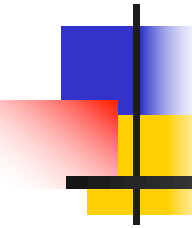
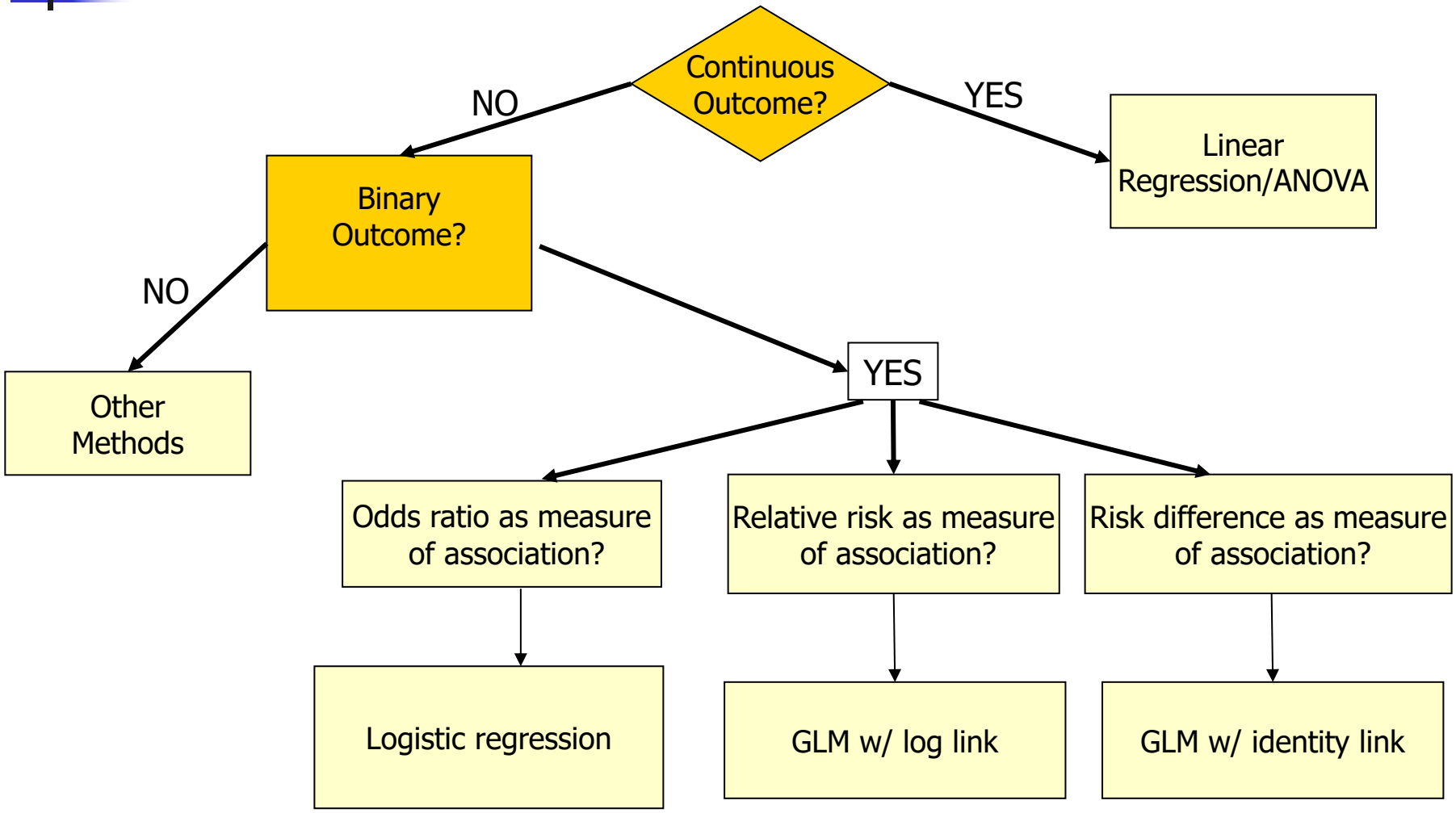
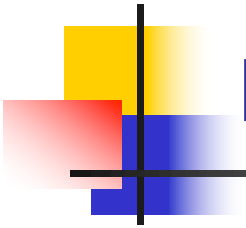


REGRESSION METHODS

Logistic regression

RECAP:


Continuous Outcome?
NO
YES
Linear Regression/ANOVA
Binary Outcome?
NO
Other Methods
YES
Odds ratio as measure of association?
Relative risk as measure of association?
Risk difference as measure of association?
Logistic regression
GLM w/ log link
GLM w/ identity link

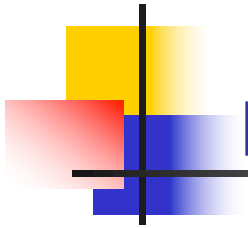


Logistic Regression: Motivation

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

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 risk of CHD
  - rs4775401 - Coded as the number of minor alleles
    - 0 = C/C, 1 = C/T, 2 = T/T.

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
```

Prevalence of CHD in C/C:
48/(48+154) = 0.238

Prevalence of CHD in C/T:
66/(66+104) = 0.388

Prevalence of CHD in T/T:
13/(13+15) = 0.464

- Does the prevalence of CHD differ across the groups?
- Without using regression, what tool could we use to look for an association between rs4755401 and CHD?

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: No	Outcome: Yes
Exposure	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) = 0.464 – 0.238 =0.226

Measures of association for binary outcomes

		Outcome	
		No	Yes
Exposure	Yes	a	b
	No	c	d

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

Measures of association for binary outcomes

		Outcome	
		No	Yes
Exposure	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)) = 0.464 / 0.238 =1.95

Measures of association for binary outcomes

		Outcome	
		No	Yes
Exposure	Yes	a	b
	No	c	d

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

Measures of association for binary outcomes

- The *odds* is the ratio of the risk of having an outcome to the risk of not having the outcome
- If p is the risk of an outcome, then the odds of the outcome are p/(1-p)
- The odds ratio (OR) is the ratio of the odds of the outcome in the "exposed" to the odds of the outcome in the "unexposed":

$$OR = [p_1 /(1- p_1)]/ [p_0 /(1- p_0)] = \text{odds ratio}$$

where p_1=risk in exposed and p_0=risk in unexposed

- Like the relative risk, the odds ratio provides a measure of association in a ratio (rather than a difference)
- The odds ratio is the ratio of two ratios (i.e. the ratio of odds)
- The OR approximates RR for rare events
- The OR is more complicated to interpret than the RR (except for rare events), but there are some study designs (namely, case-control studies) where it is not possible to directly estimate the risk ratio, but one can always estimate the odds ratio

Measures of association for binary outcomes

- Say the chance of "disease" (D) if you're "exposed" (E) = 0.25
- Then the ***odds*** of getting D (for those who are exposed) are

  0.25/0.75 = 1/3 or 1:3

- Say the chance of "disease" if you're "not exposed" =0.1
- Then the ***odds*** of getting D (for those who are not exposed) are

  0.1/0.9 = 1/9 or 1:9

- Then the disease odds ratio (ratio of the odds of disease in the exposed to the odds of disease in the unexposed) is

  (1/3)/(1/9) = 3

Q: What is the risk ratio here? 2.5

Measures of association for binary outcomes

		Outcome: No	Outcome: Yes
Exposure	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

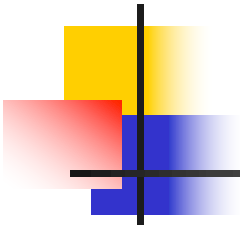
Measures of association for binary outcomes

		Outcome	
		No	Yes
Exposure	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 ≈ RR



Logistic Regression: Motivation

- The chi-squared test is adequate for investigating the association between two categorical variables
- But what if we want to investigate the association between a continuous predictor like cholesterol and a binary outcome like CHD?
- Or what if we want to adjust for potential confounders?
- 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 = binary response (0=no, 1=yes)
  X = continuous exposure
  p = E(Y | X) = P(Y = 1|X)
- One solution – fit a linear model

$$E(Y \mid X) = P(Y = 1 \mid X) = \beta_0 + \beta_1 X$$

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

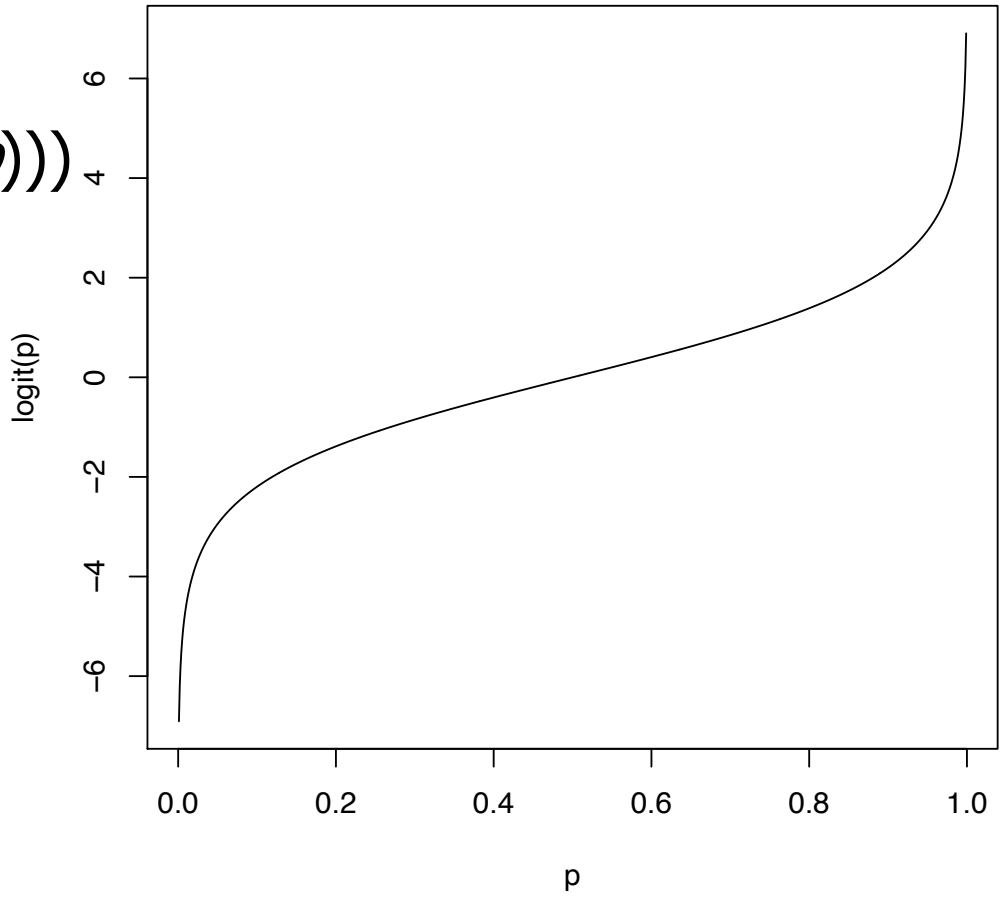
Motivating example: CHD and cholesterol

```
> 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

- Alternative: use a transformation that maps P(Y = 1|X) to the real line
- Let logit(p) = log(p/ (1 - p)))
- $p \in (0, 1)$
- $p/(1 - p) \in (0, \infty)$
- $\log(p/(1 - p)) \in (-\infty, \infty)$



Logistic regression

- logit(p) = log(p / (1 - p)))

  … this ensures that p lies between 0 and 1

- Regress logit(p) on X

$$\text{logit}[E(Y|X)] = \log[P(Y=1|X)/(1- P(Y=1|X))] = \beta_0 + \beta_1 X$$

- It turns out that the slope coefficients in logistic regression are readily interpretable: they are just log odds ratios!

Interpretation of logistic regression parameters

- On the log-odds scale

  $\log[\text{odds}(Y=1|X = (c+1))] = \beta_0 + \beta_1(c+1)$

  – $\log[\text{odds}(Y=1|X = c)] = \beta_0 + \beta_1 c$

$\log[\text{odds}(Y=1|X = (c+1))] - \log[\text{odds}(Y=1|X = c)] = \beta_1$

$\log[\text{odds}(Y=1|X = (c+1))/\text{odds}(Y=1|X = c)] = \beta_1$

$\log[OR] = \beta_1$

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

Motivating example: CHD and cholesterol

- A 1 mg/dl difference is very small, so we might be interested in estimating the OR associated with a larger difference such as 10 mg/dl
- In this case, just as in linear regression we just need to multiply our coefficient by the appropriate factor

```
> exp(10*glm.mod1$coef)
 (Intercept)          chol
6.466861e-49 1.732831e+00
```

- Comparing two people whose cholesterol levels differ by 10 mg/dl, the person with the higher cholesterol has 1.73 times higher odds of CHD compared to the person with lower 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

$$\text{logit}[E(Y|X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- $\exp(\beta_j)$ interpreted as the OR associated with a one unit change in the *j′*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
```

Motivating example

- As we have seen before, exponentiating the coefficients gives us odds ratios

```
> exp(glm.mod2$coef)
      (Intercept)                chol factor(rs4775401)1 factor(rs4775401)2
     8.937908e-06        1.057009e+00        2.212697e+00        3.121483e+00
```

- A one mg/dl increase in cholesterol is associated with 1.06 times higher odds of CHD after adjusting for genotype
- We can also obtain confidence intervals for the odds ratios

```
> exp(confint(glm.mod2))

                              2.5 %       97.5 %
(Intercept)           5.776075e-07 0.0001096301
chol                  1.043422e+00 1.0719733312
factor(rs4775401)1    1.336145e+00 3.6998174205
factor(rs4775401)2    1.250542e+00 7.8187264825
```

Hypothesis testing for logistic regression

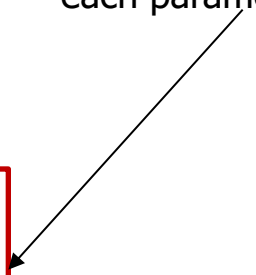
- Maximum likelihood is the standard method of estimating parameters from logistic models and is based on finding the estimates which maximize the joint probability for the observed data under the chosen model.
- The Wald test uses maximum likelihood estimates (MLE) and their standard errors to conduct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

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

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

  Model 2: logit $p_i = \beta_0 + \beta_1 chol_i + \beta_2 SNP_{1i} + \beta_3 SNP_{2i}$
- Example: Compare model 1 to model 2

  H_0: $\beta_2 = \beta_3 = 0$

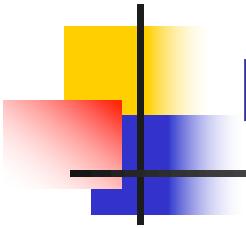
  $LRT = -2\,[L_1 - L_2] \sim \chi^2_2$

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



Logistic Regression: Assumptions

1. Logit(E[Y|x]) is related linearly to x
2. Y's are independent of each other

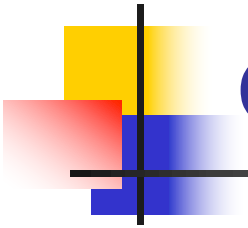
Summary

We have considered:

- Measures of association for binary outcomes
- Logistic regression
  - Interpretation
  - Estimation
  - Hypothesis testing

REGRESSION METHODS

Generalized linear models



Generalized linear models

- So far we have considered :
  - Continuous outcomes – linear regression/ANOVA
  - Binary outcomes – logistic regression

- Generalized linear models (GLMs) provide a way to model
  - Continuous and binary outcomes
  - Additional types of outcome variables (e.g. counts)
  - Additional functional forms for the relationship between outcomes and predictors

Generalized Linear Models

- GLMs allow us to estimate regression models for outcomes arising from *exponential family distributions.* This family includes many familiar distributions including Normal, Binomial and Poisson.

- A GLM is specified based on three components:
  - Outcome distribution
  - Linear predictor
  - Link function

- We will see that linear and logistic regression are both GLMs with specific choice of outcome and link function!

Outcome distribution

- The first step in fitting a GLM is to choose an appropriate distribution for your outcome

- Examples
  - Continuous outcome – Normal
  - Binary outcome – Binomial
  - Count outcome – Poisson

Linear predictor

- After specifying a distribution for the outcome, we specify the linear predictor,

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- This is just the systematic piece of our regression model

- As in other regression models we have seen, we need to identify the set of covariates to be included

Link function

- Finally, we specify a link function, g[E(Y)]:

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- This describes the functional form of the relationship between E(Y) and the linear predictor
- In linear regression, we use the *identity link function* g[E(Y)] = E(Y)
- In logistic regression, we use the *logit link function* g[(E(Y)] = log[E(Y)/(1-E(Y))]

Generalized linear models

A few example GLMS:

Distribution	Link function		Model
Normal	Identity	g[E(Y)]=E(Y)	Linear regression
Binomial	Logit	g[E(Y)]= log[E(Y)/(1-E(Y))]	Logistic regression
Poisson	Log	g[E(Y)]=log[E(Y)]	Poisson GLM
Gamma	Log	g[E(Y)]=log[E(Y)]	Gamma GLM

Alternatives to logistic regression

- Odds ratio is limited by difficulty of interpretation
- Relative risk is more interpretable
- To estimate a relative risk using regression we can use the log linear model:

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

- This is sometimes referred to as "relative risk regression"
- $\exp(\beta_1)$ is the relative risk associated with a one-unit increase in x

Modified Poisson regression

- To estimate the relative risk, we could use a binomial GLM with log link.
  - It turns out that estimation for this model is very challenging and results are sensitive to outliers in X
- An alternative approach that performs better in practice is *modified Poisson regression*
- This method uses a Poisson GLM with log link
- Using a Poisson model for binary data will give incorrect standard errors because the variance for binary outcomes differs from the variance for Poisson outcomes
- We can combine the Poisson GLM with a robust variance estimator to account for this violation of the model's assumptions

Modified Poisson regression

```
> glm.rr <- glm(chd ~ chol+factor(rs4775401), family = "poisson", data = cholesterol)
> coeftest(glm.rr, vcov = sandwich)

 z test of coefficients:

                     Estimate Std. Error  z value  Pr(>|z|)
(Intercept)        -7.0649420  0.5860408 -12.0554 < 2.2e-16 ***
chol                0.0296341  0.0027524  10.7668 < 2.2e-16 ***
factor(rs4775401)1  0.4151094  0.1444449   2.8738  0.004055 **
factor(rs4775401)2  0.6384162  0.2000234   3.1917  0.001414 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modified Poisson regression

- Relative risk of CHD associated with 1 mg/dl increase in cholesterol is 1.03.

```
> exp(glm.rr$coef)
      (Intercept)                chol factor(rs4775401)1 factor(rs4775401)2
     0.0008545444        1.0300775972        1.5145364657        1.8934796543
```

- Compare this to the odds ratio we obtained earlier using logistic regression

```
> exp(glm.mod2$coef)
      (Intercept)                chol factor(rs4775401)1 factor(rs4775401)2
    8.937908e-06        1.057009e+00        2.212697e+00        3.121483e+00
```

Relative risk regression: Assumptions

1. log(E[Y|x]) = log(P(Y=1 |x) is related linearly to x
   Warning: this can lead to predicted probabilities > 1

2. Y's are independent of each other

Risk difference regression

- Recall, we also considered fitting a linear model to binary outcome data
- This allows us to estimate differences in risk associated with a 1 unit difference in the predictor
- By using robust standard errors, we can account for violation of the assumptions of normality and equal variance

```
> glm.rd <- glm(chd ~ chol+factor(rs4775401), family = "gaussian", data = cholesterol)
> coeftest(glm.rd, vcov = sandwich)

 z test of coefficients:

                      Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)        -1.48541657  0.13724141 -10.8234 < 2.2e-16 ***
chol                0.00939240  0.00076156  12.3331 < 2.2e-16 ***
factor(rs4775401)1  0.14274314  0.04231723   3.3732 0.0007431 ***
factor(rs4775401)2  0.21210838  0.08223706   2.5792 0.0099020 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Risk difference regression

- A 1 mg/dl difference is very small, so we might be interested in estimating the RD associated with a larger difference such as 10 mg/dl

- Comparing two people with the same rs4775401 genotype whose cholesterol levels differ by 10 mg/dl, the risk of CHD for the person with the higher cholesterol is 9.4% higher (in absolute terms) compared to the person with lower cholesterol

- Comparing two people with the same cholesterol level, a person with rs4775401 C/T is estimated to have risk of CHD 14.3% higher (in absolute terms) than a person with rs4775401 C/C

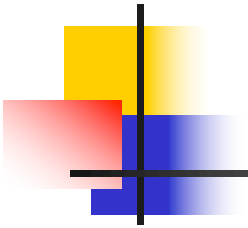
- Comparing two people with the same cholesterol level, a person with rs4775401 T/T is estimated to have risk of CHD 21.2% higher (in absolute terms) than a person with rs4775401 C/C

Risk difference regression: Assumptions

1. $E[Y|x] = P(Y=1|x)$ is related linearly to x

Warning: this can lead to predicted probabilities > 1 or < 0

2. Y's are independent of each other



Summary

We have considered:

- Logistic regression
  - Interpretation
  - Estimation
- Generalized linear models
  - Relative risk regression
  - Risk difference regression

Exercise

- Work on Exercise 13-17
  - Try each exercise on your own
  - Make note of any questions or difficulties you have
  - At **10:15PT** we will meet as a group to go over the solutions and discuss your questions

Module summary

- In this module we have covered a variety of regression methods that can be used to analyze continuous and binary outcomes:

- Continuous outcomes
  - Simple linear regression
  - Multiple linear regression
  - ANOVA

- Binary outcomes
  - Logistic regression
  - Relative risk regression
  - Risk difference regression

- These methods are foundational for many statistical analyses, and we hope you will be able to apply them to your future research!

Everything is regression!

(Professor Scott Emerson)