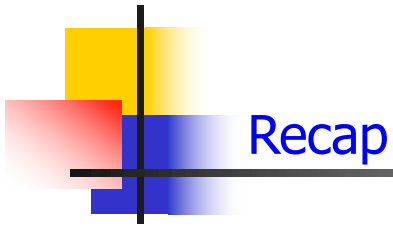


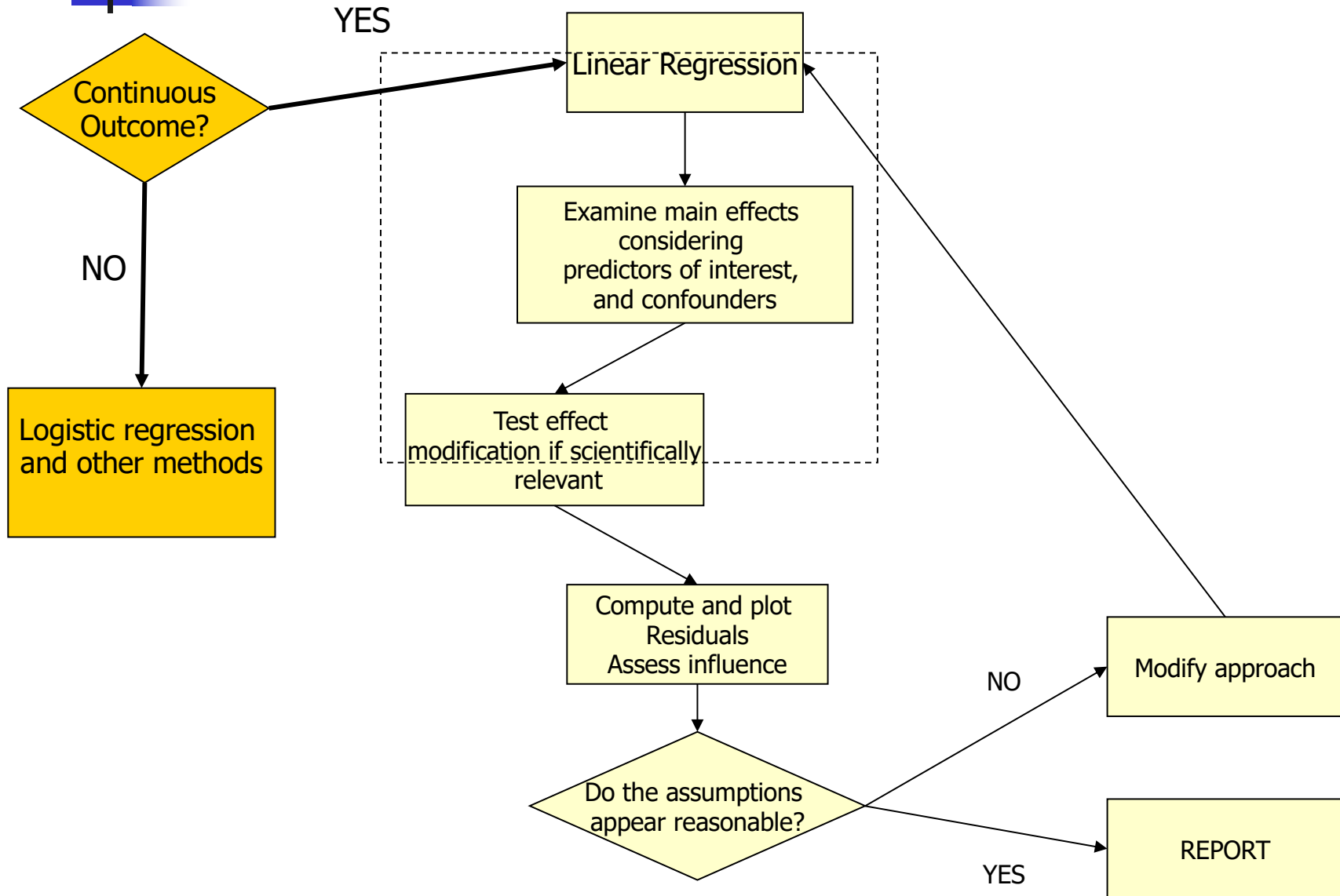


REGRESSION MODELS

ANOVA



Recap



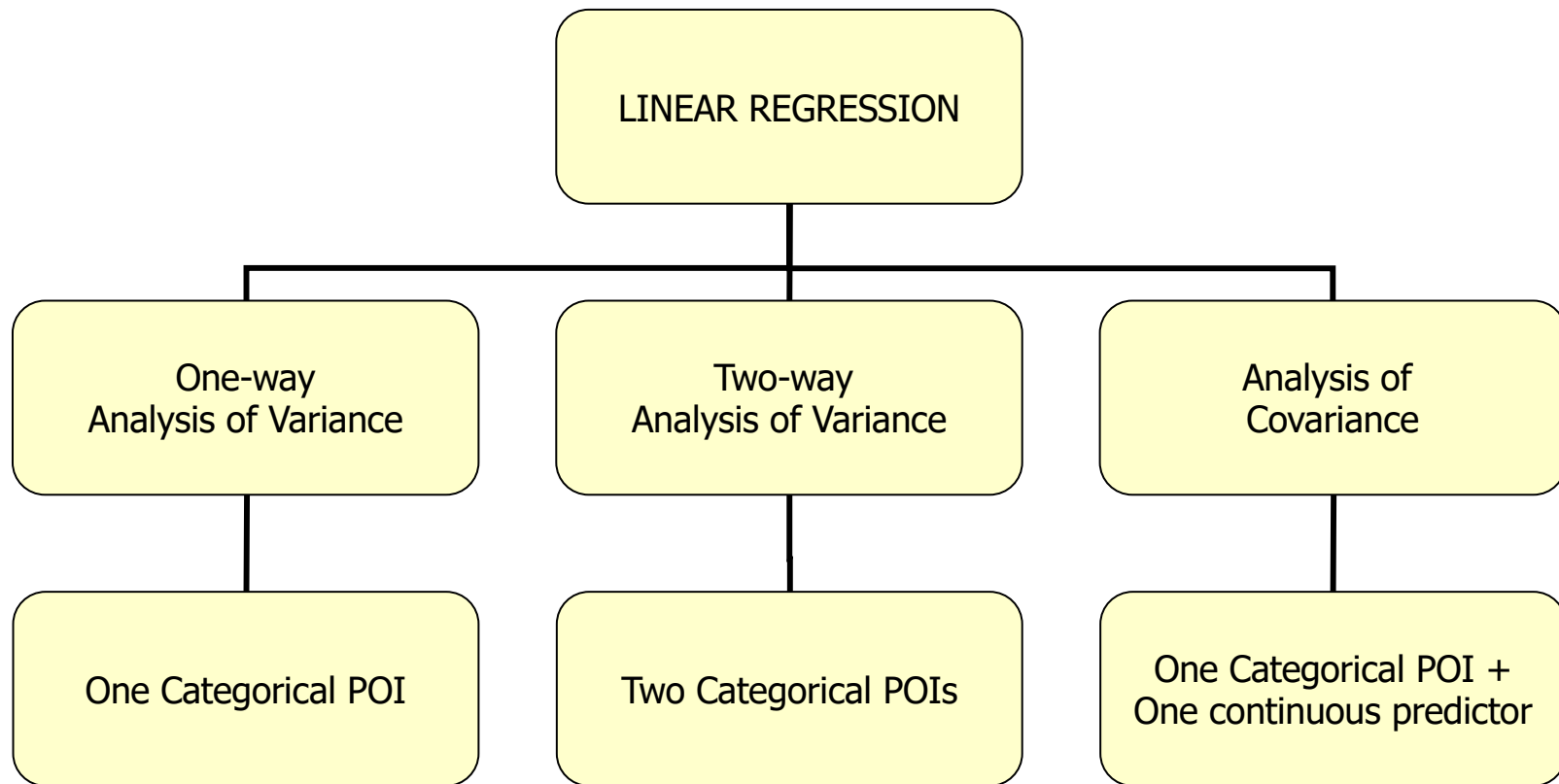


COMING UP NEXT: ANOVA – a special case of linear regression

- What if the independent variables of interest are categorical?
- In this case, comparing the mean of the continuous outcome in the different categories may be of interest
- This is what is called **AN**alysis **Of** **VA**riance
- We will show that it is just a special case of linear regression



ANOVA – a special case of linear regression



Uses dummy variables to represent categorical variables!



Outline

- Motivation: We will consider some examples of ANOVA and show that they are special cases of linear regression
- ANOVA as a regression model
 - Dummy variables
- One-way ANOVA models
 - Contrasts
 - Multiple comparisons
- Two-way ANOVA models
 - Interactions
- ANCOVA models



ANOVA/ANCOVA: 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 diabetes on cholesterol levels
 - Does the effect of rs174548 on cholesterol differ between people with and without diabetes?
 - 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

- Scientific hypotheses of interest:
 - Assess the effect of rs174548 on cholesterol levels.
 - Assess the effect of rs174548 and diabetes on cholesterol levels
 - Does the effect of rs174548 on cholesterol differ between people with and without diabetes?
 - Assess the effect of rs174548 and age on cholesterol levels
 - Does the effect of rs174548 on cholesterol differ depending on subject's age?



Motivation: Example

Here are some descriptive summaries:

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

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



Motivation: Example

Another way of getting the same results:

```
> by(chol, factor(rs174548), mean)
```

```
  factor(rs174548): 0
```

```
[1] 181.0617
```

```
-----  
  factor(rs174548): 1
```

```
[1] 187.8639
```

```
-----  
  factor(rs174548): 2
```

```
[1] 186.5
```

```
> by(chol, factor(rs174548), sd)
```

```
  factor(rs174548): 0
```

```
[1] 21.13998
```

```
-----  
  factor(rs174548): 1
```

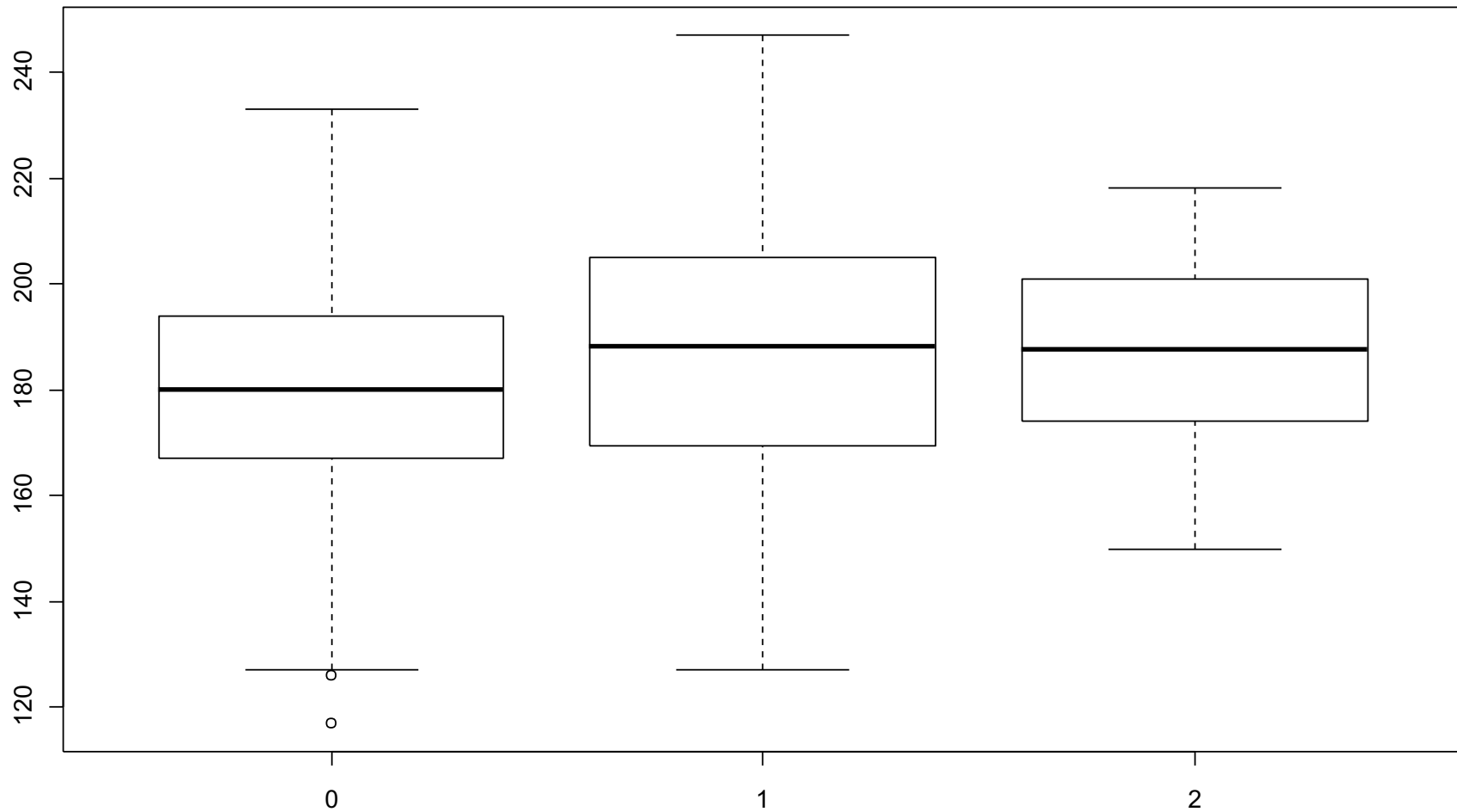
```
[1] 23.74541
```

```
-----  
  factor(rs174548): 2
```

```
[1] 17.38333
```

Motivation: Example

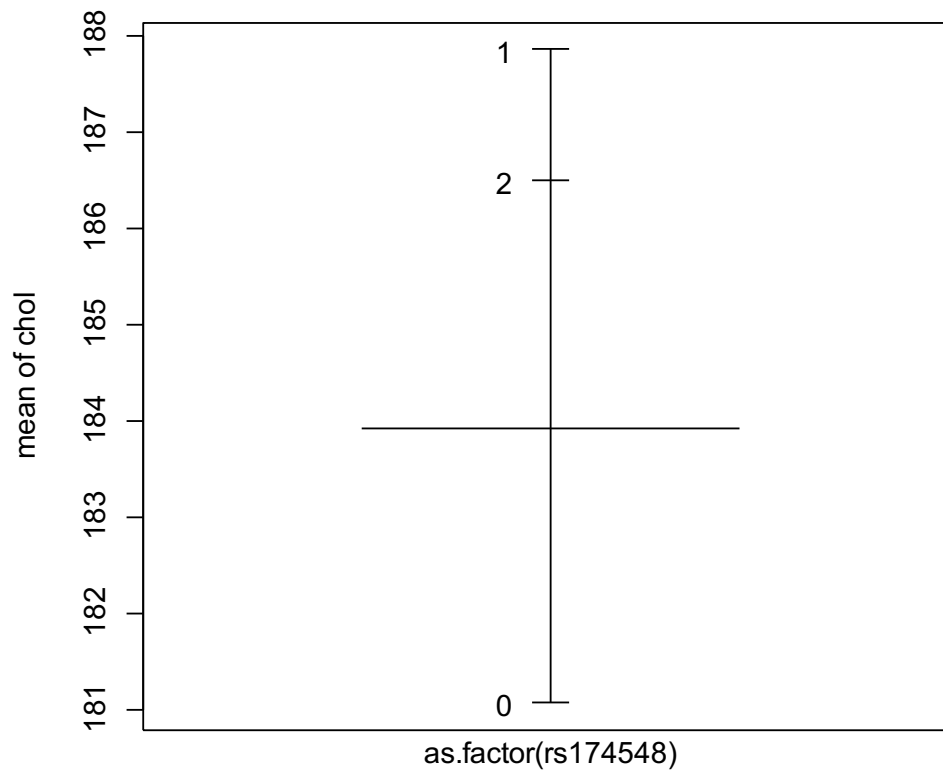
Is rs174548 associated with cholesterol?



R command: `boxplot(chol ~ factor(rs174548))` 11

Motivation: Example

Another graphical display:



Factors

R command:

```
plot.design(chol ~ factor(rs174548))
```



Motivation: Example

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

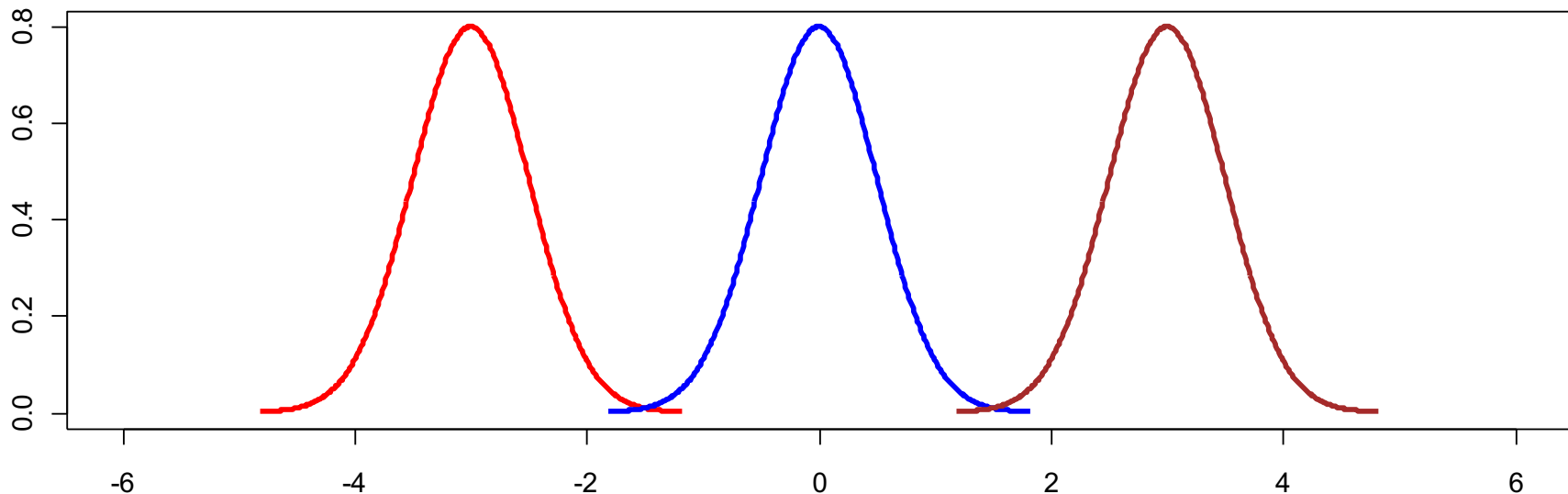


REGRESSION MODELS

One-way ANOVA as a regression model

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations

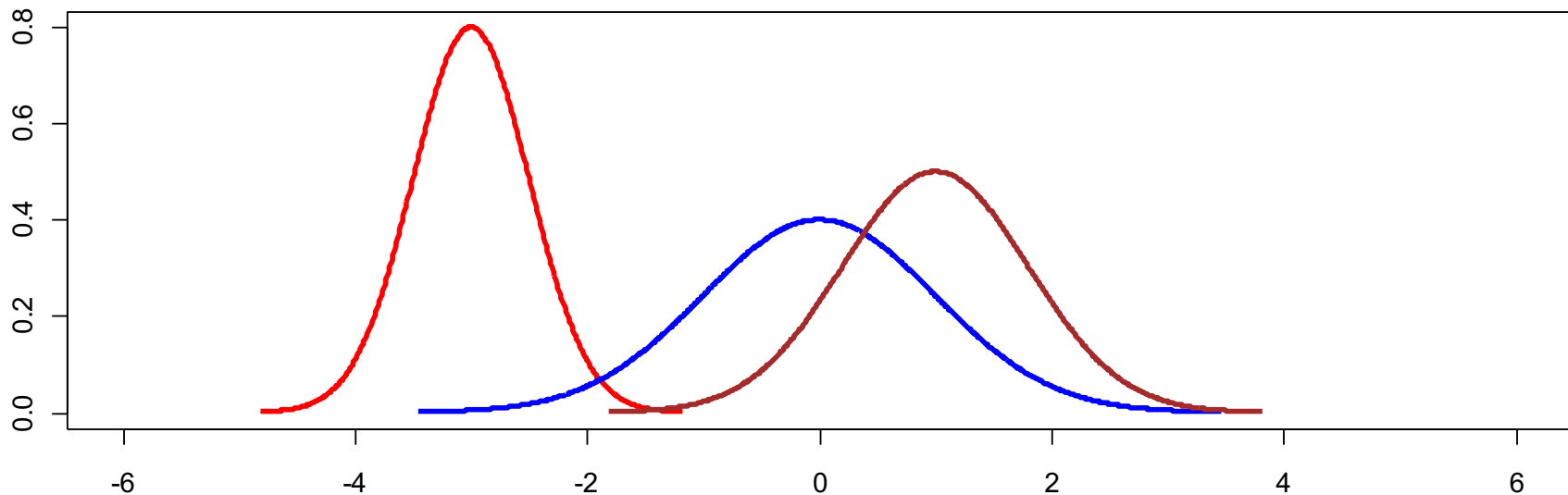


Assumptions for Classical ANOVA Framework:

Independence
Normality
Equal variances

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations





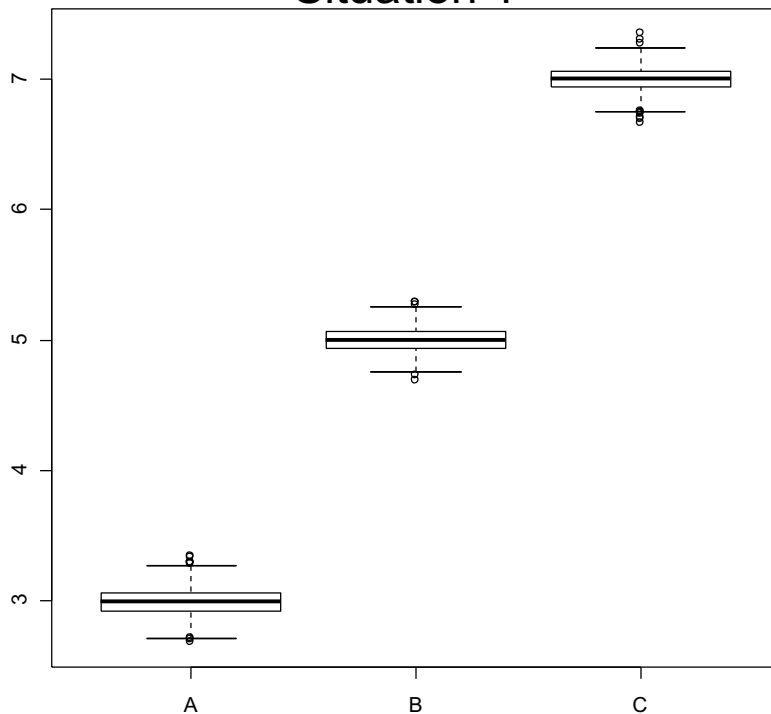
ANalysis Of VAriance Models (ANOVA)

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

ANalysis Of VAriance Models (ANOVA)

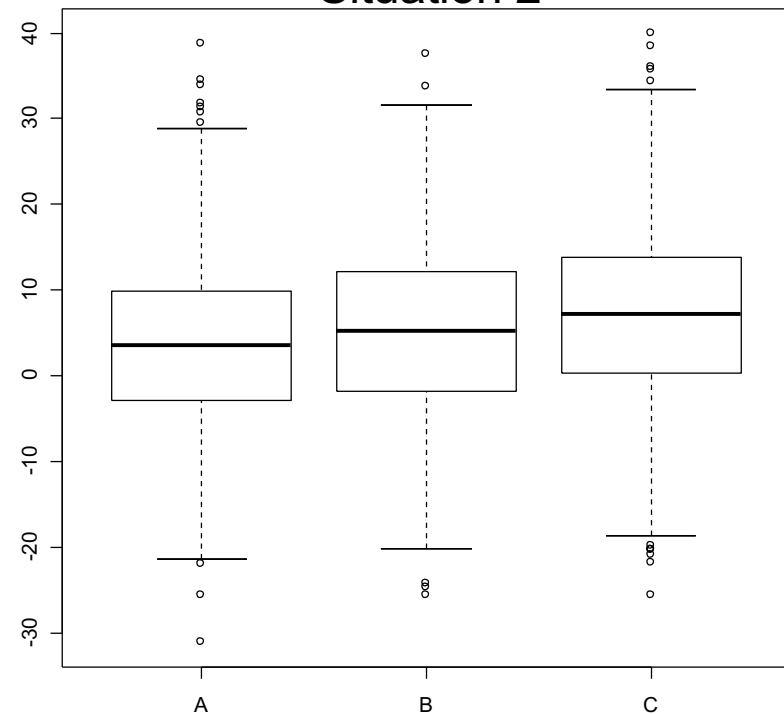
In both data sets, the true population means are: 3 (A), 5 (B), 7(C)

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 population means!
 - The ratio MSR/MSE is the F-statistic.
- We can make these comparisons with multiple linear regression: the different groups are represented with “dummy” variables



ANOVA as a multiple regression model

- **Dummy Variables:**

- Suppose you have a categorical variable C with k categories 0,1, 2, ..., k-1. To represent that variable, we can construct k-1 dummy variables of the form

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

$$x_2 = \begin{cases} 1, & \text{if subject is in category 2} \\ 0, & \text{otherwise} \end{cases}$$

...

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

The omitted category (here category 0) 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	Mean cholesterol	X_1	X_2
C/C	μ_0	0	0
C/G	μ_1	1	0
G/G	μ_2	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

Mean	Regression Model
μ_0	β_0
μ_1	$\beta_0 + \beta_1$
μ_2	$\beta_0 + \beta_2$



ANOVA as a multiple regression model

- Regression with Dummy Variables:

- Example:

- Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1x_1 + \beta_2x_2$

- Interpretation of model parameters?

- $\mu_0 = \beta_0$: mean cholesterol when rs174548 is C/C
 - $\mu_1 = \beta_0 + \beta_1$: mean cholesterol when rs174548 is C/G
 - $\mu_2 = \beta_0 + \beta_2$: mean cholesterol when rs174548 is G/G



ANOVA as a multiple regression model

- Regression with Dummy Variables:

- Example:

- Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1x_1 + \beta_2x_2$

- Interpretation of model parameters?

- $\mu_0 = \beta_0$: mean cholesterol when rs174548 is C/C
 - $\mu_1 = \beta_0 + \beta_1$: mean cholesterol when rs174548 is C/G
 - $\mu_2 = \beta_0 + \beta_2$: 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 ($\mu_1 - \mu_0$).
 - β_2 : difference in mean cholesterol levels between groups with rs174548 equal to G/G and C/C ($\mu_2 - \mu_0$).



ANOVA: One-Way Model

- Goal:

- Compare the means of K independent groups (defined by a categorical predictor)

- Statistical Hypotheses:

- (Global) Null Hypothesis:

$$H_0: \mu_0 = \mu_1 = \dots = \mu_{K-1} \text{ or, equivalently,}$$

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{K-1} = 0$$

- 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

$H_0: \mu_0 = \mu_1 = \dots = \mu_{K-1}$ vs. $H_1: \text{not all means are equal}$

$H_0: \beta_1 = \beta_2 = \dots = \beta_{K-1} = 0$

- 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 as long as you state you have a categorical variable (may need to use: 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 = lm(chol ~ factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ 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 ***
factor(rs174548)1     6.802     2.321  2.930  0.00358 **
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)
Analysis of Variance Table

Response: chol
              Df Sum Sq Mean Sq F value  Pr(>F)
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)
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
```

```
> 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
```

```
> fit1 = lm(chol ~ factor(rs174548))
> summary(fit1)
```

```
Call:
lm(formula = chol ~ 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 ***
factor(rs174548)1     6.802     2.321   2.930  0.00358 **
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)
Analysis of Variance Table
```

```
Response: chol
      Df Sum Sq Mean Sq F value    Pr(>F)
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 = lm(chol ~ factor(rs174548))
> summary(fit1)
```

```
Call:
lm(formula = chol ~ 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 ***
factor(rs174548)1     6.802     2.321   2.930  0.00358 **
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)
Analysis of Variance Table
```

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

```
> 1-pf(4.4865, 2, 397)
[1] 0.01183671
> 1-pf(((3624+690)/2)/481, 2, 397)
[1] 0.01186096
```

ANOVA: One-Way Model

```
> fit1 = lm(chol ~ factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ 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
factor(rs174548)1     6.802     2.321   2.930  0.00358
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)
Analysis of Variance Table

Response: chol

              Df Sum Sq Mean Sq F value Pr(>F)
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 = lm(chol ~ factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ 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
factor(rs174548)1     6.802     2.321   2.930  0.00358
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)
Analysis of Variance Table

Response: chol
              Df Sum Sq Mean Sq F value Pr(>F)
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 = lm(chol ~ factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ 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
factor(rs174548)1     6.802      2.321   2.930  0.00358
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)
Analysis of Variance Table

Response: chol
          Df Sum Sq Mean Sq F value Pr(>F)
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)

```
> fit2 = lm(chol ~ -1 + factor(rs174548))
> summary(fit2)
Call:
lm(formula = chol ~ -1 + 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|)
factor(rs174548)0    181.062      1.455  124.41 <2e-16 ***
factor(rs174548)1    187.864      1.809  103.88 <2e-16 ***
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(fit2)
Analysis of Variance Table
Response: chol

              Df    Sum Sq Mean Sq F value    Pr(>F)
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

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

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

```
> fit1.1 = lm(chol ~ rs174548)
> summary(fit1.1)

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(fit1.1)
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 39
```



ANOVA: One-Way Model

```
> fit1.1 = lm(chol ~ rs174548)
> summary(fit1.1)

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(fit1.1)
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 = lm(chol ~ factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ 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
factor(rs174548)1     6.802     2.321   2.930  0.00358
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)
Analysis of Variance Table

Response: chol
              Df Sum Sq Mean Sq F value Pr(>F)
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
(coming up)



One-Way ANOVA allowing for unequal variances

We can also perform one-way ANOVA allowing for unequal variances (Welch's ANOVA):

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

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

```
data: chol and factor(rs174548)
```

```
F = 4.3258, num df = 2.000, denom df = 73.284, p-value = 0.01676
```

- We reject the null hypothesis that the mean cholesterol levels are the same across groups defined by rs174548 ($p=0.01676$).
 - What are the groups with differences in means?

MULTIPLE COMPARISONS (coming up)



One-Way ANOVA with robust standard errors

We can also use robust standard errors to get correct variance estimates:

```
> fit1 = lm(chol ~ factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ 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
factor(rs174548)1     6.802     2.321   2.930  0.00358
factor(rs174548)2     5.438     4.540   1.198  0.23167

> lmtest::coefest(fit1, vcov = sandwich::sandwich)
t test of coefficients:

              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    181.0617     1.4000 129.3283 < 2.2e-16 ***
factor(rs174548)1     6.8023     2.4020   2.8319  0.004863 **
factor(rs174548)2     5.4383     3.6243   1.5005  0.134272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Kruskal-Wallis Test

- Non-parametric analogue to the one-way ANOVA
 - Based on ranks; does not require normality
- In our example:

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

```
      Kruskal-Wallis rank sum test
```

```
data:  chol by 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!



REGRESSION METHODS

MULTIPLE COMPARISONS



ANOVA: One-Way Model

- What are the groups with differences in means?

MULTIPLE COMPARISONS:

$$\left. \begin{array}{l} \mu_0 = \mu_1? \\ \mu_0 = \mu_2? \\ \mu_1 = \mu_2? \end{array} \right\} \text{Pairwise comparisons}$$

$$(\mu_1 + \mu_2)/2 = \mu_0? \longrightarrow \text{Non-pairwise comparison}$$



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 $C = K(K-1)/2$	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 C tests performed, use level α/C (or multiply p -values by C).
 - Simple
 - Conservative
 - Must decide on number of tests beforehand
 - Widely applicable
 - Can be done without software!



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).

Multiple Comparisons

This option considers all pairwise comparisons

```
> ## call library for multiple comparisons
> library(multcomp)
>
> ## fit model
> fit2 = lm(chol ~ -1 + 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(fit2, 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 + 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 + 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

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

```
Simultaneous Tests for General Linear Hypotheses
```

```
Multiple Comparisons of Means: Tukey Contrasts
```

```
Fit: lm(formula = chol ~ -1 + 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

- What about using other adjustment methods?
 - For example, we used:
 - > `summary(mc, test=adjusted("bonferroni"))`
(all pairwise comparisons, with Bonferroni adjustment)
 - > `summary(mc, test=adjusted("fdr"))`
(all pairwise comparisons, with FDR adjustment)
 - Other options 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"))`

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

Summary

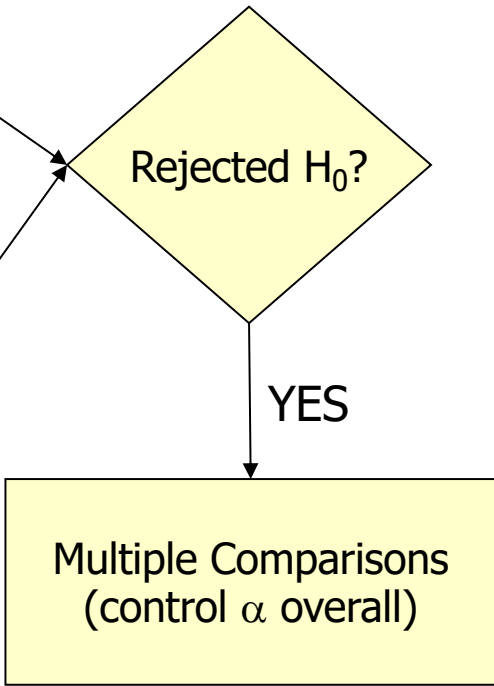
GOAL: Comparison of means across K groups

One-way ANOVA:
 $H_0: \mu_0 = \mu_1 = \dots = \mu_{K-1}$
 $H_1: \text{not all means are equal}$

Multiple Regression:
Model: $E[Y|\text{groups}] = \beta_0 + \beta_1 \text{group}_2 + \dots + \beta_{k-1} \text{group}_k$
where group_1 is the reference group
 $H_0: \beta_1 = \beta_2 = \dots = \beta_{k-1} = 0$
 $H_1: \text{not all } \beta_i \text{ are equal to zero}$

Relationships:

$$\begin{aligned} \mu_0 &= \beta_0 \\ \mu_1 &= \beta_0 + \beta_1 \\ \mu_2 &= \beta_0 + \beta_2 \\ &\dots \\ \mu_{K-1} &= \beta_0 + \beta_{K-1} \end{aligned}$$



e.g. Bonferroni: $\alpha/\#\text{comparisons}$



REGRESSION METHODS

Two-way ANOVA models



ANOVA: Two-Way Model

Motivation:

- Scientific question:
 - Assess the effect of rs174548 and diabetes 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



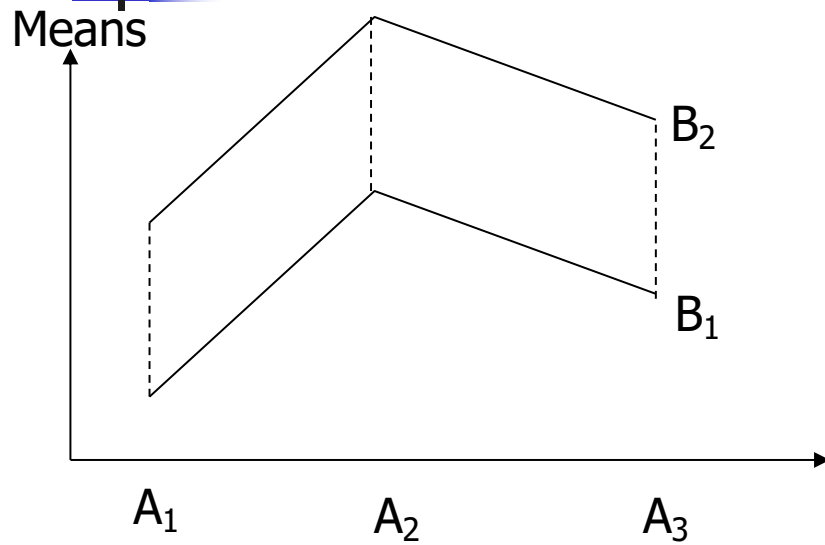
ANOVA: Two-Way Model

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

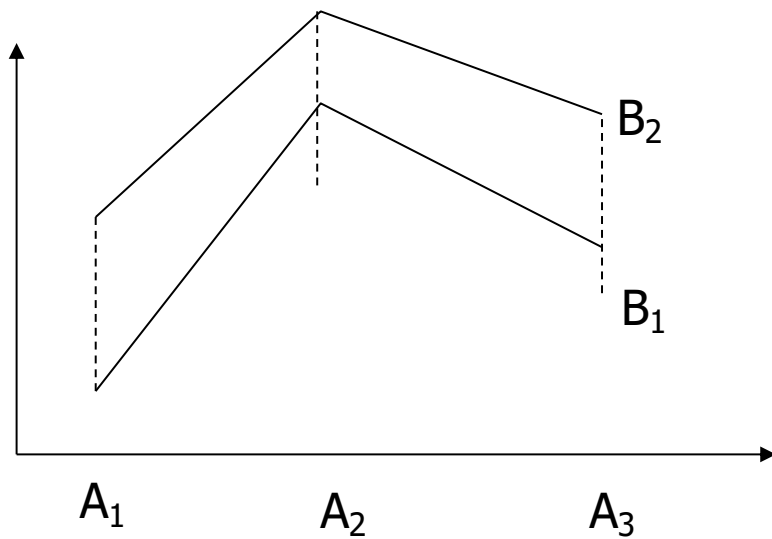
Factor A

	A_1	A_2	A_3
Factor B 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?

$$H_0: \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_1: \beta_1 \text{ or } \beta_2 \text{ not zero}$$

How do we test for main effect of factor B?

$$H_0: \beta_3 = 0 \quad \text{vs.} \quad H_1: \beta_3 \text{ not zero}$$



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?

$$\left\{ \begin{array}{l} H_0: \beta_4 = \beta_5 = 0 \quad \text{vs.} \\ H_1: \beta_4 \text{ or } \beta_5 \text{ not zero} \end{array} \right.$$

IMPORTANT:

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

ANOVA: Two-Way Model (without interaction)

```
> fit1 = lm(chol ~ factor(DM) + factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ factor(DM) + 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 ***
factor(DM)1     11.053      2.126    5.199 3.22e-07 ***
factor(rs174548)1  7.236      2.250    3.215 0.00141 **
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

> fit0 = lm(chol ~ factor(DM))
> anova(fit0, fit1)
Analysis of Variance Table

Model 1: chol ~ factor(DM)
Model 2: chol ~ factor(DM) + 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 ~ factor(DM) + factor(rs174548))
> summary(fit1)
Call:
lm(formula = chol ~ factor(DM) + factor(rs174548))

Residuals:
    Min       1Q   Median       3Q      Max
-66.653 -14.463  -0.601  15.445  57.635

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    175.365      1.786   98.208 < 2e-16 ***
factor(DM)1     11.053      2.126    5.199 3.22e-07 ***
factor(rs174548)1  7.236      2.250    3.215 0.00141 **
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 ~ factor(DM)
Model 2: chol ~ factor(DM) + 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 people without diabetes in C/C group: 175.365 mg/dl
 - Estimated difference in mean cholesterol levels between people with and without diabetes adjusted by genotype: 11.053 mg/dl
 - Estimated difference in mean cholesterol levels between C/G and C/C groups adjusted by diabetes status: 7.236 mg/dl
 - Estimated difference in mean cholesterol levels between G/G and C/C groups adjusted by diabetes status: 5.184 mg/dl
- There is evidence that cholesterol is associated with diabetes ($p < 0.001$).
- There is evidence that cholesterol is associated with genotype ($p = 0.005$)



ANOVA: Two-Way Model (without interaction)

- In words:
 - Adjusting for diabetes status, 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 diabetes status
 - (this is because the model does not have an interaction between diabetes and genotype!)

ANOVA: Two-Way Model (with interaction)

```
> fit2 = lm(chol ~ factor(DM) * factor(rs174548))
> summary(fit2)
```

Call:

```
lm(formula = chol ~ factor(DM) * factor(rs174548))
```

Residuals:

Min	1Q	Median	3Q	Max
-70.529	-13.604	-0.974	14.171	54.882

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	178.1182	2.0089	88.666	< 2e-16	***
factor(DM) 1	5.7109	2.7982	2.041	0.04192	*
factor(rs174548) 1	0.9597	3.1306	0.307	0.75933	
factor(rs174548) 2	-0.2015	6.4053	-0.031	0.97492	
factor(DM) 1:factor(rs174548) 1	12.7398	4.4650	2.853	0.00456	**
factor(DM) 1: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: 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: Model comparison

```
> anova(fit1,fit2)
```

```
Analysis of Variance Table
```

```
Model 1: chol ~ factor(DM) + factor(rs174548)
```

```
Model 2: chol ~ factor(DM) * 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)

■ Interpretation of results:

```
> fit2 = lm(chol ~ factor(DM) * factor(rs174548))
> summary(fit2)
```

```
Call:
lm(formula = chol ~ factor(DM) * factor(rs174548))
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-70.529 -13.604  -0.974  14.171  54.882
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      178.1182    2.0089  88.666 < 2e-16 ***
factor(DM) 1         5.7109    2.7982   2.041  0.04192 *
factor(rs174548) 1    0.9597    3.1306   0.307  0.75933
factor(rs174548) 2   -0.2015    6.4053  -0.031  0.97492
factor(DM) 1:factor(rs174548) 1 12.7398    4.4650   2.853  0.00456 **
factor(DM) 1: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 ~ factor(DM) + factor(rs174548)
Model 2: chol ~ factor(DM) * 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
```

- Estimated mean cholesterol for people without diabetes in C/C group:

178.12 mg/dl

- Estimated mean cholesterol for people with diabetes in C/C group:

(178.12 + 5.7109) mg/dl

- Estimated mean cholesterol for people without diabetes in C/G group:

(178.12 + 0.9597) mg/dl

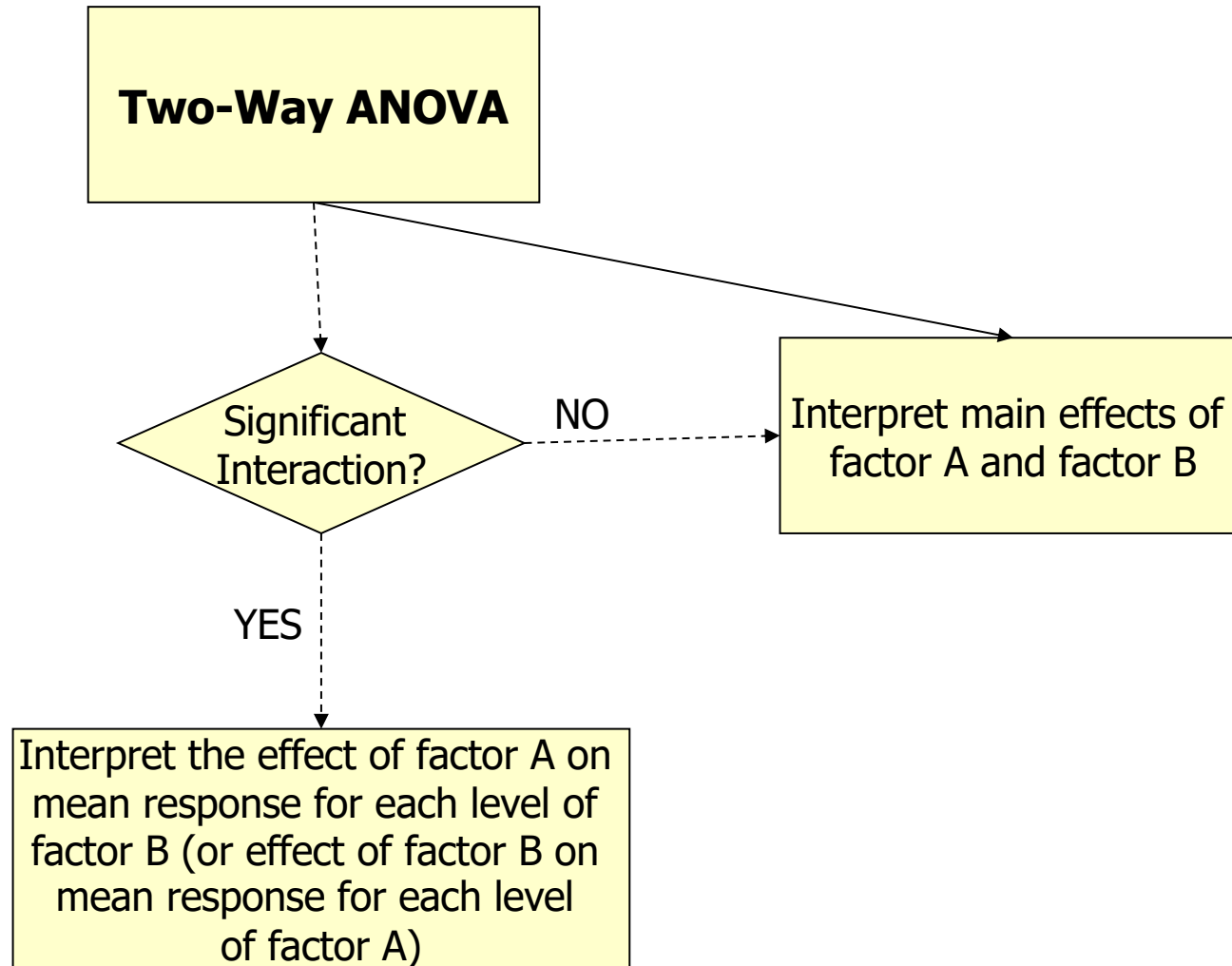
- Estimated mean cholesterol for people with diabetes in C/G group:

(178.12 + 5.7109 + 0.9597 + 12.7398) mg/dl

- ...

- There is evidence for an interaction between diabetes and genotype (p= 0.015)

SUMMARY:





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: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$

Interaction term

Note that:

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

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

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 ~ factor(rs174548))
> summary(fit0)
Call:
lm(formula = chol ~ 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 ***
factor(rs174548)1    6.802      2.321   2.930  0.00358 **
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)
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 ~ factor(rs174548) + age)
> summary(fit1)
Call:
lm(formula = chol ~ factor(rs174548) + age)

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

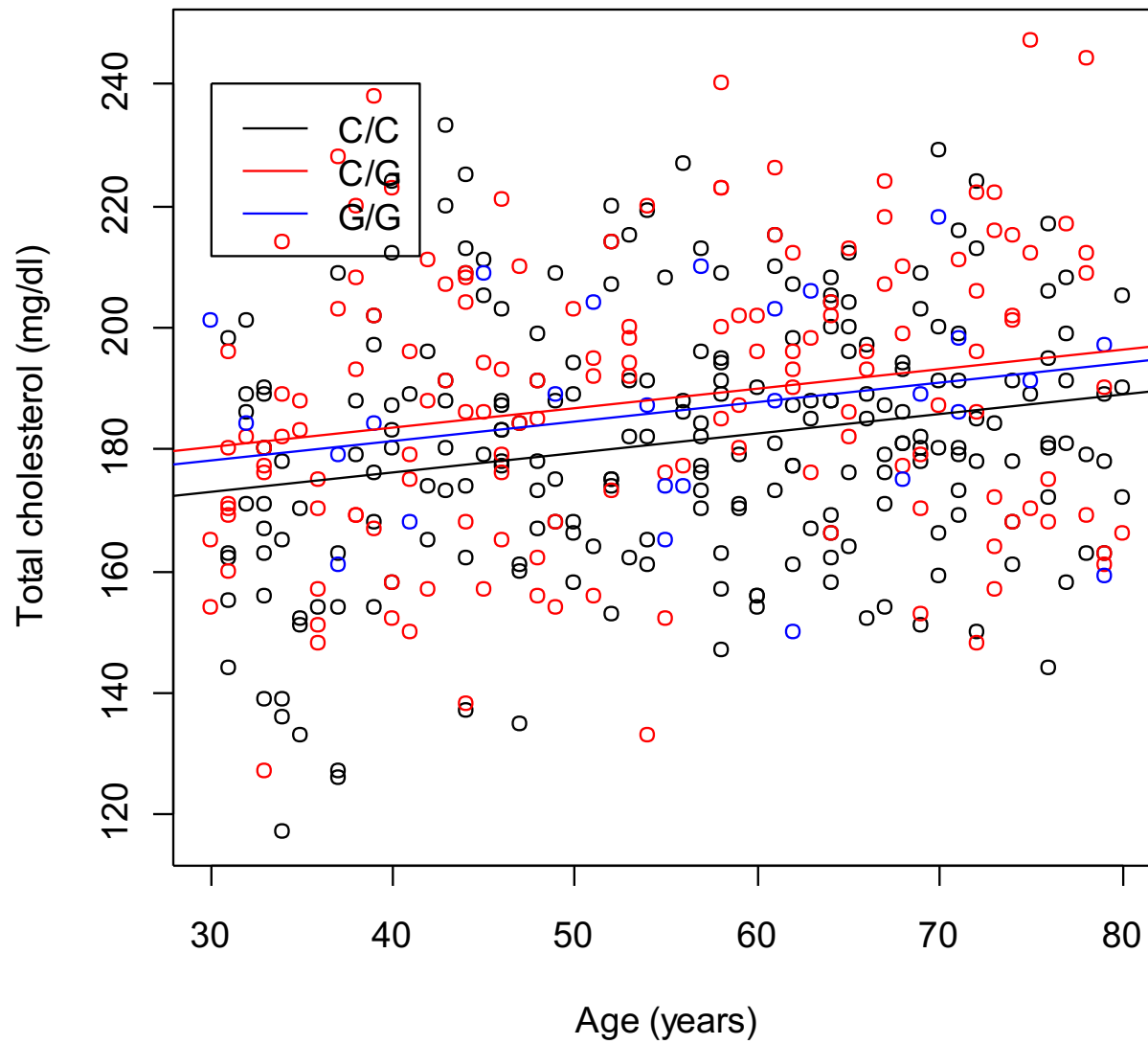
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      163.28125    4.36422   37.414 < 2e-16 ***
factor(rs174548)1    7.30137    2.27457    3.210  0.00144 **
factor(rs174548)2    5.08431    4.44331    1.144  0.25321
age                0.32140    0.07457    4.310  2.06e-05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

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 ~ factor(rs174548)
Model 2: chol ~ 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 ~ factor(rs174548) * age)
> summary(fit2)
Call:
lm(formula = chol ~ 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 ***
factor(rs174548)1     3.42799     8.79946   0.390  0.69707
factor(rs174548)2    16.53004    18.28067   0.904  0.36642
age                0.30576     0.10154   3.011  0.00277 **
factor(rs174548)1:age  0.07159     0.15617   0.458  0.64692
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
```

ANCOVA

```
> fit0 = lm(chol ~ age)
> summary(fit0)

Call:
lm(formula = chol ~ age)

Residuals:
    Min       1Q   Median       3Q      Max
-60.453 -14.643  -0.022   14.659   58.995

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
```

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

Model 1: chol ~ age
Model 2: chol ~ factor(rs174548) * age
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     398 187187
2     394 181961    4   5226.6 2.8293 0.02455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test of
coincident
lines



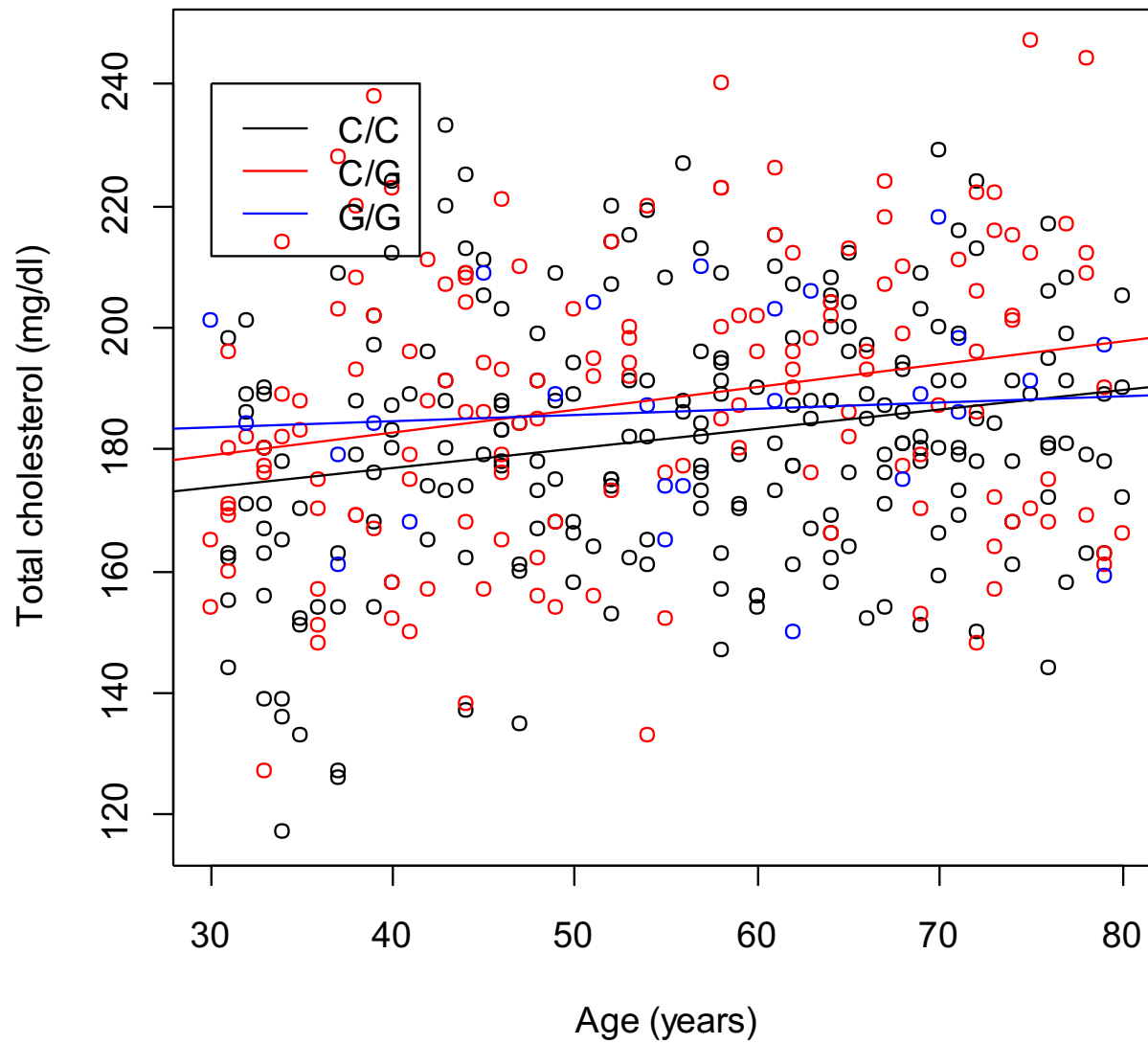
ANCOVA

Test of
parallel lines

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

Model 1: chol ~ factor(rs174548) + age
Model 2: chol ~ 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

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

- If the slopes are the same,

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



ANCOVA

- If the slopes are the same,

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

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

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

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

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



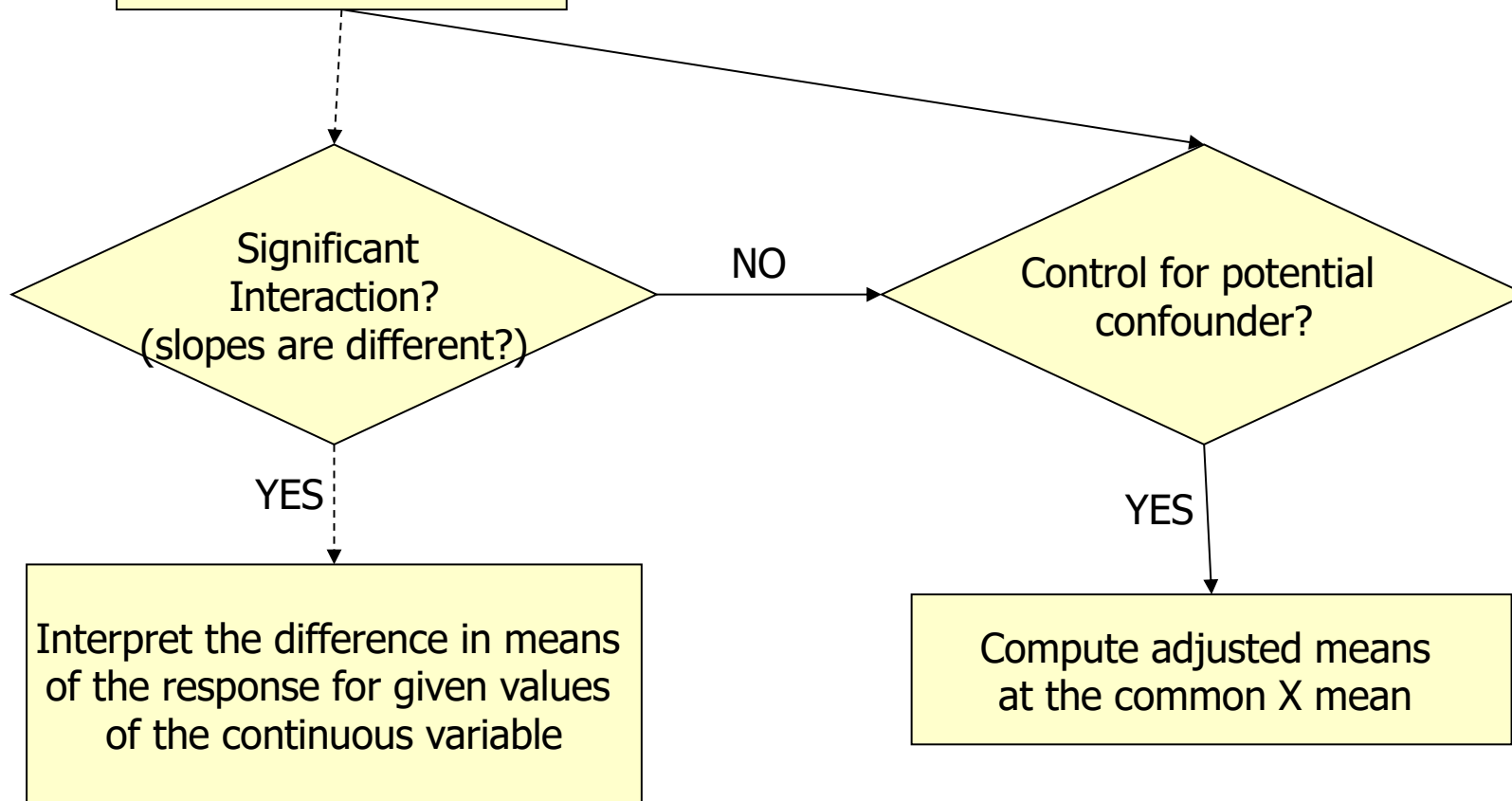
ANCOVA

```
> ## 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
```

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

SUMMARY:

ANCOVA





Summary

We have considered:

- ANOVA and ANCOVA
 - Interpretation
 - Estimation
 - Interaction

- Multiple comparisons



Lab

- Let's work on Exercises 9-12