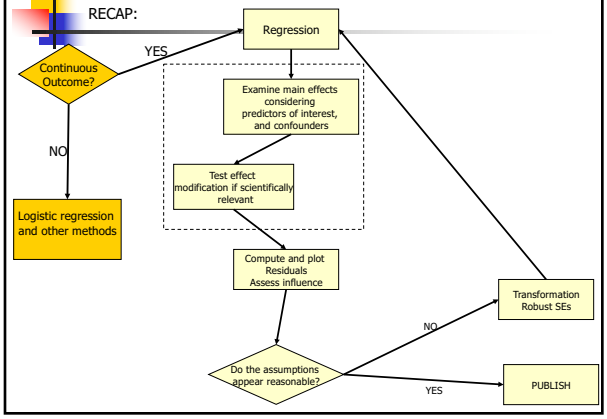
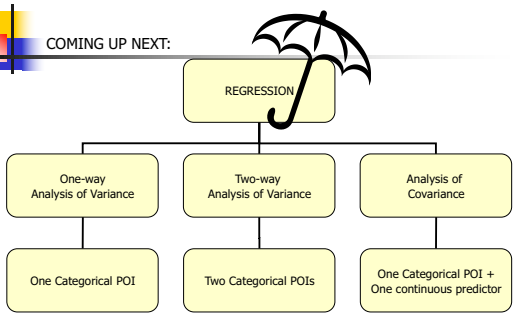
REGRESSION MODELS

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
- Logistic regression

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 sex 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, 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
```

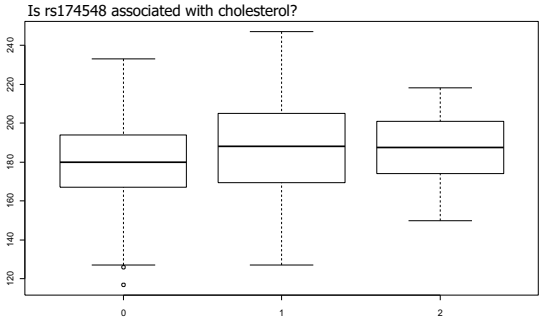
Motivation: Example

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

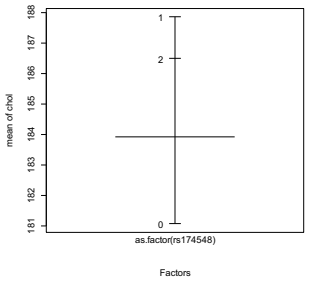
Is rs174548 associated with cholesterol?



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

Motivation: Example

Another graphical display:



R command:
`plot.design(chol ~ factor(rs174548))`

Motivation: Example

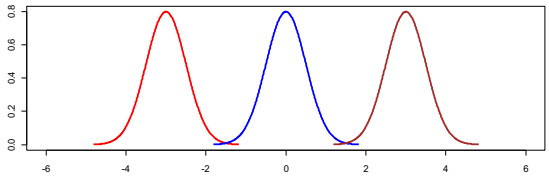
- Feature:
  - How do the mean responses compare across different groups?
    - Categorical/qualitative predictor

ANOVA

As a regression model

ANalysis Of VAriance Models (ANOVA)

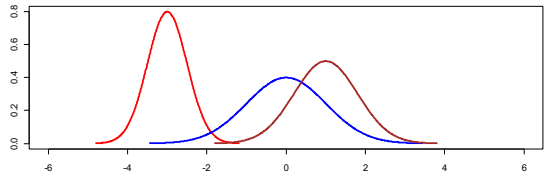
- Compares the means of several populations



Assumptions for Classical ANOVA Framework:
Independence
Normality
Equal variances

ANalysis Of VAriance Models (ANOVA)

- Compares the means of several populations

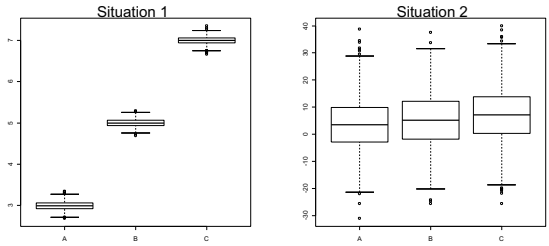


ANalysis Of VAriance Models (ANOVA)

- 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)



Low variance within groups

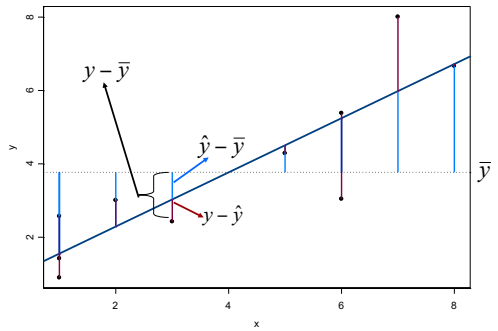
High variance within groups

Where do you expect to detect difference between population means?

ANalysis Of VAriance Models (ANOVA)

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

Decomposition of sum of squares



ANalysis Of VAriance Models (ANOVA)

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

ANOVA as a multiple regression model

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

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

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

$$\cdots$$

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

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: factor)

ANOVA: One-Way Model

By hand:
Creating "dummy" variables:

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

Fitting the ANOVA model:

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

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

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

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

> anova(fit0)
Analysis of Variance Table

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

ANOVA: One-Way Model

Better:
Let R do it for you!

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

```
> fit1.1 = lm(chol ~ factor(rs174548))
> summary(fit1.1)
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    598     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)
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 ~ factor(rs174548))
> summary(fit1.1)
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(fit1.1)
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

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

```
> fit1.1 = lm(chol ~ factor(rs174548))
> summary(fit1.1)
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: 0.02501,    Adjusted R-squared: 0.01314
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)
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 + factor(rs174548))
> summary(fit1.2)
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(fit1.2)
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

Alternative form:
- Different command!

```
> fit1.3 = aov(chol ~ factor(rs174548))
> summary(fit1.3)
                     Df Sum Sq Mean Sq F value  Pr(>F)
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)
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) factor(rs174548)1    factor(rs174548)2
          181.061674             6.802272            5.438326
```

ANOVA: One-Way Model

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

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

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

Call:
lm(formula = chol ~ rs174548)

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

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

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

> anova(fit2)
Analysis of Variance Table

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


ANOVA: One-Way Model

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

Call:
lm(formula = chol ~ rs174548)

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

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

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

> anova(fit2)
Analysis of Variance Table

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

- Model: $E[Y|x] = \beta_0 + \beta_1 x$
where Y: cholesterol, x: rs174548
- Interpretation of model parameters?
  - β_0: mean cholesterol in the C/C group [estimate: 181.575 mg/dl]
  - β_1: mean cholesterol difference between C/G and C/C – or – between G/G and C/G groups [estimate: 4.703 mg/dl]
- This model presumes differences between "consecutive" groups are the same (in this example, linear dose effect of allele) – more restrictive than the ANOVA model!

Back to the ANOVA model…

ANOVA: One-Way Model

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

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) $=1-(1-0.05)^r$	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 + 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 + factor(rs174548))

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


Multiple Comparisons

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

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


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

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


Multiple Comparisons

- 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 + 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 + 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 + 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 + 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).

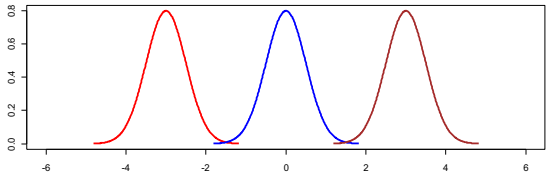
ANOVA

MODEL CHECKING

ANOVA Assumptions

- Recall the assumptions for classical ANOVA are:

Independence
Normality
Equal variance



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 ~ factor(rs174548))

        Bartlett test of homogeneity of variances

data:  chol by 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 ~ 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
```

One-Way ANOVA with robust standard errors

```
> summary(gee(chol ~ 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) factor(rs174548)1     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 ~ 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
factor(rs174548)1     6.802272   2.321365   2.930290    2.402005   2.831914
factor(rs174548)2     5.438326   4.539833   1.197913    3.624271   1.500530

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


Kruskal-Wallis Test

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

```
> kruskal.test(chol ~ 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!

Multiple Comparisons (following Kruskal-Wallis Test)

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

        Wilcoxon rank sum test with continuity correction

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


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

        Wilcoxon rank sum test with continuity correction

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


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

        Wilcoxon rank sum test with continuity correction

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

Summary:

GOAL: Comparison of Means across K groups

Relationships:

$\mu_1 = \beta_0$

$\mu_2 = \beta_0 + \beta_1$

$\mu_3 = \beta_0 + \beta_2$

...

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

One-way ANOVA:

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

H_1: not all means are equal

Multiple Regression:

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

where $group_1$ is the reference group

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

H_1: not all β_i are equal to zero

Rejected H_0?

YES

Multiple Comparisons (control α overall)

e.g. Bonferroni: α/#comparisons

ANOVA

Two-way ANOVA models

ANOVA: Two-Way Model Motivation:

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

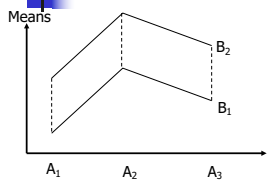
ANOVA: Two-Way Model

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

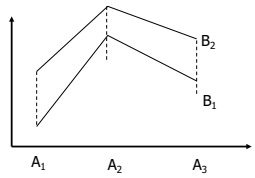
Factor A

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

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

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

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

> anova(fit0,fit1)
Analysis of Variance Table

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


ANOVA: Two-Way Model (without interaction)

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

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

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

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

> anova(fit0,fit1)
Analysis of Variance Table

Model 1: chol ~ factor(sex)
Model 2: chol ~ factor(sex) + 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 sex: 7.236 mg/dl
  - Estimated difference in mean cholesterol levels between G/G and C/C groups adjusted by sex: 5.184 mg/dl
  - There is evidence that cholesterol is associated with sex ($p < 0.001$).
  - 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 ~ factor(sex) * factor(rs174548))
> summary(fit2)

Call:
lm(formula = chol ~ factor(sex) * 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 ***
factor(sex)1                              5.7109     2.7982   2.041  0.04192 *
factor(rs174548)1                         0.9597     3.1306   0.307  0.75933
factor(rs174548)2                        -0.2015     6.4053  -0.031  0.97492
factor(sex)1:factor(rs174548)1           12.7398     4.4650   2.853  0.00456 **
factor(sex)1:factor(rs174548)2           10.2296     8.7482   1.169  0.24297
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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


ANOVA: Model comparison

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

Model 1: chol ~ factor(sex) + factor(rs174548)
Model 2: chol ~ factor(sex) * 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)

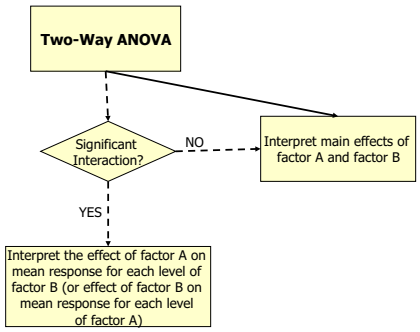
```
> fit2 = lm(chol ~ factor(sex) * factor(rs174548))
> summary(fit2)

Call:
lm(formula = chol ~ factor(sex) * 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 ***
factor(sex)1                            5.7109     2.7982   2.041  0.04192 *
factor(rs174548)1                       0.9597     3.1306   0.307  0.75933
factor(rs174548)2                      -0.2015     6.4053  -0.031  0.97492
factor(sex)1:factor(rs174548)1         12.7398     4.4650   2.853  0.00456 **
factor(sex)1:factor(rs174548)2         10.2296     8.7482   1.169  0.24297
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

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

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

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

Interpretation of results:

- Estimated mean cholesterol for male C/C group: 178.12 mg/dl
- Estimated mean cholesterol for female C/C group? (178.12 + 5.7109) mg/dl
- Estimated mean cholesterol for male C/G group: (178.12 +0.9597) mg/dl
- Estimated mean cholesterol for female C/G group: (178.12 + 5.7109 + 0.9597 + 12.7398) mg/dl
- …

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



ANCOVA MODELS

(aka ANACOVA)

ANalysis of COVAriance Models (ANCOVA) Motivation:

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

ANalysis of COVAriance Models (ANCOVA)

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

ANalysis of COVAriance Models (ANCOVA)

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

Interaction term

Note that:

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

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

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

ANCOVA

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

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

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

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

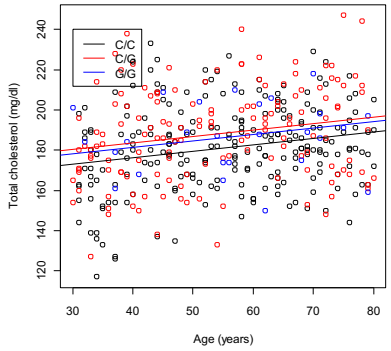
ANCOVA

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

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

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

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

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


ANCOVA

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

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

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

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

> anova(fit0,fit1)
Analysis of Variance Table

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

ANCOVA



ANCOVA

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

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

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

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


ANCOVA

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

Call:
lm(formula = chol ~ age)

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

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

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

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

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

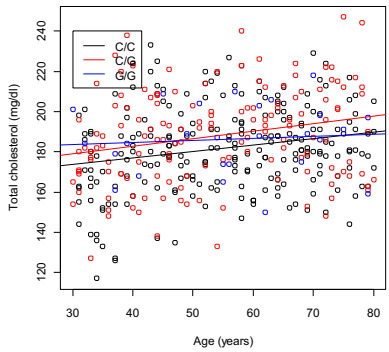
Test of coincident lines

ANCOVA

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

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

Test of parallel lines

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,

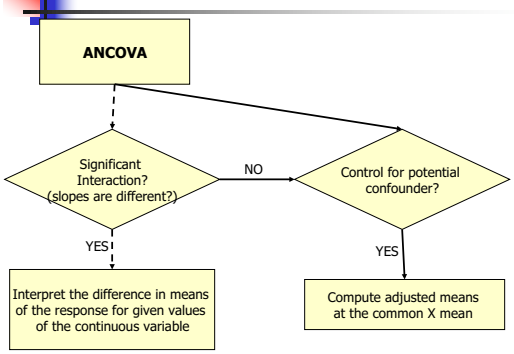
$$E[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 (CG) + \beta_3 (GG)$$

  - 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

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

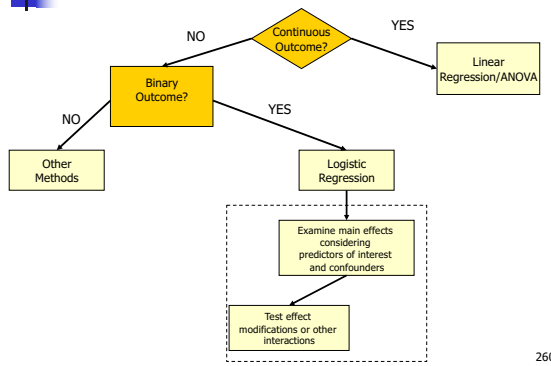
ANCOVA

```
> ## 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
```

Logistic Regression

RECAP:



Logistic Regression: Motivation

- Many scientific questions of interest involve a binary outcome (e.g. disease/no disease)
- Let's investigate if genetic factors are associated with coronary heart disease.

Logistic Regression: Motivation

- Scientific questions of interest:
  - Assess the effect of rs4775401 on CHD
  - Assess the effect of cholesterol on CHD
  - Assess the effect of rs4775401 on CHD after accounting for cholesterol

Logistic Regression: Motivation

- Scientific question:
  - Assess the effect of rs4775401 on odds of CHD

Motivation: rs4755401 and CHD

Here is a contingency table for the SNP and CHD:

```
> table(rs4775401,chd)
          chd
rs4775401   0   1
        0 154  48
        1 104  66
        2  15  13
```

Without using regression, what tool could we use to look for an association?

Motivation: rs4755401 and CHD

Here is a contingency table for the SNP and CHD:

```
> table(rs4775401,chd)

      0   1
  0 154  48
  1 104  66
  2  15  13
```

Without using regression, what tool could we use to look for an association?

```
> chisq.test(rs4775401,chd)

Pearson's Chi-squared test
data:  rs4775401 and chd
X-squared = 12.657, df = 2, p-value = 0.001785
```

In addition to hypothesis testing, we need to summarize the strength of association between the two variables

Measures of association for binary outcomes

Outcome

Exposure		No	Yes
	Yes	a	b
	No	c	d

- Risk difference (RD) = P(outcome|exposed) - P(outcome|not exposed)
  = (b/(a+b)) - (d/(c+d))

```
> table(rs4775401,chd)

      0   1
  0 154  48
  1 104  66
  2  15  13
```

- RD(T/T vs C/C) = 13/(13+15) – 48/(48+154) = 0.23

Measures of association for binary outcomes

Outcome

Exposure		No	Yes
	Yes	a	b
	No	c	d

- Risk difference interpretation
  - Additive difference in probability (risk) between exposed and unexposed
  - Also called *excess risk*
  - -1 < RD < 1
  - RD = 0 ⇒ no association; risk of outcome same for exposed and unexposed

Measures of association for binary outcomes

Exposure	Outcome: No	Outcome: Yes
Yes	a	b
No	c	d

- Relative risk (RR) = P(outcome|exposed)/P(outcome|not exposed)
  = (b/(a+b))/(d/(c+d))

```
> table(rs4775401,chd)

      0   1
  0 154  48
  1 104  66
  2  15  13
```

- RR(T/T vs C/C) = (13/(13+15)) / (48/(48+154)) = 1.95

Measures of association for binary outcomes

Exposure	Outcome: No	Outcome: Yes
Yes	a	b
No	c	d

- Relative risk interpretation
  - Multiplicative difference in probability (risk) of outcome among exposed compared to unexposed
  - $0 < RR < \infty$
  - RR = 1 ⇒ no association; risk of outcome same for exposed and unexposed

Measures of association for binary outcomes

Exposure	Outcome: No	Outcome: Yes
Yes	a	b
No	c	d

- Odds = P/(1-P)
- Odds ratio (OR) = Odds(outcome|exposed)/Odds(outcome|not exposed)
  = ((b/(a+b))/(a/(a+b)))/((d/(c+d))/(c/(c+d)))
  = (b/a)/(d/c) = (bc)/(ad)

```
> table(rs4775401,chd)

      0   1
  0 154  48
  1 104  66
  2  15  13
```

- OR(T/T vs C/C) = (13/15) / (48/154) = 2.78

Measures of association for binary outcomes

Exposure	Outcome: No	Outcome: Yes
Yes	a	b
No	c	d

- Odds ratio interpretation
  - Multiplicative difference in odds of outcome between exposed and unexposed
  - $0 < OR < \infty$
  - $OR = 1 \Rightarrow$ no association; odds of outcome same for exposed and unexposed

Pros and cons of measures of association

- RD is appealing because it directly communicates absolute increase in risk
  - Often more policy relevant than relative measures
- RR more directly interpretable than OR (most people don't have an intuitive understanding of odds)
- OR estimable in case-control studies where RR and RD are not
- For rare outcomes, $OR \approx RR$

Logistic Regression: Motivation

- The chi-squared test is adequate for investigating the association between two categorical predictors
- But what if we want to investigate the association between a continuous predictor like cholesterol and a binary outcome like CHD?
- Logistic regression will provide us with a tool for this

Binary outcome and continuous exposure

- Objective: Estimate association between binary outcome and continuous exposure
- Y_i = binary response
  X_i = continuous exposure
  $p_i = E(Y_i|X_i) = P(Y_i = 1|X_i)$
- One solution – fit a linear model

$$E(Y_i \mid X_i) = P(Y_i = 1 \mid X_i) = \beta_0 + \beta_1 X_i$$

- This is just a standard linear model except our outcome is binary
- Interpretation of β_1?
- Problems with this approach?

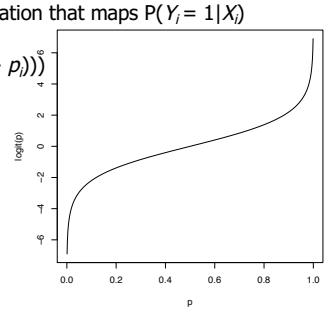
Motivating example: CHD and cholesterol

```
> lm.mod1  <- lm(chd ~ chol, data = cholesterol)
> summary(lm.mod1)

Call:
lm(formula = chd ~ chol, data = cholesterol)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7067 -0.3301 -0.1289  0.3975  1.0227

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.4245087  0.1747852   -8.15 4.77e-15 ***
chol         0.0094718  0.0009436   10.04  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4169 on 398 degrees of freedom
Multiple R-squared:  0.202,       Adjusted R-squared:    0.2
F-statistic: 100.8 on 1 and 398 DF,  p-value: < 2.2e-16
```

What is the interpretation of the cholesterol parameter estimate?

Binary outcome and continuous exposure

- Solution: use a transformation that maps $P(Y_i = 1|X_i)$ to the real line
- Let $\text{logit}(p_i) = \log(p_i / (1 - p_i))$
- $p_i \in (0, 1)$
- $p_i/(1 - p_i) \in (0, \infty)$
- $\log(p_i/(1 - p_i)) \in (-\infty, \infty)$
- Regress $\text{logit}(p_i)$ on X_i



$$\text{logit}(E(Y_i \mid X_i)) = \log(P(Y_i = 1 \mid X_i) / (1 - P(Y_i = 1 \mid X_i))) = \beta_0 + \beta_1 X_i$$

Interpretation of logistic regression parameters

- On the log-odds scale

$$\log(\text{odds}(Y_i = 1 \mid X_i = (c+1))) = \beta_0 + \beta_1(c+1)$$
$$- \log(\text{odds}(Y_i = 1 \mid X_i = c)) = \beta_0 + \beta_1 c$$
$$\log(\text{odds}(Y_i = 1 \mid X_i = (c+1))) - \log(\text{odds}(Y_i = 1 \mid X_i = c)) = \beta_1$$
$$\log\left(\boxed{\text{odds}(Y_i = 1 \mid X_i = (c+1)) / \text{odds}(Y_i = 1 \mid X_i = c)}\right) = \beta_1$$
$$\log(\text{OR}) = \beta_1$$

Odds Ratio (OR)

- That is, for two observations that differ by one unit in X there is a difference of β_1 in their log odds of Y = 1
- Or, equivalently, the log of the ratio of the odds of Y = 1 (i.e. the log OR) for two units that differ in X by one unit is β_1

Interpretation of logistic regression parameters

- By exponentiating we arrive at a simpler interpretation

$$\exp(\log(\text{OR})) = \exp(\beta_1)$$
$$\text{OR} = \exp(\beta_1)$$

- So for two observations that differ in X by one unit there is a multiplicative difference in their odds of Y = 1 of $\exp(\beta_1)$
- Or, equivalently, the ratio of the odds of Y = 1 (i.e., the odds ratio) for two observations that differ in X by one unit is $\exp(\beta_1)$

Motivating example: CHD and cholesterol

```
> glm.mod1 <- glm(chd ~ chol, family = "binomial")
> summary(glm.mod1)

Call:
glm(formula = chd ~ chol, family = "binomial", data = cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7437  -0.8219  -0.4852   0.9096   2.4536

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.09600    1.29881  -8.543  < 2e-16 ***
chol          0.05498    0.00678   8.109 5.12e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 409.71  on 398  degrees of freedom
AIC: 413.71

Number of Fisher Scoring iterations: 4
```

- What do these results tell us about the relationship between cholesterol and CHD?

Motivating example: CHD and cholesterol

```
> glm.mod1 <- glm(chd ~ chol, family = "binomial")
> summary(glm.mod1)

Call:
glm(formula = chd ~ chol, family = "binomial", data = cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7437  -0.8219  -0.4852   0.9096   2.4536

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.09600    1.29881  -8.543  < 2e-16 ***
chol          0.05498    0.00678   8.109 5.12e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 409.71  on 398  degrees of freedom
AIC: 413.71

Number of Fisher Scoring iterations: 4
```

- Comparing two people who differ in cholesterol by 1 mg/dl, the log odds of CHD are higher by 0.055 for the individual with higher cholesterol

Motivating example: CHD and cholesterol

- Differences in log odds are pretty spectacularly difficult to interpret!
- It would be much better to exponentiate the coefficients and report odds ratios

```
> exp(glm.mod1$coef)
 (Intercept)        chol
1.517293e-05 1.056515e+00
> exp(confint(glm.mod1))
Waiting for profiling to be done...
                  2.5 %       97.5 %
(Intercept) 1.061838e-06 0.0001744859
chol        1.043101e+00 1.0712556915
```

- Comparing two people who differ in cholesterol by 1 mg/dl, the odds of CHD are higher by a factor of 1.06 (95% CI: 1.04, 1.07) for the individual with higher cholesterol

Multivariable logistic regression

- Often we are interested in examining associations between multiple predictors simultaneously and a binary outcome
- Multiple logistic regression follows same pattern as linear regression

$$\text{logit}(E(Y_i \mid X_{1i},\ldots,X_{pi})) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$$

- $\exp(\beta_\kappa)$ interpreted as the OR associated with a one unit change in the *k*th predictor, among individuals with other predictors at same levels (or holding other predictors constant/controlling for/adjusting for etc.)

Motivating example

```
> glm.mod2 <- glm(chd ~ chol+factor(rs4775401), family = "binomial", data = cholesterol)
> summary(glm.mod2)

Call:
glm(formula = chd ~ chol + factor(rs4775401), family = "binomial",
    data = cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5528  -0.7810  -0.4585   0.8037   2.6275

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -11.625209   1.335335  -8.706  < 2e-16 ***
chol                0.055443   0.006872   8.069 7.11e-16 ***
factor(rs4775401)1  0.794212   0.259257   3.063  0.00219 **
factor(rs4775401)2  1.138308   0.464317   2.452  0.01422 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 397.27  on 396  degrees of freedom
AIC: 405.27

Number of Fisher Scoring iterations: 4
```


Hypothesis testing for logistic regression

- Maximum likelihood is the standard method of estimating parameters from logistic models and is based on finding the estimates which maximize the joint probability for the observed data under the chosen model.
- The Wald test uses maximum likelihood estimates (MLE) and their standard errors to conduct hypothesis tests
- Test: H_0: $\beta_j = 0$ (no association) vs. H_A: $\beta_j \neq 0$
- Construct a z-score:

$$z = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim N(0, 1) \Rightarrow \text{Wald Test}$$

Motivating example

```
> glm.mod2 <- glm(chd ~ chol+factor(rs4775401), family = "binomial", data = cholesterol)
> summary(glm.mod2)

Call:
glm(formula = chd ~ chol + factor(rs4775401), family = "binomial",
    data = cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5528  -0.7810  -0.4585   0.8037   2.6275

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -11.625209   1.335335  -8.706  < 2e-16 ***
chol                0.055443   0.006872   8.069 7.11e-16 ***
factor(rs4775401)1  0.794212   0.259257   3.063  0.00219 **
factor(rs4775401)2  1.138308   0.464317   2.452  0.01422 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 397.27  on 396  degrees of freedom
AIC: 405.27

Number of Fisher Scoring iterations: 4
```

Wald statistics and p-values for each parameter

Likelihood ratio test

- The likelihood ratio statistic is useful in comparing nested models. (LRT = likelihood ratio test)
- This allows us to test hypotheses about multiple parameters simultaneously such as

  H_0: $\beta_1 = \beta_2 = 0$ vs

  H_A: at least one parameter not equal to 0
- In order to use the LRT we must fit a nested hierarchy of models
- For example:

  Model 1: logit $p_i = \beta_0 + \beta_1 chol_i$

  Model 2: logit $p_i = \beta_0 + \beta_1 chol_i + \beta_2 SNP_{1i} + \beta_3 SNP_{2i}$

Likelihood ratio test

- The LRT allows us to test the significance of the additional parameters in the larger model.
- Example: Compare model 2 to model 3

  H_0: $\beta_2 = \beta_3 = 0$

  $LRT = -2\,[L_1 - L_2] \sim \chi^2_2$

  df = # parameters being tested

Example: Likelihood ratio test

```
> lrtest(glm.mod1,glm.mod2)
Likelihood ratio test

Model 1: chd ~ chol
Model 2: chd ~ chol + factor(rs4775401)
  #Df  LogLik Df Chisq Pr(>Chisq)
1   2 -204.85
2   4 -198.63  2 12.44   0.001989 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- After accounting for cholesterol, there is a statistically significant association between rs4775401 and CHD

Summary

We have considered:

- ANOVA and ANCOVA
  - Interpretation
  - Estimation
  - Interaction
- Logistic regression
  - Interpretation
  - Estimation

Everything is regression!

(Professor Scott Emerson)