

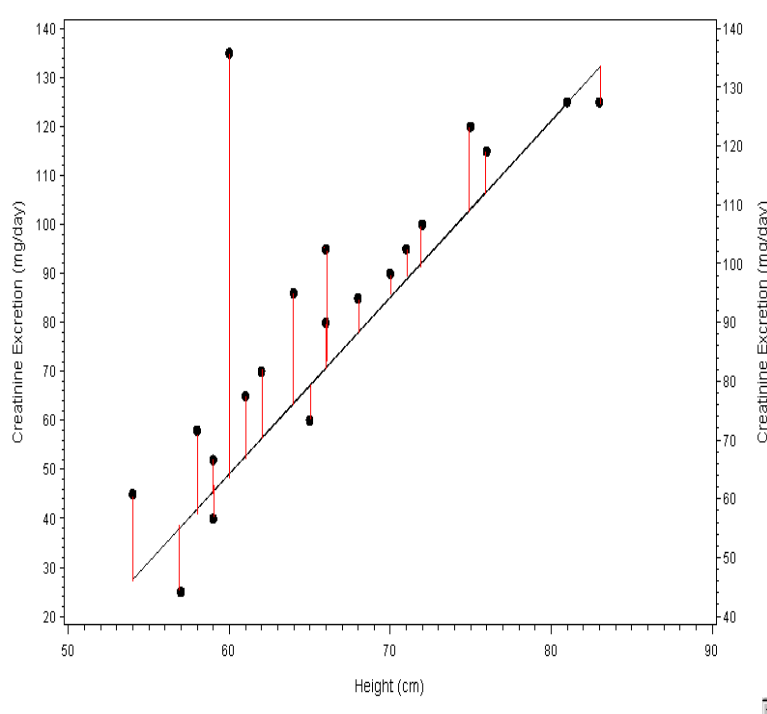
Topics to be Covered

- **Logistic regression**
 - A single predictor
 - Multiple predictors
 - Confounder and effect Modifier
 - ROC curve
 - Model diagnosis

Borrow Ideas from Linear Regression

- Recall that linear regression allowed us to assess the association of multiple variables (predictors), each of which was continuous, categorical, or binary, simultaneously with a continuous outcome variable
- We would like to apply such an approach to our situation in which the outcome is now binary
 - Regression used with a binary outcome is called **logistic regression**
- However, there are some modifications we have to make in order for logistic regression to “make sense”

Review: Simple Linear Regression



Given a scatterplot of x vs. y

$$y = b_0 + b_1x + \text{Error}$$

Find the **best-fit line**,

$$\mu = b_0 + b_1x$$

Choose b_0 and b_1 that can minimize the sum of errors

Simple Logistic Regression

- Outcome, y , takes 0 or 1
- What's the distribution of y ?

$$y \sim \text{Binomial}(1, p)$$

If we model $p = \beta_0 + \beta_1 x$

- We fit the following model for $p = \text{probability of death}$:

$$\begin{aligned} p &= \beta_0 + \beta_1 FTB \\ &= 0.50 + 0.01(FTB) \end{aligned}$$

- If $FTB = 20$, then

$$p = 0.5 + 0.01(20) = 0.7 \quad \checkmark$$

- If $FTB = 60$, then

$$p = 0.5 + 0.01(60) = 1.1 \quad X$$

Solution to the Challenge

- **Solution:** We can transform the probability p and use this transformation as the dependent variable so that the transformed p can take values beyond $[0,1]$.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log \text{ odds}$$

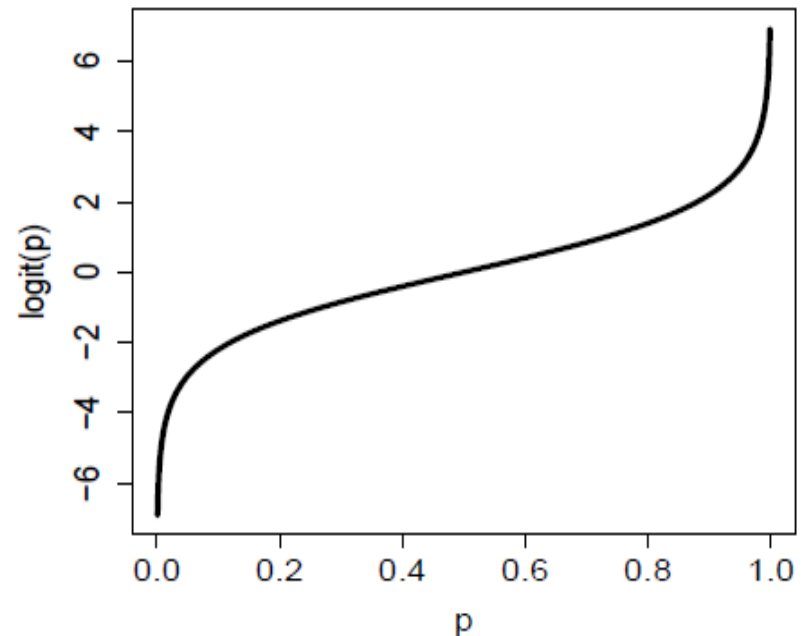
$$-\infty < \text{logit}(p) < \infty$$

- The logit transformation is the most commonly used transformation in binary data regression, because it leads to a model for the odds ratio (at least on the log scale).

Probability, Odds, Logit Transformation

Probability (p)	Odds (p/(1-p))	Log(odds)
0.1	0.11	-2.2
0.25	0.33	-1.1
0.5	1	0
0.75	3	1.1
0.9	9	2.2

p increases → odds increases
→ logit p increases



Linear vs. Logistic Regression

- Linear regression model:

1. $Y \sim N(\mu, \sigma^2)$
2. $\mu = \beta_0 + \sum_{j=1}^r \beta_j X_j$: mean of Y (μ) is a linear function of β 's and X 's

- Logistic regression model:

1. $Y \sim \text{Bin}(1, p)$
2. $\text{logit}(p) = \beta_0 + \sum_{j=1}^r \beta_j X_j$: logit transformation of the disease probability p is a linear function of β 's and X 's

Logistic Regression with a Binary Predictor

Re-analyze it Using Logistic Regression

- A study by Pauling (1971) randomized subjects to either Vitamin C or placebo
- Subjects were followed for a specific period of time to see if they developed a cold
- The results of the study were:

Vitamin C	Developed Cold		Total
	Yes	No	
Yes	17	122	139
No	31	109	140
Total	48	231	279

Here, Cold is the Disease and Vitamin C is the Exposure

Logistic Regression with a Binary Predictor

- If we let

X_E = 1 if exposed, 0 otherwise

Y = 1 if diseased, 0 otherwise

$$p = P(Y = 1 | X_E)$$

id	X_E	y
1	1	1
2	0	1
3	1	0
...
50	0	1

then our regression model is:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_E$$

Interpret β for a Binary Predictor

The regression model is:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_E$$

- From this model, we see:

$$\log\left(\frac{p_0}{1-p_0}\right) = \beta_0 \quad \text{when } X_E = 0 \text{ (non-exposed)}$$

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 \quad \text{when } X_E = 1 \text{ (exposed)}$$

✓ β_0 is log odds for non-exposed

Let's take a difference between exposed and non-exposed:

$\bullet \log\left(\frac{x}{y}\right) = \log(x) - \log(y)$

$e^{\log(x)} = x$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}\right) = \cancel{\beta_0} + \beta_1 - \cancel{\beta_0}$$

✓ β_1 is log OR for exposed vs non-exposed (e^{β_1} is the OR)

Test β in Logistic Regression

- To assess whether or not exposure is related to disease, we have the hypothesis test

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

because when $\beta_1 = 0$, the odds ratio of disease for exposed relative to non-exposed is 1.

- As stated several times, our test statistic takes the familiar form

$$t^2 = \frac{(\hat{\beta}_1 - 0)^2}{\widehat{\text{Var}}(\hat{\beta}_1)},$$

which is also known as a **Wald test**, and has a chi-squared test with 1 df under the null hypothesis

Analysis Using R

```
> res=glm(outcome ~ trmt, family=binomial, data=cold)
> summary(res)
```

Call:

```
glm(formula = outcome ~ trmt, family = binomial, data = col
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.7075	-0.7075	-0.5108	-0.5108	2.0500

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.2574	0.2035	-6.177	6.53e-10 ***
trmt1	-0.7134	0.3293	-2.166	0.0303 *

$$OR = \exp(-0.7134) = 0.49$$

```
> # For odds ratio
> exp(res$coefficients)
```

(Intercept)	trmt1
0.2844037	0.4899524

```
>
```

```
> # for confidence intervals
```

```
> exp(confint(res))
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	0.1877551	0.4181998
trmt1	0.2523178	0.9240466

Vitamin C	Developed Cold		Total
	Yes	No	
Yes	17	122	139
No	31	109	140
Total	48	231	279

OR using 2 by 2 table:

$$(17 \times 109) / (31 \times 122) = 0.49$$

Estimate of trmt1 (β_1) > 0 then OR > 1
 Estimate of trmt1 (β_1) < 0 then OR < 1
 Estimate of trmt1 (β_1) = 0 then OR = 1

Logistic Regression with a Continuous Predictor

Interpret β for a Continuous Predictor

- Linear regression model :

$$\mu = \beta_0 + \beta_1 X$$

- Interpretation of β_1 : the change in mean response associated with a one-unit change in X

- Logistic regression model:

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

- Interpretation of β_1 : the change in $\text{logit}(p)$ associated with a one-unit change in X

Example- Coronary Heart Disease

- Consider the data in Hosmer and Lemeshow's text collected on 100 subjects examining the association of age with coronary heart disease (CHD)
- Here is an excerpt of the data:

id	age	chd
1	20	0
2	23	0
3	24	0
4	25	0
.	.	.
.	.	.
.	.	.
97	64	0
98	64	1
99	65	1
100	69	1

Example: Interpret β

- $\log \text{Odds}(\text{chd} | \text{age}_i) = -5.3095 + 0.1109 \times \text{age}_i$
- Compute the log odds for a subject with age= 25 and 26:
- $\log \text{Odds}(\text{chd} | \text{age}_i = 25) =$
- $\text{Log Odds}(\text{chd} | \text{age}_i = 26) =$
- $\text{LogOdds}(\text{chd} | \text{age}_i = 26) - \text{LogOdds}(\text{chd} | \text{age}_i = 25) =$
- $OR = \frac{\text{Odds}(\text{chd} | \text{age}_i=26)}{\text{Odds}(\text{chd} | \text{age}_i=25)} =$

Conclusion:

Example: Interpret β

- $\log \text{Odds}(\text{chd} | \text{age}_i) = -5.3095 + 0.1109 \times \text{age}_i$
- Compute the log odds for a subject with age= 25 and 26:
- $\log \text{Odds}(\text{chd} | \text{age}_i = 25) = -5.3095 + 0.1109 \times 25$
- $\log \text{Odds}(\text{chd} | \text{age}_i = 26) = -5.3095 + 0.1109 \times 26$
- $\log \text{Odds}(\text{chd} | \text{age}_i = 26) - \log \text{Odds}(\text{chd} | \text{age}_i = 25) = 0.1109$
- $OR = \frac{\text{Odds}(\text{chd} | \text{age}_i=26)}{\text{Odds}(\text{chd} | \text{age}_i=25)} = \exp(0.1109) = 1.117$

Conclusion: the odds of CHD was **increased** by **11.7%** for every one year increase in age.

Summary: Logistic Regression with One Predictor X

Continuous

id	age	y
1	10	1
2	20	1
3	31	0
...
50	60	1

β_1 is the change in $\text{logit}(p)$ (i.e., **logOR**) for every unit increase in X;
 β_0 is log odds for age=0

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

Binary

id	Smoker	y
1	Y	1
2	N	1
3	Y	0
...
50	N	1

1) Set the reference category
 2) β_1 is the **logOR** comparing the non-reference to the reference category; β_0 is log odds for the reference category

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

Categorical (4 levels)

id	stage	y
1	I	1
2	II	1
3	III	0
...
50	IV	1

1) Create 3 dummy variables
 2) Set the reference category
 3) β_1, β_2 and β_3 represent the **logOR**, each compares a category to the reference category; β_0 is log odds for the reference category

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Multiple Logistic Regression

Two binary Predictors - Framingham study

- We focus upon data collected from the Framingham Study, a well-known observational study of the epidemiology of cardiovascular disease
- Outcome variable: $Y = \begin{cases} 1 & \text{CHD} \\ 0 & \text{no CHD} \end{cases}$
- Exposure of interest: $X_{SBP} = \begin{cases} 1 & SBP \geq 165 \\ 0 & SBP < 165 \end{cases}$
- Potential confounder: $X_{Age} = \begin{cases} 1 & Age \geq 55 \\ 0 & Age < 55 \end{cases}$

Interpret Coefficients in Framingham Study

- Logistic regression model:

$$\text{logit}(p) = \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age}$$

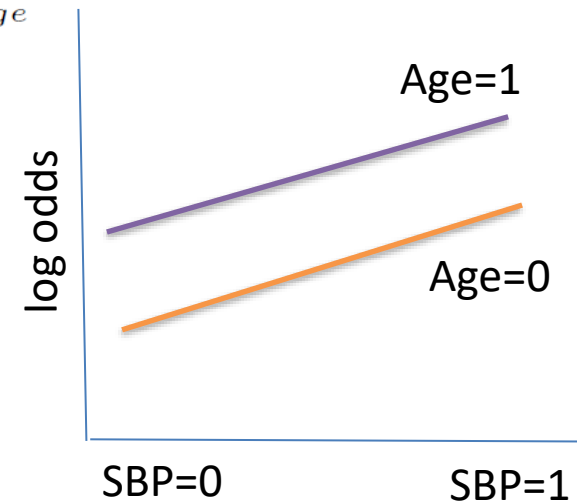
- How to interpret β_1 ?

Given $X_{sbp}=0$ and $X_{age}=x_2$. The **log odds** is:

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_2 x_2$$

Given $X_{sbp}=1$ and $X_{age}=x_2$. The **log odds** is:

$$\log\left(\frac{p_2}{1-p_2}\right) = \beta_0 + \beta_1 + \beta_2 x_2$$



- Thus, β_1 is the log OR while holding age as constant (**adjusting** for age in the model),

(Same interpretation when age has >2 categories or is continuous)

Analysis Results

Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Standard Error	Wald Chi-Square	P value
Intercept	-0.6758	0.1404	23.1851	<.0001
age	0.3257	0.2118	2.3641	0.1242
sbp	0.8850	0.2619	11.4156	0.0007

$$\text{logit}(p) = -0.68 + 0.89 \times \text{sbp} + 0.33 \times \text{age}$$

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age 1 vs 0	1.385	0.914	2.098
sbp 1 vs 0	2.423	1.450	4.049

Conclusion: sbp>165 has higher odds of CHD compared to sbp<165 (OR=2.42, 95% CI: 1.45, 4.05, p=0.0007)

Summary: Interpret Coefficients in Regression with Multiple Variables

- In general, a model with multiple predictors (binary, categorical or continuous) , β_j for predictor j is the **log OR** while holding all other predictors as constant (**adjusting** for all other predictors in the model)

Assessing Confounding - Framingham study

Is age a confounder? We compare β_1 in (1) vs in (2)

$$(1) \quad \text{logit}(p) = \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age} \quad \text{Adjusted estimate}$$

$$(2) \quad \text{logit}(p) = \beta_0 + \beta_1 X_{SBP} \quad \text{Crude estimate}$$

- Estimates for β_1 are **0.885** in (1) vs **0.905** in (2)
- OR are **2.423** in (1) vs **2.472** in (2)
- Is age a confounder? use the 10% rule for the OR

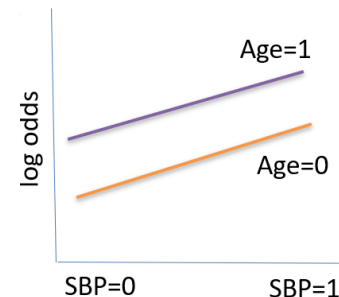
Interaction/Effect Modification (hard; optional)

Framingham Study with Two Binary Predictors

- Logistic regression models for $p = Pr(Y = 1)$

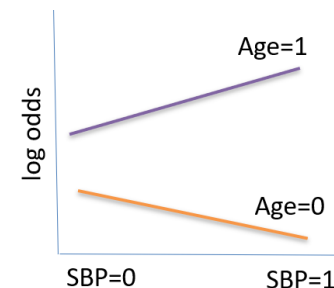
- Without interaction:

$$\text{logit}(p) = \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age}$$



- With interaction:

$$\text{logit}(p) = \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age} + \beta_3 X_{SBP} X_{Age}$$



Why β_3 measures the effect of the interaction?

$$X_{SBP} = \begin{cases} 1 & SBP \geq 165 \\ 0 & SBP < 165 \end{cases}$$

$$X_{Age} = \begin{cases} 1 & Age \geq 55 \\ 0 & Age < 55 \end{cases}$$

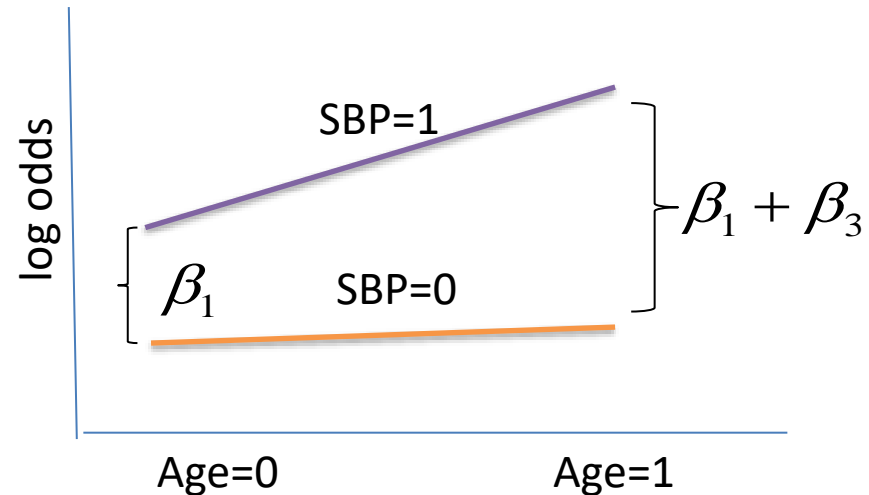
$$\begin{aligned} \text{logit}(p) = & \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age} \\ & + \beta_3 X_{SBP} X_{Age} \end{aligned}$$

Given $X_{sbp}=0$ and $X_{age}=0$, the **log odds** is

Given $X_{sbp}=1$ and $X_{age}=0$, the **log odds** is:

Given $X_{sbp}=0$ and $X_{age}=1$, the **log odds** is:

Given $X_{sbp}=1$ and $X_{age}=1$, the **log odds** is:



$$\log\left(\frac{p_{00}}{1-p_{00}}\right) = \beta_0$$

$$\log\left(\frac{p_{10}}{1-p_{10}}\right) = \beta_0 + \beta_1$$

$$\log\left(\frac{p_{01}}{1-p_{01}}\right) = \beta_0 + \beta_2$$

$$\log\left(\frac{p_{11}}{1-p_{11}}\right) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

Interpret Regression Coefficients in the Framingham Study

- Logistic regression model w/o interaction:

$$\text{logit}(p) = \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age}$$

β_1 : log OR for SBP ≥ 165 vs SBP < 165 after adjusting for age

- Logistic regression model with interaction:

$$\text{logit}(p) = \beta_0 + \beta_1 X_{SBP} + \beta_2 X_{Age} + \beta_3 X_{SBP} X_{Age}$$

β_1 : log OR for SBP ≥ 165 vs SBP < 165 among men with age < 55

- What's the log OR for SBP ≥ 165 vs SBP < 165 among men > 55 ?

$$\beta_1 + \beta_3$$

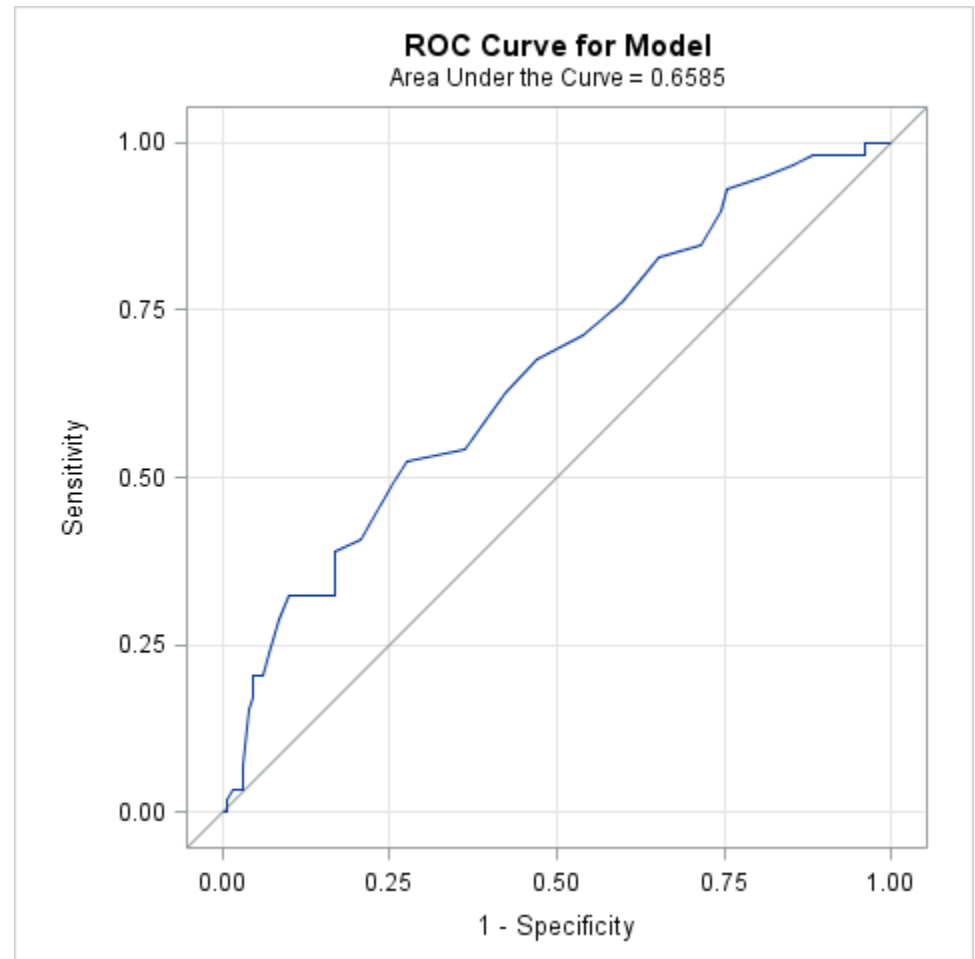
Remarks on the Interaction/Effect Modification

- If the interaction is statistically significant, must keep the main effects (terms that make up the interaction) in the model as well
- For an interaction between categorical variable and continuous variable, we would need terms in the model for the interaction between the dummy variables with the continuous variable.
- When there is an interaction in the model, there is no real useful interpretation for the corresponding main effects

ROC Curve in Logistic Regression

ROC Curve from Logistic Regression

- **ROC** (receiver operating characteristic), or ROC curve, illustrates the performance of a binary classifier system as its discrimination threshold is varied.
- It is between 0.5 and 1
- The closer it is to 1, the better the discrimination ability




ROC Curve from Logistic Regression

$$\text{sensitivity(TPF)} = \text{Prob}(\textcolor{teal}{T} = 1 \mid \textcolor{red}{D} = 1)$$

$$\text{specificity(1-FPF)} = \text{Prob}(\textcolor{teal}{T} = 0 \mid \textcolor{red}{D} = 0)$$

Dichotomize phat based on
threshold = 0.32358



id	AGE	WT	LOW	phat	Test
1	19	1	0	0.32358	1
2	33	1	1	0.13571	0
3	20	0	0	0.47076	1
4	21	0	0	0.48389	1
5	18	0	0	0.44464	1
6	21	1	0	0.28977	0
7	22	1	1	0.27367	0
8	17	0	0	0.43169	0
9	29	1	1	0.17754	0
10	26	1	1	0.21511	0

Calculate
TPF and
FPF

ROC curve is TPF vs FPF
across all thresholds

Threshold	TPF	FPF
0.32358	0.32	0.17
0.47076	0.36	0.17
0.48389	0.39	0.17
0.44464	0.41	0.21
0.28977	0.49	0.25
0.27367	0.53	0.28
0.43169	0.63	0.42
0.47076	0.68	0.47

$$\text{logit}(p) = \beta_0 + \sum_{j=1}^r \beta_j X_j \longrightarrow \hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^r \hat{\beta}_j X_j}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^r \hat{\beta}_j X_j}}$$

Model Diagnosis

Hosmer-Lemeshow Test

- First, divide the data into ~ 10 groups based on the percentiles of the fitted values \hat{p}_i , i.e. the first group is subject with $\hat{p}_i \leq 0.10$, etc.
 - For each of the 10 groups, compute
 - O_i , the number of observed events
 - E_i , the number of expected events under H_0
- E_i is the sum of the predicted probabilities for all the individuals in group i .

- Our test statistic is the usual Pearson χ^2 statistic

$$T = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{\text{Var}(O_i)}$$

- Under the null hypothesis that the fitted model is correct, T has a chi-squared distribution with $10-2=8$ df

Some Notes about Hosmer-Lemeshow Test

- Note that the use of 10 groups is arbitrary; any number of groups N could be used, i.e. quintiles ($N=5$); the appropriate df is always $N - 2$
- There must be at least three groups in order for the Hosmer-Lemeshow statistic to be computed.
- The validity of the H-L test has been questioned by some authors, but still appears to be used frequently