.009221
alternative hypothesis: true location shift is not equal to 0
```


Summary:

GOAL: Comparison of Means across K groups

Relationships:

$\mu_1 = \beta_0$

$\mu_2 = \beta_0 + \beta_1$

$\mu_3 = \beta_0 + \beta_2$

...

$\mu_K = \beta_0 + \beta_{K-1}$

One-way ANOVA:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: not all means are equal

Multiple Regression:

Model: $E[Y|groups] = \beta_0 + \beta_1 group_2 + \ldots + \beta_{k-1} group_k$

where $group_1$ is the reference group

H_0: $\beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$

H_1: not all β_i are equal to zero

Rejected H_0?

YES

Multiple Comparisons (control α overall)

e.g. Bonferroni: α/#comparisons

ANOVA

Two-way ANOVA models

ANOVA: Two-Way Model
Motivation:

- Scientific question:
  - Assess the effect of rs174548 and gender on cholesterol levels.

ANOVA: Two-Way Model

- Factors: A and B
- Goals:
  - Test for main effect of A
  - Test for main effect of B
  - Test for interaction effect of A and B

ANOVA: Two-Way Model

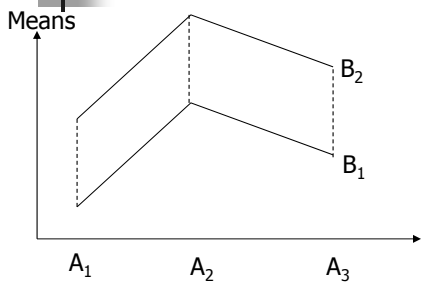
- To simplify discussion, assume that factor A has three levels, while factor B has two levels

Factor A

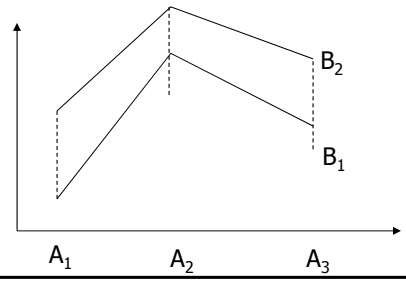
Factor B

	A_1	A_2	A_3
B_1	μ_{11}	μ_{21}	μ_{31}
B_2	μ_{12}	μ_{22}	μ_{32}

ANOVA: Two-Way Model



Parallel lines = No interaction



Lines are not parallel = Interaction

ANOVA: Two-Way Model

- Recall:
  - Categorical variables can be represented with “dummy” variables
  - Interactions are represented with “cross-products”

ANOVA: Two-Way Model

- Model 1:
  $E[Y|A_2, A_3, B_2] = \beta_0 + \beta_1 A_2 + \beta_2 A_3 + \beta_3 B_2$.
  - What are the means in each combination-group?

	A_1	A_2	A_3
B_1	$\mu_{11}=\beta_0$	$\mu_{21}=\beta_0+\beta_1$	$\mu_{31}=\beta_0+\beta_2$
B_2	$\mu_{12}=\beta_0+\beta_3$	$\mu_{22}=\beta_0+\beta_1+\beta_3$	$\mu_{32}=\beta_0+\beta_2+\beta_3$

ANOVA: Two-Way Model

- Model 1:
  $E[Y|A_2, A_3, B_2] = \beta_0 + \beta_1 A_2 + \beta_2 A_3 + \beta_3 B_2$.

	A_1	A_2	A_3
B_1	$\mu_{11}=\beta_0$	$\mu_{21}=\beta_0+\beta_1$	$\mu_{31}=\beta_0+\beta_2$
B_2	$\mu_{12}=\beta_0+\beta_3$	$\mu_{22}=\beta_0+\beta_1+\beta_3$	$\mu_{32}=\beta_0+\beta_2+\beta_3$

Model with no interaction:
- Difference in means between groups defined by factor B does not depend on the level of factor A.
- Difference in means between groups defined by factor A does not depend on the level of factor B.

ANOVA: Two-Way Model

- Model 2:
  $E[Y|A_2, A_3, B_2] = \beta_0 + \beta_1 A_2 + \beta_2 A_3 + \beta_3 B_2 + \beta_4 A_2 B_2 + \beta_5 A_3 B_2$
  - What are the means in each combination-group?

	A_1	A_2	A_3
B_1	$\mu_{11}=\beta_0$	$\mu_{21}=\beta_0+\beta_1$	$\mu_{31}=\beta_0+\beta_2$
B_2	$\mu_{12}=\beta_0+\beta_3$	$\mu_{22}=\beta_0+\beta_1+\beta_3+\beta_4$	$\mu_{32}=\beta_0+\beta_2+\beta_3+\beta_5$

ANOVA: Two-Way Model

- Three (possible) tests
  - Interaction of A and B (may want to start here)
    - Rejection would imply that differences between means of A depends on the level of B (and vice-versa) so stop
  - Main effect of A
    - Test only if no interaction
  - Main effect of B
    - Test only if no interaction

[Note: If you have one observation per cell, you cannot test interaction!]

ANOVA: Two-Way Model

- Model without interaction
  $E[Y|A_2, A_3, B_2] = \beta_0 + \beta_1 A_2 + \beta_2 A_3 + \beta_3 B_2$.

  How do we test for main effect of factor A?

  H_0: $\beta_1 = \beta_2 = 0$ vs. H_1: β_1 or β_2 not zero

  How do we test for main effect of factor B?

  H_0: $\beta_3 = 0$ vs. H_1: β_3 not zero

ANOVA: Two-Way Model

- Model with interaction:
  $E[Y|A_2, A_3, B_2] = \beta_0 + \beta_1 A_2 + \beta_2 A_3 + \beta_3 B_2 + \beta_4 A_2 B_2 + \beta_5 A_3 B_2$

  How do we test for interactions?

  H_0: $\beta_4 = \beta_5 = 0$ vs.

  H_1: β_4 or β_5 not zero

IMPORTANT:

If you reject the null, do not test main effects!!!

ANOVA: Two-Way Model (without interaction)

```
> fit1 = lm(chol ~ as.factor(sex) + as.factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ as.factor(sex) + as.factor(rs174548))

Residuals:
     Min       1Q   Median       3Q      Max
-66.6534 -14.4633  -0.6008  15.4450  57.6350

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           175.365      1.786  98.208  < 2e-16 ***
as.factor(sex)1        11.053      2.126   5.199 3.22e-07 ***
as.factor(rs174548)1    7.236      2.250   3.215  0.00141 **
as.factor(rs174548)2    5.184      4.398   1.179  0.23928
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.24 on 396 degrees of freedom
Multiple R-squared: 0.08458,    Adjusted R-squared: 0.07764
F-statistic:  12.2 on 3 and 396 DF,  p-value: 1.196e-07

> anova(fit0,fit1)
Analysis of Variance Table

Model 1: chol ~ as.factor(sex)
Model 2: chol ~ as.factor(sex) + as.factor(rs174548)
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1    398 183480
2    396 178681  2    4799.1 5.318 0.005259 **
```


ANOVA: Two-Way Model (without interaction)


```
> fit1 = lm(chol ~ as.factor(sex) + as.factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ as.factor(sex) + as.factor(rs174548))

Residuals:
     Min       1Q   Median       3Q      Max
-66.6534 -14.4633  -0.6008  15.4450  57.6350

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           175.365      1.786  98.208  < 2e-16 ***
as.factor(sex)1        11.053      2.126   5.199 3.22e-07 ***
as.factor(rs174548)1    7.236      2.250   3.215  0.00141 **
as.factor(rs174548)2    5.184      4.398   1.179  0.23928

Residual standard error: 21.24 on 396 degrees of freedom
Multiple R-squared: 0.08458,    Adjusted R-squared: 0.07764
F-statistic:  12.2 on 3 and 396 DF,  p-value: 1.196e-07

> anova(fit0,fit1)
Analysis of Variance Table

Model 1: chol ~ as.factor(sex)
Model 2: chol ~ as.factor(sex) + as.factor(rs174548)
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1    398 183480
2    396 178681  2    4799.1 5.318 0.005259 **
```


- Interpretation of results:
  - Estimated mean cholesterol for male C/C group: 175.37 mg/dl
  - Estimated difference in mean cholesterol levels between females and males adjusted by genotype: 11.053 mg/dl
  - Estimated difference in mean cholesterol levels between C/G and C/C groups adjusted by gender: 7.236 mg/dl
  - Estimated difference in mean cholesterol levels between G/G and C/C groups adjusted by gender: 5.184 mg/dl
  - There is evidence that cholesterol is associated with gender ($p < 0.001$).
  - There is evidence that cholesterol is associated with genotype ($p=0.005$)

ANOVA: Two-Way Model (without interaction)

- In words:
  - Adjusting for sex, the difference in mean cholesterol comparing C/G to C/C is 7.236 and comparing G/G to C/C is 5.184.
    - This difference does not depend on sex
      - (this is because the model does not have an interaction between sex and genotype!)

ANOVA: Two-Way Model (with interaction)

```
> fit2 = lm(chol ~ as.factor(sex) * as.factor(rs174548))
> summary(fit2)

Call:
lm(formula = chol ~ as.factor(sex) * as.factor(rs174548))

Residuals:
     Min       1Q   Median       3Q      Max
-70.5286 -13.6037  -0.9736  14.1709  54.8818

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           178.1182     2.0089  88.666  < 2e-16 ***
as.factor(sex)1                         5.7109     2.7982   2.041  0.04192 *
as.factor(rs174548)1                    0.9597     3.1306   0.307  0.75933
as.factor(rs174548)2                   -0.2015     6.4053  -0.031  0.97492
as.factor(sex)1:as.factor(rs174548)1  12.7398     4.4650   2.853  0.00456 **
as.factor(sex)1:as.factor(rs174548)2  10.2296     8.7482   1.169  0.24297
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.07 on 394 degrees of freedom
Multiple R-squared: 0.1039,     Adjusted R-squared: 0.09257
F-statistic:  9.14 on 5 and 394 DF,  p-value: 3.062e-08
```

ANOVA: Model comparison

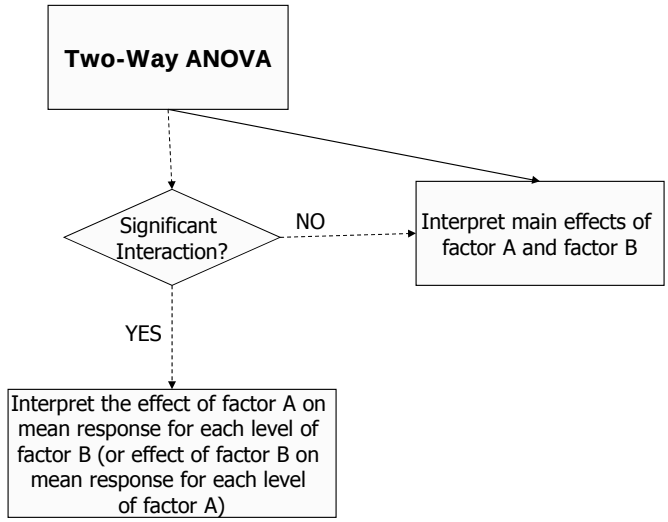
```
> anova(fit1,fit2)
Analysis of Variance Table

Model 1: chol ~ as.factor(sex) + as.factor(rs174548)
Model 2: chol ~ as.factor(sex) * as.factor(rs174548)
  Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
1    396 178681
2    394 174902   2      3779 4.2564 0.01483 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


ANOVA: Two-Way Model (with interaction)

```
> fit2 = lm(chol ~ as.factor(sex) * as.factor(rs174548))
> summary(fit2)

Call:
lm(formula = chol ~ as.factor(sex) * as.factor(rs174548))

Residuals:
     Min       1Q   Median       3Q      Max
-70.5286 -13.6037  -0.9736  14.1709  54.8818

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         178.1182     2.0089  88.666  < 2e-16 ***
as.factor(sex)1                       5.7109     2.7982   2.041  0.04192 *
as.factor(rs174548)1                  0.9597     3.1306   0.307  0.75933
as.factor(rs174548)2                 -0.2015     6.4053  -0.031  0.97492
as.factor(sex)1:as.factor(rs174548)1  12.7398     4.4650   2.853  0.00456 **
as.factor(sex)1:as.factor(rs174548)2  10.2296     8.7482   1.169  0.24297
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.07 on 394 degrees of freedom
Multiple R-squared: 0.1039,     Adjusted R-squared: 0.09257
F-statistic:  9.14 on 5 and 394 DF,  p-value: 3.062e-08
```

```
> anova(fit1,fit2)
Analysis of Variance Table

Model 1: chol ~ as.factor(sex) + as.factor(rs174548)
Model 2: chol ~ as.factor(sex) * as.factor(rs174548)
  Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
1    396 178681
2    394 174902   2      3779 4.2564 0.01483 *
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interpretation of results:
  - Estimated mean cholesterol for male C/C group: 178.12 mg/dl
  - Estimated mean cholesterol for female C/C group? (178.12 + 5.7109) mg/dl
  - Estimated mean cholesterol for male C/G group: (178.12 +0.9597) mg/dl
  - Estimated mean cholesterol for female C/G group: (178.12 + 5.7109 + 0.9597 + 12.7398) mg/dl
  - …
- There is evidence for an interaction between sex and genotype (p= 0.015)



ANCOVA MODELS

(aka ANACOVA)

ANalysis of COVAriance Models (ANCOVA) Motivation:

- Scientific question:
  - Assess the effect of rs174548 on cholesterol levels adjusting for age

ANalysis of COVAriance Models (ANCOVA)

- ANOVA with one or more continuous variables
  - Equivalent to regression with “dummy” variables and continuous variables
  - Primary comparison of interest is across k groups defined by a categorical variable, but the k groups may differ on some other potential predictor or confounder variables [also called covariates].

ANalysis of COVAriance Models (ANCOVA)

- To facilitate discussion assume
  - Y: continuous response (e.g. cholesterol)
  - X: continuous variable (e.g. age)
  - Z: dummy variable (e.g. indicator of C/G or G/G versus C/C)
- Model: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 \boxed{XZ} + \varepsilon$

Interaction term

Note that:

$$Z = 0 \Rightarrow E[Y \mid X, Z = 0] = \beta_0 + \beta_1 X$$

$$Z = 1 \Rightarrow E[Y \mid X, Z = 1] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X$$

This model allows for different intercepts/slopes for each group.

ANCOVA

- Testing coincident lines: $H_0 : \beta_2 = 0, \beta_3 = 0$
  - Compares overall model with reduced model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Testing parallelism: $H_0 : \beta_3 = 0$
  - Compares overall model with reduced model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

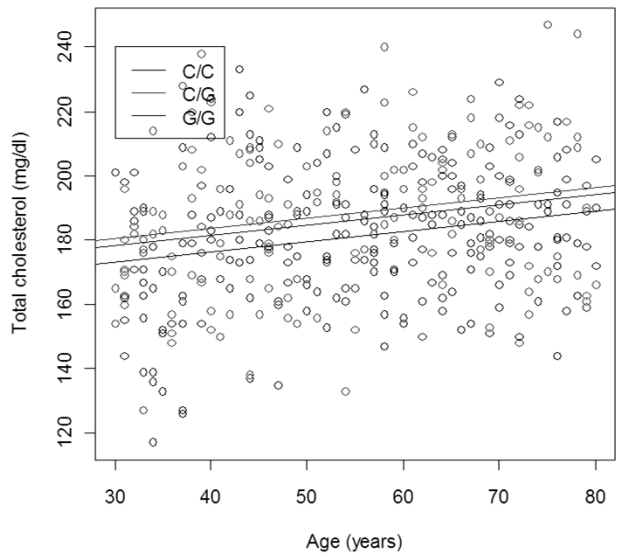
ANCOVA

```
> fit0 = lm(chol ~ as.factor(rs174548))
> summary(fit0)
Call:
lm(formula = chol ~ as.factor(rs174548))

Residuals:
      Min        1Q    Median        3Q       Max
-64.06167 -15.91338  -0.06167  14.93833  59.13605

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           181.062      1.455 124.411  < 2e-16 ***
as.factor(rs174548)1    6.802      2.321   2.930  0.00358 **
as.factor(rs174548)2    5.438      4.540   1.198  0.23167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.93 on 397 degrees of freedom
Multiple R-squared: 0.0221,     Adjusted R-squared: 0.01718
F-statistic: 4.487 on 2 and 397 DF,  p-value: 0.01184

> anova(fit0)
Analysis of Variance Table
Response: chol
                     Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(rs174548)   2   4314    2157  4.4865 0.01184 *
Residuals           397 190875     481
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


ANCOVA

```
> fit1 = lm(chol ~ as.factor(rs174548) + age)
> summary(fit1)
Call:
lm(formula = chol ~ as.factor(rs174548) + age)

Residuals:
     Min       1Q   Median       3Q      Max
-57.2089 -14.4293   0.4443  14.2652  55.8985

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          163.28125    4.36422  37.414  < 2e-16 ***
as.factor(rs174548)1   7.30137    2.27457   3.210  0.00144 **
as.factor(rs174548)2   5.08431    4.44331   1.144  0.25321
age                    0.32140    0.07457   4.310 2.06e-05 ***

Residual standard error: 21.46 on 396 degrees of freedom
Multiple R-squared: 0.06592,    Adjusted R-squared: 0.05884
F-statistic: 9.316 on 3 and 396 DF,  p-value: 5.778e-06

> anova(fit0,fit1)
Analysis of Variance Table

Model 1: chol ~ as.factor(rs174548)
Model 2: chol ~ as.factor(rs174548) + age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    397 190875
2    396 182322  1    8552.9 18.577 2.062e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

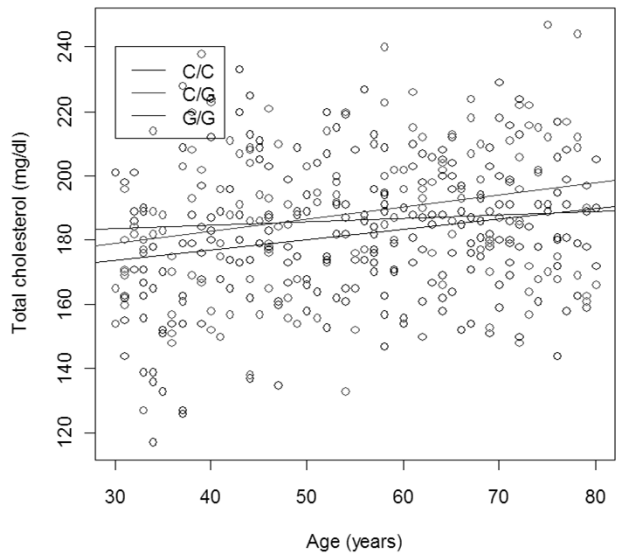
ANCOVA



ANCOVA

```
> fit2 = lm(chol ~ as.factor(rs174548) * age)
> summary(fit2)
Call:
lm(formula = chol ~ as.factor(rs174548) * age)

Residuals:
     Min       1Q   Median       3Q      Max
-57.5425 -14.3002   0.7131  14.2138  55.7089

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              164.14677    5.79545  28.323  < 2e-16 ***
as.factor(rs174548)1       3.42799    8.79946   0.390  0.69707
as.factor(rs174548)2      16.53004   18.28067   0.904  0.36642
age                        0.30576    0.10154   3.011  0.00277 **
as.factor(rs174548)1:age   0.07159    0.15617   0.458  0.64692
as.factor(rs174548)2:age  -0.20255    0.31488  -0.643  0.52043

Residual standard error: 21.49 on 394 degrees of freedom
Multiple R-squared: 0.06777,    Adjusted R-squared: 0.05594
F-statistic: 5.729 on 5 and 394 DF,  p-value: 4.065e-05

> anova(fit1,fit2)
Analysis of Variance Table

Model 1: chol ~ as.factor(rs174548) + age
Model 2: chol ~ as.factor(rs174548) * age
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    396 182322
2    394 181961  2    361.11 0.391 0.6767
```

ANCOVA



ANCOVA

- In summary:
  - If the slopes are not equal, then age is an effect modifier

$$E[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 (CG) + \beta_3 (GG) + \beta_4 (x * CG) + \beta_5 (x * GG)$$

  - If the slopes are the same,

$$E[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 (CG) + \beta_3 (GG)$$

ANCOVA

- If the slopes are the same,

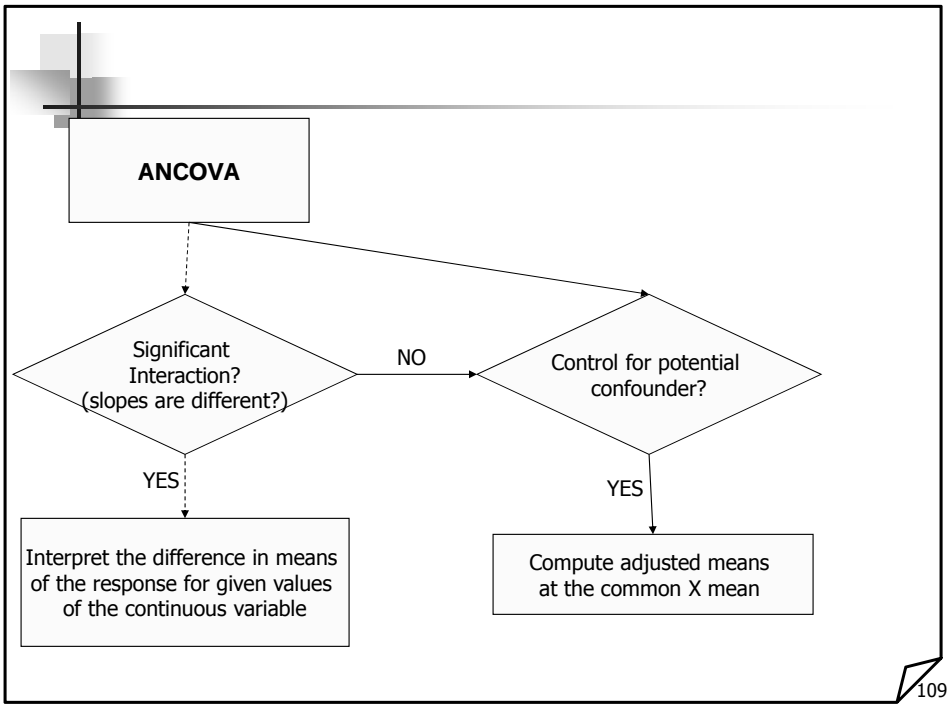
$$E[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 (CG) + \beta_3 (GG)$$

  - then one can obtain adjusted means for the three genotypes using the mean age over all groups
    - For example, the adjusted means for the four groups would be

$$\overline{Y}_1(\text{adj}) = \hat{\beta}_0 + \bar{x}\hat{\beta}_1$$
$$\overline{Y}_2(\text{adj}) = (\hat{\beta}_0 + \hat{\beta}_2) + \bar{x}\hat{\beta}_1$$
$$\overline{Y}_3(\text{adj}) = (\hat{\beta}_0 + \hat{\beta}_3) + \bar{x}\hat{\beta}_1$$

ANCOVA

```
> ## unadjusted mean cholesterol levels for different genotypes
> predict(fit0, new=data.frame(rs174548=0))
       1
181.0617
> predict(fit0, new=data.frame(rs174548=1))
       1
187.8639
> predict(fit0, new=data.frame(rs174548=2))
    1
186.5

> ## mean cholesterol for different genotypes adjusted by age
> predict(fit1, new=data.frame(age=mean(age),rs174548=0))
       1
180.9013
> predict(fit1, new=data.frame(age=mean(age),rs174548=1))
       1
188.2026
> predict(fit1, new=data.frame(age=mean(age),rs174548=2))
       1
185.9856
```



Experimental Designs & ANOVA

- This section is not intended to be comprehensive
- No endorsement for any of the articles cited here

Tool Kit

- Controls and Placebos:
  - Provides a baseline comparison with test groups
- Blinding:
  - When successfully applied, it eliminates the possibility that the end comparison measures expectations rather than real treatment differences
- Blocking:
  - Arranges units into homogeneous subgroups so that treatments can be randomly assigned to units within each block
    - Improves precision for treatment comparisons
    - Controls for confounding variables by grouping experimental units into blocks with similar values of the variable

Tool Kit

- Stratification
  - Involves partitioning of population units into homogeneous subgroups – called strata – and performing random sampling of population units in each strata
  - (stratification pertains to random sampling; blocking pertains to random assignment)
- Covariates
  - Inclusion may control for potentially confounding factors
  - Inclusion may improve precision in treatment comparisons
- Randomization
  - Allows for controlling for factors not explictly controlled for in the design (by blocking) or in the analysis (by covariates)
  - Enables causal inferences

Tool Kit

- Random Sampling
  - Means employing a random procedure to select units from a population
    - To ensure that sample is representative of the population
    - To permit an inference that patterns observed in the sample are characteristic of patterns in the population as a whole
- Replication
  - It refers to assigning one treatment to multiple units within each block.
    - Increases precision for treatment effects (increased sample size)
    - Allows for model assessment
- Balance
  - Same number of units to each treatment
    - Optimizes precision for treatment comparisons

Terminology

- Treatments
  - A factor level in a single-factor study or a combination of factor levels in a multi-factor study
    - How many factors should be examined?
    - How many levels should each factor have?
- Experimental units
  - Smallest unit of the experiment such that any two different experimental units may receive different treatments

One-Way Data Patterns

Factor

YYYY	YYYY	YYYY

Equal number of replicates per treatment

Factor

YY	YYYYY	YYY

Unequal number of replicates per treatment

"Dictionary":
- Factor: categorical predictor
- Levels: categories of the predictor variable

Two-Way Data Patterns

Factor 1

Factor 2

Y	Y	Y
Y	Y	Y
Y	Y	Y
Y	Y	Y

Single observation per cell

Factor 1

Factor 2

YYY	YYY	YYY
YYY	YYY	YYY
YYY	YYY	YYY
YYY	YYY	YYY

Equal replication per cell

Factor 1

Factor 2

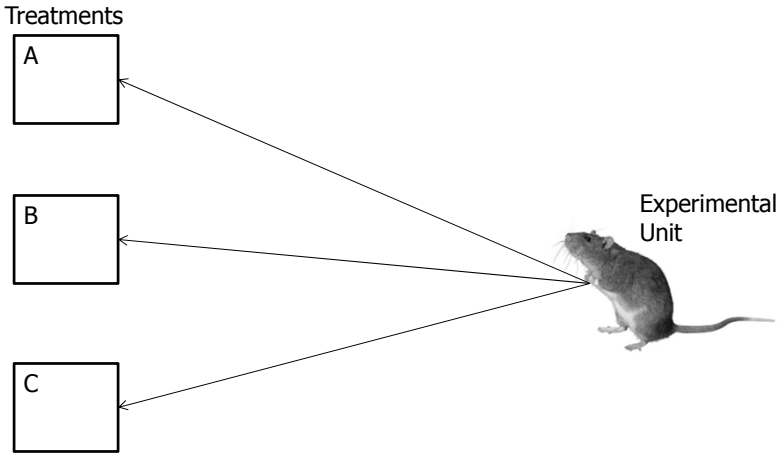
YY	YYY	YYYYYY
YYY	YYYY	YY
Y	YYY	YYYY
YYYYY	YY	Y

Non-systematic replications

Completely Randomized Design

- Treatments are allocated to the experimental units completely at random
  - Every experimental unit has an equal chance of receiving any of the treatments
- Simple & flexible
  - Allows for any number of treatments
  - Sample sizes can vary from treatment to treatment
- Inefficient when the experimental units are heterogeneous

Completely Randomized Design



Statistical model? <u>One-way ANOVA model</u>

Completely Randomized Design: an Example

- **Title: "Hepatocyte growth factor incorporated chitosan nanoparticles augment the differentiation of stem cell into hepatocytes for the recovery of liver cirrhosis in mice."**
  - Authors: Pulavendran S, Rose C, Mandal AB. J Nanobiotechnology. 2011 Apr 28;9:15.

- **Abstract [partial]:**
  - **BACKGROUND:** Short half-life and low levels of growth factors in the niche of injured microenvironment necessitates the exogenous and sustainable delivery of growth factors along with stem cells to augment the regeneration of injured tissues.
  - **METHODS:** Recombinant human hepatocyte growth factor (HGF) was incorporated into chitosan nanoparticles (CNP) by ionic gelation method and studied for its morphological and physiological characteristics. Cirrhotic mice received either hematopoietic stem cells (HSC) or mesenchymal stemcells (MSC) with or without HGF incorporated chitosan nanoparticles (HGF-CNP) and saline as control. Biochemical, histological, immunostaining and gene expression assays were carried out using serum and liver tissue samples […].
  - **RESULTS:** Serum levels of selected liver protein and enzymes were significantly increased in the combination of MSC and HGF-CNP (MSC+HGF-CNP) treated group.
  - **CONCLUSION:** […] Transplantation of bone marrow MSC in combination with HGF-CNP could be an ideal approach for the treatment of liver cirrhosis.

Completely Randomized Design: Exercise

- What is the goal of the experiment?
- What is(are) the response variables?
- What are the factors?
- How many levels?
- **Statistical model?**

Factorial Design

- A factorial design is used to evaluate <u>two or more factors</u> simultaneously.
- Factorial designs are more efficient than one-factor-at-a-time designs
- Factorial designs allow for investigations of interactions.

Factorial Design: an example

- **Title: "Fermentable fiber ameliorates fermentable protein-induced changes in microbial ecology, but not the mucosal response, in the colon of piglets".**
  - Pieper R, Kröger S, Richter JF, Wang J, Martin L, Bindelle J, Htoo JK, von Smolinski D, Vahjen W, Zentek J, Van Kessel AG. J Nutr. 2012 Apr;142(4):661-7. Epub 2012 Feb 22.
- **Abstract (partial):** Dietary inclusion of fermentable carbohydrates (fCHO) is reported to reduce large intestinal formation of putatively toxic metabolites derived from fermentable proteins (fCP). However, the influence of diets high in fCP concentration on epithelial response and interaction with fCHO is still unclear. Thirty-two weaned piglets were fed 4 diets in a **2 × 2 factorial design** with low fCP/low fCHO [14.5% crude protein (CP)/14.5% total dietary fiber (TDF)]; low fCP/high fCHO (14.8% CP/16.6% TDF); high fCP low fCHO (19.8% CP/14.5% TDF); and high fCP/high fCHO (20.1% CP/18.0% TDF) as dietary treatments. After 21-23 d, pigs were killed and colon digesta and tissue samples analyzed for indices of microbial ecology, tissue expression of genes for cell turnover, cytokines, mucus genes (MUC), and oxidative stress indices. Pig performance was unaffected by diet. […] High dietary fCP increased ($P < 0.05$) expression of PCNA, IL1β, IL10, TGFβ, MUC1, MUC2, and MUC20, irrespective of fCHO concentration.

Factorial Design: Exercise

- What is the goal of the experiment?
- What is(are) the response variables?
- What are the factors?
- For each factor, how many levels?
- How many treatments?
- **Statistical model?**

Factorial Design: an example

TABLE 3 Relative mRNA abundance of proliferating cell nuclear antigen, caspase 3, pro- and antiinflammatory cytokines, and mucus genes in the colon of piglets fed diets containing a low or high concentration of fCHO or fCP[1,2]

	Low fCP		High fCP		*P* value[3]		
Gene	Low fCHO	High fCHO	Low fCHO	High fCHO	fCHO	fCP	fCHO x fCP
PCNA	0.81 ± 0.05	0.79 ± 0.04	0.89 ± 0.08	0.90 ± 0.04	0.94	<0.05	0.76
CASP	0.80 ± 0.04	0.85 ± 0.06	0.88 ± 0.06	0.85 ± 0.04	0.83	0.46	0.37
IL1β	0.87 ± 0.11	0.89 ± 0.07	1.01 ± 0.10	1.05 ± 0.07	0.71	<0.05	0.89
IL6	0.76 ± 0.13	0.81 ± 0.15	1.04 ± 0.19	1.01 ± 0.15	0.96	0.07	0.77
IL10	0.92 ± 0.07	0.90 ± 0.09	1.09 ± 0.08	1.05 ± 0.04	0.61	<0.05	0.86
TGFβ	0.88 ± 0.09	0.85 ± 0.10	1.11 ± 0.09	1.07 ± 0.05	0.61	<0.01	0.93
MUC1	0.71 ± 0.11	0.73 ± 0.09	0.89 ± 0.09	0.87 ± 0.08	0.83	0.05	0.61
MUC2	0.84 ± 0.14	0.82 ± 0.09	1.05 ± 0.10	1.00 ± 0.08	0.97	0.05	0.79
MUC20	0.81 ± 0.05	0.79 ± 0.04	0.89 ± 0.08	0.90 ± 0.04	0.72	<0.05	0.85

[1] Data are mean ± SE, n = 8/group. fCHO, fermentable carbohydrate; fCP, fermentable crude protein.
[2] Values are given as arbitrary values based on standard curves using pooled RNA samples. The mRNA abundance was normalized using *18S* rRNA, 60S ribosomal protein L19 (*RPL19*), hypoxanthine phosphoribosyltransferase I (*HPRTI*), and β-Actin as housekeeping genes.
[3] The *P* values indicate main effects for fCP and fCHO, respectively.

Are these results unexpected?
Any concerns?

Randomized Complete Block Designs

- Experimental units are assigned to homogeneous groups (aka "blocks").
  - Reduces the variation and increases the precision of treatment comparisons
- Members of each block are randomly assigned to different treatments.
  - Randomized complete block design: each block contains all treatment combinations
  - Randomized incomplete block design: number of treatments exceeds the number of units in each block

Randomized Complete Block Designs

Large N-S variability

Small E-W variability

Within each block, a separate randomization allocates treatments to experimental units

Block 1	C	A	B	D
Block 2	D	B	A	C
Block 3	A	B	D	C

Randomized Complete Block Designs

- Factors:
  - Block (control factor)
  - Treatment (factor of interest)

- **Statistical Model**
  - Two-way ANOVA model
    - (additive model with single replication)

Randomized Complete Block Designs: An example

A researcher studied the effects of three experimental diets with varying fat contents on the total lipid (fat) level in plasma. Total lipid level is a widely used predictor of coronary heart disease. Fifteen male subjects who were within 20% of their ideal body weight were grouped into five blocks according to age. Within each block, the three experimental diets were randomly assigned to three subjects. Data on reduction in lipid level (in grams per liter) after the subjects were on the diet for a fixed period of time were recorded.

Randomized Complete Block Designs: An example

	Fat Content of Diet		
Age Group	Extremely Low	Fairly Low	Moderately Low
Ages 15-24	0.73	0.67	0.15
Ages 25-34	0.86	0.75	0.21
Ages 35-44	0.94	0.81	0.26
Ages 45-54	1.4	1.32	0.75
Ages 55-64	1.62	1.41	0.78

Randomized Complete Block Designs: Exercise

- What is the goal of the experiment?
- What is (are) the response variables?
- What is the factor of interest? What is the blocking factor? For each factor, how many levels?
- How many treatments?
- **Statistical model?**

Randomized Complete Block Designs: Another example

TITLE: "UV REPAIR AND RESISTANCE TO SOLAR UV-B IN AMPHIBIAN EGGS - A LINK TO POPULATION DECLINES"

- **Author(s)**: BLAUSTEIN, AR (BLAUSTEIN, AR); HOFFMAN, PD (HOFFMAN, PD); HOKIT, DG (HOKIT, DG); KIESECKER, JM (KIESECKER, JM); WALLS, SC (WALLS, SC); HAYS, JB (HAYS, JB) Source: PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA Volume: 91 Issue: 5 Pages: 1791-1795
- **Abstract [partial]**: The populations of many amphibian species, in widely scattered habitats, appear to be in severe decline; other amphibians show no such declines. There is no known single cause for the declines, but their widespread distribution suggests involvement of global agents-increased UV-B radiation, for example. We addressed the hypothesis that differential sensitivity among species to UV radiation contributes to these population declines. We focused on species-specific differences in the abilities of eggs to repair UV radiation damage to DNA and differential hatching success of embryos exposed to solar radiation at natural oviposition sites. Quantitative comparisons of activities of a key UV-damage-specific repair enzyme, photolyase, among oocytes and eggs from 10 amphibian species were reproducibly characteristic for a given species but varied > 80-fold among the species. Levels of photolyase generally correlated with expected exposure of eggs to sunlight. Among the frog and toad species studied, the highest activity was shown by the Pacific treefrog (Hyla regilla), whose populations are not known to be in decline. The Western toad (Bufo boreas) and the Cascades frog (Rana cascadae), whose populations have declined markedly, showed significantly lower photolyase levels. […] These observations are thus consistent with the UV-sensitivity hypothesis.

Randomized Complete Block Designs: Another example

- Goal: Is the failure rate different for species with different levels of activity of photolyase?
- Factors:
  - UV-B Filter:
    - UV-B blocking filter
    - UV-B transmitting filter
    - No Filter
  - Species:
    - Toad (Bufo boreas)
    - Tree frog (Hyla regilla)
    - Cascade frog (Rana cascadae)
- Randomization:
  - Filtering treatments and egg species randomly assigned to enclosures constructed to contain clusters of 150 eggs

Randomized Complete Block Designs: Another example

- Four sites: [three with single species]
  - Sparks Lake (tree frog)
  - Small Lake (Cascade frog)
  - Lost Lake (toad)
  - Three Creeks (all three species)
- Only eggs of naturally occurring species were assigned to enclosures at each site
- Blocking factor: Amphibian species/sites
  - At Three Creeks: experiment is a 3 by 3 factorial design
  - At other sites: single factor experiment

Randomized Complete Block Designs

Large N-S variability

Small E-W variability

Within each block, a separate randomization allocates treatments to experimental units

Block 1	C	A	B	D
Block 2	D	B	A	C
Block 3	A	B	D	C

What if need to control (large) variability in both N-S and E-S directions???

Latin Square Designs

- Employs two blocking variables ("row" and "column")
  - Allows for better control of experimental variation
- Features:
  - There are r treatments
  - There are two blocking variables; each with r categories
  - Each row and each column in the design contains all treatments
  - Only one treatment per combination block

Latin Square Designs

Latin square for 3 treatments

A	B	C
C	A	B
B	C	A

Each treatment appears exactly once in each column and in each row.

Latin square for 4 treatments

A	B	D	C
D	C	A	B
B	D	C	A
C	A	B	D

Latin Square Designs: An example

- In a study of chemotherapy treatments for breast cancer, researchers wanted to control for the effects of age and BMI.

		Age (years)			
		[40,50)	[50,60)	[60,70)	70+
BMI	<20	A	B	C	D
	[20,25)	B	C	D	A
	[25,30)	C	D	A	B
	30+	D	A	B	C

Latin Square Designs: randomization

- Randomization is a bit complex because there are multiple possible Latin squares.
  - Example:
    - For r = 4, there are 576 possible Latin squares (4 are of standard form).
    - A Latin square is said to be in standard form (also, normalized or reduced) if both its first row and its first column are in their natural order. For example, for r=4,

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Latin Square Designs: randomization

- One chooses one Latin square randomly in a particular experiment.
  - This may be done by writing down any legitimate Latin square and then randomly permuting rows and columns.
    - "Algorithm":
      - Choose a standard Latin square (may or not be at random).
      - Randomly permute all rows.
      - Randomly permute all columns.
      - Randomly assign treatments to the letters A, B, C, etc.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Rows: (2,4,1,3)

B	C	D	A
D	A	B	C
A	B	C	D
C	D	A	B

Columns: (3,4,2,1)

D	**A**	**C**	**B**
B	**C**	**A**	**D**
C	**D**	**B**	**A**
A	**B**	**D**	**C**

Latin Square Designs

- Factors:
  - Row (blocking factor 1)
  - Column (blocking factor 2)
  - Treatment (factor of interest)

- **Statistical Model**
  - Three-way ANOVA model
    - (additive model with single replication)

Everything is regression!

(Professor Scott Emerson)

Regression Lab 1

The data set cholesterol.txt available on your thumb drive contains the following variables:

Field Descriptions

`ID`: Subject ID

`sex`: Sex: 0 = male, 1 = female

`age`: Age in years

`chol`: Serum total cholesterol, mg/dl

`BMI`: Body-mass index, kg/m^2

`TG`: Serum triglycerides, mg/dl

`apoE`: Apolipoprotein E genotype, with six genotypes coded 1-6: 1 = e2/e2, 2 = e2/e3, 3 = e2/e4, 4 = e3/e3, 5 = e3/e4, 6 = e4/e4

`rs174548`: Candidate SNP 1 genotype, chromosome 11, physical position 61,327,924. Coded as the number of minor alleles: 0 = C/C, 1 = C/G, 2 = G/G.

`rs4775401`: Candidate SNP 2 genotype, chromosome 15, physical position 59,476,915. Coded as the number of minor alleles: 0 = C/C, 1 = C/T, 2 = T/T.

The goal of the regression labs will be to use the data set to explore the relationship between triglycerides and several predictor variables. The objective of this first lab will be

- Become familiar with R and
- Begin to explore the cholesterol dataset.
- Use graphical methods to investigate associations between triglycerides and BMI

1. Start your R session.

2. Create a script file to record your R code. Open a script file by clicking on File -> New Script (for PC) or File-> New Document (for Mac).

3. Load the cholesterol data set.

4. Compute the sample mean, median and standard deviation of triglycerides.

5. View the boxplot, stem-and-leaf displays and histograms for triglycerides.

6. Create a variable called IBMI that takes the value 1 if BMI > 25 and 0 if BMI<=25 .

7. Compute summary measures of triglycerides for the two groups of subjects defined by IBMI.

8. Plot boxplots for triglycerides separately for the two groups of subjects defined by IBMI. Does there appear to be an association between BMI and triglycerides?

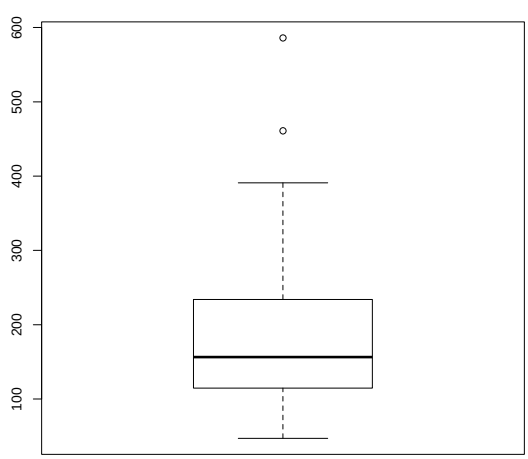
9. Plot a scatterplot of triglycerides vs BMI. Based on this plot does there appear to be an association between BMI and triglycerides? What can you additionally say about the relationship between these variables that was not possible using the boxplot?

10. Use regression to investigate the association between triglycerides and BMI. What do the linear regression model results tell us about the association?

11. Check your script file. Make sure that all important commands that you have used and any output you want to save are included in here.

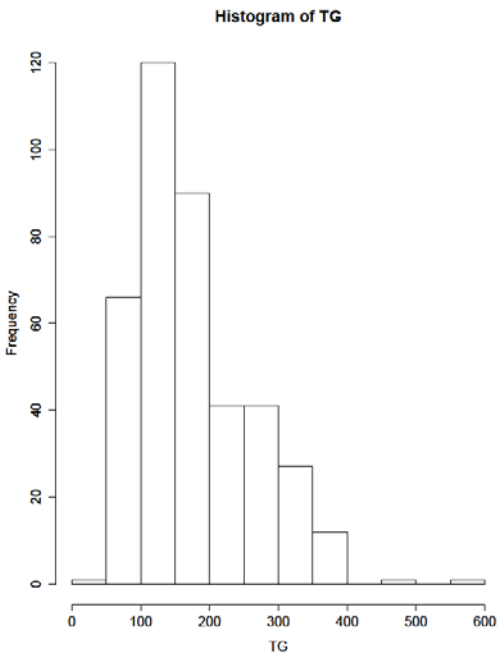
R Commands & Output:

```
> cholesterol = read.table("http://faculty.washington.edu/rhubb/sisg/SISG-Data-
cholesterol.txt", header=T)
> attach(cholesterol)
>
> # compute univariate summary statistics for triglycerides
> mean(TG)
[1] 177.44
> median(TG)
[1] 156.5
> sd(TG)
[1] 82.98323
> summary(TG)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   47.0   114.8   156.5   177.4   234.0   586.0
>
> # graphical displays for triglycerides
> boxplot(TG)
```

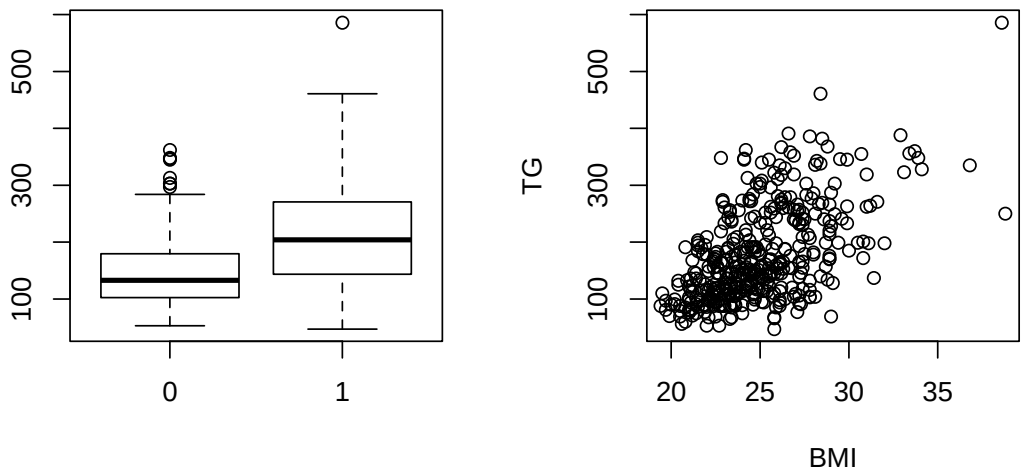



```
> stem(TG)
```

```
  The decimal point is 2 digit(s) to the right of the |

  0 |
  0 | 5556677777777788888888889999999999999999999999999999
  1 | 00000000000000000000011111111111111111111111222222222222222222222222+34
  1 | 55555555555555555555555555666666666666666666677777777777777777778888+18
  2 | 0000000000000111111111222233333333444444444444
  2 | 555555556666666666666777777777778888888899
  3 | 00000011122222334444 4
  3 | 5555555566666778999
  4 |
  4 | 6
  5 |
  5 | 9

> hist(TG)
>
```




```
> # create a binary indicator for BMI > 25
> ibmi = ifelse(BMI > 25, 1, 0)
>
> # compute univariate summary statistics for triglycerides for BMI > 25 and BMI <= 25
> tapply(TG,ibmi,mean)
       0        1
147.3839 215.6932
> tapply(TG,ibmi,median)
  0   1
133 204
> tapply(TG,ibmi,sd)
       0        1
61.70787 90.66584
>
> # plot boxplots for triglycerides separately by BMI > 25 and BMI <= 25
> par(mfrow = c(1,2))
> boxplot(TG ~ ibmi)
> # scatterplot of triglycerides vs BMI
> plot(BMI, TG)
```




```
> # fit linear regression models for the association between triglycerides and BMI
> fit1 = lm(TG ~ BMI)
> summary(fit1)

Call:
lm(formula = TG ~ BMI)

Residuals:
    Min      1Q  Median      3Q     Max 
-170.19  -45.10  -12.89   39.60  231.08 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -208.50      28.95  -7.203 2.97e-12 ***
BMI            15.44       1.15  13.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.93 on 398 degrees of freedom
Multiple R-Squared: 0.3118,     Adjusted R-squared: 0.3101 
F-statistic: 180.3 on 1 and 398 DF,  p-value: < 2.2e-16 

>
```

Regression Lab 2

The goal of this lab is to answer the following scientific questions using the cholesterol dataset.

- Are triglyceride levels associated with BMI?
- Are linear regression model assumptions satisfied for this relationship?
- Is the association between triglyceride and BMI modified by the ApoE4 allele?

1) Load the `gee` package.

2) Construct a scatterplot of triglycerides versus BMI. Are there any points that you suspect might have a large influence on the regression estimates?

3) Use regression to investigate the association between triglycerides and BMI.after removing the observations with BMI > 37. Do the points with BMI>37 appear to affect your results? How?

4) Use residuals analysis to check the linear regression model assumptions. Create a scatterplot of residuals vs fitted values and a quantile-quantile plot of residuals. Do any modeling assumptions appear to be violated? How do model results change if you use robust standard errors?

5) Investigate the association between triglycerides and BMI after log transforming triglycerides. Does this appear to correct violations of modeling assumptions?

6) Create a new binary variable indicating presence of the ApoE4 allele (apoE = 3, 5, or 6).

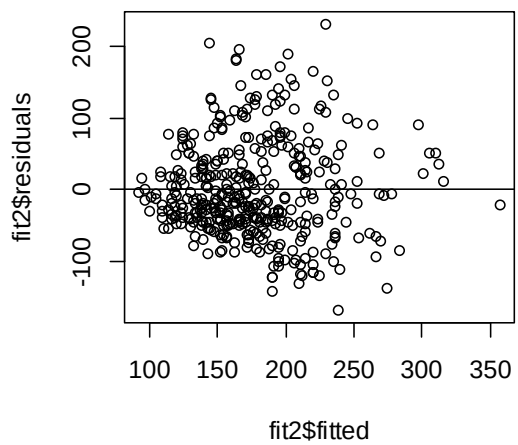
7) Plot separate scatterplots for triglycerides vs BMI for subjects in the two groups defined by presence of the ApoE4 allele. Do these plots suggest effect modification?

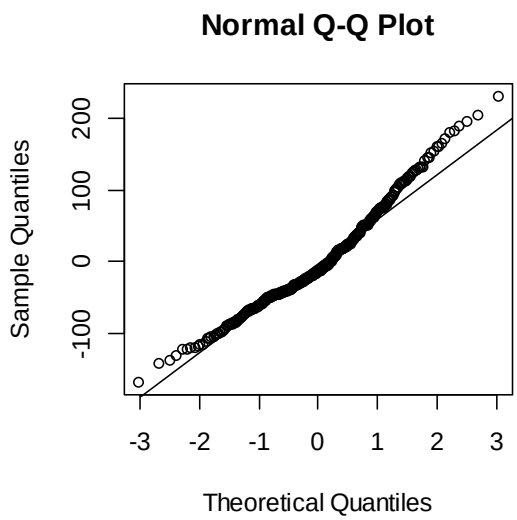
8) Fit a linear regression model that investigates whether the association between triglycerides and BMI is modified by the ApoE4 allele. Is there an association between ApoE4 and triglycerides? Is there evidence of effects modification?

R Commands & Output:

```
> # load the gee() package for robust standard errors
> library(gee)
>
> # identify outliers in scatterplot of triglycerides vs BMI
> plot(BMI,TG)
> bmi37 = which(BMI<=37)
>
> # excluding subjects with BMI > 37
> fit2 = lm(TG[bmi37] ~ BMI[bmi37])
> summary(fit2)

Call:
lm(formula = TG[bmi37] ~ BMI[bmi37])

Residuals:
    Min      1Q  Median      3Q     Max
```

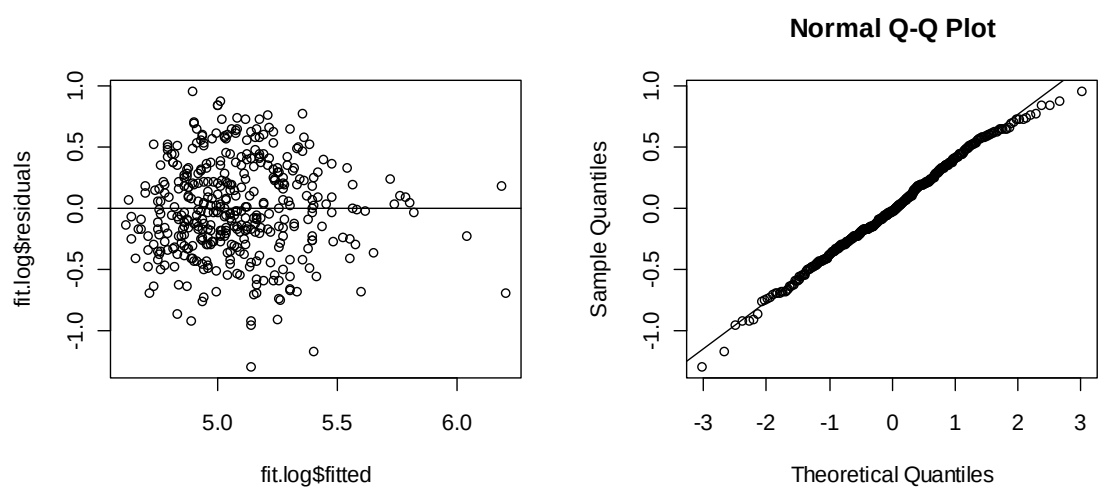
```
-169.07  -44.87  -13.22   39.45  232.05

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -202.707     30.084  -6.738 5.68e-11 ***
BMI[bmi37]      15.199      1.199  12.677  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.01 on 396 degrees of freedom
Multiple R-Squared: 0.2887,     Adjusted R-squared: 0.2869
F-statistic: 160.7 on 1 and 396 DF,  p-value: < 2.2e-16

>
> # analyze residuals from the regression analysis of triglycerides and BMI
> plot(fit2$fitted, fit2$residuals)
> abline(0,0)
> qqnorm(fit2$residuals)
> qqline(fit2$residuals)
```





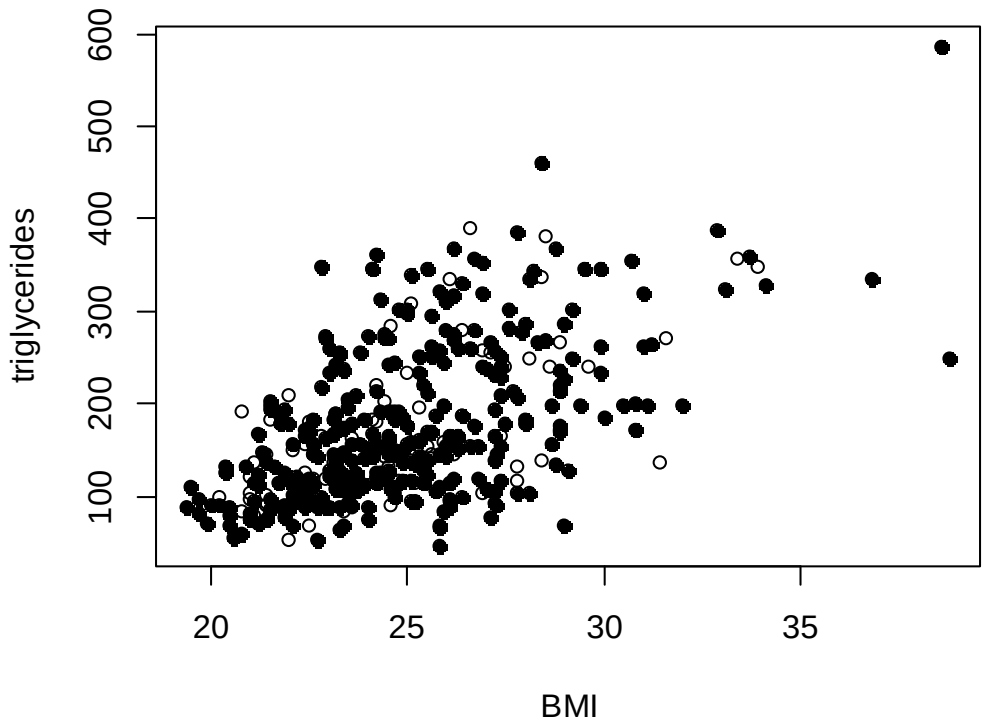

```
> # fit a linear regression model with robust standard errors
> fit.gee = gee(TG ~ BMI, id = seq(1,length(TG)))
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
[1] -208.50096   15.43748
> summary(fit.gee)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent

Call:
gee(formula = TG ~ BMI, id = seq(1, length(TG)))

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-170.18608  -45.09554  -12.88618   39.60133  231.07641


Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -208.50096  28.946250 -7.203039   32.021396 -6.511301
BMI           15.43748   1.149603 13.428538    1.322308 11.674646
```

```
Estimated Scale Parameter:  4750.958
Number of Iterations:  1

Working Correlation
     [,1]
[1,]    1
# calculate p-values for robust regression
> z = abs(fit.gee$coef/sqrt(diag(fit.gee$robust)))
> 2*(1-pnorm(z))
 (Intercept)          BMI
7.450263e-11 0.000000e+00
>
> # fit a regression model for log transformed triglycerides and BMI
> fit.log = lm(log(TG) ~ BMI)
> summary(fit.log)

Call:
lm(formula = log(TG) ~ BMI)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29019 -0.25303 -0.01692  0.26530  0.95800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.023584   0.162175   18.64   <2e-16 ***
BMI         0.082045   0.006441   12.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3862 on 398 degrees of freedom
Multiple R-Squared: 0.2896,     Adjusted R-squared: 0.2878
F-statistic: 162.3 on 1 and 398 DF,  p-value: < 2.2e-16

>
> # analyze residuals from the regression analysis of log transformed
> # triglycerides and BMI
> par(mfrow = c(1,2))
> plot(fit.log$fitted, fit.log$residuals)
> abline(0,0)
> qqnorm(fit.log$residuals)
> qqline(fit.log$residuals)
```




```
> #
binary variable indicating presence of ApoE4
> apoe4 = ifelse(apoE %in% c(3,5,6), 1, 0)
>
> # scatterplot with subjects stratified by ApoE4
> par(mfrow = c(1,1))
```

```
> plot(BMI[apoe4 == 0], TG[apoe4 == 0], pch = 19, xlab = "BMI", ylab = "triglycerides")
> points(BMI[apoe4 == 1], TG[apoe4 == 1], pch = 1)
>
```




```
> # multiple linear regression of triglycerides on BMI, ApoE4, and interaction
> fit3 = lm(TG ~ BMI + apoe4 + BMI*apoe4)
> summary(fit3)

Call:
lm(formula = TG ~ BMI + apoe4 + BMI * apoe4)

Residuals:

    Min      1Q  Median      3Q     Max
-170.04  -45.72  -13.03   38.88  231.12

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -204.0193    32.4558  -6.286  8.6e-10 ***
BMI             15.2780     1.2857  11.883  < 2e-16 ***
apoe4           -20.9439    72.6801  -0.288    0.773
BMI:apoe4        0.7464     2.9088   0.257    0.798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.09 on 396 degrees of freedom
Multiple R-Squared: 0.3121,   Adjusted R-squared: 0.3068
F-statistic: 59.88 on 3 and 396 DF,  p-value: < 2.2e-16
```

ANOVA Lab 1

The goal of this lab is to answer the following scientific questions using the cholesterol dataset:

- Is rs4775401 associated with cholesterol levels?
- Is ApoE associated with cholesterol levels?

2. Set your working directory as appropriate and read in the cholesterol data set.

3. Load packages "multcomp" and "gee"

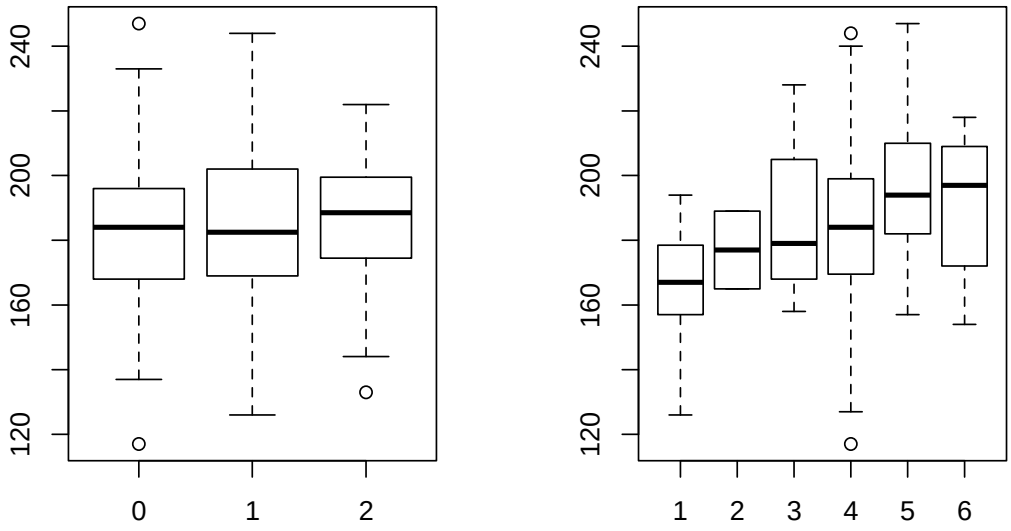
4. Perform a descriptive analysis to investigate the scientific questions of interest using numeric and graphical methods.

5. Compare the mean cholesterol levels between genotype groups defined by rs4775401.
   a. Perform the one-way ANOVA using the regression approach.
   b. Compare the above results with those obtained when
      i. allowing for unequal variances
      ii. using robust standard errors
      iii. using a nonparametric test
   c. Is there evidence that mean cholesterol levels between genotype groups are different? If so, perform all pairwise multiple comparisons using Bonferroni's adjustment. Try out different adjustment methods too.
   d. Interpret your results

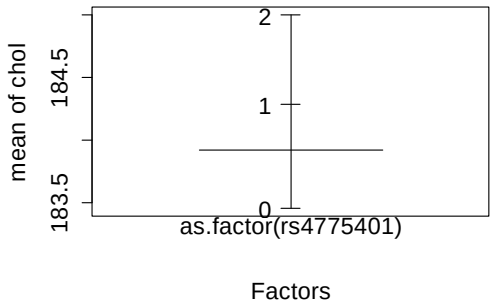
6. Repeat the steps described in problem 4 to compare the mean cholesterol levels between genotype groups defined by ApoE.

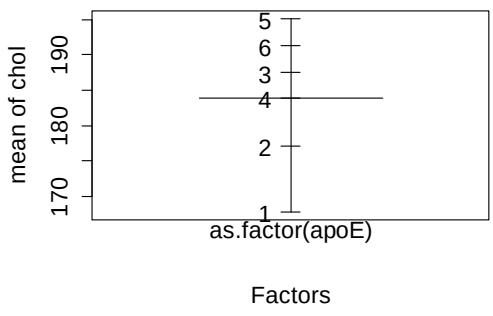
R Commands & Output:

```
> library(multcomp)
> library(gee)
>
> ## read data set ----------------------------------------------------------------
> cholesterol = read.table("http://faculty.washington.edu/rhubb/sisg/SISG-Data-cholesterol.txt",
header=T)
> attach(cholesterol)
>
> ## Exploratory data analysis ----------------------------------------------------
> ## graphical display: boxplot
> par(mfrow = c(1,2))
> boxplot(chol ~ as.factor(rs4775401))
> boxplot(chol ~ as.factor(apoE))
```




```
> ## alternative graphical display: graph of means
> par(mfrow = c(2,1))
> plot.design(chol ~ as.factor(rs4775401))
> plot.design(chol ~ as.factor(apoE))
```






```
> ## numeric descriptives
> tapply(chol, as.factor(rs4775401), mean)
       0        1        2
183.4505 184.2882 185.0000
> tapply(chol, as.factor(rs4775401), sd)
       0        1        2
20.70619 23.85693 21.70851
>
> tapply(chol, as.factor(apoE), mean)
       1        2        3        4        5        6
167.7843 177.0000 187.6000 183.9551 195.2000 191.3000
> tapply(chol, as.factor(apoE), sd)
       1        2        3        4        5        6
15.70008 16.97056 28.58846 22.08829 18.65493 23.56575
>
> ## Inferential data analysis ----------------------------------------------------
> fit1 = lm(chol ~ as.factor(rs4775401))
> summary(fit1)

Call:
lm(formula = chol ~ as.factor(rs4775401))

Residuals:
     Min       1Q   Median       3Q      Max
-66.4505 -15.4505  -0.2882  15.5495  63.5495

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           183.4505     1.5597 117.618   <2e-16 ***
as.factor(rs4775401)1   0.8377     2.3072   0.363    0.717
as.factor(rs4775401)2   1.5495     4.4702   0.347    0.729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.17 on 397 degrees of freedom
Multiple R-squared: 0.0005135,  Adjusted R-squared: -0.004522
F-statistic: 0.102 on 2 and 397 DF,  p-value: 0.903
```

```
> anova(fit1)
Analysis of Variance Table

Response: chol
                      Df Sum Sq Mean Sq F value Pr(>F)
as.factor(rs4775401)   2    100      50   0.102  0.903
Residuals            397 195089     491
>
> fit2 = lm(chol ~ as.factor(apoE))
> summary(fit2)

Call:
lm(formula = chol ~ as.factor(apoE))

Residuals:
   Min     1Q Median     3Q    Max
-66.95 -13.96  -0.37  15.04  60.05

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       167.784      2.934  57.194  < 2e-16 ***
as.factor(apoE)2    9.216     15.102   0.610  0.54205
as.factor(apoE)3   19.816      9.818   2.018  0.04423 *
as.factor(apoE)4   16.171      3.202   5.051 6.74e-07 ***
as.factor(apoE)5   27.416      3.919   6.996 1.14e-11 ***
as.factor(apoE)6   23.516      7.246   3.246  0.00127 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.95 on 394 degrees of freedom
Multiple R-squared: 0.114,    Adjusted R-squared: 0.1028
F-statistic: 10.14 on 5 and 394 DF,  p-value: 3.755e-09

> anova(fit2)
Analysis of Variance Table

Response: chol
                 Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(apoE)   5  22257  4451.5  10.142 3.755e-09 ***
Residuals       394 172932   438.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
> ## all pairwise comparisons with different methods for adjustment
> M2 = contrMat(table(apoE), type="Tukey")
> fit3 = lm(chol ~ -1 + as.factor(apoE))
> mc2 = glht(fit3, linfct =M2)
> summary(mc2, test=adjusted("none"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610 0.542055
3 - 1 == 0   19.816      9.818   2.018 0.044232 *
4 - 1 == 0   16.171      3.202   5.051 6.74e-07 ***
5 - 1 == 0   27.416      3.919   6.996 1.14e-11 ***
6 - 1 == 0   23.516      7.246   3.246 0.001272 **
3 - 2 == 0   10.600     17.528   0.605 0.545701
4 - 2 == 0    6.955     14.869   0.468 0.640228
5 - 2 == 0   18.200     15.040   1.210 0.226971
6 - 2 == 0   14.300     16.228   0.881 0.378751
4 - 3 == 0   -3.645      9.457  -0.385 0.700119
5 - 3 == 0    7.600      9.723   0.782 0.434885
```

```
6 - 3 == 0    3.700     11.475   0.322 0.747289
5 - 4 == 0   11.245      2.898   3.881 0.000122 ***
6 - 4 == 0    7.345      6.748   1.088 0.277055
6 - 5 == 0   -3.900      7.116  -0.548 0.583984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- none method)

> summary(mc2, test=adjusted("bonferroni"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  1.00000
3 - 1 == 0   19.816      9.818   2.018  0.66348
4 - 1 == 0   16.171      3.202   5.051 1.01e-05 ***
5 - 1 == 0   27.416      3.919   6.996 1.71e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.01909 *
3 - 2 == 0   10.600     17.528   0.605  1.00000
4 - 2 == 0    6.955     14.869   0.468  1.00000
5 - 2 == 0   18.200     15.040   1.210  1.00000
6 - 2 == 0   14.300     16.228   0.881  1.00000
4 - 3 == 0   -3.645      9.457  -0.385  1.00000
5 - 3 == 0    7.600      9.723   0.782  1.00000
6 - 3 == 0    3.700     11.475   0.322  1.00000
5 - 4 == 0   11.245      2.898   3.881  0.00183 **
6 - 4 == 0    7.345      6.748   1.088  1.00000
6 - 5 == 0   -3.900      7.116  -0.548  1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- bonferroni method)

> summary(mc2, test=adjusted("holm"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  1.00000
3 - 1 == 0   19.816      9.818   2.018  0.48655
4 - 1 == 0   16.171      3.202   5.051 9.43e-06 ***
5 - 1 == 0   27.416      3.919   6.996 1.71e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.01527 *
3 - 2 == 0   10.600     17.528   0.605  1.00000
4 - 2 == 0    6.955     14.869   0.468  1.00000
5 - 2 == 0   18.200     15.040   1.210  1.00000
6 - 2 == 0   14.300     16.228   0.881  1.00000
4 - 3 == 0   -3.645      9.457  -0.385  1.00000
5 - 3 == 0    7.600      9.723   0.782  1.00000
6 - 3 == 0    3.700     11.475   0.322  1.00000
5 - 4 == 0   11.245      2.898   3.881  0.00159 **
6 - 4 == 0    7.345      6.748   1.088  1.00000
6 - 5 == 0   -3.900      7.116  -0.548  1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- holm method)

> summary(mc2, test=adjusted("hochberg"))

         Simultaneous Tests for General Linear Hypotheses
```

```
Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  0.74729
3 - 1 == 0   19.816      9.818   2.018  0.48655
4 - 1 == 0   16.171      3.202   5.051 9.43e-06 ***
5 - 1 == 0   27.416      3.919   6.996 1.71e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.01527 *
3 - 2 == 0   10.600     17.528   0.605  0.74729
4 - 2 == 0    6.955     14.869   0.468  0.74729
5 - 2 == 0   18.200     15.040   1.210  0.74729
6 - 2 == 0   14.300     16.228   0.881  0.74729
4 - 3 == 0   -3.645      9.457  -0.385  0.74729
5 - 3 == 0    7.600      9.723   0.782  0.74729
6 - 3 == 0    3.700     11.475   0.322  0.74729
5 - 4 == 0   11.245      2.898   3.881  0.00159 **
6 - 4 == 0    7.345      6.748   1.088  0.74729
6 - 5 == 0   -3.900      7.116  -0.548  0.74729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- hochberg method)

> summary(mc2, test=adjusted("hommel"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  0.74729
3 - 1 == 0   19.816      9.818   2.018  0.48655
4 - 1 == 0   16.171      3.202   5.051 9.43e-06 ***
5 - 1 == 0   27.416      3.919   6.996 1.71e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.01527 *
3 - 2 == 0   10.600     17.528   0.605  0.74729
4 - 2 == 0    6.955     14.869   0.468  0.74729
5 - 2 == 0   18.200     15.040   1.210  0.74729
6 - 2 == 0   14.300     16.228   0.881  0.74729
4 - 3 == 0   -3.645      9.457  -0.385  0.74729
5 - 3 == 0    7.600      9.723   0.782  0.74729
6 - 3 == 0    3.700     11.475   0.322  0.74729
5 - 4 == 0   11.245      2.898   3.881  0.00159 **
6 - 4 == 0    7.345      6.748   1.088  0.74729
6 - 5 == 0   -3.900      7.116  -0.548  0.74729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- hommel method)

> summary(mc2, test=adjusted("BH"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  0.72998
3 - 1 == 0   19.816      9.818   2.018  0.13270
4 - 1 == 0   16.171      3.202   5.051 5.05e-06 ***
```

```
5 - 1 == 0   27.416      3.919   6.996 1.71e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.00477 **
3 - 2 == 0   10.600     17.528   0.605  0.72998
4 - 2 == 0    6.955     14.869   0.468  0.73872
5 - 2 == 0   18.200     15.040   1.210  0.56743
6 - 2 == 0   14.300     16.228   0.881  0.71016
4 - 3 == 0   -3.645      9.457  -0.385  0.74729
5 - 3 == 0    7.600      9.723   0.782  0.72481
6 - 3 == 0    3.700     11.475   0.322  0.74729
5 - 4 == 0   11.245      2.898   3.881  0.00061 ***
6 - 4 == 0    7.345      6.748   1.088  0.59369
6 - 5 == 0   -3.900      7.116  -0.548  0.72998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- BH method)

> summary(mc2, test=adjusted("BY"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  1.00000
3 - 1 == 0   19.816      9.818   2.018  0.44032
4 - 1 == 0   16.171      3.202   5.051 1.68e-05 ***
5 - 1 == 0   27.416      3.919   6.996 5.68e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.01583 *
3 - 2 == 0   10.600     17.528   0.605  1.00000
4 - 2 == 0    6.955     14.869   0.468  1.00000
5 - 2 == 0   18.200     15.040   1.210  1.00000
6 - 2 == 0   14.300     16.228   0.881  1.00000
4 - 3 == 0   -3.645      9.457  -0.385  1.00000
5 - 3 == 0    7.600      9.723   0.782  1.00000
6 - 3 == 0    3.700     11.475   0.322  1.00000
5 - 4 == 0   11.245      2.898   3.881  0.00203 **
6 - 4 == 0    7.345      6.748   1.088  1.00000
6 - 5 == 0   -3.900      7.116  -0.548  1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- BY method)

> summary(mc2, test=adjusted("fdr"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = chol ~ -1 + as.factor(apoE))

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    9.216     15.102   0.610  0.72998
3 - 1 == 0   19.816      9.818   2.018  0.13270
4 - 1 == 0   16.171      3.202   5.051 5.05e-06 ***
5 - 1 == 0   27.416      3.919   6.996 1.71e-10 ***
6 - 1 == 0   23.516      7.246   3.246  0.00477 **
3 - 2 == 0   10.600     17.528   0.605  0.72998
4 - 2 == 0    6.955     14.869   0.468  0.73872
5 - 2 == 0   18.200     15.040   1.210  0.56743
6 - 2 == 0   14.300     16.228   0.881  0.71016
4 - 3 == 0   -3.645      9.457  -0.385  0.74729
5 - 3 == 0    7.600      9.723   0.782  0.72481
6 - 3 == 0    3.700     11.475   0.322  0.74729
5 - 4 == 0   11.245      2.898   3.881  0.00061 ***
6 - 4 == 0    7.345      6.748   1.088  0.59369
```

```
6 - 5 == 0   -3.900      7.116  -0.548  0.72998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- fdr method)

>
> ## One-way (not assuming equal variances)
> oneway.test(chol ~ as.factor(rs4775401))

        One-way analysis of means (not assuming equal variances)

data:  chol and as.factor(rs4775401)
F = 0.1046, num df = 2.000, denom df = 75.608, p-value = 0.9008

>
> oneway.test(chol ~ as.factor(apoE))

        One-way analysis of means (not assuming equal variances)

data:  chol and as.factor(apoE)
F = 11.8601, num df = 5.00, denom df = 8.56, p-value = 0.001177

>
> ## Using robust standard errors
> summary(gee(chol ~ as.factor(rs4775401), id=seq(1,length(chol))))
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
          (Intercept) as.factor(rs4775401)1 as.factor(rs4775401)2
          183.4504950             0.8377402             1.5495050

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent

Call:
gee(formula = chol ~ as.factor(rs4775401), id = seq(1, length(chol)))

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-66.4504950 -15.4504950  -0.2882353  15.5495050  63.5495050


Coefficients:
                        Estimate Naive S.E.     Naive z Robust S.E.
(Intercept)           183.4504950   1.559715 117.6179395    1.453272
as.factor(rs4775401)1   0.8377402   2.307238   0.3630923    2.332437
as.factor(rs4775401)2   1.5495050   4.470234   0.3466273    4.282708
                        Robust z
(Intercept)           126.2327489
as.factor(rs4775401)1   0.3591694
as.factor(rs4775401)2   0.3618049

Estimated Scale Parameter:  491.4078
Number of Iterations:  1

Working Correlation
     [,1]
[1,]    1
>
> summary(gee(chol ~ as.factor(apoE), id=seq(1,length(chol))))
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
     (Intercept) as.factor(apoE)2 as.factor(apoE)3 as.factor(apoE)4 as.factor(apoE)5
      167.784314         9.215686        19.815686        16.170742        27.415686
as.factor(apoE)6
       23.515686
```

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent

Call:
gee(formula = chol ~ as.factor(apoE), id = seq(1, length(chol)))

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-66.955056 -13.955056  -0.369685  15.044944  60.044944


Coefficients:
                    Estimate Naive S.E.    Naive z Robust S.E.  Robust z
(Intercept)      167.784314   2.933622 57.1935612    2.176791 77.078744
as.factor(apoE)2   9.215686  15.101746  0.6102398    8.760047  1.052013
as.factor(apoE)3  19.815686   9.817778  2.0183473   11.640722  1.702273
as.factor(apoE)4  16.170742   3.201564  5.0508889    2.561033  6.314149
as.factor(apoE)5  27.415686   3.919011  6.9955617    3.163857  8.665273
as.factor(apoE)6  23.515686   7.245513  3.2455516    7.397258  3.178974

Estimated Scale Parameter:  438.9132
Number of Iterations:  1

Working Correlation
     [,1]
[1,]    1
> ## non-parametric ANOVA
> kruskal.test(chol ~ as.factor(rs4775401))

        Kruskal-Wallis rank sum test

data:  chol by as.factor(rs4775401)
Kruskal-Wallis chi-squared = 0.5761, df = 2, p-value = 0.7497


> kruskal.test(chol ~ as.factor(apoE))

        Kruskal-Wallis rank sum test

data:  chol by as.factor(apoE)
Kruskal-Wallis chi-squared = 48.246, df = 5, p-value = 3.164e-09
```

ANOVA Lab 2

The goal of this lab is to answer the following scientific questions using the cholesterol dataset.

- Are rs174548 and apoE associated with cholesterol levels?
- Does the effect of apoE on cholesterol levels depend on rs174548?

1. Obtain a cross-tabulation of the groups defined by rs174548 and apoE.

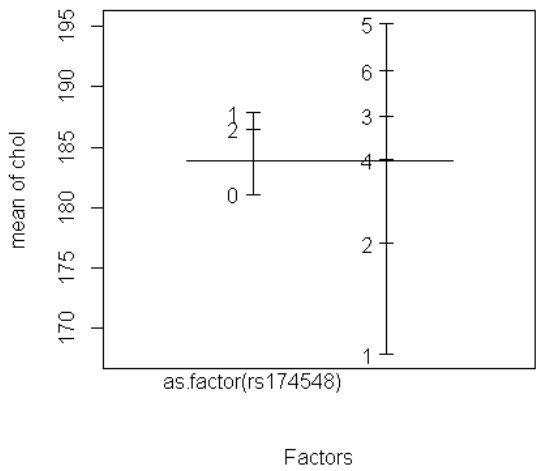
2. Perform a descriptive analysis to investigate the scientific questions of interest using numeric and graphical methods.

3. Fit a two-way ANOVA model with an interaction between rs174548 and apoE. Test the interaction. What do you conclude?

4. Fit a two-way ANOVA model without the interaction between rs174548 and apoE. Test the main effects of rs174548 and apoE. What do you conclude?

R Commands & Output:

```
> ## Two-way ANOVA ------------------------------------------------------------
> ## exploratory data analysis
> table(rs174548, apoE)
        apoE
rs174548   1   2   3   4   5   6
       0  33   2   2 144  40   6
       1  17   0   3  99  24   4
       2   1   0   0  24   1   0
>
> tapply(chol, list(as.factor(rs174548), as.factor(apoE)), mean)
         1   2        3        4        5        6
0 168.0909 177 192.0000 180.4653 193.6250 180.6667
1 167.7059  NA 184.6667 187.9192 199.0833 207.2500
2 159.0000  NA       NA 188.5417 165.0000       NA
> tapply(chol, list(as.factor(rs174548), as.factor(apoE)), sd)
          1        2        3        4        5        6
0 17.39318 16.97056 18.38478 21.00646 18.07773 23.04488
1 12.65783       NA 37.85939 24.03810 18.82856 14.68276
2       NA       NA       NA 16.46598       NA       NA
>
> plot.design(chol ~ as.factor(rs174548) + as.factor(apoE))
```




```
> ## model with interaction
> fit1 = lm(chol ~ as.factor(rs174548)*as.factor(apoE))
> summary(fit1)

Call:
lm(formula = chol ~ as.factor(rs174548) * as.factor(apoE))

Residuals:
    Min      1Q  Median      3Q     Max 
-63.465 -13.021  -0.042  13.671  56.081 

Coefficients: (4 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          168.091      3.609  46.577  < 2e-16 ***
as.factor(rs174548)1                  -0.385      6.189  -0.062  0.95043    
as.factor(rs174548)2                  -9.091     21.043  -0.432  0.66598    
as.factor(apoE)2                       8.909     15.097   0.590  0.55546    
as.factor(apoE)3                      23.909     15.097   1.584  0.11409    
as.factor(apoE)4                      12.374      4.001   3.093  0.00213 ** 
as.factor(apoE)5                      25.534      4.875   5.237 2.68e-07 ***
as.factor(apoE)6                      12.576      9.201   1.367  0.17249    
as.factor(rs174548)1:as.factor(apoE)2      NA         NA      NA       NA    
as.factor(rs174548)2:as.factor(apoE)2      NA         NA      NA       NA    
as.factor(rs174548)1:as.factor(apoE)3  -6.948     19.912  -0.349  0.72731    
as.factor(rs174548)2:as.factor(apoE)3      NA         NA      NA       NA    
as.factor(rs174548)1:as.factor(apoE)4   7.839      6.755   1.160  0.24659    
as.factor(rs174548)2:as.factor(apoE)4  17.167     21.534   0.797  0.42582    
as.factor(rs174548)1:as.factor(apoE)5   5.843      8.183   0.714  0.47560    
as.factor(rs174548)2:as.factor(apoE)5 -19.534     29.722  -0.657  0.51142    
as.factor(rs174548)1:as.factor(apoE)6  26.968     14.744   1.829  0.06816 .  
as.factor(rs174548)2:as.factor(apoE)6      NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.73 on 386 degrees of freedom
Multiple R-squared:  0.15,    Adjusted R-squared: 0.1214 
F-statistic: 5.241 on 13 and 386 DF,  p-value: 1.169e-08 

> 
> ## model without interaction
> fit2 = lm(chol ~ as.factor(rs174548) + as.factor(apoE))
> summary(fit2)

Call:
lm(formula = chol ~ as.factor(rs174548) + as.factor(apoE))
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-64.074 -13.074  -0.328  14.390  56.507 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           165.535      3.005  55.082  < 2e-16 ***
as.factor(rs174548)1    6.419      2.208   2.907  0.00385 ** 
as.factor(rs174548)2    5.575      4.348   1.282  0.20060    
as.factor(apoE)2       11.465     14.990   0.765  0.44483    
as.factor(apoE)3       18.213      9.749   1.868  0.06249 .  
as.factor(apoE)4       15.539      3.191   4.869 1.63e-06 ***
as.factor(apoE)5       27.209      3.886   7.002 1.10e-11 ***
as.factor(apoE)6       23.197      7.184   3.229  0.00135 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.77 on 392 degrees of freedom
Multiple R-squared: 0.1338,	Adjusted R-squared: 0.1183 
F-statistic:  8.65 on 7 and 392 DF,  p-value: 6.989e-10

## compare models with and without interaction
Analysis of Variance Table

Model 1: chol ~ as.factor(rs174548) + as.factor(apoE)
Model 2: chol ~ as.factor(rs174548) * as.factor(apoE)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    392 169074                           
2    386 165903  6    3170.5 1.2294 0.2901
```

ANOVA Lab 3

The goal of this lab is to answer the following scientific questions using the cholesterol dataset.

- Controlling for age, is apoE associated with cholesterol levels?
- Does age modify the association between apoE and cholesterol levels?

1. Perform a descriptive analysis to investigate the scientific questions of interest using numeric and graphical methods.

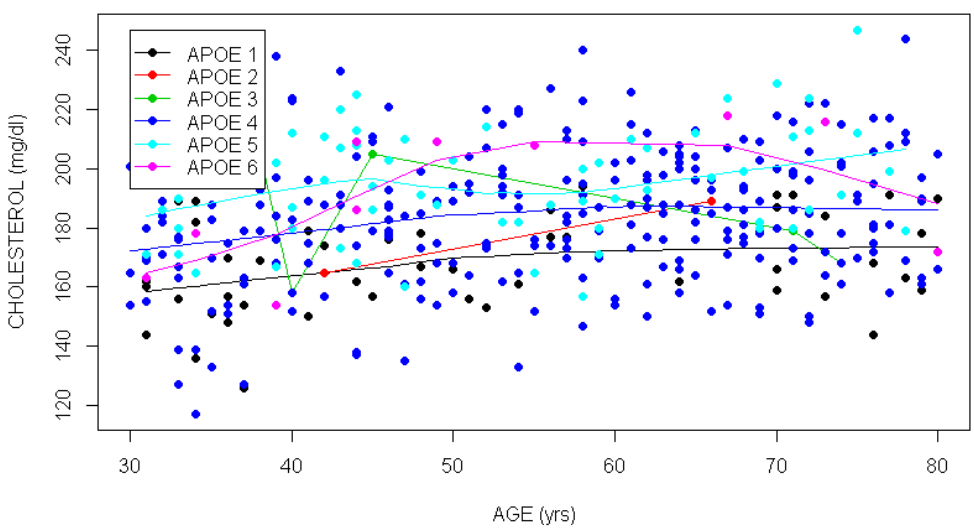
2. Fit an ANCOVA model with an interaction between apoE and age. Test the interaction. What do you conclude?

3. Fit an ANCOVA model without an interaction between apoE and age. Compare the results with the one-way ANOVA model that compares mean cholesterol levels among genotypes defined by apoE. What can you say about the role of age? [Is it an effect modifier? Or is it a confounder? Or is it a precision variable?]

R Commands & Output:

```
> by(cbind(chol,age), apoE, cor, method="pearson")
INDICES: 1
          chol       age
chol 1.0000000 0.3120186
age  0.3120186 1.0000000
------------------------------------------------------------------
INDICES: 2
     chol age
chol    1   1
age     1   1
-----------------------------------------------------------------------------------------
INDICES: 3
           chol        age
chol  1.0000000 -0.4778431
age  -0.4778431  1.0000000
-----------------------------------------------------------------------------------------
INDICES: 4
          chol       age
chol 1.0000000 0.2032922
age  0.2032922 1.0000000
-----------------------------------------------------------------------------------------
INDICES: 5
          chol       age
chol 1.0000000 0.2265928
age  0.2265928 1.0000000
-----------------------------------------------------------------------------------------
INDICES: 6
          chol       age
chol 1.0000000 0.4487348
age  0.4487348 1.0000000

> by(cbind(chol,age), apoE, cor, method="spearman")
INDICES: 1
```

```
          chol       age
chol 1.0000000 0.3139938
age  0.3139938 1.0000000
-----------------------------------------------------------------
INDICES: 2
     chol age
chol    1   1
age     1   1
-----------------------------------------------------------------------------------------
INDICES: 3
     chol  age
chol  1.0 -0.4
age  -0.4  1.0
-----------------------------------------------------------------------------------------
INDICES: 4
          chol       age
chol 1.0000000 0.1728862
age  0.1728862 1.0000000
-----------------------------------------------------------------------------------------
INDICES: 5
          chol       age
chol 1.0000000 0.1832910
age  0.1832910 1.0000000
-----------------------------------------------------------------------------------------
INDICES: 6
          chol       age
chol 1.0000000 0.5457317
age  0.5457317 1.0000000

>
> plot(age, chol, xlab="AGE (yrs)", ylab="CHOLESTEROL (mg/dl)", type="n")
> for (i in 1:6){
+  lines(lowess(age[apoE==i], chol[apoE==i]), col=i)
+  points(age[apoE==i], chol[apoE==i], col=i, pch=16)
+ }
> legend(min(age), max(chol), legend=paste("APOE", seq(1,6)), col=seq(1,6), pch=16,
lty=1)
>
```




```
> ## ANCOVA Model with an interaction
> fit1 = lm(chol ~ as.factor(apoE) * age)
> summary(fit1)

Call:
lm(formula = chol ~ as.factor(apoE) * age)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-59.979 -13.752   0.249  13.407  59.430 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          151.941019   9.809703  15.489   <2e-16 ***
as.factor(apoE)2     -28.941019  67.563772  -0.428   0.6686    
as.factor(apoE)3      76.872259  33.724424   2.279   0.0232 *  
as.factor(apoE)4      14.219079  11.074130   1.284   0.1999    
as.factor(apoE)5      26.898046  14.182657   1.897   0.0586 .  
as.factor(apoE)6       6.803879  24.107708   0.282   0.7779    
age                    0.302625   0.179168   1.689   0.0920 .  
as.factor(apoE)2:age   0.697375   1.221654   0.571   0.5684    
as.factor(apoE)3:age  -1.074409   0.606386  -1.772   0.0772 .  
as.factor(apoE)4:age   0.015568   0.200101   0.078   0.9380    
as.factor(apoE)5:age   0.006881   0.259484   0.027   0.9789    
as.factor(apoE)6:age   0.328288   0.445469   0.737   0.4616    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.51 on 388 degrees of freedom
Multiple R-squared: 0.164,	Adjusted R-squared: 0.1403 
F-statistic: 6.918 on 11 and 388 DF,  p-value: 1.081e-10

> ## ANCOVA Model without an interaction
> fit2 = lm(chol ~ as.factor(apoE) + age)
> summary(fit2)

Call:
lm(formula = chol ~ as.factor(apoE) + age)

Residuals:
    Min      1Q  Median      3Q     Max 
-60.162 -14.070   0.099  13.674  59.289 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      151.56390    4.71788  32.125  < 2e-16 ***
as.factor(apoE)2   8.70538   14.77291   0.589 0.556012    
as.factor(apoE)3  19.49128    9.60399   2.029 0.043081 *  
as.factor(apoE)4  15.06399    3.14216   4.794 2.32e-06 ***
as.factor(apoE)5  27.25811    3.83373   7.110 5.51e-12 ***
as.factor(apoE)6  23.74897    7.08773   3.351 0.000884 ***
age                0.30983    0.07153   4.331 1.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.49 on 393 degrees of freedom
Multiple R-squared: 0.1544,	Adjusted R-squared: 0.1415 
F-statistic: 11.96 on 6 and 393 DF,  p-value: 2.409e-12

## compare models with and without interaction
> anova(fit2, fit1)
Analysis of Variance Table

Model 1: chol ~ as.factor(apoE) + age
Model 2: chol ~ as.factor(apoE) * age
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    393 165052                           
2    388 163184  5    1868.3 0.8885 0.4887

>
> ## ONE-WAY ANOVA model
> fit3 = lm(chol ~ as.factor(apoE))
> summary(fit3)

Call:
lm(formula = chol ~ as.factor(apoE))
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-66.95 -13.96  -0.37  15.04  60.05 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        167.784      2.934  57.194  < 2e-16 ***
as.factor(apoE)2     9.216     15.102   0.610  0.54205    
as.factor(apoE)3    19.816      9.818   2.018  0.04423 *  
as.factor(apoE)4    16.171      3.202   5.051 6.74e-07 ***
as.factor(apoE)5    27.416      3.919   6.996 1.14e-11 ***
as.factor(apoE)6    23.516      7.246   3.246  0.00127 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.95 on 394 degrees of freedom
Multiple R-squared: 0.114,	Adjusted R-squared: 0.1028 
F-statistic: 10.14 on 5 and 394 DF,  p-value: 3.755e-09

> anova(fit3, fit2)
Analysis of Variance Table

Model 1: chol ~ as.factor(apoE)
Model 2: chol ~ as.factor(apoE) + age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    394 172932                                  
2    393 165052  1    7879.5 18.762 1.883e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
> ## mean cholesterol for different genotypes
> predict(fit3, new=data.frame(apoE=1))
       1 
167.7843 

> predict(fit3, new=data.frame(apoE=2))
  1 
177 
> predict(fit3, new=data.frame(apoE=3))
    1 
187.6 
> predict(fit3, new=data.frame(apoE=4))
       1 
183.9551 
> predict(fit3, new=data.frame(apoE=5))
    1 
195.2 
> predict(fit3, new=data.frame(apoE=6))
    1 
191.3 
> 
> ## mean cholesterol for different genotypes adjusted by age
> predict(fit2, new=data.frame(age=mean(age),apoE=1))
       1 
168.5495 
> predict(fit2, new=data.frame(age=mean(age),apoE=2))
       1 
177.2548 
> predict(fit2, new=data.frame(age=mean(age),apoE=3))
       1 
188.0407 
> predict(fit2, new=data.frame(age=mean(age),apoE=4))
       1 
183.6134 
> predict(fit2, new=data.frame(age=mean(age),apoE=5))
       1 
195.8076 
> predict(fit2, new=data.frame(age=mean(age),apoE=6))
       1 
192.2984 
```