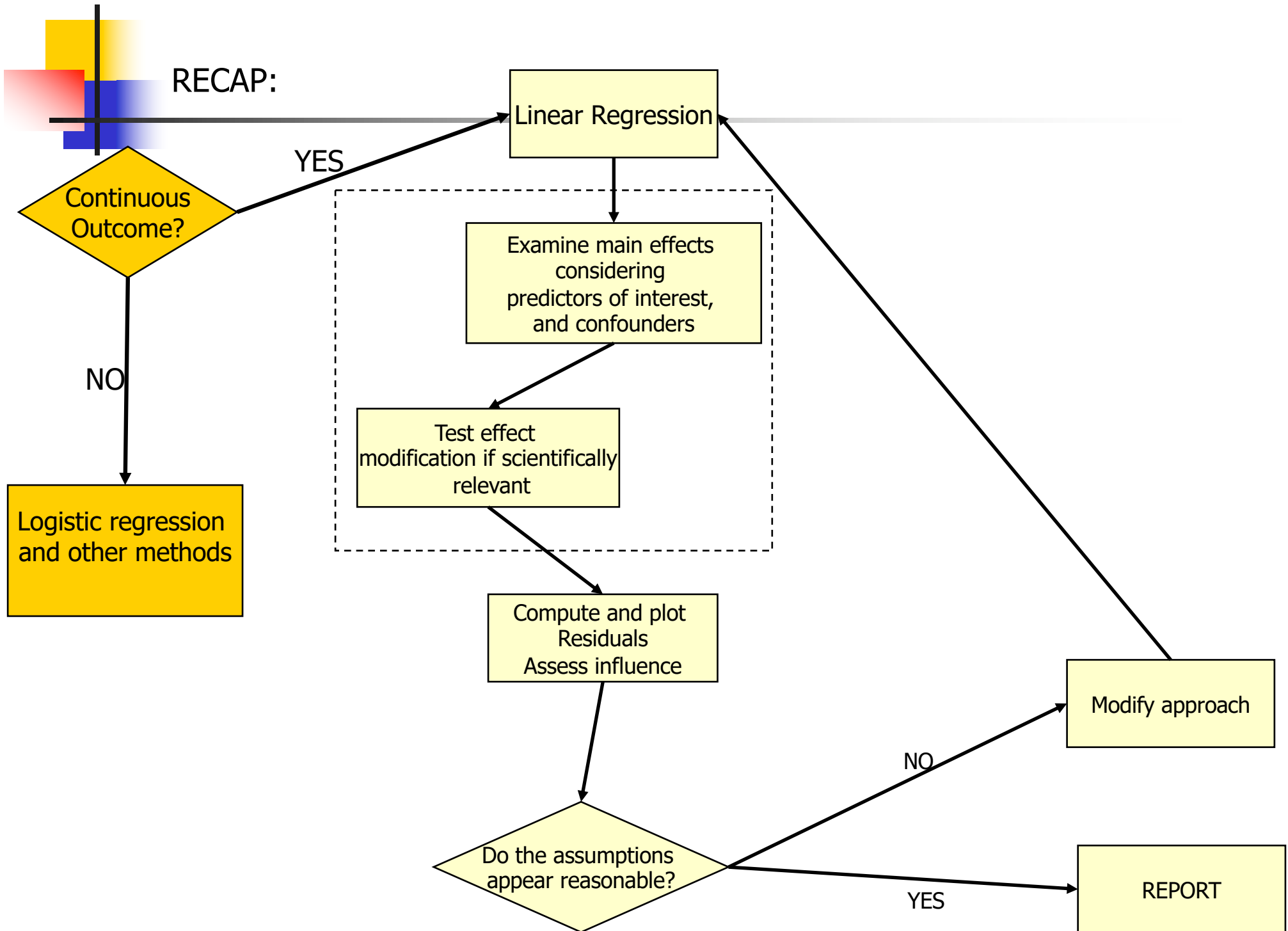




REGRESSION MODELS

ANOVA



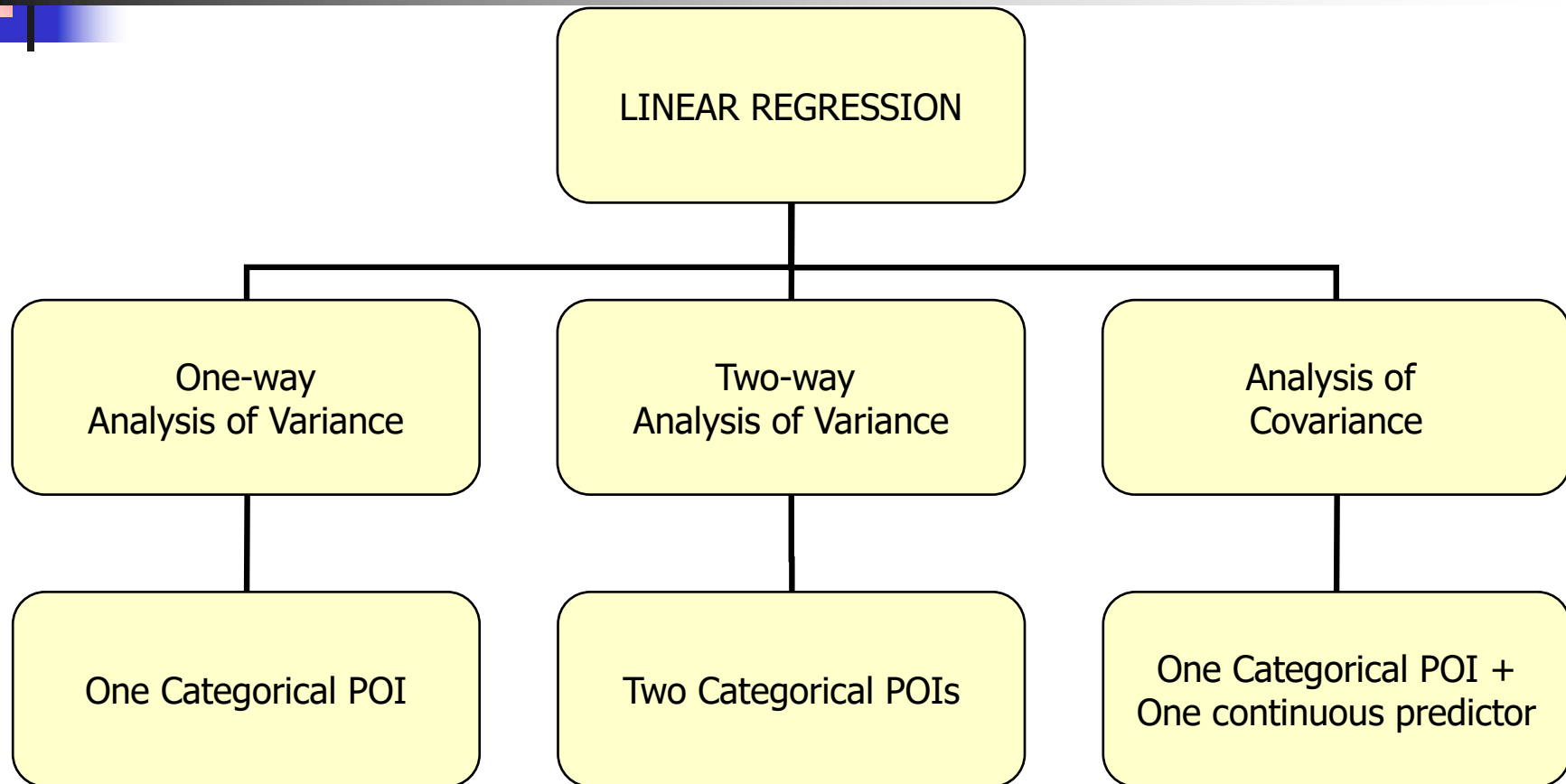


COMING UP NEXT: ANOVA – a special case of linear regression

- 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

Motivation:

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



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
```

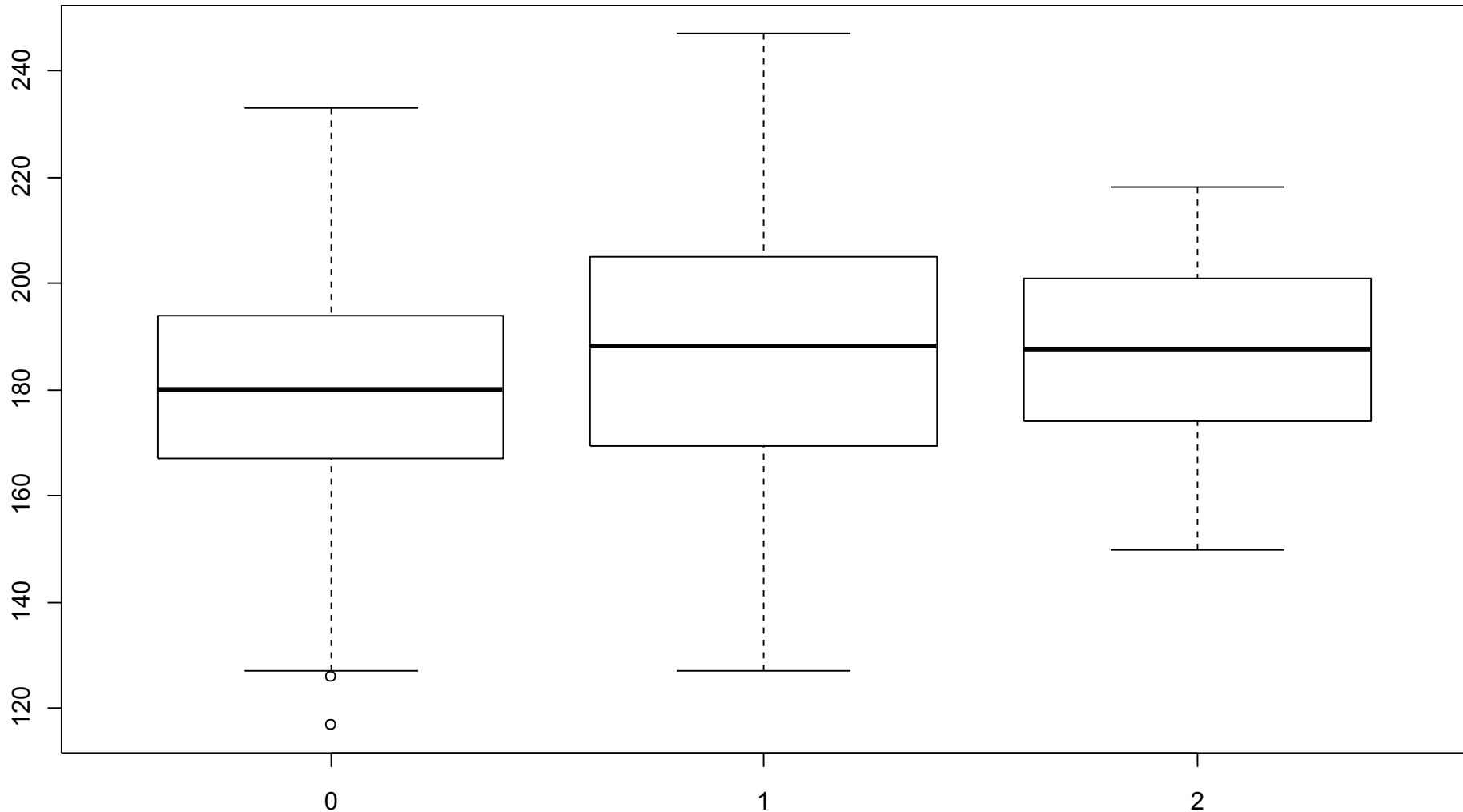
```
[1] 23.74541
```

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

```
[1] 17.38333
```

Motivation: Example

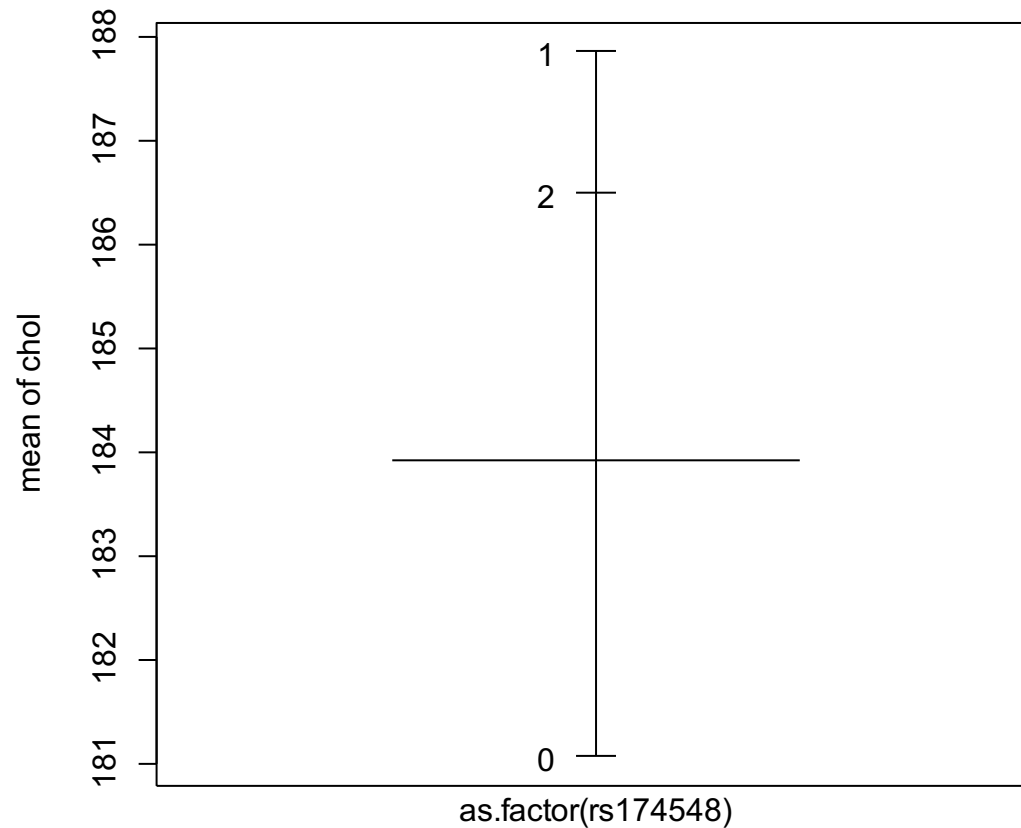
Is rs174548 associated with cholesterol?



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

Motivation: Example

Another graphical display:



R command:

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

Factors



Motivation: Example

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

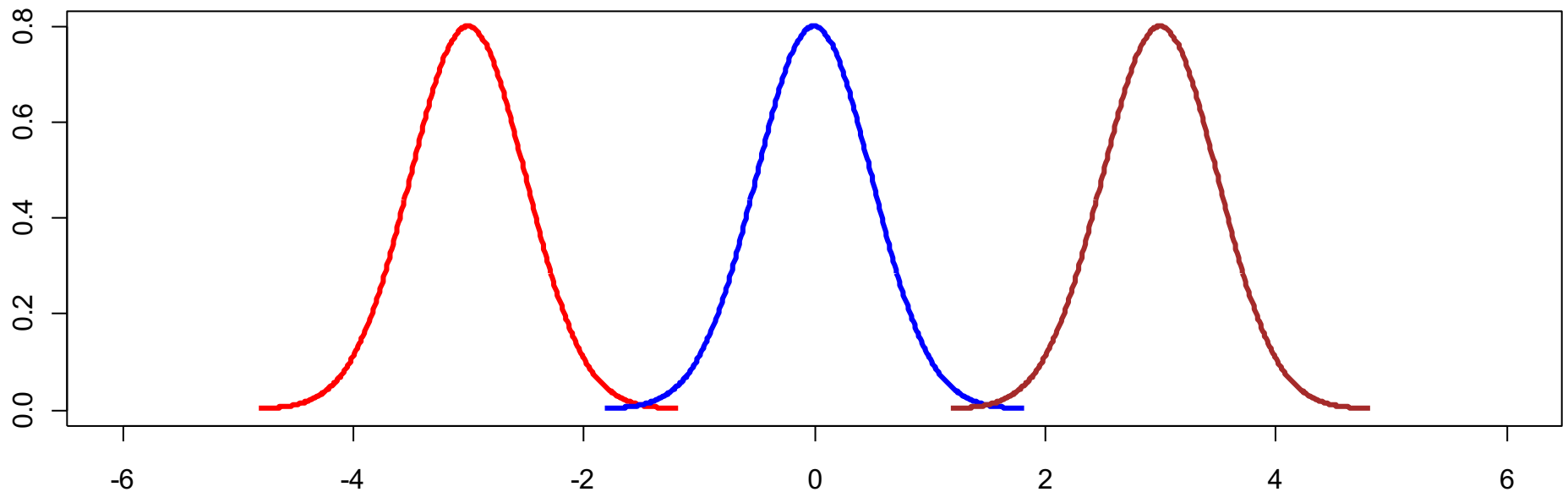


REGRESSION MODELS

One-way ANOVA as a regression model

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations

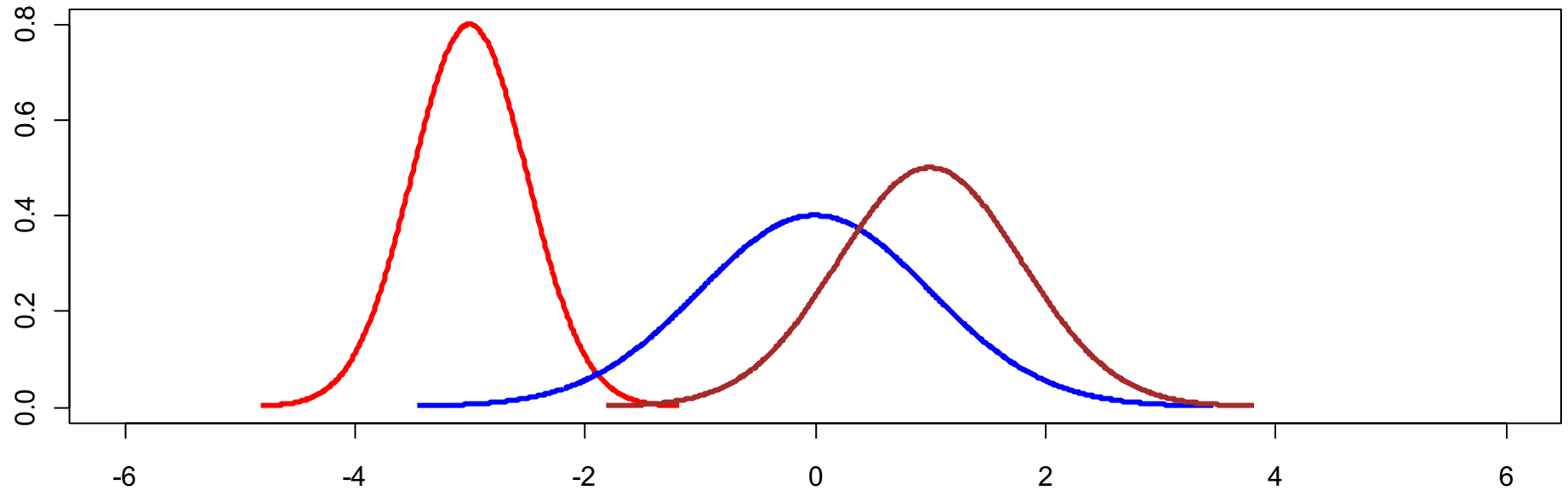


Assumptions for Classical ANOVA Framework:

Independence
Normality
Equal variances

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations





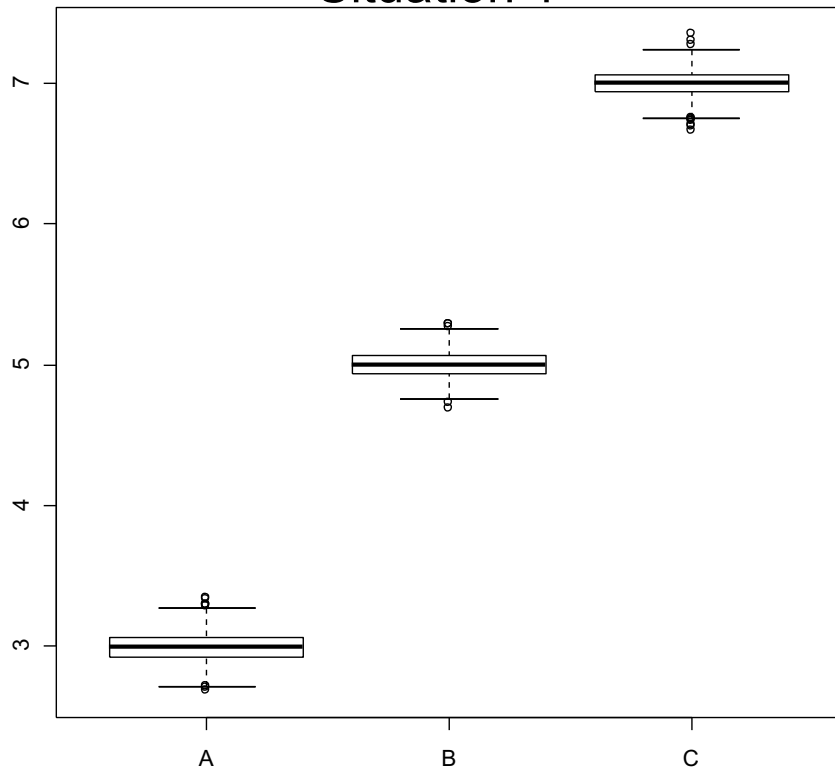
ANalysis Of VAriance Models (ANOVA)

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

ANalysis Of VAriance Models (ANOVA)

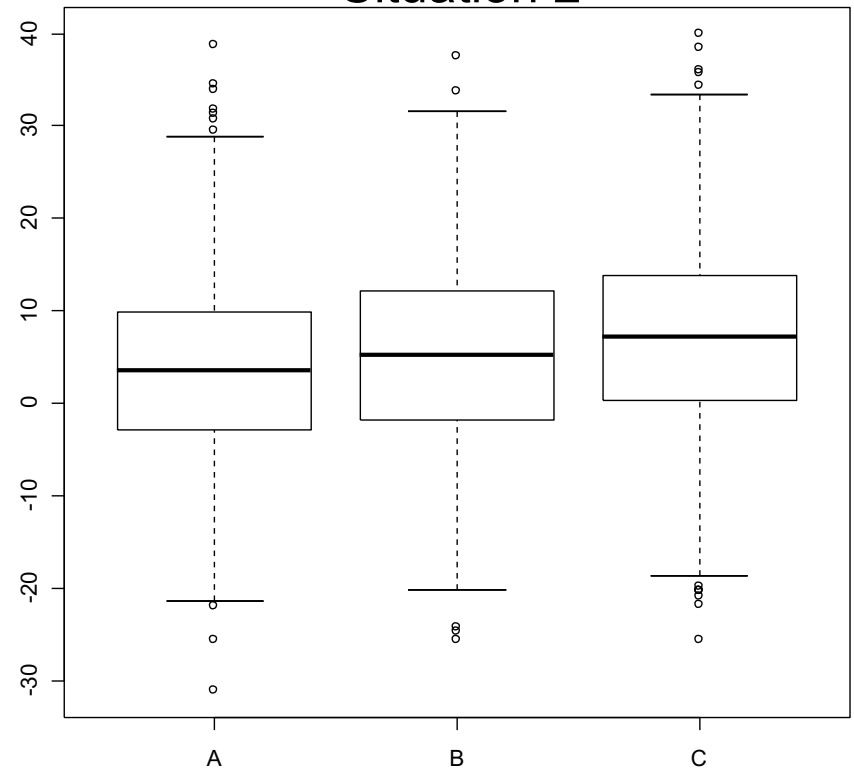
In both data sets, the true population means are: 3 (A), 5 (B), 7(C)

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

- 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

$H_0: \mu_1 = \mu_2 = \dots = \mu_K$ 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	$SSR = \sum_i (\bar{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{K-1}$	$\frac{MSR}{MSE}$
Residual	n-K	$SSE = \sum_{i,j} (y_{ij} - \bar{y}_i)^2$	$MSE = \frac{SSE}{n-K}$	
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: 21.92 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::coefstest(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!