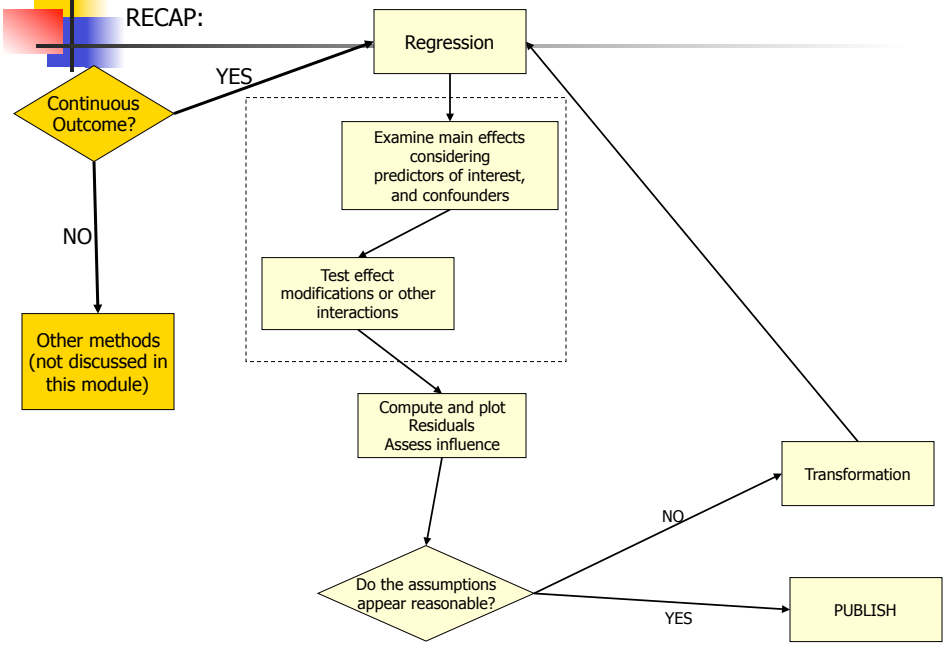
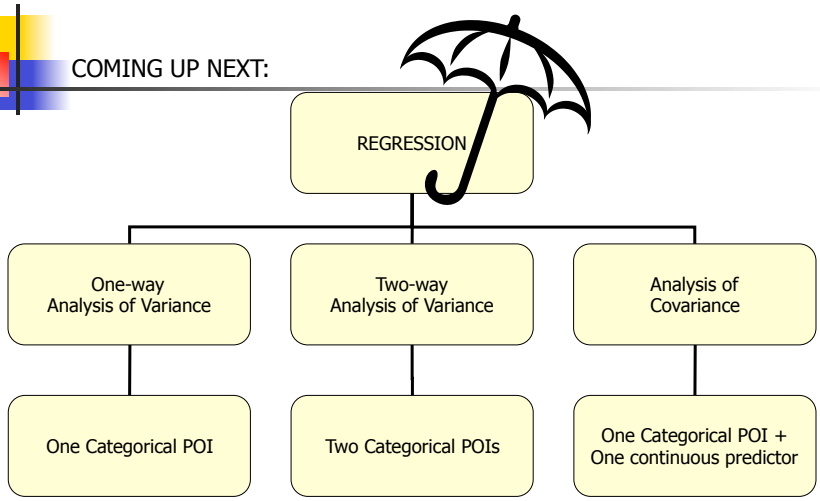
REGRESSION MODELS

ANOVA MODELS

RECAP:



COMING UP NEXT:



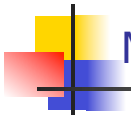
Uses dummy variables to represent categorical variables!

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

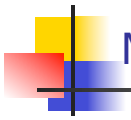
- Let's investigate if genetic factors are associated with cholesterol levels.
  - Ideally, you would have a confirmatory analysis of scientific hypotheses formulated prior to data collection
  - Alternatively, you could consider an exploratory analysis – hypotheses generation for future studies

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:

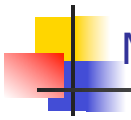
- Scientific question:
  - Assess the effect of rs174548 on cholesterol levels.



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
```

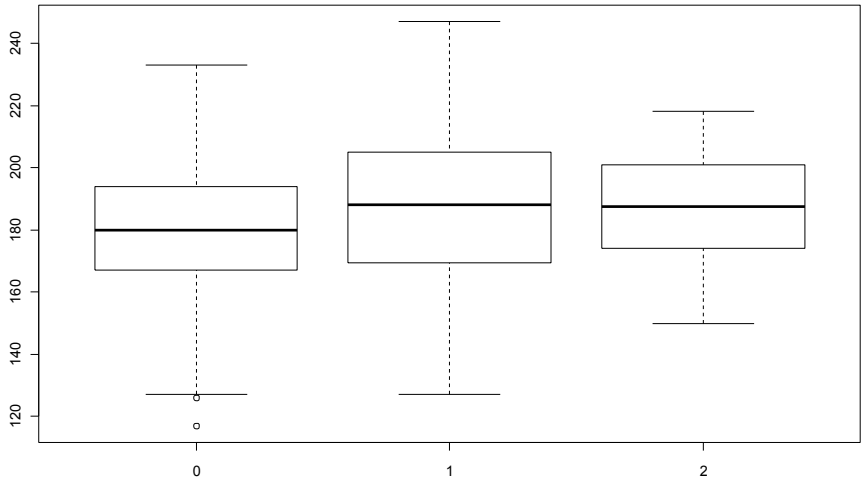
Motivation: Example

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

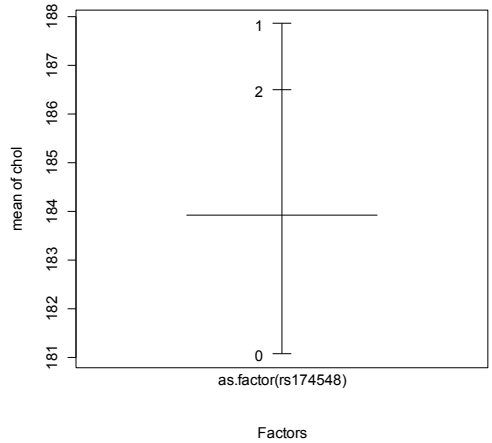
Is rs174548 associated with cholesterol?



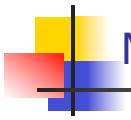
R command: boxplot(chol ~ as.factor(rs174548))

Motivation: Example

Another graphical display:



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



Motivation: Example

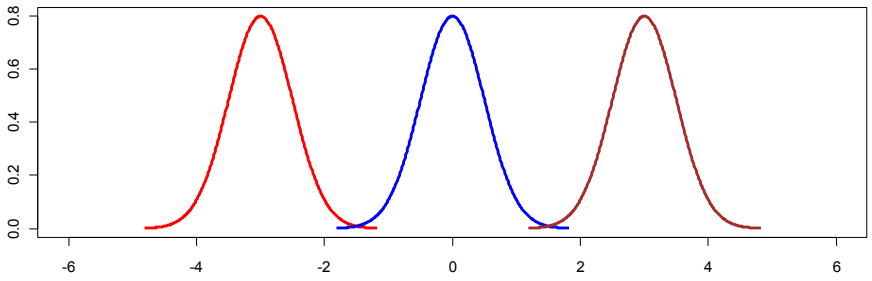
- Feature:
  - How do the mean responses compare across different groups?
    - Categorical/qualitative predictor

ANOVA

As a regression model

ANalysis Of VAriance Models (ANOVA)

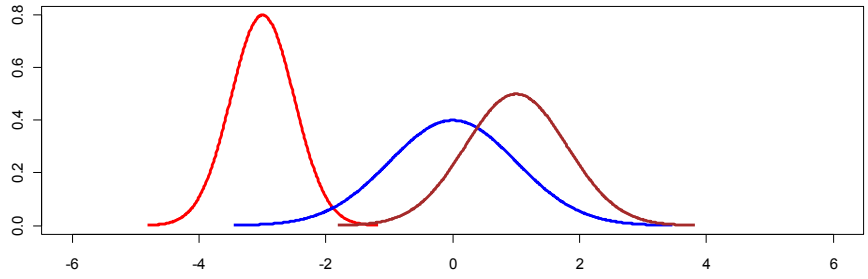
- Compares the means of several populations



Assumptions for Classical ANOVA Framework:
Independence
Normality
Equal variances

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations



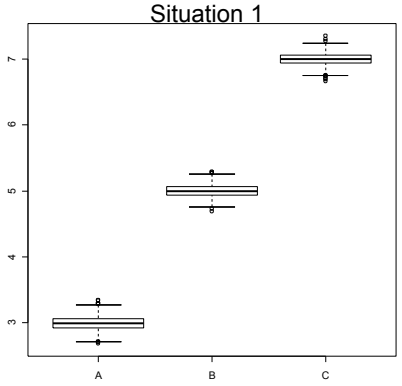
ANalysis Of VAriance Models (ANOVA)

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

ANalysis Of VAriance Models (ANOVA)

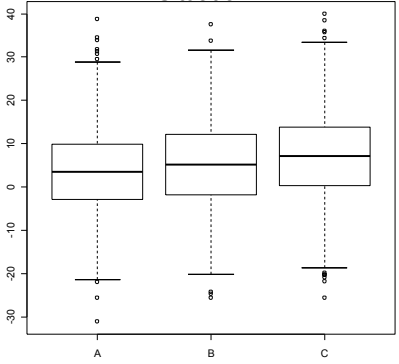
In both data sets, the true population means are: 3 (A), 5 (B), 7(C)

Situation 1



Low variance within groups

Situation 2



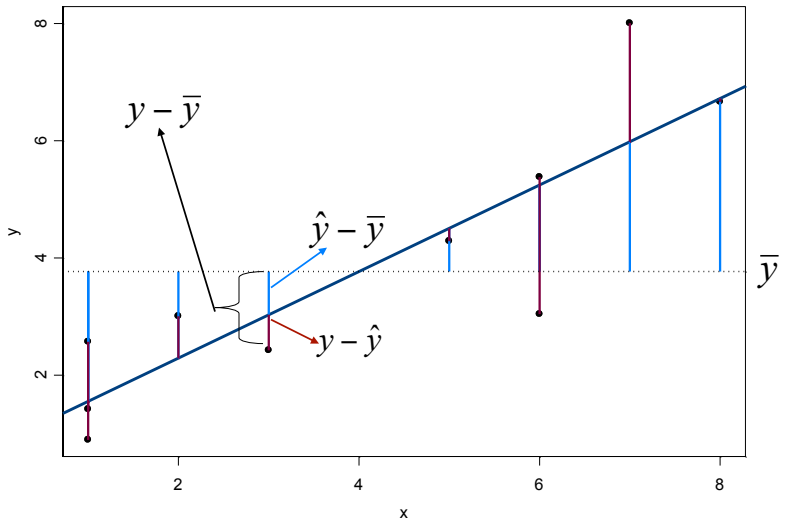
High variance within groups

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 population means!
      - 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: 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 as long as 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)
```

partial f-test

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!

global f-test???

```
> 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
```

```
> 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
---
```

```
> fit1.1 = 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
---
```

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


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)) "-1" removes intercept
> 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!

aov() just test whether the mean is difference across different subgroups

```
> 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).

the "4.703" here means adding one more "rs174548" increase your chr… level by 4.703 ??? I don't think I get it!!!

```
> 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
```

treat the number of "rs174548" as continous variable

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:

$\mu_0 = \mu_1$?

$\mu_0 = \mu_2$? } Pairwise comparisons

$\mu_1 = \mu_2$?

$(\mu_1 + \mu_2)/2 = \mu_0$? ⟶ 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

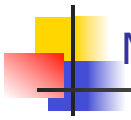
This option considers all pairwise comparisons

```
> ## 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

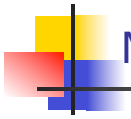
Multiple Comparisons

```
> ## -- 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.

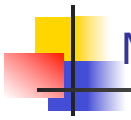
$$\mu_0 = (\mu_1 + \mu_2)/2$$

Or equivalently,

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

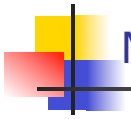
Or equivalently,

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



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

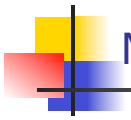
Using R for multiple comparisons with "user-defined" contrasts:

```
> 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

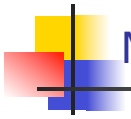
- What about using other adjustment methods?
  - For example, we used:

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

    (all pairwise comparisons, with Bonferroni adjustment)

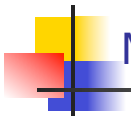
  - Other options, in place of "bonferroni", are:
    - summary(mc, test=adjusted("holm"))
    - summary(mc, test=adjusted("hochberg"))
    - summary(mc, test=adjusted("hommel"))
    - summary(mc, test=adjusted("BH"))
    - summary(mc, test=adjusted("BY"))
    - summary(mc, test=adjusted("fdr"))

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).

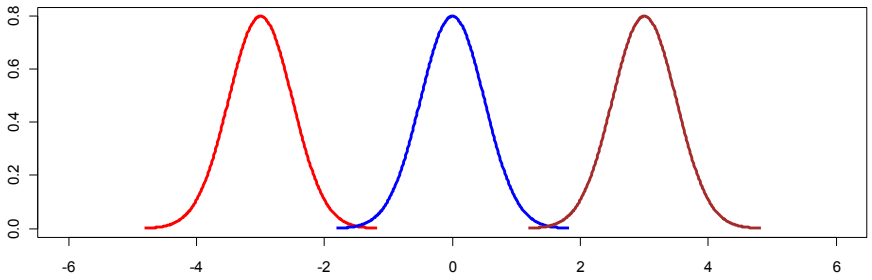
ANOVA

MODEL CHECKING

ANOVA Assumptions

- Recall the assumptions for classical ANOVA are:

  Independence
  Normality
  Equal variance

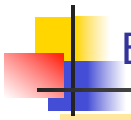


this test is not that powerful

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

for non-normal distribution

- 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

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

Relationships:

$\mu_1 = \beta_0$
$\mu_2 = \beta_0 + \beta_1$
$\mu_3 = \beta_0 + \beta_2$
...
$\mu_K = \beta_0 + \beta_{K-1}$

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

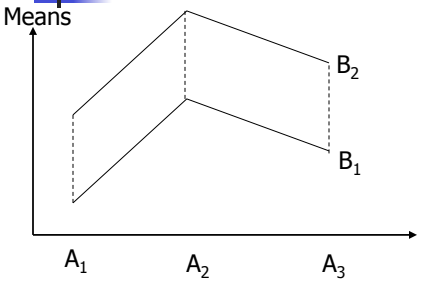
ANOVA: Two-Way Model

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

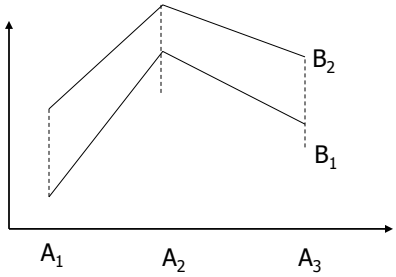
Factor A

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 **
```

f-test
null-hypothesis:
B1=B2=B3=0

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
```

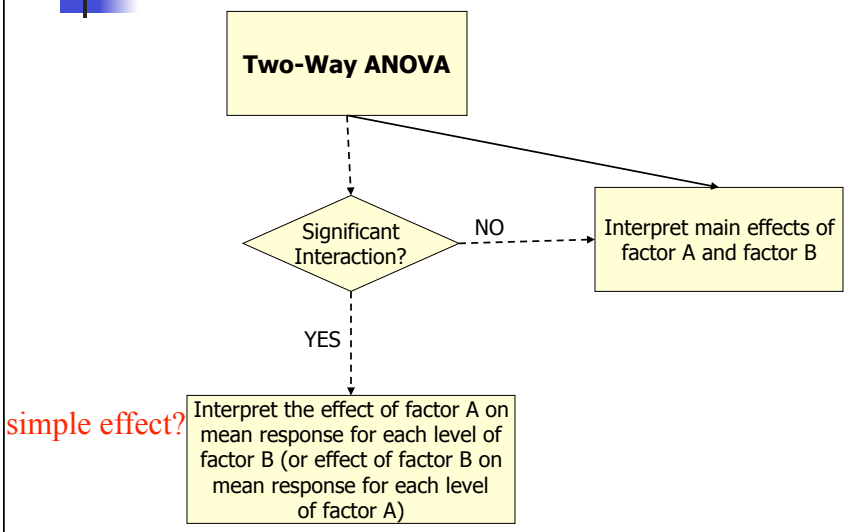

ANOVA: Two-Way Model (with interaction)

```
> 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
```

- 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
  - …

```
> 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
```

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



simple effect?

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

Interaction term

Note that:

no error term anymore because this is the mean

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

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

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

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

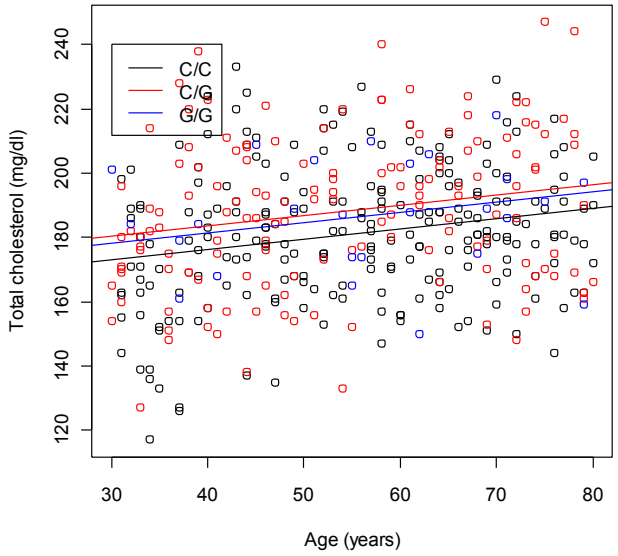
ANCOVA

```
> 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
```

ANCOVA

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

Residuals:
    Min      1Q  Median      3Q     Max
-64.062 -15.913  -0.062  14.938  59.136

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,fit2)
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    394 181961  3      8914 6.4339 0.0002912 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

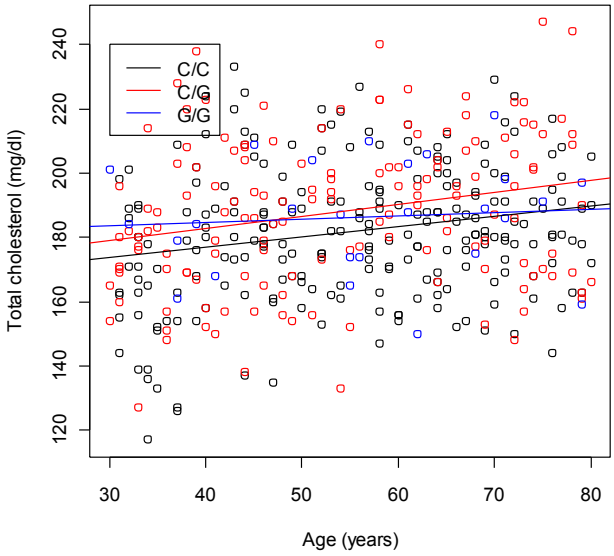
Test of coincident lines

ANCOVA

Test of parallel lines

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

Model 1: chol ~ 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,

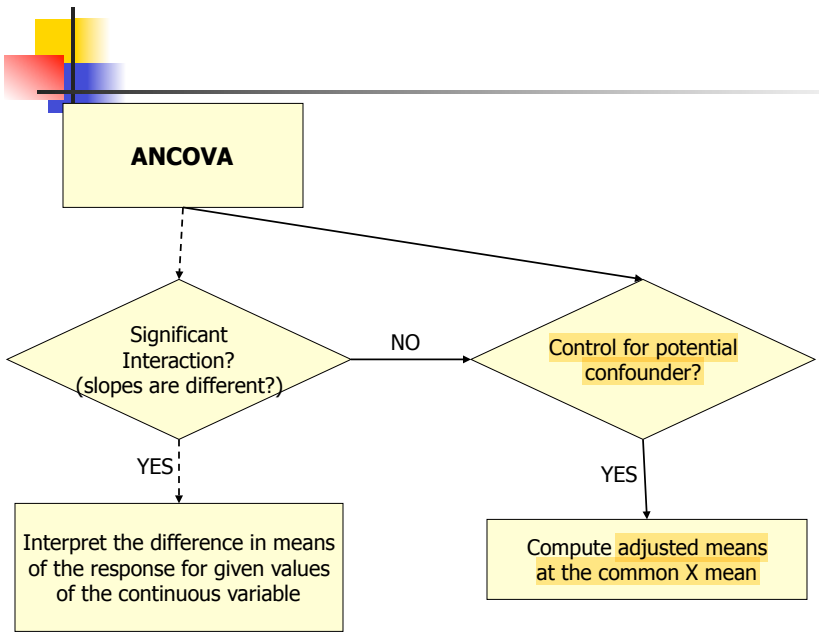
$$E[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 (CG) + \beta_3 (GG)$$

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

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

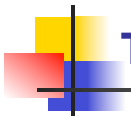
ANCOVA

```
> ## 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
```



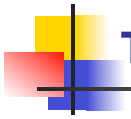
Experimental Designs & ANOVA

- 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

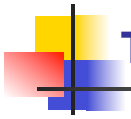


Tool Kit

- 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

balanced design

Factor		
YY	YYYYY	YYY

Unequal number of replicates per treatment

unbalanced design, OK now, but still better to have balanced design

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

Two-Way Data Patterns

Factor 2 \ Factor 1			
	Y	Y	Y
	Y	Y	Y
	Y	Y	Y
	Y	Y	Y

Single observation per cell

Factor 2 \ Factor 1			
	YYY	YYY	YYY
	YYY	YYY	YYY
	YYY	YYY	YYY
	YYY	YYY	YYY

Equal replication per cell

can test for main & interaction effect for the right and bottom designs

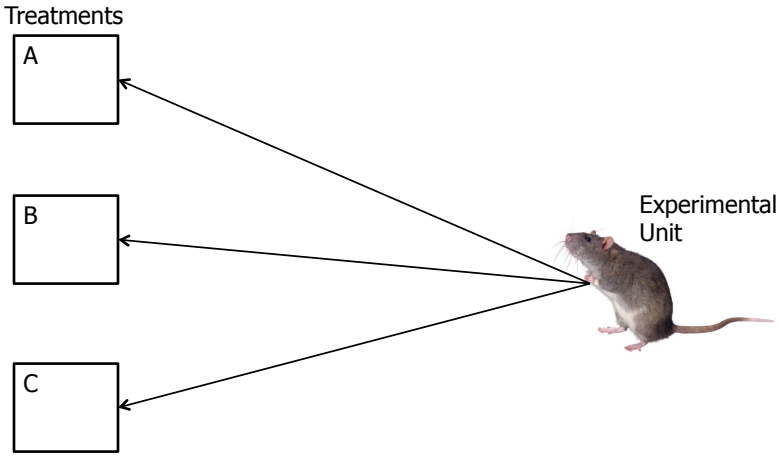
Factor 2 \ Factor 1			
	YY	YYY	YYYYYY
	YYY	YYYY	YY
	Y	YYY	YYYY
	YYYYY	YY	Y

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? One-way ANOVA model

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 two or more factors 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

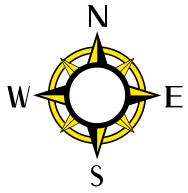
[1] Data are mean ± SE, n = 8/group. fCHO, fermentable carbohydrate; fCP, fermentable crude protein.
[2] Values are given as arbitrary values based on standard curves using pooled RNA samples. The mRNA abundance was normalized using *18S* rRNA, 60S ribosomal protein L19 (*RPL19*), hypoxanthine phosphoribosyltransferase I (*HPRT1*), 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

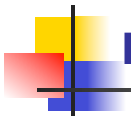


Large N-S variability

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.

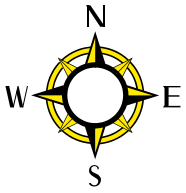
Randomized Complete Block Designs: Another example

- 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



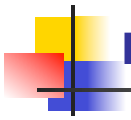
Large N-S variability

Small E-W variability

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

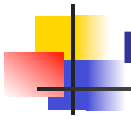
Block 1	C	A	B	D
Block 2	D	B	A	C
Block 3	A	B	D	C

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)