

Lesson 6

Logistic Regression

BIS 505b

Yale University
Department of Biostatistics

Pagano Chapter 20

Date Modified: 03/28/2021

Goals for this Lesson

Addressing a Research Question

- ① How to interpret an **odds ratio** as a measure of association
- ② How to investigate the **association of risk factors** (continuous or categorical) with a **dichotomous outcome**
- ③ Calculate and interpret **predicted probabilities**
- ④ Assess model **goodness-of-fit**

Contents

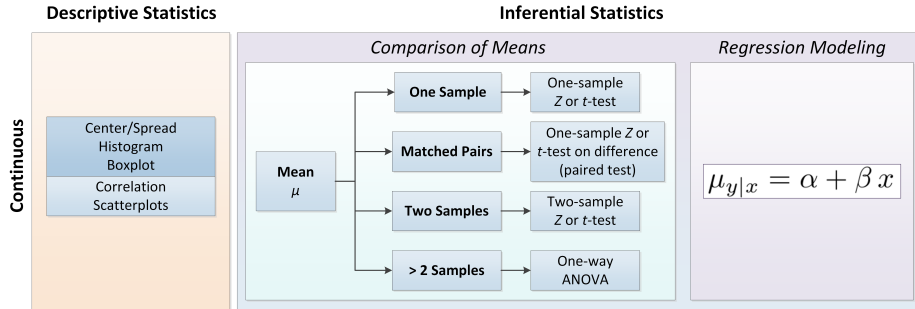
- 1 Binary Outcome
 - Estimating p
- 2 Simple Logistic Regression
 - Motivation
 - Interpretation of the Model
 - Fitting the Model
 - Inference
- 3 Multiple Logistic Regression
 - Continuous and Binary Regressors
 - Categorical Regressors
 - Interaction Terms
 - Diagnostics

Progress this Unit

- 1 Binary Outcome
 - Estimating p
- 2 Simple Logistic Regression
 - Motivation
 - Interpretation of the Model
 - Fitting the Model
 - Inference
- 3 Multiple Logistic Regression
 - Continuous and Binary Regressors
 - Categorical Regressors
 - Interaction Terms
 - Diagnostics

Lessons 3 and 4: Linear Regression

- When Y is **continuous**, interested in estimating the **mean** of Y , μ



- Linear Regression:** Identify explanatory variables that help predict the mean of the Y by explaining the observed variation in the outcome

Binary Outcome

- Suppose the response is **binary** or **dichotomous** (i.e., can take two possible levels)

$$Y_i = \begin{cases} 1 & \text{Yes, Success, Event, Disease} \\ 0 & \text{No, Failure, No Event, No Disease} \end{cases}$$

- Each individual is either a “success” or a “failure”
 - **1** represents a success (the outcome most interested in)
 - **0** represents a failure



Binary Outcome

- $Y \sim B(p)$ is a **Bernoulli** random variable with probability of success, p
- p is the proportion of times the random variable takes value 1; mean of Bernoulli random variable

$$Y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

Y Variable	Mean
Continuous	μ
Dichotomous	p

Sample Estimate, \hat{p}

- p : True population proportion; fraction of the *population* experiencing the event of interest. Probability of success in any Bernoulli trial.
- \hat{p} : Sample proportion; sample mean of 0/1 variable; fraction of the *sample* experiencing the event of interest

Population Parameter	Point Estimate*
p	$\hat{p} = \frac{\sum_{i=1}^n y_i}{n}$

*Based on sample of size n

- **Note:** $\sum_{i=1}^n y_i \sim \text{Bin}(n, p)$, total number of successes out of n trials
- The **proportion** serves as the building block for other useful numerical measures

Odds of Success

- **Odds:** Ratio of the probability that the event of interest *will* happen to the probability that the event *will not* happen

Odds

$$\text{Odds} = \frac{p}{1 - p}$$

- $0 \leq p \leq 1$, giving odds ≥ 0
- Odds > 1 : When $p > 1 - p$; $p > 0.5$, chance of success is greater than chance of failure
- Odds < 1 : When $1 - p > p$; $p < 0.5$, chance of failure is greater than chance of success

Estimating the Odds

Estimate of Odds

$$\widehat{\text{Odds}} = \frac{\hat{p}}{1 - \hat{p}}$$

\hat{p}	$1 - \hat{p}$	$\widehat{\text{Odds}}$
0	1	0/1 = 0
0.1	0.9	1/9 = 0.11
0.2	0.8	1/4 = 0.25
0.3	0.7	3/7 = 0.43
0.4	0.6	2/3 = 0.67
0.5	0.5	1/1 = 1
0.6	0.4	3/2 = 1.5
0.7	0.3	7/3 = 2.33
0.8	0.2	4/1 = 4
0.9	0.1	9/1 = 9



Odds Ratio

- **Odds ratio:** Ratio of the odds of success (e.g., disease) of *one group* (exposed) to the odds of success in *another group* (unexposed); **relative odds**

Odds Ratio

$$\widehat{OR} = \frac{\text{Odds of disease in exposed}}{\text{Odds of disease in unexposed}} = \frac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_0}{1 - \hat{p}_0}}$$



Odds Ratio

- In a 2×2 setting (binary exposure variable and binary response variable), the estimated **odds ratio** is:

Exposure	Disease		n_j	\hat{p}_j	$1 - \hat{p}_j$
	Yes	No			
Present	a	b	$a + b$	$\hat{p}_1 = \frac{a}{a + b}$	$1 - \hat{p}_1 = \frac{b}{a + b}$
Absent	c	d	$c + d$	$\hat{p}_0 = \frac{c}{c + d}$	$1 - \hat{p}_0 = \frac{d}{c + d}$

$$\widehat{OR} = \frac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_0}{1 - \hat{p}_0}} = \frac{\frac{\frac{a}{a+b}}{\frac{b}{a+b}}}{\frac{\frac{c}{c+d}}{\frac{d}{c+d}}} = \frac{ad}{bc}$$

Progress this Unit

- 1 Binary Outcome
 - Estimating p
- 2 Simple Logistic Regression
 - Motivation
 - Interpretation of the Model
 - Fitting the Model
 - Inference
- 3 Multiple Logistic Regression
 - Continuous and Binary Regressors
 - Categorical Regressors
 - Interaction Terms
 - Diagnostics

Modeling

- In **linear regression**, utilized the relationship between X and Y to better estimate μ
 - Lessons 3 and 4 provided the foundation for investigating multiple variable relationships, although the linear regression framework is specific to a **continuous outcome**
- Would like to examine multiple variable associations when the outcome Y is **dichotomous**
 - Interested in estimating the **mean of Y** , in this case, the proportion p , and determine whether any explanatory variables help estimate its value

R Code

```
# Response (Y) of interest (0/1)
> table(fhssrs$CVD)
 0  1
77 23
```

R Code

```
# Overall phat
> prop.table(table(fhssrs$CVD))
 0  1
0.77 0.23
```

Utilizing Covariates

Table: Proportion developing CVD by baseline systolic blood pressure category

j	Systolic BP Category	CVD		n_j	\hat{p}_j
		Yes	No		
2	> 150	7	12	19	$0.368 = \frac{7}{19}$
1	$(120, 150]$	12	39	51	$0.235 = \frac{12}{51}$
0	≤ 120	4	26	30	$0.133 = \frac{4}{30}$
	Overall	23	77	100	$0.23 = \frac{23}{100}$

- Overall, 23% of individuals in our sample are diagnosed with CVD
- Suspect there are certain factors that affect the likelihood of CVD
- If can classify an individual according to these characteristics, can:
 1. Estimate his/her probability of developing CVD with greater precision than that afforded by the single value \hat{p} and
 2. Take measures to decrease this probability

Odds Ratio: Example

- Computing odds ratios for 2×2 tables

Systolic BP Category	CVD		\hat{p}_j
	Yes	No	
> 150	7	12	0.368
≤ 120	4	26	0.133

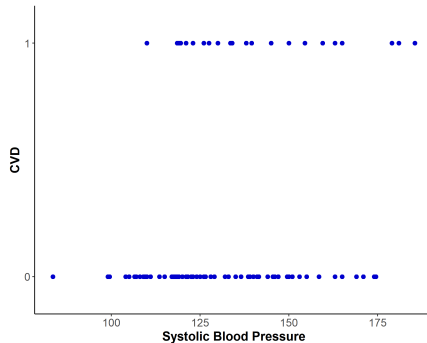
$$\widehat{OR} = \frac{\frac{\hat{p}_2}{1 - \hat{p}_2}}{\frac{\hat{p}_0}{1 - \hat{p}_0}} = \frac{\frac{0.368}{0.632}}{\frac{0.133}{0.867}} = \frac{7 \times 26}{12 \times 4} = 3.792$$

Systolic BP Category	CVD		\hat{p}_j
	Yes	No	
$(120, 150]$	12	39	0.235
≤ 120	4	26	0.133

$$\widehat{OR} = \frac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_0}{1 - \hat{p}_0}} = \frac{\frac{0.235}{0.765}}{\frac{0.133}{0.867}} = \frac{12 \times 26}{39 \times 4} = 2$$

Scatterplot of the Relationship

- Would like to represent systolic BP as a continuous variable



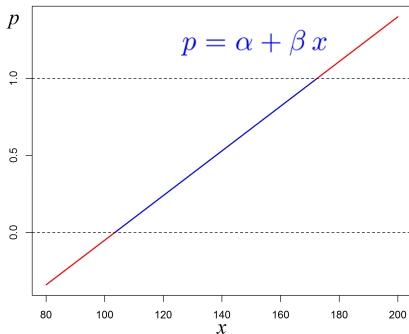
- Tendency for individuals who develop CVD to have higher systolic blood pressure, on average

Modeling Strategy 1

- Fit a model of the form:

$$p = \alpha + \beta x$$

- Where x represents systolic blood pressure
- This is a standard linear regression model
- **Problem:** Since p is a probability, it is restricted to take values between 0 and 1. However, $\alpha + \beta x$ can take values outside this range.

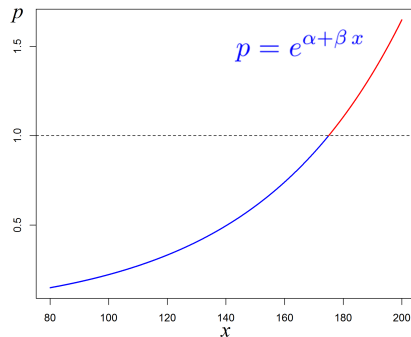


Modeling Strategy 2

- Fit a model of the form:

$$p = e^{\alpha + \beta x}$$

- This is a log-linear model
- **Problem:** p is guaranteed to be positive, however this model can give an estimated value of $p > 1$.

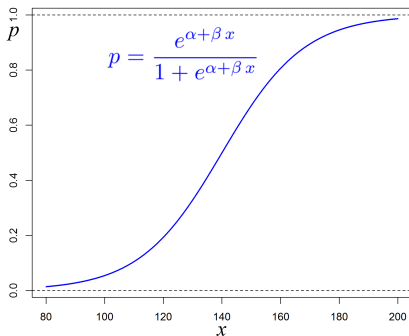


Modeling Strategy 3

- Fit a model of the form:

$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

- This is the **logistic function**
- Logistic function cannot yield a value < 0 or > 1
- Restricts the estimated value of p to the required range



Simple Logistic Regression Model

- In this setting where $p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$, the odds of a success (i.e., $Y = 1$) is:

$$\frac{p}{1-p} = \frac{\frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}}{\frac{1}{1 + e^{\alpha+\beta x}}} = e^{\alpha+\beta x}$$

- Taking the (natural) log of both sides:

$$\log \equiv \ln$$

$$\log \left(\frac{p}{1-p} \right) = \log (e^{\alpha+\beta x})$$

$$\text{log-odds of success} = \alpha + \beta x$$

Simple Logistic Regression Model

- Re-writing gives the logistic regression model:

Logistic Function, Probability of Success

$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Simple Logistic Regression Model

$$\log \left(\frac{p}{1 - p} \right) = \alpha + \beta x$$

Binary response variable → Logistic regression

The Logit

- The log-odds is also called the **logit**
- p is the true probability of a success

$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right) = \alpha + \beta x$$

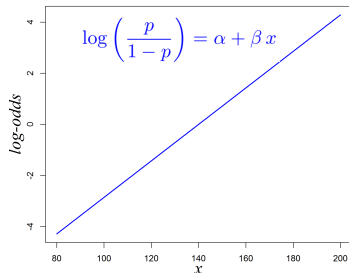
The **logit function** is the log of the odds of a success (log-odds)

$\frac{p}{1-p}$ is the odds of a success

- The **logit link** function maps the $(0, 1)$ range of probabilities onto $(-\infty, +\infty)$, the range of the linear predictor

Logistic Regression

- Resembles a linear regression model: $\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$
- Logistic regression belongs to a family of models called **generalized linear models (glm)**
- Mean of Y (i.e., p) is “**linked**” to $\alpha + \beta x$ through the **logit function**



- Instead of assuming the relationship between p and x is linear (Strategy 1), we assume the relationship between $\log\left(\frac{p}{1-p}\right)$ and x is linear
- This technique of fitting a model of this form is **logistic regression**

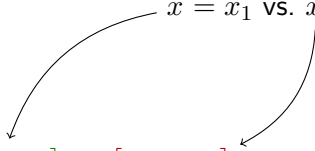


Interpretation of β : Change in Log-Odds

- Can compare the log-odds of a success under two conditions by subtracting the equation for the log-odds ($\log\left(\frac{p}{1-p}\right)$) under the two conditions, e.g., when

- $\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$

$$\begin{aligned}
 \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) &= [\alpha + \beta x_1] - [\alpha + \beta x_0] \\
 &= \alpha + \beta x_1 - \alpha - \beta x_0 \\
 &= (x_1 - x_0) \beta
 \end{aligned}$$

$x = x_1 \text{ vs. } x_0$


Interpretation of β : Change in Log-Odds

- For example, a 1-unit increase in x gives an expected difference of β in the log-odds

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) &= [\alpha + \beta(x+1)] - [\alpha + \beta x] \\ &= \alpha + \beta x + \beta - \alpha - \beta x \\ &= \beta\end{aligned}$$

Interpretation of β : Log Odds Ratio

- Impact on log-odds might not seem like a meaningful outcome
- However, the difference between two log-odds is a log odds ratio

Rule of Logs

$$\log(a) - \log(b) = \log\left(\frac{a}{b}\right)$$

$$\begin{aligned}\beta &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) \\ &= \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}\right) \\ &= \log(\text{OR})\end{aligned}$$

Slope Parameter from Logistic Regression Model: $\log(\text{OR})$

Slope β from a logistic regression model is the log odds ratio associated with a 1-unit increase in risk factor x

Interpretation of e^β : Odds Ratio

- Exponentiating both sides,

$$\beta = \log(\text{OR})$$

$$e^\beta = e^{\log(\text{OR})}$$

$$e^\beta = \text{OR}$$

e^β from Logistic Regression Model: OR

e^β from a logistic regression model is the **odds ratio** associated with a 1-unit increase in risk factor x

Interpretation of Slope: Continuous x

- Equation when $x = x + 1$:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= \alpha + \beta(x+1) \\ &= \alpha + \beta x + \beta\end{aligned}$$

- Equation when $x = x$:

$$\begin{aligned}\log\left(\frac{p_0}{1-p_0}\right) &= \alpha + \beta(x) \\ &= \alpha + \beta x\end{aligned}$$

$$\begin{aligned}\log(\text{OR}) &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) \\ &= [\alpha + \beta x + \beta] - [\alpha + \beta x] \\ &= \beta\end{aligned}$$

Odds Ratio

e^β = Odds Ratio for a 1-unit increase in x

Interpretation of Slope: c -unit Increase in x

- Equation when $x = x + c$:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= \alpha + \beta(x + c) \\ &= \alpha + \beta x + c\beta\end{aligned}$$

- Equation when $x = x$:

$$\begin{aligned}\log\left(\frac{p_0}{1-p_0}\right) &= \alpha + \beta(x) \\ &= \alpha + \beta x\end{aligned}$$

$$\begin{aligned}\log(\text{OR}) &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) \\ &= [\alpha + \beta x + c\beta] - [\alpha + \beta x] \\ &= c\beta\end{aligned}$$

Odds Ratio

$e^{c\beta}$ = Odds Ratio for a c -unit increase in x

Interpretation of Slope: Binary x

- Equation for **exposed** ($x = 1$):

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= \alpha + \beta(1) \\ &= \alpha + \beta\end{aligned}$$

- Equation for **unexposed** ($x = 0$):

$$\begin{aligned}\log\left(\frac{p_0}{1-p_0}\right) &= \alpha + \beta(0) \\ &= \alpha\end{aligned}$$

$$\begin{aligned}\log(\text{OR}) &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) \\ &= [\alpha + \beta] - [\alpha] \\ &= \beta\end{aligned}$$

Odds Ratio

e^β = Odds Ratio comparing exposed to unexposed

Multiplicative Effect

- A change in x has a **multiplicative effect** on the odds

$$\frac{p_1}{1 - p_1} = \frac{p_0}{1 - p_0} \times e^{\beta}$$

- If $\beta = 0$, then **OR** = $e^{\beta} = 1$

- $\frac{p_1}{1 - p_1} = \frac{p_0}{1 - p_0} = e^{\alpha}$, p is not related to x

- If $\beta > 0$, then **OR** = $e^{\beta} > 1$

- If $\beta < 0$, then **OR** = $e^{\beta} < 1$

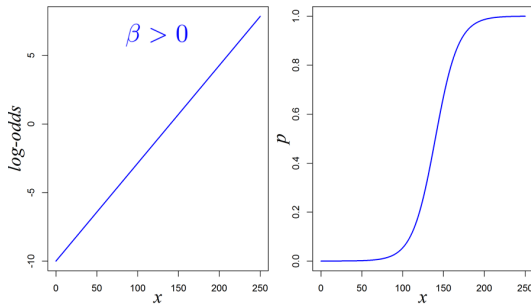
Interpretation of α

- The intercept α is the value of the log-odds when $x = 0$
- In the case of a binary covariate x , α equals the log-odds of disease in the unexposed

- $\log\left(\frac{p}{1-p}\right) = \alpha + \beta(0) = \alpha$

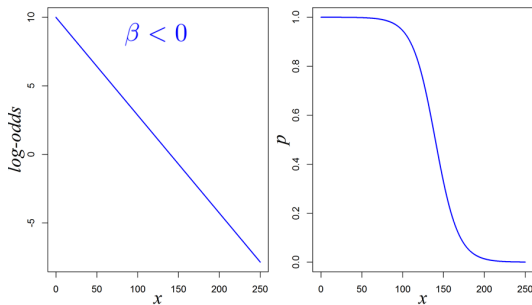
Relationship Between Log-Odds and p

- A positive slope, β , indicates that larger values of x are related to a larger log-odds
- When the log-odds increase, the probability p increases



Relationship Between Log-Odds and p

- A negative slope, β , indicates that larger values of x are related to a smaller log-odds
- When the log-odds decrease, the probability p decreases



Maximum Likelihood Estimation

- Logistic regression uses **maximum likelihood estimation** to find estimates of the population parameters, α and β
- The **likelihood** for a given model is the joint probability of the observed outcomes expressed as a function of the chosen regression model (and of the unknown model coefficients α , β)
- The estimates for α and β are obtained by choosing the a and b that maximize the numerical value of the (log)-likelihood function for the observed sample, hence the name maximum-likelihood estimation
- Important property of likelihoods from nested models is that maximized likelihood from larger model will always be at least as large as that for smaller model
- The numerical value of the likelihood does not have a useful interpretation, but is used in the **likelihood ratio test**, which compares the likelihoods from two nested models (analogous to the F -tests in linear regression)

Maximum Likelihood Estimation

- Estimates of the population parameters, α and β , are a and b , respectively

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b x$$

- a : Estimated log odds of success (logit) when $x = 0$
- b : Estimated log odds ratio for a 1-unit increase in x
- $a + b x$: Estimated logit used to calculate \hat{p} for given x

Interpretation of Estimated Slope

$$\log(\widehat{\text{OR}}) = b \qquad \widehat{\text{OR}} = e^b$$

Estimated Probability of Success

$$\hat{p} = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Binary Predictor

- Consider the model containing a single binary predictor (z)

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = a + b z \quad \text{where } z = \begin{cases} 1 & \text{if Exposed} \\ 0 & \text{if Unexposed} \end{cases}$$

- $\log(\widehat{OR})$ and \widehat{OR} of the response for exposed vs. unexposed are:

$$\log(\widehat{OR}) = b \quad \widehat{OR} = e^b$$

- Example:** Modeling incidence of CVD using current smoking status

$$z = \begin{cases} 1 & \text{if Current smoker} \\ 0 & \text{if Non-smoker} \end{cases}$$

Binary Predictor: Example

R Code, Logistic Regression

```
# Checking reference category of x, No=reference
> contrasts(fhssrs$CURSMOKE_factor)
      Yes
No      0
Yes     1
# Fitting logistic regression model, modeling event CVD = 1
> mod1 <- glm(CVD ~ CURSMOKE_factor, data = fhssrs, family = binomial(link="logit"))
> summary(mod1)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.4917    0.3689  -4.043 0.0000527
CURSMOKE_factorYes  0.5198    0.4843   1.073  0.283
```

$$\log\left(\frac{\hat{p}}{1 - \hat{p}}\right) = -1.49 + 0.52 \text{ Cursmoke}$$

Binary Predictor: Example

R Code, Logistic Regression

```
# "a" and "b"
> coef(mod1)
      (Intercept) CURSMOKE_factorYes
      -1.4916549      0.5197943

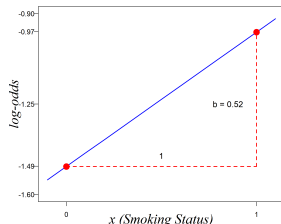
# exp(a) and exp(b)
> exp(coef(mod1))
      (Intercept) CURSMOKE_factorYes
      0.225000      1.681682
```

- Log odds of CVD are higher by 0.52 in current smokers vs. non-smokers
- Relative odds (OR) of developing CVD in current smokers vs. non-smokers is $\widehat{OR} = e^b = e^{0.52} = 1.68$

Interpretation of Slope, Binary Predictor: Example

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + bz = -1.49 + 0.52z$$

- Log-odds for **non-smokers** ($z = 0$): $\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = -1.49 + 0.52(0) = -1.49 = a$
- Log-odds for **current smokers** ($z = 1$): $\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = -1.49 + 0.52(1) = -0.97 = a + b$



- Difference in log-odds = $b = 0.52$
- $\widehat{OR} = e^b = e^{0.52} = 1.68$
- **Interpretation:** Odds of CVD in current smokers is **1.68** times that in non-smokers

Fitted Probability of Success: Example

$$\hat{p} = \frac{e^{a+bz}}{1 + e^{a+bz}} = \frac{e^{-1.49+0.52z}}{1 + e^{-1.49+0.52z}}$$

- Estimated probability of CVD for **non-smokers** ($z = 0$):

- $a + bz = -1.49 + 0.52(0) = -1.49$

$$\hat{p}_0 = \frac{e^a}{1 + e^a} = \frac{e^{-1.49}}{1 + e^{-1.49}} = 0.1837$$

- Estimated probability of CVD for **current smokers** ($z = 1$):

- $a + bz = -1.49 + 0.52(1) = -1.49 + 0.52 = -0.97$

$$\hat{p}_1 = \frac{e^{a+b}}{1 + e^{a+b}} = \frac{e^{-1.49+0.52}}{1 + e^{-1.49+0.52}} = \frac{e^{-0.97}}{1 + e^{-0.97}} = 0.2745$$

Fitted Probability of Success: Example

R Code, Predicted Probabilities

```
# Predicted p when CURSMOKE_factor = "No"
> pred.x <- data.frame(CURSMOKE_factor = factor("No", levels = c("No", "Yes")))
> predict(mod1, newdata = pred.x, type = "response")
      1
0.1836735

# Predicted p when CURSMOKE_factor = c("No", "Yes")
> pred.x <- data.frame(CURSMOKE_factor = factor(c("No", "Yes"), levels = c("No", "Yes")))
> predict(mod1, newdata = pred.x, type = "response")
      1      2
0.1836735 0.2745098
```

- $\hat{p}_0 = 0.1837$

- $\hat{p}_1 = 0.2745$

Logistic Regression, 2×2 Table Connection: Example

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = -1.49 + 0.52 z$$

Current Smoker	CVD		n_j	\hat{p}_j	$\frac{\hat{p}_j}{1 - \hat{p}_j}$
	Yes	No			
Yes ($z = 1$)	14	37	51	$\hat{p}_1 = \frac{14}{51} = 0.2745$	$\frac{0.2745}{0.7255} = 0.378$
No ($z = 0$)	9	40	49	$\hat{p}_0 = \frac{9}{49} = 0.1837$	$\frac{0.1837}{0.8163} = 0.225$

- $\widehat{OR} = \frac{14 \times 40}{37 \times 9} = 1.68$
- $b = \log(\widehat{OR}) = \log(1.682) = 0.52$
- $\frac{\hat{p}_0}{1 - \hat{p}_0} = \frac{0.1837}{0.8163} = 0.225$
- $a = \log \left(\frac{\hat{p}_0}{1 - \hat{p}_0} \right) = \log(0.225) = -1.49$
- Estimated probabilities from our model (\hat{p}_1, \hat{p}_0) equal the proportion of individuals with disease in the exposed and unexposed groups

Benefits of Logistic Regression

- With logistic regression, x can be continuous
- Can predict the probability of disease p and estimate odds ratios for different risk factors
- Can extend the simple logistic regression model to include additional covariates, building a multiple logistic regression model to give adjusted odds ratios after exponentiation

Logistic Regression Model, Continuous Predictor: Example

- **Example:** Modeling incidence of CVD using systolic BP

- x : SysBP, continuous

$$\log\left(\frac{\hat{p}}{1 - \hat{p}}\right) = a + b \text{ SysBP}$$

- $\log(\widehat{\text{OR}})$ and $\widehat{\text{OR}}$ of CVD associated with a 1-mmHg increase in systolic BP are:

$$\log(\widehat{\text{OR}}) = b \qquad \widehat{\text{OR}} = e^b$$

Continuous Predictor: Example

R Code, Logistic Regression

```
# Fitting logistic regression model, modeling event CVD = 1
> mod2 <- glm(CVD ~ SYSBP, data = fhssrs, family = binomial(link="logit"))
> summary(mod2)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.89564      1.67336  -2.926  0.00344
SYSBP        0.02716      0.01197   2.269  0.02325
...
# exp(a) and exp(b)
> exp(coef(mod2))
(Intercept)      SYSBP
0.007479097 1.027534000
```

$$\log\left(\frac{\hat{p}}{1 - \hat{p}}\right) = -4.90 + 0.027 \text{ SysBP}$$

Interpretation of Slope, Continuous Predictor: Example

- For the data examining the relationship between systolic blood pressure and CVD, the log-odds of CVD is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -4.90 + 0.027x$$

- Interpretation:** 1-mmHg increase in blood pressure is associated with increase in log odds of CVD of 0.027

$$\log(\widehat{OR}) = 0.027$$

- Exponentiate b to give the estimated odds ratio associated with a 1-mmHg increase in systolic BP

$$\widehat{OR} = e^{0.027} = 1.028$$

- One-unit increase in systolic BP increases odds of CVD by 2.8%

Estimating OR for c -unit Increase in x : Example

- The effect of a c -unit (e.g., 10-unit) increase in systolic blood pressure on the odds of CVD is:

$$\widehat{\text{OR}} = e^{cb} = e^{10 \times 0.0272} = 1.31$$

- Interpretation:** 10-unit increase in systolic BP increases odds of CVD by 31%



Fitted Probability of Success: Example

- Estimated probability of CVD for individual with systolic blood pressure of $x = 100$ mmHg:

- $$a + bx = -4.896 + 0.0272(100) = -4.896 + 2.72 = -2.176$$

$$\hat{p} = \frac{e^{a+bx}}{1 + e^{a+bx}} = \frac{e^{-2.176}}{1 + e^{-2.176}} = 0.10$$

R Code, Predicted Probabilities

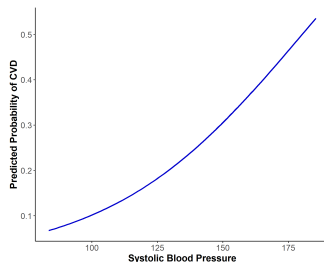
```
# Predicted p when SYSBP = 100, 120, and 150
> pred.x <- data.frame(SYSBP = c(100, 120, 150))
> predict(mod2, newdata = pred.x, type = "response")
      1      2      3
0.1016095 0.1629784 0.3054709
```

Probability of Success: Example

- Slope $b = 0.027$ is positive, indicating that as x increases, log-odds increase
- When the log-odds increase, the probability p increases
 - More likely to experience CVD if systolic BP is higher
 - Less likely to experience CVD if systolic BP is lower

x (SysBP)	$a + b x$	\hat{p}
100	-2.176	0.10
120	-1.632	0.16
150	-0.816	0.31

Figure: Relationship between \hat{p} and x not linear



Hypothesis Testing

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

- After estimating coefficients, the next step is the assessment of the significance of the parameters in the model through hypothesis testing
- **Question:** Is x linearly related to the log-odds of the event of interest?
- If β is significantly different from 0, then x is important in predicting disease (success)

Hypothesis Test for β

- **Hypothesis test** for the slope parameter β (Wald Test)
 - $H_0: \beta = 0$, No relationship between odds of disease and x
 - Under H_0 , OR = 1
 - $H_0: \beta \neq 0$, There is a relationship between odds of disease and x

Wald Test Statistic for Slope

$$Z = \frac{b}{s_b} \sim N(0, 1)$$

Hypothesis Test for β , Continuous Predictor: Example

- **Example:** Is there an association between *systolic BP* and *CVD*? $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$

R Code, Logistic Regression

```
> mod2 <- glm(CVD ~ SYSBP, data = fhssrs, family = binomial(link="logit"))
```

```
> summary(mod2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.89564	1.67336	-2.926	0.00344
SYSBP	0.02716	0.01197	2.269	0.02325

- $z = \frac{b}{s_b} = \frac{0.0272}{0.012} = 2.269$ compared to $N(0, 1)$ $pval = 2*(1-pnorm(2.269))$
- Reject H_0 if $|z| \geq z_{1-\frac{\alpha}{2}} = z_{.975} = z^* = 1.96$
- $|z| \geq z^* \rightarrow$ Reject H_0
- $p = 2 \times P(Z \geq 2.269) = 0.023$
- $p > 0.05 \rightarrow$ Reject H_0
- **Conclusion:** There is an association between systolic BP and CVD. 1-unit increase in systolic BP increases odds of CVD by 2.8%; $\widehat{OR} = 1.028$

Hypothesis Test for β , Binary Predictor: Example

- **Example:** Is there an association between *smoking status* and *CVD*? $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$

R Code, Logistic Regression

```
> mod1 <- glm(CVD ~ CURSMOKE_factor, data = fhssrs, family = binomial(link="logit"))
> summary(mod1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.4917	0.3689	-4.043	0.0000527
CURSMOKE_factorYes	0.5198	0.4843	1.073	0.283

- $z = \frac{b}{s_b} = \frac{0.5198}{0.4843} = 1.073$ compared to $N(0, 1)$ $pval = 2*(1-pnorm(1.073))$
- Reject H_0 if $|z| \geq z_{1-\frac{\alpha}{2}} = z_{0.975} = z^* = 1.96$
- $|z|$ not $\geq z^* \rightarrow$ Fail to Reject H_0
- $p = 2 \times P(Z \geq 1.073) = 0.283$
- $p \leq 0.05 \rightarrow$ Fail to Reject H_0
- **Conclusion:** Current smoking status does not have a statistically significant effect on the odds of developing CVD (OR not significantly different from 1)



Confidence Interval for $\log(\text{OR})$ and OR

- **Confidence interval** for the slope β , $\log(\text{OR})$, has the form:

100(1 - α)% Confidence Interval for $\log(\text{OR}), \beta$

$$b \pm z_{1-\frac{\alpha}{2}} s_b = (c_L, c_U)$$

- To find the confidence interval for the **odds ratio**, exponentiate the lower and upper bounds of the CI for the $\log(\text{OR})$

100(1 - α)% Confidence Interval for OR

$$(e^{c_L}, e^{c_U})$$

- **Note:** The confidence interval for the OR will *exclude* 1 if we rejected H_0 . The CI will *include* 1 if we failed to reject H_0 .

Confidence Interval for $\log(\text{OR})$ and OR: Example

- **Example:** Confidence interval of OR for *systolic BP*

R Code, Confidence Intervals

```
> confint.default(mod2) # CI for beta (log OR)
              2.5 %      97.5 %
(Intercept) -8.175374367 -1.61591204
SYSBP        0.003703049  0.05062046

> exp(cbind(OR = coef(mod2), confint.default(mod2))) # OR and CI for OR
              OR      2.5 %      97.5 %
(Intercept) 0.007479097 0.0002815011 0.1987094
SYSBP        1.027534000 1.0037099141 1.0519236
```

- 95% CI for $\log(\text{OR})$, β :

$$\hat{b} \pm z_{1-\frac{\alpha}{2}} s_b = 0.0272 \pm 1.96 \times 0.0120 = (c_L = 0.0037, c_U = 0.0506)$$
- 95% CI for OR: $(e^{0.0037}, e^{0.0506}) = (1.004, 1.052)$, which excludes 1

Progress this Unit

- 1 Binary Outcome
 - Estimating p
- 2 Simple Logistic Regression
 - Motivation
 - Interpretation of the Model
 - Fitting the Model
 - Inference
- 3 Multiple Logistic Regression
 - Continuous and Binary Regressors
 - Categorical Regressors
 - Interaction Terms
 - Diagnostics

Multiple Logistic Regression Model

- As in the case of linear regression, logistic regression models can include more than one independent variable

Multiple Logistic Regression Model

$$\log \left(\frac{p}{1-p} \right) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

Probability of Success

$$p = \frac{e^{\alpha + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Interpreting Coefficients from Multiple Regression Model

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

- Multiple regression analysis statistically adjusts the estimated effect of each variable in the model for other variables included in the model
- For example, the estimated slope b_2 is the $\log(\widehat{OR})$ associated with a 1-unit increase in x_2 , holding all other x variables in the model constant/controlling for all other x variables

Multiple Logistic Regression Model: Example

- **Example:** Systolic BP is an important predictor of CVD when considered alone. Age is a potential confounder of this relationship. After controlling for age, is *systolic BP* still an important predictor of *CVD*?
 - x_1 : SysBP, continuous
 - x_2 : Age, continuous

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 \text{ SysBP} + b_2 \text{ Age}$$

- b_1 : log-OR associated with 1-unit increase in systolic BP, holding age constant (controlling for age)
- b_2 : log-OR associated with 1-unit increase in age, holding systolic BP constant (controlling for systolic BP)

Multiple Logistic Regression Model: Example

R Code, Logistic Regression

```
> mod3 <- glm(CVD ~ SYSBP + AGE, data = fhssrs, family = binomial(link="logit"))
> summary(mod3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.98985      2.09384  -3.338 0.000843
SYSBP        0.01568      0.01346   1.165 0.244041
AGE          0.07170      0.03656   1.961 0.049867
```

- Holding age constant, a 1-mmHg increase in **systolic BP** increases the *log-odds* of developing CVD by $b_1 = 0.0157$ (increases *odds* of CVD by $e^{0.0157} = 1.0158$ times). However, this effect is not statistically significant (p -value = 0.244).
- After age is controlled for, systolic BP is no longer significantly associated with CVD
- Holding systolic BP constant, a 1-year increase in **age** significantly increases *odds* of CVD by $e^{0.0717} = 1.074$ times (p -value = 0.0499)

Public Health Application: Logistic Regression

Public Health Application

Association Between Maternal Use of Folic Acid Supplements and Risk of Autism Spectrum Disorders in Children

Pål Surén, MD, MPH

Christine Roth, MSc

Michaeline Bresnahan, PhD

Margaretha Haugen, PhD

Mady Hornig, MD

Deborah Hirtz, MD

Kari Kveim Lie, MD

W. Ian Lipkin, MD

Per Magnus, MD, PhD

Ted Reichborn-Kjennerud, MD, PhD

Synnve Schjølberg, MSc

George Davey Smith, MD, DSc

Importance Prenatal folic acid supplements reduce the risk of neural tube defects in children, but it has not been determined whether they protect against other neurodevelopmental disorders.

Objective To examine the association between maternal use of prenatal folic acid supplements and subsequent risk of autism spectrum disorders (ASDs) (autistic disorder, Asperger syndrome, pervasive developmental disorder—not otherwise specified [PDD-NOS]) in children.

Design, Setting, and Patients The study sample of 85 176 children was derived from the population-based, prospective Norwegian Mother and Child Cohort Study (MoBa). The children were born in 2002-2008; by the end of follow-up on March 31, 2012, the age range was 3.3 through 10.2 years (mean, 6.4 years). The exposure of primary interest was use of folic acid from 4 weeks before to 8 weeks after the start of pregnancy, defined as the first day of the last menstrual period before conception. Relative risks of ASDs were estimated by odds ratios (ORs) with 95% CIs in a logistic regression analysis. Analyses were adjusted for maternal education level, year of birth, and parity. *JAMA.* 2013;309(6):570-577

- Multiple logistic regression model used to examine the association between autism spectrum disorders in children and maternal use of folic acid supplements, adjusted for year of birth, maternal education level, and parity

[Link to article](#)

Public Health Application: Logistic Regression

Public Health Application

Table 2. Risk of Autistic Disorder According to Maternal Folic Acid Use

Folic Acid Use	No. (%)		OR (95% CI)	
	Total	Autistic Disorder	Unadjusted	Adjusted ^a
No	24 134 (28.3)	50 (0.21)	1 [Reference]	1 [Reference]
Yes	61 042 (71.7)	64 (0.10)	0.51 (0.35-0.73)	0.61 (0.41-0.90)

Abbreviation: OR, odds ratio.

^aAdjusted for year of birth, maternal education level, and parity. For maternal education, missing data were included as a separate category in the logistic regression model.

Statistical Analyses

Analyses were performed using SPSS version 19.0 (SPSS Inc). Odds ratios (ORs) with 95% CIs for the association between folic acid use and risk of each ASD were estimated from logistic regression models. The adjusted models included adjustment for year of birth, maternal education level, and parity, because these were the only covariates that had any influence on the OR estimates.

Results of the logistic regression analysis for autistic disorder are reported in TABLE 2. There was an inverse association between folic acid use and subsequent risk of autistic disorder. Autistic disorder was present in 0.10% (64/61 042) of children whose mothers took folic acid, compared with 0.21% (50/24 134) in children whose mothers did not take folic acid. The adjusted OR of autistic disorder was 0.61 (95% CI, 0.41-0.90) in children of folic acid users. Adjustment for maternal illness and medication use did not affect the OR (eTable 2).

Continuous and Binary Predictors: Example

- **Example:** Controlling for age, is presence of *elevated diastolic BP* (> 90 mmHg) an important predictor of *CVD*?

- x_1 : HighDiastolic, dichotomous $= \begin{cases} 1 & \text{if Diastolic BP} > 90 \\ 0 & \text{if Diastolic BP} \leq 90 \end{cases}$
- x_2 : Age, continuous

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 \text{ HighDiastolic} + b_2 \text{ Age}$$

Continuous and Binary Predictors: Example

R Code, Logistic Regression

```
> mod4 <- glm(CVD ~ HIGHDIABP_factor + AGE, data = fhssrs, family = binomial(link="logit"))
> summary(mod4)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.65317	1.80805	-3.127	0.00177
HIGHDIABP_factorYes	1.15737	0.55875	2.071	0.03833
AGE	0.08153	0.03442	2.368	0.01786

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.653 + 1.157 \text{ HighDiastolic} + 0.0815 \text{ Age}$$

- Holding age constant, the presence of elevated diastolic BP is associated with CVD
- Age-adjusted estimated OR for presence vs. absence of elevated diastolic BP = $e^{1.1573} = 3.181$ ($p\text{-value} = 0.038$)



Continuous and Binary Predictors: Example

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = -5.653 + 1.157 \text{ HighDiastolic} + 0.0815 \text{ Age}$$

- Elevated diastolic ($x_1 = 1$):

$$\begin{aligned} \log \left(\frac{\hat{p}_1}{1 - \hat{p}_1} \right) &= -5.653 + 1.157 (1) + 0.0815 \text{ Age} \\ &= -4.496 + 0.0815 \text{ Age} \end{aligned}$$

$$\hat{p}_1 = \frac{e^{-4.496 + 0.0815 \text{ Age}}}{1 + e^{-4.496 + 0.0815 \text{ Age}}}$$

- No elevated diastolic ($x_1 = 0$):

$$\begin{aligned} \log \left(\frac{\hat{p}_0}{1 - \hat{p}_0} \right) &= -5.653 + 1.157 (0) + 0.0815 \text{ Age} \\ &= -5.653 + 0.0815 \text{ Age} \end{aligned}$$

$$\hat{p}_0 = \frac{e^{-5.653 + 0.0815 \text{ Age}}}{1 + e^{-5.653 + 0.0815 \text{ Age}}}$$

Tabular Setup to Aid in Calculation of $\widehat{\text{logit}}$: Example

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = -5.653 + 1.157 \text{ HighDiastolic} + 0.0815 \text{ Age}$$

- Elevated diastolic, age 50 ($x_1 = 1$, $x_2 = 50$):

Parameter	Estimate		Scenario		Contribution to the $\widehat{\text{logit}}$
Intercept	-5.6528	×	1	=	-5.6528
HighDiastolic	1.1573	×	1	=	1.1573
Age	0.0815	×	50	=	4.075
					$\sum = -0.4205$

- Estimated probability that a 50-year old with elevated diastolic BP develops CVD:

$$\hat{p} = \frac{e^{-0.4205}}{1 + e^{-0.4205}} = \frac{0.657}{1.657} = 0.396$$

Continuous and Binary Predictors: Example

Figure: Estimated Log-odds of CVD

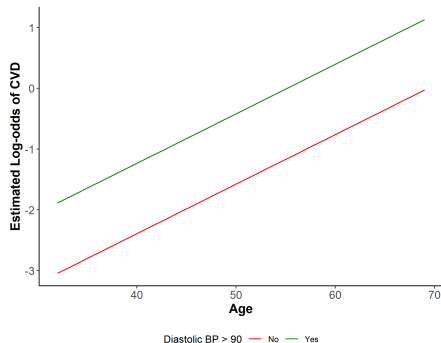
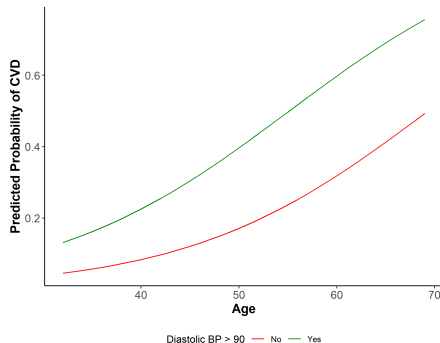


Figure: Fitted Probability of CVD



Categorical Variables

- As in linear regression, we can represent a categorical variable with C levels as $C - 1$ dichotomous (dummy) variables in the model

Systolic BP Category	z_1	z_2
≤ 120	0	0
$(120, 150]$	1	0
> 150	0	1

- Lowest systolic BP category is the reference group ($z_1 = z_2 = 0$)

Categorical Variables: Example

Systolic BP Category	SYSBPGRP	z_1	z_2
≤ 120 (ref.)	0	0	0
$(120, 150]$	1	1	0
> 150	2	0	1

$$\log\left(\frac{\hat{p}}{1 - \hat{p}}\right) = a + b_1 z_1 + b_2 z_2$$

- b_1 is $\log(\widehat{OR})$ for $(120, 150]$ vs. ≤ 120 (reference)
- b_2 is $\log(\widehat{OR})$ for > 150 vs. ≤ 120 (reference)

Categorical Variables: Example

R Code, Categorical Predictor

```
# Grouping SYSBP
> fhssrs$SYSBPGRP[fhssrs$SYSBP <= 120] = 0
> fhssrs$SYSBPGRP[fhssrs$SYSBP > 120 & fhssrs$SYSBP <= 150] = 1
> fhssrs$SYSBPGRP[fhssrs$SYSBP > 150] = 2

> fhssrs$SYSBPGRP_factor = factor(fhssrs$SYSBPGRP,
+   levels = 0:2,
+   labels = c("<= 120", "(120, 150]", "> 150"))

> contrasts(fhssrs$SYSBPGRP_factor)
      (120, 150] > 150
<= 120           0      0
(120, 150]       1      0
> 150            0      1
```

- $\text{SYSBPGRP} = 0$:
Systolic BP ≤ 120
- $\text{SYSBPGRP} = 1$:
Systolic BP (120, 150]
- $\text{SYSBPGRP} = 2$:
Systolic BP > 150

Systolic BP Category	SYSBPGRP	z_1	z_2
≤ 120 (ref.)	0	0	0
(120, 150]	1	1	0
> 150	2	0	1

- R automatically creates dummy variables for factor variables

Categorical Variables: Example

R Code, Logistic Regression

```
> mod5 <- glm(CVD ~ SYBPGRP_factor, data = fhssrs, family = binomial(link="logit"))
```

```
> summary(mod5)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.8718	0.5371	-3.485	0.000492
SYBPGRP_factor(120, 150]	0.6931	0.6304	1.099	0.271553
SYBPGRP_factor> 150	1.3328	0.7174	1.858	0.063191

OR and CI for OR

```
> exp(cbind(OR = coef(mod5), confint.default(mod5))) # OR and CI for OR
```

	OR	2.5 %	97.5 %
(Intercept)	0.1538462	0.05369311	0.4408134
SYBPGRP_factor(120, 150]	2.0000000	0.58131513	6.8809494
SYBPGRP_factor> 150	3.7916667	0.92936197	15.4694689

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = -1.872 + 0.693 \text{SYS}_{120-150} + 1.333 \text{SYS}_{>150}$$

Categorical Variables: Example

R Code, Logistic Regression

```
# logOR, OR and CI for OR
> cbind(b = coef(mod5), OR = exp(coef(mod5)), exp(confint.default(mod5)))
```

	b	OR	2.5 %	97.5 %
(Intercept)	-1.8718022	0.1538462	0.05369311	0.4408134
SYSBGRP_factor(120, 150]	0.6931472	2.0000000	0.58131513	6.8809494
SYSBGRP_factor> 150	1.3328057	3.7916667	0.92936197	15.4694689

- \widehat{OR} comparing middle to lowest systolic BP category, $e^{b_1} = e^{0.6931} = 2$ (p -value = 0.27)
- \widehat{OR} comparing highest to lowest systolic BP category, $e^{b_2} = e^{1.3328} = 3.792$ (p -value = 0.063)

Categorical Variables: Example

- Look familiar?

Systolic BP Category	CVD		\hat{p}_j
	Yes	No	
> 150	7	12	0.368
≤ 120	4	26	0.133

$$\widehat{OR} = \frac{\frac{\hat{p}_2}{1 - \hat{p}_2}}{\frac{\hat{p}_0}{1 - \hat{p}_0}} = \frac{\frac{0.368}{0.632}}{\frac{0.133}{0.867}} = \frac{7 \times 26}{12 \times 4} = \boxed{3.792}$$

Systolic BP Category	CVD		\hat{p}_j
	Yes	No	
$(120, 150]$	12	39	0.235
≤ 120	4	26	0.133

$$\widehat{OR} = \frac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_0}{1 - \hat{p}_0}} = \frac{\frac{0.235}{0.765}}{\frac{0.133}{0.867}} = \frac{12 \times 26}{39 \times 4} = \boxed{2}$$

Categorical Variables: Example

- Benefit of regression modeling is that we can easily control for factors by including them in the model, yielding **adjusted odds ratios**
- Example:** Controlling for age, what is the effect of *systolic BP category* on *CVD*?
 - x_1 : Age, continuous
 - z_1 : $SYS_{120-150}$, dichotomous $= \begin{cases} 1 & \text{if Systolic BP} > 150 \\ 0 & \text{otherwise} \end{cases}$
 - z_2 : $SYS_{>150}$, dichotomous $= \begin{cases} 1 & \text{if Systolic BP } (120, 150] \\ 0 & \text{otherwise} \end{cases}$

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 \text{Age} + b_2 \text{SYS}_{120-150} + b_3 \text{SYS}_{>150}$$

Categorical Variables: Example

R Code, Logistic Regression

```
> mod6 <- glm(CVD ~ AGE + SYSBPGRP_factor, data = fhssrs, family = binomial(link="logit"))
> summary(mod6)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.54606	1.79406	-3.091	0.00199
AGE	0.07834	0.03534	2.217	0.02665
SYSBPGRP_factor(120, 150]	0.36243	0.66043	0.549	0.58316
SYSBPGRP_factor> 150	0.69868	0.77823	0.898	0.36931

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.546 + 0.078 \text{ Age} + 0.362 \text{ SYS}_{120-150} + 0.699 \text{ SYS}_{>150}$$

- Age-adjusted $\widehat{\text{OR}}$ comparing middle to lowest systolic BP category, $e^{b_2} = e^{0.362} = 1.437$ (p -value = 0.583)
- Age-adjusted $\widehat{\text{OR}}$ comparing highest to lowest systolic BP category, $e^{b_3} = e^{0.699} = 2.011$ (p -value = 0.369)

Inference for Categorical Variables

- The **Likelihood Ratio Test** can be used to simultaneously test the significance of more than 1 parameter
- Useful in determining if a categorical variable, which is represented by a group of dummy variables, is an important predictor in the logistic regression model (**Partial F -Test** analogue)
- Likelihood Ratio Test compares the value of the maximized log-likelihood under the **full model** (all predictors included) to the log-likelihood under the **reduced model** (model under H_0)
- For example, to test overall effect of systolic BP in the model including age, compare:
 - Full model containing **age** and **systolic BP** to
 - Reduced model containing only **age**

Likelihood Ratio Test

- **Full (F) model:** $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2$
- **Reduced (R) model:** $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1$
 - $H_0: \beta_2 = \beta_3 = 0$
 - $H_1: \beta_2$ and β_3 are not both 0

Likelihood Ratio Test Statistic

$$\begin{aligned} G &= -2 \log\text{-likelihood}(R) - (-2 \log\text{-likelihood}(F)) \\ &= -2 [\log\text{-likelihood}(R) - \log\text{-likelihood}(F)] \end{aligned}$$

- $G \sim \chi^2_{df}$, where df = Number of parameters tested under H_0
- For logistic regression, the log-likelihood is always negative because the likelihood contribution from each observation is a probability between 0 and 1

Likelihood Ratio Test: Example

R Code, LRT

```
# Full Model
> mod.full <- glm(CVD ~ AGE + SYSBPGRP_factor, data = fhssrs, family = binomial(link="logit"))
> logLik(mod.full)                # log-likelihood(F)
'log Lik.' -49.52189 (df=4)
> neg2LL.full <- as.numeric(-2*logLik(mod.full))      # -2*log-likelihood(F)

# Reduced Model (does not include SYSBPGRP_factor)
> mod.reduced <- glm(CVD ~ AGE, data = fhssrs, family = binomial(link="logit"))
> logLik(mod.reduced)             # log-likelihood(R)
'log Lik.' -49.93036 (df=2)
> neg2LL.reduced <- as.numeric(-2*logLik(mod.reduced)) # -2*log-likelihood(R)
> G <- neg2LL.reduced - neg2LL.full   # Test statistic
> G
[1] 0.8169461
> qchisq(0.95, df = 2)              # Critical value
[1] 5.991465
> 1 - pchisq(G, df = 2)             # P-value (upper tail area only)
[1] 0.6646644
```


Likelihood Ratio Test: Example

R Code, LRT

```
# Likelihood Ratio Test, H0: beta2 = beta3 = 0
> anova(mod.reduced, mod.full, test = "Chisq")
Analysis of Deviance Table
Model 1: CVD ~ AGE
Model 2: CVD ~ AGE + SYSBPGRP_factor
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      98    99.861
2      96    99.044  2   0.81695   0.6647
```

Model	-2 log L	Parameters
Reduced (x_1)	99.861	2
Full (x_1, z_1, z_2)	99.044	4
Difference	$G = 0.817$	$df = 2$

- **Example:** Is *systolic BP category* an important predictor of *CVD* in a model containing age?
 - **Step 1:** State the hypotheses
 - $H_0 : \beta_2 = \beta_3 = 0$ vs.
 - $H_1 : \beta_2$ and β_3 not both 0
 - **Step 2:** Specify significance level
 - $\alpha = 0.05$
 - **Step 3:** Compute the appropriate test statistic
 - $G = -2 \log L(R) - (-2 \log L(F)) = 99.861 - 99.044 = 0.817$

$$G \sim \chi_2^2$$

Likelihood Ratio Test: Example

- **Step 4:** Generate the decision rule

`x95 = qchisq(0.95, df = 2)`

- Reject H_0 if $G \geq \chi^2_{1-\alpha}(2) = \chi^2_{.95}(2) = 5.99$: Critical value

- **Step 5:** Draw a conclusion about H_0

`pval = 1 - pchisq(0.817, df = 2)`

- $G = 0.817$

- $p = P(G \geq 0.817) = 0.665$

- G not $\geq 5.994 \rightarrow$ Fail to reject H_0

- p not $\leq 0.05 \rightarrow$ Fail to reject H_0

- **Conclusion:** Fail to reject the null hypothesis that $\beta_2 = \beta_3 = 0$. Thus, we cannot conclude systolic BP category is contributing significantly to this model containing age.



Categorizing Data

- Although we like to use all of the available information contained in a predictor variable, it is sometimes more intuitive to **categorize** data, particularly when reporting odds ratios
- Doing this may result in loss of precision and loss of power to detect differences. Arbitrary cut-points may be criticized.

- x_1 : HighDiastolic, dichotomous = $\begin{cases} 1 & \text{if Diastolic BP} \geq 90 \\ 0 & \text{if Diastolic BP} < 90 \end{cases}$

- x_2 : AgeOver50, dichotomous = $\begin{cases} 1 & \text{if Age} \geq 50 \\ 0 & \text{if Age} < 50 \end{cases}$

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 \text{HighDiastolic} + b_2 \text{AgeOver50}$$

Logistic Regression Model: Example

R Code, Logistic Regression

```
> mod7 <- glm(CVD ~ HIGHDIABP_factor + AgeOver50_factor, data = fhssrs,
               family = binomial(link="logit"))
> summary(mod7)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.0942     0.4506  -4.648 0.00000336
HIGHDIABP_factor>=90  0.7939     0.5417   1.466  0.1428
AgeOver50_factor>=50  1.1152     0.5421   2.057  0.0397
> exp(coef(mod7))
              (Intercept) HIGHDIABP_factor>=90 AgeOver50_factor>=50
                0.1231651             2.2120198             3.0500843
```

- Holding age constant, those who have high diastolic BP have 2.21 times the odds of CVD than those who do not, $e^{0.7939} = 2.21$ (p -value = 0.14)
- Holding diastolic BP constant, those with age ≥ 50 have 3.05 times the odds of CVD than those who are under 50, $e^{1.1152} = 3.05$ (p -value = 0.04)

Estimated Probabilities, Binary Variables: Example

R Code, Fitted Probabilities

```
# Values of AgeOver50_factor and HIGHDIABP_factor in the data
> x.age <- levels(fhssrs$AgeOver50_factor)
> x.age
[1] "<50" ">=50"
> x.diabp <- levels(fhssrs$HIGHDIABP_factor)
> x.diabp
[1] "<90" ">=90"

# Creates a data frame from all combinations of the supplied vectors or factors
> pred.x <- expand.grid(AgeOver50_factor = x.age, HIGHDIABP_factor = x.diabp)

# Fitted probabilities
> pred.x$phat.mod7 <- predict(mod7, newdata = pred.x, type = "response")
> pred.x
```

	AgeOver50_factor	HIGHDIABP_factor	phat.mod7
1	<50	<90	0.1096590
2	>=50	<90	0.2730783
3	<50	>=90	0.2141106
4	>=50	>=90	0.4538433

Table of Estimated Probabilities

- Categorizing continuous variables simplifies estimating probabilities
- However, when including continuous variables, can estimate probabilities for fixed values of the predictors within the range of the data
- To aid in interpretation of results, can create a table of estimated probabilities of CVD for specific levels of the risk factor(s)
 - x_1 : Diastolic BP, continuous
 - x_2 : Age, continuous

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 \text{ DiastolicBP} + b_2 \text{ Age}$$

Logistic Regression Model: Example

R Code, Logistic Regression

```
> mod8 <- glm(CVD ~ DIABP + AGE, data = fhssrs, family = binomial(link="logit"))
> summary(mod8)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.53276      2.40042  -3.555 0.000378
DIABP         0.03859      0.02014   1.916 0.055345
AGE           0.08017      0.03451   2.323 0.020174
> exp(coef(mod8))
(Intercept)      DIABP      AGE
0.0001969102 1.0393404724 1.0834674462
```

- Holding age constant, relative odds of CVD associated with a 1-year increase in diastolic blood pressure is $e^{0.039} = 1.039$ (p -value = 0.055)
- Holding diastolic blood pressure constant, relative odds of CVD associated with a 1-unit increase in age is $e^{0.080} = 1.083$ (p -value = 0.020)

Table of Estimated Probabilities: Example

$$\hat{p} = \hat{P}(Y = 1 | \text{DIABP}, \text{AGE}) = \frac{e^{-8.53+0.039 \text{ DIABP}+0.080 \text{ AGE}}}{1 + e^{-8.53+0.039 \text{ DIABP}+0.080 \text{ AGE}}}$$

Table: Estimated probability of CVD

	Diastolic BP		
	70	90	110
Age 40	0.0675	0.1355	0.2532
Age 50	0.1390	0.2589	0.4305
Age 60	0.2647	0.4378	0.6276
Age 70	0.4452	0.6345	0.7897

Interaction: Example

- **Example:** As we *age*, the odds of *CVD* increase. Is this effect the same for *males* and *females*?
- To answer this question, include the **interaction** of age and sex in the model:
 - x_1 : Age, continuous
 - x_2 : Sex, dichotomous = $\begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$
 - x_3 : Age \times Sex

$$\log \left(\frac{\hat{p}}{1 - \hat{p}} \right) = a + b_1 \text{ Age} + b_2 \text{ Sex} + b_3 \text{ Age} \times \text{Sex}$$

Interaction: Example

R Code, Logistic Regression

```
> mod9 <- glm(CVD ~ AGE + SEX_factor + AGE*SEX_factor, data=fhssrs, family=binomial(link="logit"))
> summary(mod9)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.397010	1.854607	-0.753	0.4513
AGE	0.002337	0.038170	0.061	0.9512
SEX_factorMale	-4.854374	3.090217	-1.571	0.1162
AGE:SEX_factorMale	0.110642	0.059713	1.853	0.0639

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.397 + 0.002 \text{ Age} - 4.854 \text{ Male} + 0.111 \text{ Age} \times \text{Male}$$

- As in linear regression, to test if the effect of age differs in males and females, test if the interaction term (β_3) is significantly different from 0

$$z = \frac{b_3}{s_{b_3}} = \frac{0.111}{0.0597} = 1.853$$

- Do not have evidence to reject $H_0: \beta_3 = 0$ (p -value = 0.064)

Interaction: Example

- Log-odds for **females** ($x_2 = 0$):

$$\begin{aligned}\log\left(\frac{\hat{p}_0}{1 - \hat{p}_0}\right) &= -1.397 + 0.002 \text{ Age} - 4.854(0) + 0.111 \text{ Age} \times (0) \\ &= -1.397 + 0.002 \text{ Age} = a + b_1 \text{ Age}\end{aligned}$$

- Log-odds for **males** ($x_2 = 1$):

$$\begin{aligned}\log\left(\frac{\hat{p}_1}{1 - \hat{p}_1}\right) &= -1.397 + 0.002 \text{ Age} - 4.854(1) + 0.111 \text{ Age} \times (1) \\ &= -6.251 + 0.113 \text{ Age} = (a + b_2) + (b_1 + b_3) \text{ Age}\end{aligned}$$

- $\widehat{\text{OR}}$ for 1-year increase in age in **females**: $e^{0.002} = 1.002$ (p -value = 0.9512 from test of β_1)
- $\widehat{\text{OR}}$ for 1-year increase in age in **males**: $e^{0.113} = 1.120$ (p -value = 0.0139 from test of $\beta_1 + \beta_3$)

Interaction: Example

R Code, OR for Age in Males

```
> library(multcomp)      # Load required package
# Vector that specifies linear combination of coefficients interested in
> K <- rbind(c(0, 1, 0, 1)) # 1 = coefficients "on" when estimating slope in Males
> rownames(K) <- "b1+b3 (slope in Males)"
> summary(glht(mod9, linfct = K))
Simultaneous Tests for General Linear Hypotheses
Fit: glm(formula = CVD ~ AGE + SEX_factor + AGE * SEX_factor, family = binomial(link = "logit"),
  data = fhssrs2)
Linear Hypotheses:

                Estimate Std. Error z value Pr(>|z|)
b1+b3 (slope in Males) == 0  0.11298    0.04592    2.46   0.0139
(Adjusted p values reported -- single-step method)

# OR and CI of OR of effect of Age in Males
> exp(confint(glht(mod9, linfct = K))$confint[1,])
Estimate      lwr      upr
1.119609 1.023244 1.225049
```

Interaction: Example

Figure: Estimated Log-odds of CVD

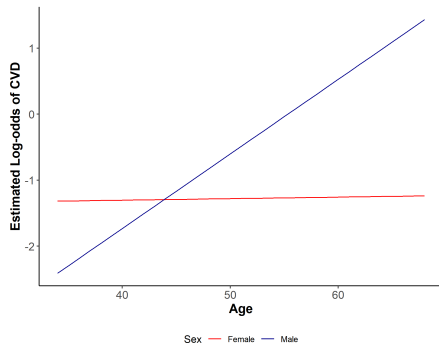
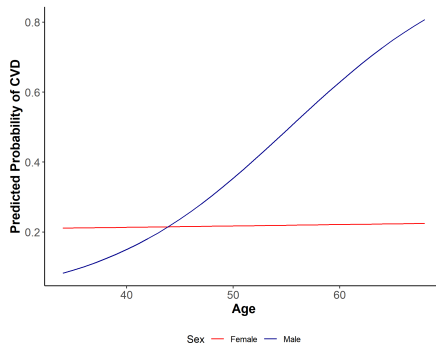


Figure: Fitted Probability of CVD



McFadden's R^2

- Unlike linear regression, there is no R^2 statistic that explains the proportion of variance in the dependent variable that is explained by the model in logistic regression
- There are a number of **pseudo R^2** measures that are often reported
- A common metric is **McFadden's R^2**

McFadden's R^2

$$R_m^2 = 1 - \frac{\log\text{-likelihood}(Mod)}{\log\text{-likelihood}(Null)}$$

- Ranges from 0 to just under 1: values closer to zero indicate model has no predictive power
- Values between 0.2-0.4 indicate (in McFadden's words) excellent model fit.
- Where $\log\text{-likelihood}(Mod)$: log-likelihood of the model being evaluated
 $\log\text{-likelihood}(Null)$: log-likelihood of the null model containing the intercept only

McFadden's R^2

R Code, McFadden's R^2

```
# Model being evaluated
> mod8 <- glm(CVD ~ DIABP + AGE, data = fhssrs, family = binomial(link="logit"))

# Null model
> mod.null <- glm(CVD ~ 1, data = fhssrs, family = binomial(link="logit"))

# McFadden's R2
> R2m <- 1 - as.numeric(logLik(mod8))/as.numeric(logLik(mod.null))
> R2m
[1] 0.1090154
```

- If model does not predict outcome better than the null model, log-likelihood will not be much larger than the log-likelihood of the null model, and so $\log\text{-likelihood}(Mod)/\log\text{-likelihood}(Null) \approx 1$, and $R_m^2 \approx 0$ (model has no predictive value)

Hosmer-Lemeshow Test

- The **Hosmer-Lemeshow goodness-of-fit test** can be used to determine how well a logistic regression model fits
 - H_0 : The model **does** fit the data well
 - H_1 : The model **does not** fit the data well
- A significant Hosmer-Lemeshow test result indicates the model is not correctly specified and doesn't fit the data well
- A non-significant Hosmer-Lemeshow test result is consistent with the null hypothesis that our model is correctly specified and fits the data well. Remember, though, that we do not **prove** H_0 , only fail to reject H_0 .

Basic Idea of the Hosmer-Lemeshow Test

- Groups are created based on **predicted probability of the event** ($Y = 1$) for each observation, \hat{p}
- Generally, observations are divided into deciles or 10 groups based on the range of \hat{p} . For example, $n = 100$ observations are divided into $g = 10$ groups each containing $n_j = 10$ observations, $j = 1, \dots, g$.
- Compare observed and expected counts of successes and failures in those groups using a chi-squared statistic

R Code, Details of Hosmer-Lemeshow Test

```
> HL <- hoslem.test(fhssrs$CVD, fitted(mod8))
# Obs. and expected no. of failures/successes
> cbind(HL$observed, HL$expected)
```

	y0	y1	yhat0	yhat1
[0.0178,0.0799]	9	1	9.449264	0.5507364
(0.0799,0.099]	10	0	9.091278	0.9087216
(0.099,0.123]	9	1	8.892376	1.1076236
(0.123,0.163]	8	2	8.559780	1.4402201
(0.163,0.193]	9	1	8.184674	1.8153257
(0.193,0.233]	7	3	7.833206	2.1667936
(0.233,0.278]	6	4	7.429639	2.5703605
(0.278,0.333]	8	2	6.998297	3.0017027
(0.333,0.448]	6	4	6.199088	3.8009116
(0.448,0.691]	5	5	4.362396	5.6376043

Hosmer-Lemeshow Test: Example

R Code, Hosmer-Lemeshow Test

```
# Model being evaluated
mod8 <- glm(CVD ~ DIABP + AGE, data = fhssrs, family = binomial(link="logit"))

# Load required package
> library(ResourceSelection)

# Run Hosmer-Lemeshow test: input Y-variable and fitted probabilities from mod8
> HL <- hoslem.test(fhssrs$CVD, fitted(mod8))
> HL

Hosmer and Lemeshow goodness of fit (GOF) test

data: fhssrs$CVD, fitted(mod8)
X-squared = 2.377, df = 8, p-value = 0.9672
```

- Do not have evidence to reject H_0 that the model fits the data well (p -value = 0.97)

Calibration vs. Discrimination

- Hosmer-Lemeshow test is used to determine if the fitted model is well **calibrated** (i.e., good agreement between observed risk and predicted (fitted) probability)
- A **Receiver Operating Characteristic (ROC) curve** is used to assess the **discrimination** of the fitted model (i.e., how well the model is able to distinguish between those who experience a success/positives (develop CVD) and those who do not/negatives)

ROC Curve

- Basic idea of ROC curve:
 - Fit the logistic regression model
 - Calculate all fitted probabilities
 - Choose a cutoff probability value
 - Classify all predicted values above the cutoff as predicting an event (predicted positive) and below the cutoff as not predicting the event (predicted negative)
 - Compare observed positive/negatives to predicted positives/negatives

Sensitivity and Specificity: ROC Curve

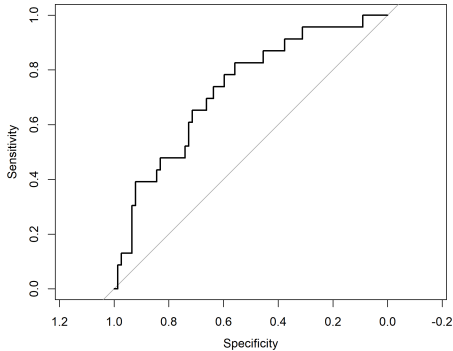
- ROC curves use **sensitivity** and **specificity**
- **Sensitivity** is proportion of truly positive observations that are classified as a positive by the model (probability model predicts positive when the observation is positive ($Y = 1$))
- **Specificity** is the proportion of truly negative observations that are classified as negative by the model (probability model predicts negative when the observation is negative ($Y = 0$))

Sensitivity and Specificity: ROC Curve

	Observed Positive	Observed Negative
Predicted Positive ($>$ cutoff)	a	b
Predicted Negative ($<$ cutoff)	c	d

- **Sensitivity** = $\frac{a}{a + c}$: True positive rate
- **Specificity** = $\frac{d}{b + d}$: True negative rate
- High sensitivity and specificity indicate better fit of the model
- Each predicted probability in the data is used as a cutoff value and each sensitivity vs. 1-specificity is plotted, giving a **ROC curve**

ROC Curve: Example



R Code, ROC Curve

```
# Load required package
> library(pROC)

# Fitted model probabilities for each observation
> pred.p <- predict(mod8, type = "response")

# Response variable ~ predicted probabilities
> roccurve <- roc(fhssrs$CVD ~ pred.p)

# Plot ROC curve
> plot(roccurve)
```

ROC Curve: Example

R Code, Underlying Sensitivities and Specificities

```
> cbind(cutpoints = roccurve$thresholds, specificity = roccurve$specificities,
        sensitivity = roccurve$sensitivities)
      cutpoints specificity sensitivity
[1,]      -Inf  0.00000000  1.00000000
[2,]  0.03044901  0.01298701  1.00000000
[3,]  0.04507084  0.02597403  1.00000000
[4,]  0.04809755  0.03896104  1.00000000
...
[50,]  0.19285279  0.57142857  0.78260870
[51,]  0.19305031  0.58441558  0.78260870
...
[98,]  0.59581603  0.98701299  0.08695652
[99,]  0.65046382  0.98701299  0.04347826
[100,] 0.67422798  0.98701299  0.00000000
[101,]      Inf  1.00000000  0.00000000
```


AUC of the ROC Curve

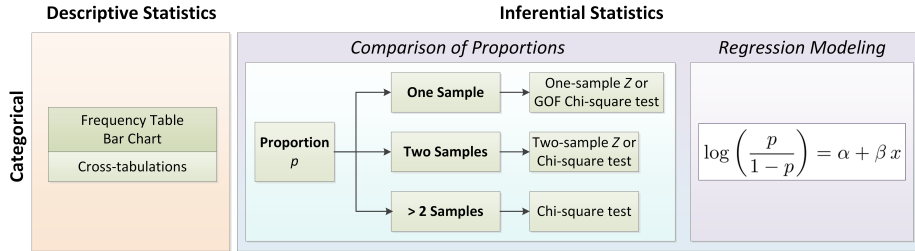
- The area under the ROC curve (AUC of the ROC) can be used to evaluate the discrimination of the model; this value is called the *c-statistic*
- *Discrimination* is a measure of how well the model is able to distinguish between those who experience the success (develop CVD) and those who do not
 - $c = 0.5$: no discrimination
 - $0.7 \leq c < 0.8$: acceptable discrimination
 - $0.8 \leq c < 0.9$: excellent discrimination
 - $0.9 \leq c < 1$: outstanding discrimination
 - $c = 1$: perfect discrimination

R Code, C-statistic

```
# C-statistic  
> auc(roccurve)  
Area under the curve: 0.7307
```

Lesson Summary

- When Y is **binary**, interested in estimating **mean** of the response, p
- Interested in the relationship between regressors and the log-odds



Estimated Probability of Success

$$\hat{p} = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Simple Logistic Regression Model

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = a + bx$$