Biost 517 / Biost 514
Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 19:

Regression with Binary Outcomes: Introduction to Logistic Regression

Binary Random Variables

- Many variables of interest can take on only two values.
- Clinical and epidemiological studies often generate outcomes that are dichotomous
  - Presence/absence of a condition or characteristic at a particular time
  - Indication of whether a response occurred within a defined period of observations
  - For convenience, often coded as 0 or 1 “indicator” variable
    - Vital Status: “Dead” coded 0= alive 1= dead
    - Prostate Cancer: “In Remission” coded 0=no 1= yes
    - Sex: “Female” coded 0= male 1= female
    - Intervention: “Treatment x” coded 0= control 1= new therapy

Binary Random Variables

- Sometimes continuous variables are dichotomized
  - For scientific reasons (statistically less precise)
    - Systolic Blood pressure: greater than 160 mm Hg (Hypertension)
    - Prostate Specific Antigen (PSA) : normal is less than 4 ng/ml
    - Serum glucose: normal range is less than 120 mg/dl (important for diabetics)

Bernoulli Probability Distribution

- A binary variable Y_i must have a Bernoulli probability distribution
  - A single parameter: $p = Pr(Y_i = 1)$ with $0 \leq p \leq 1$.
  - We write $Y \sim B(1, p)$ and the prrobability mass function is

$$P(Y_i = y) = p^y (1-p)^{1-y}$$
where y can be either 0 or 1

- Mean: $E[Y_i] = Pr(Y_i = 1) = p$
- Variance: $Var(Y_i) = p(1-p)$
  - A "mean – variance" relationship
    - If the mean is different between two groups, the variance must also be different. Maximum variance of 0.25 when $p = 0.5$
- The sum of n independent Bernoulli random variables has a binomial distribution: $S_n = Y_1 + \ldots + Y_n \sim B(n, p)$
  - Mean: $E[S_n] = np$
  - Variance: $Var(S_n) = np(1-p)$

Statistical Hypotheses

- Summary measures of interest for a Bernoulli random variable are pretty much limited to either
  - The proportion *p* (a mean), or
  - The odds *o* = *p / (1-p)*

- Contrasts used to compare the distribution of a Bernoulli random variable across subpopulations thus include
  - Difference in proportions: $p_1 - p_0$ (risk difference (RD))
  - Ratio of proportions : p_1 / p_0 (risk ratio (RR))
  - Odds ratio : $o_1 / o_0 = \dfrac{p_1 / (1 - p_1)}{p_0 / (1 - p_0)}$ (odds ratio (OR))

Linear Regression with Binary Response

- Conceptually, there should be no problem modeling the proportion (which is the mean of the distribution)
- Consider the linear regression model for an outcome Y and a predictor X. If Y is a binary response, then

$$E[Y|X=x]=Pr(Y=1|X=x)=\beta_0+\beta_1 x$$

- This model provides an estimate of the proportion (mean) of the subpopulation with a predictor value of $X = x$ that have the binary outcome $(Y = 1)$.
- β_1 is the "**Risk Difference (RD)**" or difference in proportions between two subpopulations who differ in the predictor by 1 unit:

$$Pr(Y=1|X=x+1)-Pr(Y=1|X=x)=\beta_1$$

Limitations with using Linear Regression with Binary Response

- Linear regression models are not widely used for binary response data due to some important limitations:
- Classical linear requires equal variances in each predictor group in order for CI and p values to be valid
- But with binary *Y*, the variance within a group depends on the mean
  - Mean: $E[Y] = p$ Variance: $Var(Y) = p(1-p)$
  - In the presence of an association between response and POI, we will definitely have heteroscedasticity

- When using the Huber-White sandwich estimate of robust standard errors, this problem is not as big of a limitation
  - However, moderate sample sizes are still needed

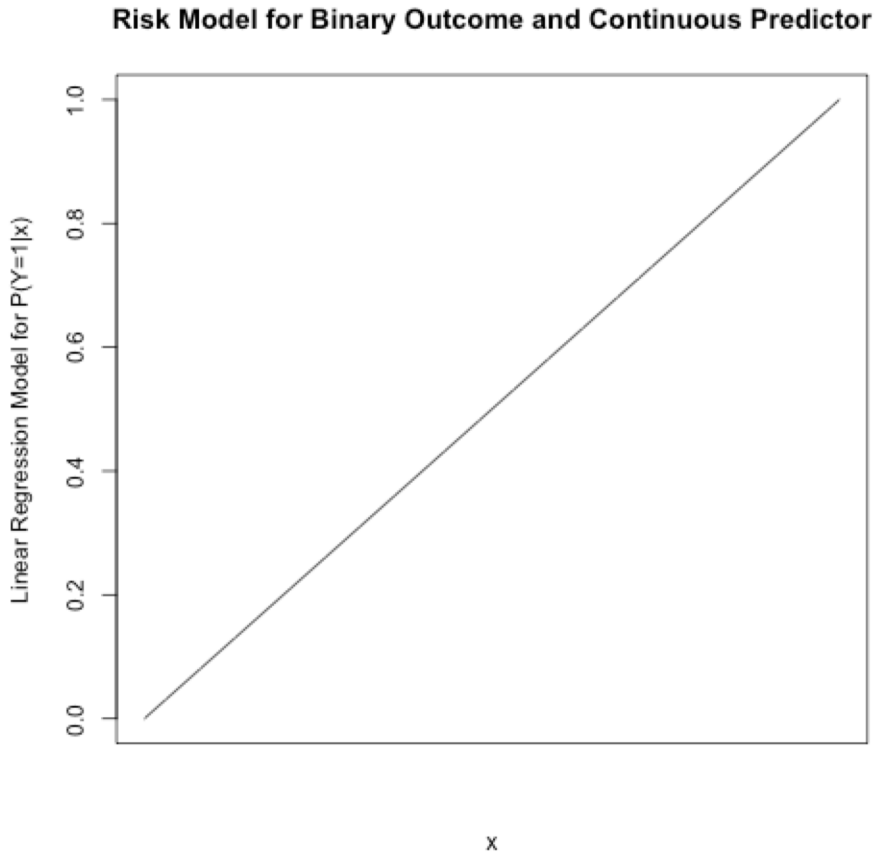
Limitations with using Linear Regression with Binary Response

- The outcome is a probability of risk. Any reasonable estimates of the regression coefficients must constrain the estimated probability to lie between 0 and 1. This imposes constraints on the predictors and the regression coefficient parameters, which can cause numerical problems:

$$-\frac{\beta_0}{\beta_1} \leq X \leq \frac{1-\beta_0}{\beta_1}$$

Limitations with using Linear Regression with Binary Response

– Often implausible that the outcome risk would change in a strictly linear fashion for the entire range of possible values of a continuous predictor X



Log-Linear Regression with Binary Response

- Now consider the log-linear regression model for an outcome Y and a predictor X.

$$\log(\mathrm{E}[\mathrm{Y}|\mathrm{X}=\mathrm{x}])=\log\big(\Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x})\big)=\beta_0+\beta_1 x$$

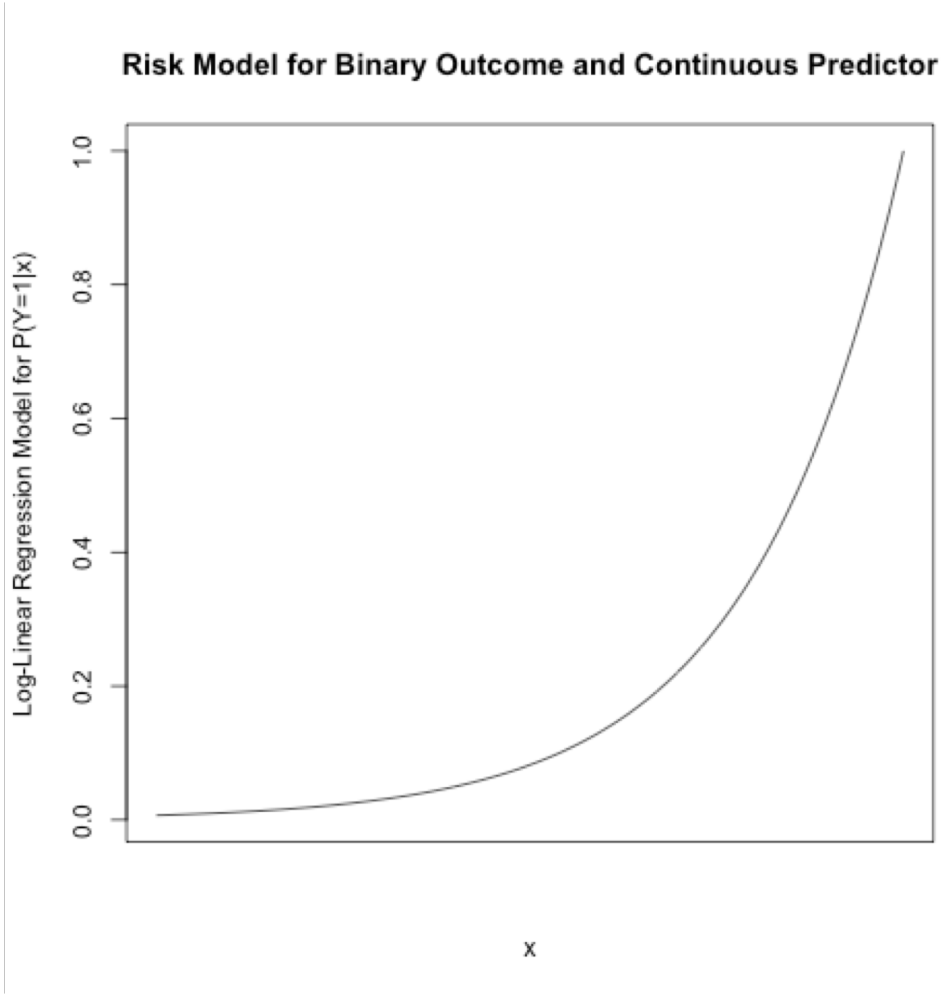
- With this model, β_1 is the logarithm of the relative risk associated with a unit increase in the predictor X, and $\exp(\beta_1)$ is often interpreted as the "**Risk Ratio (RR)**" between two subpopulations who differ in the predictor by 1 unit:

$$\log\big(\Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x}+1)\big)-\log\big(\Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x})\big)$$

$$=\log\left(\frac{\Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x}+1)}{\Pr(\mathrm{Y}=1|\mathrm{X}=\mathrm{x})}\right)=\beta_1$$

Log-Linear Regression with Binary Response Data

- The Log-linear regression model constrains risk to increase exponentially with the predictor X



Log-Linear Regression with Binary Response

- The log-linear regression model for binary data can be fit using many standard software packages. Sometimes referred to as the log binomial model
- However, numerical difficulties can arise due to constraints similar to the linear regression model for binary response data, since the probability of the response must be between 0 and 1.
- In cases when the log linear model fails to converge and relative risk estimates are desired, a Poisson regression model is recommended, where the observed binary responses are treated as if they were distributed according to the Poisson distribution.
- Poisson regression can be used instead (but not covered in this class)

Simple Logistic Regression with Binary Response

- Logistic Regression is by far the most widely used model for binary outcomes in clinical and epidemiological applications.
- Does not have numerical difficulties and does not put constraints on the values of the predictor or the parameters.
- The risk model has a logistic function

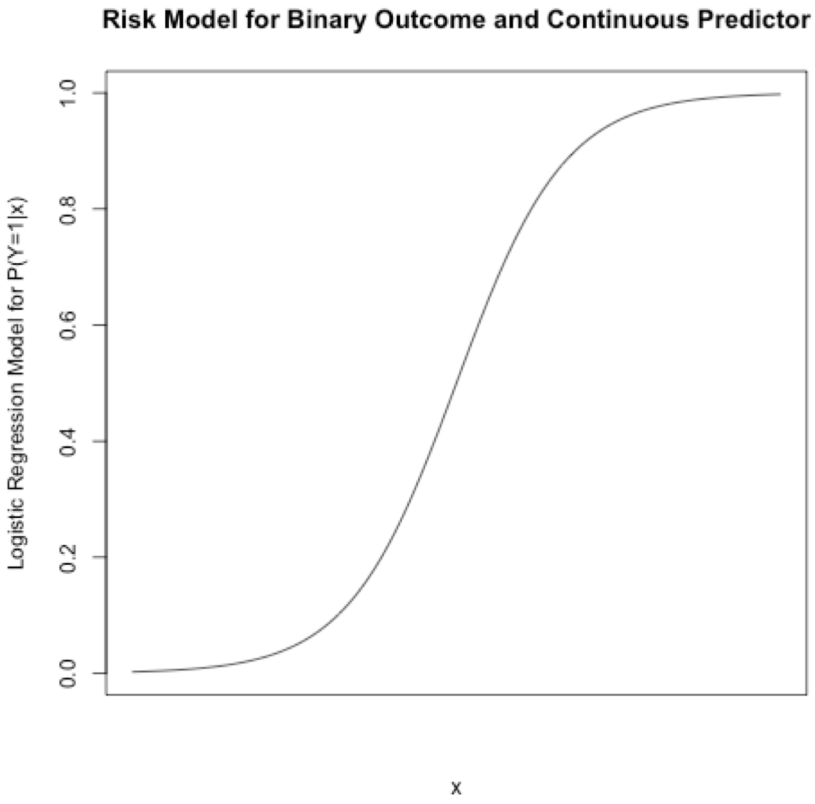
$$\Pr(Y=1|X=x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

- Using the logit link function, the model is linear in the predictor

$$\text{logit}\left(\Pr(Y=1|X=x)\right)=\log\left(\frac{\Pr(Y=1|X=x)}{1-\Pr(Y=1|X=x)}\right)=\beta_0+\beta_1 x$$

Simple Logistic Regression with Binary Response

- The risk model for logistic regression allows for a smooth change in risk throughout the range of the predictor X
- Risk increases slowly up to a "threshold" range of X followed by a more rapid increase and a subsequent leveling off of risk
- This model is consistent for many does-response relationships



Simple Logistic Regression

- Binary response variable follows a Bernoulli (or binomial distribution)
- The mean $E[Y \mid X = x] = \Pr(Y = 1 \mid X = x)$ is given by the logistic function
- Outcome values are statistically independent
- Allows continuous (or multiple) grouping variables
  - But is fine with binary grouping variable also
- Compares odds of response across groups
  - "Odds ratio"

Simple Logistic Regression

- Modeling odds of binary response Y on predictor X

$$\text{Distribution} \qquad \Pr(Y_i = 1) = p_i$$

$$\text{Model} \qquad \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \times X_i$$

$$X_i = 0 \qquad \log \text{odds} = \beta_0$$

$$X_i = x \qquad \log \text{odds} = \beta_0 + \beta_1 \times x$$

$$X_i = x + 1 \qquad \log \text{odds} = \beta_0 + \beta_1 \times x + \beta_1$$

Interpretation as Odds

- Exponentiation of regression parameters

$$\text{Distribution} \qquad \Pr(Y_i = 1) = p_i$$

$$\text{Model} \qquad \left(\frac{p_i}{1-p_i}\right) = e^{\beta_0} \times e^{\beta_1 \times X_i}$$

$$X_i = 0 \qquad \text{odds} = e^{\beta_0}$$

$$X_i = x \qquad \text{odds} = e^{\beta_0} \times e^{\beta_1 \times x}$$

$$X_i = x + 1 \qquad \text{odds} = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$$

Estimating Proportions

- Proportion = odds / (1 + odds)

Distribution $\Pr(Y_i = 1) = p_i$

Model $$p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times X_i}}{1 + e^{\beta_0} \times e^{\beta_1 \times X_i}}$$

$X_i = 0$ $$p_i = e^{\beta_0} / \left(1 + e^{\beta_0}\right)$$

$X_i = x$ $$p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x}}{1 + e^{\beta_0} \times e^{\beta_1 \times x}}$$

$X_i = x + 1$ $$p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}{1 + e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}$$

Parameter Interpretation

- Interpretation of the logistic regression parameters based on odds

- Odds when predictor is 0
  - Found by exponentiation of the intercept from the logistic regression: $\exp(\beta_0)$

- Odds ratio between groups differing in the value of the predictor by 1 unit
  - Found by exponentiation of the slope from the logistic regression: $\exp(\beta_1)$

Similarity to Other Regressions

- Logistic regression uses *maximum likelihood estimation* to find parameter estimates
- If a saturated model is fit, the estimated odds of event in each group will agree exactly with the sample odds
- In large samples, the regression parameter estimates are approximately normally distributed
  - P values and CI that are displayed for each parameter estimate are Wald- based estimates

$$95\%\,\text{CI}: \ (estimate) \pm (crit\ value) \times (std\ err) \qquad \hat{\beta} \ \pm \ z_{1-\alpha/2} \times s\hat{e}\left(\hat{\beta}\right)$$

$$\text{Test stat}: \qquad Z = \frac{(estimate) - (null)}{(std\ err)} \qquad Z = \frac{\hat{\beta} - \beta_0}{s\hat{e}\left(\hat{\beta}\right)}$$

Technical Details

- Unlike linear regression, there is no closed form expression to find the logistic regression parameter estimates
- Instead, computer programs use an iterative search
- This search may fail in saturated or nearly saturated models if some parameter corresponds to a group having all events or no events
  - In this setting, logistic regression parameters modeling the log odds are trying to estimate positive or negative infinity
  - The sample size is too small for the model
- There is little or no advantage in using “robust SE”
  - If linear model holds, the mean-variance is handled correctly
  - If a nonlinear association is true, the “model misspecification” will lead to incorrect variance estimates, but the robust SE produce very similar results to classical logistic regression

Logistic Regression in R

- In R, logistic regression can be preformed in R with the “glm” function
  - glm is ”generalized linear model”
- Commands is similar to ”lm” used for linear regression:

```
mylogit=glm(respvar~predvar,family=binomial,data=mydata)
```

Where respvar is an indicator variable for a dichotomous outcome (0 or 1)

- The default link function in R for the binomial family is “logit”
  - Provides regression parameter estimates and inference on the log odds scale
    - Intercept, slope with SE, P values

Other link functions for binomial data in R

- Different link functions are available, including the “log” link function for a log-linear regression model, or log binomial model, for inference on risk ratios.
- Use R help command for more information: `help(glm)`

```
mylogbinom=glm(respvar~predvar,family=binomial
(link="log"),data=mydata)
```

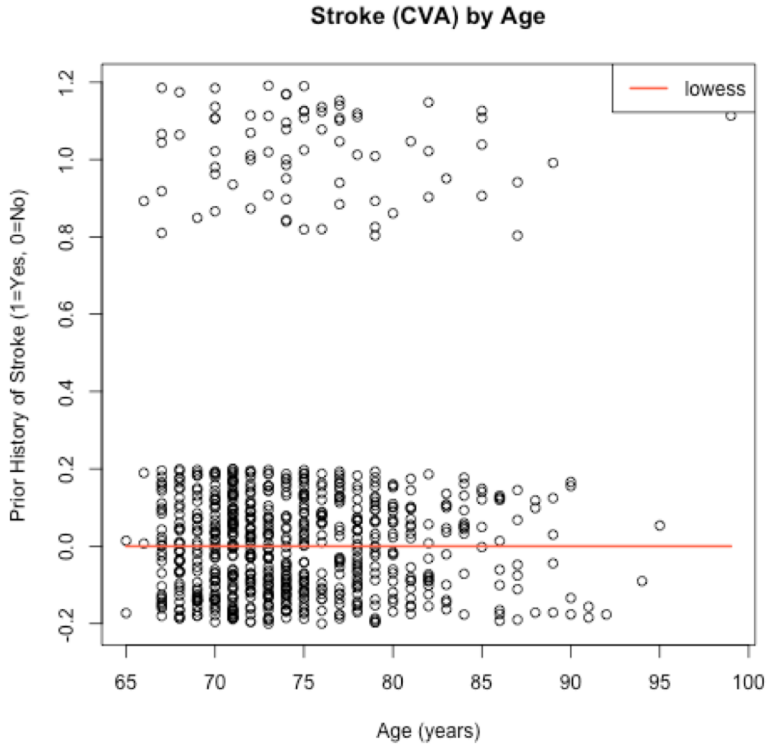
- Be careful with convergence problems when using the a log binomial model
- The default “logit” link is most widely used for inference on odds ratios.

Example: Stroke (cerebrovascular accident) and Age….

- Interested in the prevalence of stroke (cerebrovascular accident-CVA) in a subset of elderly individuals aged 65 and older who had MRI from the Cardiovascular Health Study.
- Scientific question: Is there any association between risk of stroke and age in this population of elderly individuals?
- Response variable is CVA
  - Binary variable: 0= no history of prior stroke, 1= prior history of stroke
- Predictor variable is Age
  - Continuous predictor

Example: Stroke (cerebrovascular accident) versus Age

- Scatterplot (even with superimposed smooth) is not very informative with a binary response. A "jitter" was added to the points in the plot.



- (Note that we are estimating proportions– not odds– with this plot, so we can not even judge appropriateness of linearity for logistic regression)

Example: Regression Model

- Answer scientific question of interest by assessing linear trends in log odds of stroke by age
- Estimate best fitting line for the log odds of CVA within age groups

$$\log odds\left(CVA \mid Age\right) = \beta_0 + \beta_1 \times Age$$

- An association will exist if the slope (β_1) is nonzero
  - In that case, the odds (and probability) of CVA will be different across different age groups

Logistic Regression for CVA on age: R

- Logistic regression model results with R: CVA is the binary response and Age is the predictor

```
> logistmod1=glm(cva~age,family=binomial,data=mridata)
> summary(logistmod1)

Call:
glm(formula = cva ~ age, family = binomial, data = mridata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6339  -0.4791  -0.4496  -0.4285   2.2613

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.69120    1.59128  -2.948   0.0032 **
age          0.03356    0.02104   1.595   0.1107
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of R Output

- Regression model for CVA on age
- Intercept:
  - Estimated intercept: -4.69
- Slope is labeled by variable name: “age”
  - Estimated slope: 0.0336
- Estimated linear relationship:
  - log odds CVA by age group given by

$$\log \text{odds}\, CVA = -4.69 + 0.0336 \times Age_i$$

Interpretation of Intercept

$$\log \text{odds}\, CVA = -4.69 + 0.0336 \times Age_i$$

- Estimated log odds CVA for newborns is -4.69
  - Odds of CVA for newborns is $e^{-4.69} = 0.0092$
  - Probability of CVA for newborns
    - Use prob = odds / (1+odds): .0092 / (1+.0092)= .0091

- Pretty ridiculous to try to estimate
  - We never sampled anyone less than 67
  - In this problem, the intercept is just a necessary parameter for fitting the regression model, but no meaningful scientific interpretation

Interpretation of Slope

$$\log \text{odds}\, CVA = -4.69 + 0.0336 \times Age_i$$

- Estimated difference in log odds of CVA for two groups differing by one year in age is 0.0336, with older group tending to have higher log odds
  - Odds Ratio: $e^{0.0336} = 1.034$
  - For 5 year age difference: $e^{5 \times 0.0336} = 1.034^5 = 1.183$

- (If a straight line relationship is not true, we interpret the slope as an average difference in log odds CVA per one year difference in age)

Confidence Intervals for Logistic Regression Parameters

- Confidence intervals of logistic regression parameters
- `> confint.default(logistmod1model)`

```
                  2.5 %       97.5 %
(Intercept) -7.810044604 -1.57235545
age         -0.007682691  0.07481151
```

- Need to exponentiate the confidence interval of the age slope to obtain confidence interval of the odds ratio

```
> exp(confint.default(logistmod1))
                  2.5 %    97.5 %
(Intercept) 0.00040564 0.2075557
age         0.99234675 1.0776810
```

Logistic Regression with uwIntroStats R package

- Alternatively, can just use the R package **uwIntroStats** for logistic regression analysis:
- Package will provide inference for odds ratio
- Don’t have to exponentiate the slope estimate
- Produces point estimates, confidence interval estimates, and p values for odds ratios

Logistic Regression with uwIntroStats R package

```
> library(uwIntroStats)
> logistmod2 <- regress ("odds",cva~age,data=mridata)
> logistmod2

Call:
regress(fnctl = "odds", formula = cva ~ age, data = mridata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6339  -0.4791  -0.4496  -0.4285   2.2613

Coefficients:

Raw Model:
                 Estimate  Naive SE  Robust SE       F stat    df Pr(>F)
[1] Intercept     -4.691     1.591     1.601             8.58 1    0.0035
[2] age           0.03356   0.02104   0.02117           2.51 1    0.1134

Transformed Model:
                 e(Est)     e(95%L)    e(95%H)         F stat    df Pr(>F)
[1] Intercept    9.176e-03  3.957e-04   0.2128            8.58 1    0.0035
[2] age            1.034     0.9920     1.078             2.51 1    0.1134
```

Example: Interpretation

- “From logistic regression analysis, we estimate that for two groups that differ by one year in age, the odds of stroke is 3.4% higher in the older group, though this estimate is not statistically significant. A 95% CI suggests that this observation is not unusual if a group that is one year older might have odds of stroke that was anywhere from 0.8% lower or 7.8% higher than the younger group.” A two-sided p value of 0.113 suggests that we can not with high confidence reject the null hypothesis that the odds of stroke are not associated with age.

Comments on Interpretation

- I express this as a difference between group odds rather than a change with aging
  - We did not do a longitudinal study

- To the extent that the true group log odds have a linear relationship, this interpretation applies exactly

- If the true relationship is nonlinear
  - The slope estimates the “first order trend” for the sampled age distribution
  - We should not regard the estimates of individual group probabilities / odds as accurate

Logistic Regression and χ^2 Test

- Logistic regression with a binary predictor (two groups) corresponds to familiar chi squared test

- Three possible statistics from logistic regression
  - Wald: The test based on the estimate and SE
  - Score: Corresponds to chi squared test
  - Likelihood ratio test

Example: Stroke (cerebrovascular accident) and Smoking

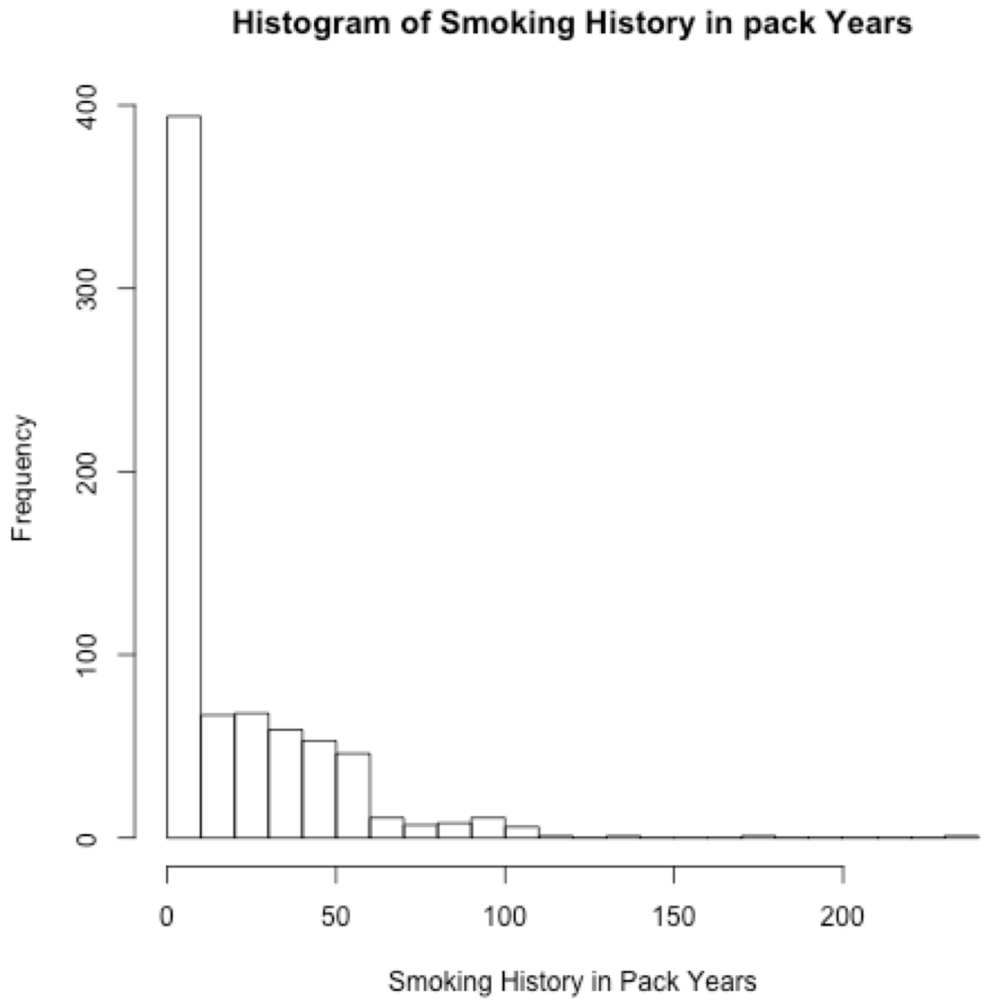
- Prevalence of stroke (cerebrovascular accident- CVA) in a subset of elderly individuals aged 65 and older who had MRI from the Cardiovascular Health Study.

- Response variable is CVA
  - Binary variable: 0= no history of prior stroke, 1= prior history of stroke

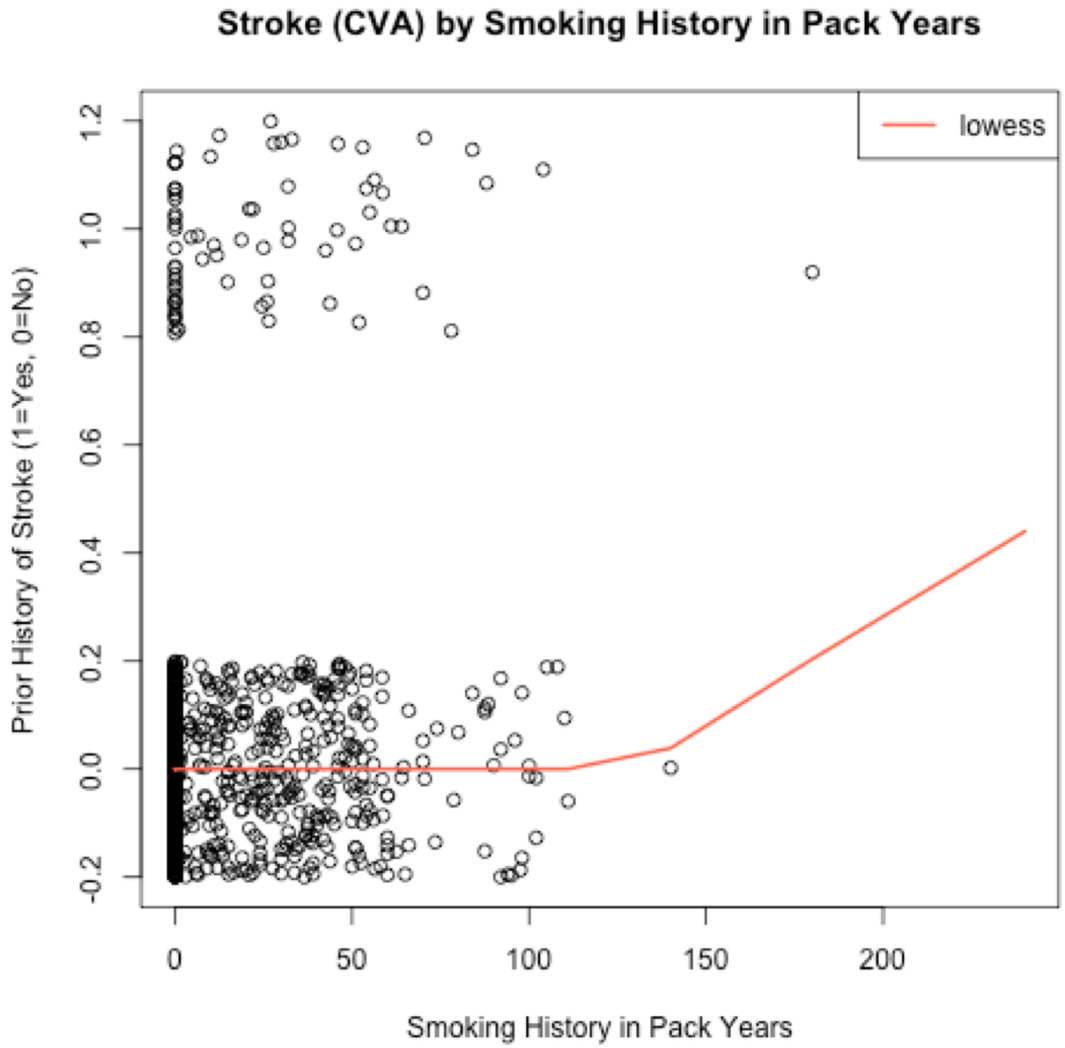
- Predictor variable is participant smoking history in pack years
  - 1 pack year = smoking 1 pack of cigarettes per day for 1 year.
  - A participant who never smoked has 0 pack years.
  - Continuous predictor

Example: Interpretation

- Smoking History in Pack Years



Example: Stroke (cerebrovascular accident) versus Smoking History



Logistic Regression for CVA on Pack Years

- CVA is the binary response; Smoking in Pack Years is the predictor

```
> library(uwIntroStats)
> logistmod2 <- regress ("odds",cva~packyrs,data=mridata)
> logistmod2
( 1  cases deleted due to missing values)


Call:
regress(fnctl = "odds", formula = cva ~ packyrs, data = mridata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7998  -0.4768  -0.4118  -0.4118   2.2403

Coefficients:

Raw Model:
                 Estimate  Naive SE   Robust SE        F stat    df Pr(>F)
[1] Intercept     -2.425     0.1576     0.1603           228.80 1  < 0.00005
[2] packyrs       0.01035   3.752e-03  3.696e-03          7.84 1    0.0052

Transformed Model:
                 e(Est)    e(95%L)   e(95%H)         F stat    df Pr(>F)
[1] Intercept     0.08850   0.06460    0.1212           228.80 1  < 0.00005
[2] packyrs        1.010     1.003     1.018             7.84 1    0.0052
```

Interpretation of Output

- Logistic Regression model for CVA on smoking history in pack years
- Intercept:
  - Estimated intercept: -2.425
- Slope is labeled by variable name: “packyrs”
  - Estimated slope: 0.0104
- Estimated linear relationship:
  - log odds CVA by pack years group given by

$$\text{Log Odds } CVA_i = -2.425 + 0.0104 \times packyears_i$$

Interpretation of Intercept

$$\text{Log Odds } CVA_i = -2.425 + 0.0104 \times packyears_i$$

- Estimated log odds CVA for individuals who never smoked is: -2.43
  - Odds of CVA for individuals who never smoked is $e^{-2.43} = 0.088$
  - Probability of CVA for individuals who never smoked
    - Use prob = odds / (1+odds): 0.088 / (1+0.088)= .081

- For this predictor, the intercept is very much of interest

Interpretation of Slope

$$\text{Log Odds } CVA_i = -2.425 + 0.0104 \times packyears_i$$

- Estimated difference in log odds CVA for two groups differing by one pack year is 0.0104, with older group tending to have higher log odds
  - Odds Ratio: $e^{0.0104} = 1.0105$
  - For 5 year age difference: $e^{5 \times 0.0104} = 1.0105^5 = 1.053$

- (If a straight line relationship is not true, we interpret the slope as an average difference in log odds CVA per one year difference in age)

Inference on Odds Ratios with uwIntroStats R package

```
> library(uwIntroStats)
> logistmod2 <- regress ("odds",cva~packyrs,data=mridata)
> logistmod2
( 1  cases deleted due to missing values)


Call:
regress(fnctl = "odds", formula = cva ~ packyrs, data = mridata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7998  -0.4768  -0.4118  -0.4118   2.2403

Coefficients:

Raw Model:
                 Estimate  Naive SE   Robust SE        F stat    df Pr(>F)
[1] Intercept     -2.425     0.1576     0.1603           228.80 1  < 0.00005
[2] packyrs      0.01035   3.752e-03  3.696e-03           7.84 1    0.0052

Transformed Model:
                 e(Est)    e(95%L)   e(95%H)         F stat    df Pr(>F)
[1] Intercept     0.08850   0.06460    0.1212           228.80 1  < 0.00005
[2] packyrs        1.010     1.003     1.018             7.84 1    0.0052
```

Example: Interpretation

- “From the logistic regression analysis, we estimate that for each one year difference in pack years of smoking, the odds of stroke is 1.0% higher in the group with more pack years of smoking. A 95% CI suggests that this observation is not unusual if the true odds of stroke for the group that has more pack years of smoking is anywhere between 0.3% and 1.8% higher for each one year difference in pack years of smoking. Because the two sided P value is P=.005, we reject the null hypothesis that the odds of stroke is not associated with pack years of smoking.

Logistic Regression with Log Transformed Predictor

- Log transformations of predictors are often used for predictors that are skewed (or that have outliers) and for improved linearity between log odds and predictor in a logistic regression analysis. As before, "log" is referring to the natural log (ln)

$$\text{Model} \qquad \log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 \times \log(X_i)$$

$$X_i = 1 \qquad \text{odds} = \mathrm{e}^{\beta_0}$$

$$X_i = x \qquad \text{odds} = \mathrm{e}^{\beta_0} \times e^{\beta_1 \times \log(x)} = \mathrm{e}^{\beta_0}\, x^{\beta_1}$$

$$X_i = kx \qquad \text{odds} = \mathrm{e}^{\beta_0} \times e^{\beta_1 \times \log(kx)} = \mathrm{e}^{\beta_0} \times e^{\beta_1 \times \log(x) + \beta_1 \times \log(k)} = \mathrm{e}^{\beta_0}\, x^{\beta_1} k^{\beta_1}$$

- If $k = 1.01$, then $1.01x$ corresponds to a 1% increase in x.
- If $k = 1.1$, then $1.1x$ corresponds to a 10% increase in x.
- If $k = 2$, then $2x$ corresponds to a 100% increase in x.

Logistic Regression with Log Transformed Predictor

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 \times \log(X_i)$$

- Intercept $\exp(\beta_0)$ is the odds for subpopulation with $X = 1$. May not be of scientific interest. Maybe outside the range of data.
- With a logistic regression model with a (natural) log transformed predictor what is the odds ratio between two populations that differ in the predictor by 1% and 10%, respectively?
  - $k = 1.01$, so the odds ratio for two subpopulation who differ in the predictor by 1% is 1.01^{β_1}
  - $k = 1.1$, so the odds ratio for two subpopulations who differ in the predictor by 10% is 1.1^{β_1}