CS555 Data Analysis and Visualization

Lecture 12

Simple Logistic Regression,

Multiple Logistic Regression, ROC Curve

Kia Teymourian

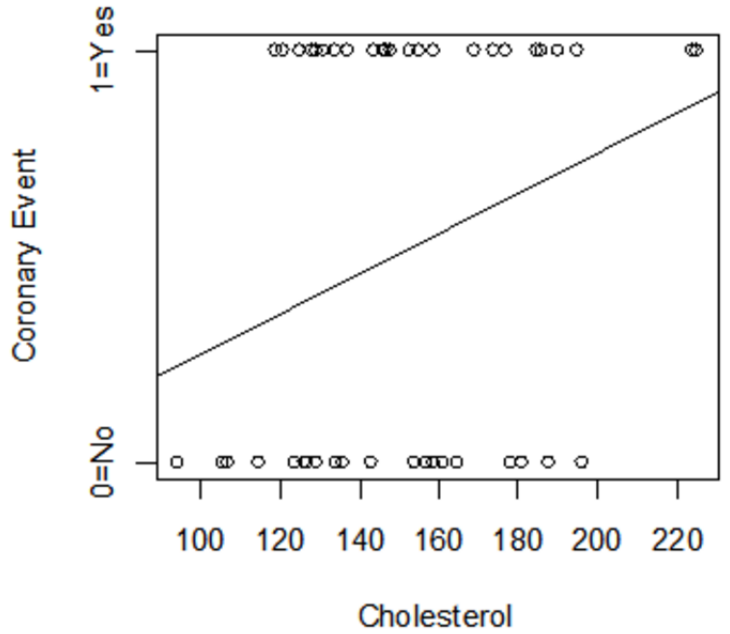
Logistic Regression

In the linear regression setting, we model the relationship between one or more explanatory variables and a continuous response variable.

If the **outcome of interest is dichotomous in nature** (taking on one of two values), then linear regression is not appropriate.

Suppose we are interested in the association between **cholesterol levels** and having a coronary event in a high risk patient population (who have had an event in the past). We collect cholesterol data for 50 subjects and then follow each for a year to see if they have another coronary event. In this case, our explanatory variable is cholesterol level and our outcome is **whether or not the subject had another coronary event**.

Many of the assumptions required for regression are not met in this setting (for example, linearity, constant variance and normality of the residuals), so inference is not valid.



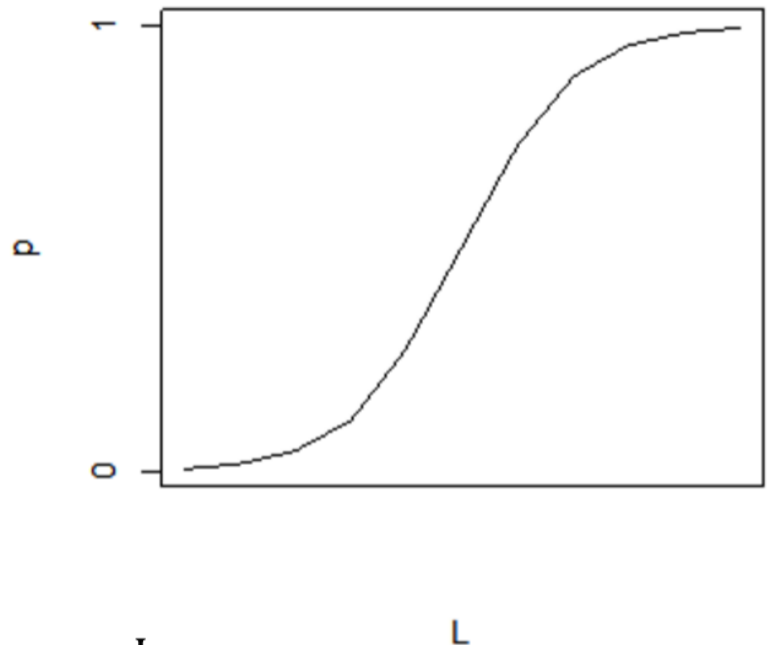
Simple Logistic Regression

The simple logistic regression model is based on a linear relationship between the natural logarithm (ln) of the odds of an event and a continuous explanatory variable.

The equation for the simple logistic regression line is

$$L = ln\frac{p}{1-p} = \beta_0 + \beta_1 x + \epsilon$$

where
L is the log odds of the event
p is the probability of a success
$\frac{p}{1-p}$ are the odds of the event
x is the explanatory variable
β_0 is the intercept
β_1 is the regression coefficient
ϵ is the random error



If we solve the above regression equation for p, we find $p = \frac{e^{\beta_0+\beta_1 x+\epsilon}}{1+e^{\beta_0+\beta_1 x+\epsilon}} = \frac{e^{\mathrm{L}}}{1+e^{\mathrm{L}}}$

This is called the logistic function.

Interpretation

Using the estimates of the regression coefficients, $\widehat{\beta_0}$ and $\widehat{\beta_1}$, we can predict the risk of the outcome of interest for a given value of x using the following equation:

$$\hat{p} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x+\epsilon}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}x+\epsilon}}$$

Interpretation

Using the estimates of the regression coefficients, $\widehat{\beta_0}$ and $\widehat{\beta_1}$, we can predict the risk of the outcome of interest for a given value of x using the following equation:

$$\hat{p} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x+\epsilon}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}x+\epsilon}}$$

Though the regression equation can be used to predict risk of the event, the interpretation of the regression coefficient(s) are generally based on odds ratios.

Consider the odds ratio of an event for a given value of $x=x_a$ versus a given value of $x=x_b$.

The estimated odds for a given value of $x=x_a$ is given by $\widehat{odds_a} = e^{\widehat{\beta_0}+\widehat{\beta_1}x_a}$
The estimated odds for a given value of $x=x_b$ is given by $\widehat{odds_b} = e^{\widehat{\beta_0}+\widehat{\beta_1}x_b}$
The odds ratio then is given by

$$\widehat{OR}\ x_a\ _{versus} x_b = \frac{\widehat{odds_a}}{\widehat{odds_b}} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x_a}}{e^{\widehat{\beta_0}+\widehat{\beta_1}x_b}} = e^{\widehat{\beta_0}+\widehat{\beta_1}x_a-\widehat{\beta_0}-\widehat{\beta_1}x_b} = e^{\widehat{\beta_1}(x_a-x_b)}$$

Interpretation: The odds of the event are $e^{\widehat{\beta_1}(x_a-x_b)}$ higher for every $x_a - x_b$ unit increase in x. Note that the interpretation here depends only on the difference in x values as opposed to their actual values.

Confidence Interval

Confidence intervals for the logistic regression setting are based on the odds ratio. The two-sided 100% × (1−α) confidence interval for $\widehat{OR}$ x_a $_{versus}$ x_b is:

$$e^{(\widehat{\beta_1} \pm z_{\frac{\alpha}{2}} \cdot SE_{\widehat{\beta_1}})(x_a - x_b)}$$

Where $SE_{\widehat{\beta_1}}$ is the standard error of the regression coefficient and $z_{\frac{\alpha}{2}}$ is the value from the standard normal distribution with a right tail probability of α/2.

An Example: logistic regression

We are interested in the association between **cholesterol levels** and **having a coronary event** in a high-risk patient population (who have had an event in the past). We collect cholesterol data for 50 subjects and then follow each for a year to see if they have another coronary event.

In this case, our explanatory variable is cholesterol level and our outcome is whether or not the subject had another coronary event.

Given the nature of our response variable, we perform a logistic regression. A summary of the beta estimates from the model are shown below.

Use these to

1) **predict the risk of another coronary event** for a high risk patient with a cholesterol level of 190.

2) **calculate the odds ratio** for a coronary event of a high-risk patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180.

3) **calculate 95% confidence interval for the odds ratio** of having a coronary event for a patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180.

An Example: logistic regression

Parameter	Estimate	Standard Error	p-value
beta_0	-3.725	1.753	0.0336
beta_1	0.024	0.012	0.0420

The risk of having a coronary event for a patient with a cholesterol level of 190 is predicted by :

$$\hat{p} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}x}} = \frac{e^{-3.725+0.024*190}}{1+e^{-3.725+0.024*190}} = \frac{e^{0.835}}{1+e^{0.835}} = 0.697$$

The odds ratio of having a coronary event for a patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180 is

$$\widehat{OR}\ x_{a\ versus} x_b = e^{\widehat{\beta_1}(x_a - x_b)} = e^{0.024\cdot(190-180)} = e^{0.24} = 1.27$$

The odds of having a coronary event are 1.27 times higher for every 10-unit increase in cholesterol level. The odds ratio comparing any two individuals with cholesterol levels which are 10 units apart are the same.

The quantity $e^{\widehat{\beta_1}}$ is the odd ratio of the event for two individuals with x values that are 1 unit apart. In other words, $e^{\widehat{\beta_1}}$ is the relative increase in odds for every 1 unit increase in x.

An Example: logistic regression

The 95% confidence interval for the odds ratio of having a coronary event for a patient with a cholesterol level of 190 versus a patient with a cholesterol level of 180 is:

$$e^{(\widehat{\beta_1} \pm z_{\frac{\alpha}{2}} \cdot SE_{\widehat{\beta_1}})(x_a - x_b)} = e^{(0.024 \pm 1.96 \cdot 0.012)(190-180)} = (e^{0.0048}, e^{0.475}) = (1.004, 1.608)$$

We are 95% confident that the odds of having a coronary event are between 1.004 and 1.608 times higher for every 10-unit increase in cholesterol level.

R commands: Generalized linear models (GLMs)

- Use the glm() function with binomial option

  - GLMs extend the linear modeling capability of R to scenarios that involve non-normal error distributions. The idea is to obtain linear functions of the predictor variables by transforming the right side of the equation by a link function.
  - "family" parameter is a simple way of specifying a choice of variance and link functions When family is set to binomial, it tells R to perform logistic regression.
  - **glm(data$event~data$explanatory1 + data$explanatory2 + …, family=binomial)**
  - Ensure that you event is coded: 1 = Event and 0= non-event (numeric, not a factor variable)
  - If one of the variables in the model is a factor variable, it is best to create dummy variables (1/0) so that you know exactly what the reference group is

Error family	Default link	Inverse of link	Use for:
gaussian	identity	1	normally distributed error
poisson	log	exp	counts (many zeros, various integers)
binomial	logit	1/(1+1/exp(x))	proportions or binary (0,1) data
Gamma	inverse	1/x	continuous data with non-constant error (constant CV)

http://plantecology.syr.edu/fridley/bio793/glm.html

Logistic Regression: R commands

- Use the glm() function with binomial option
  - **glm(data$event~data$explanatory1 + data$explanatory2 + …, family=binomial)**

- Use the **summary() function** on the saved regression result to get regression equation and associated rests for each regression coefficient

- Use the **exp() function**, which computes the exponential value of a number e^x, on the resulting coefficients to obtain odds ratios for each regression coefficient

- Use the **predict() function** on the saved regression result to get the predicted risks for each observations

- In **multiple logistic regression**, use the **wald.test()** function (from "**aod**" package) to get p value for the global test (of all beta coefficients = 0)

- Use the **roc() function (from "pROC" package"** to get c-statistic (area under the curve) to generate the ROC curve

Logistic Regression: R commands

```
> #read in data
> data <- read.csv('cevent.csv')
> #Simple logistic regression
> m <- glm(data$event ~ data$chol, family=binomial)
> summary(m)

Call:
glm(formula = data$event ~ data$chol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5752  -0.9629  -0.7217   1.1418   2.1732

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.72518    1.75307  -2.125   0.0336 *
data$chol    0.02359    0.01160   2.034   0.0420 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 68.593  on 49  degrees of freedom
Residual deviance: 63.859  on 48  degrees of freedom
AIC: 67.859
Number of Fisher Scoring iterations: 4
```

Logistic Regression: R commands

```
> #OR per 1 unit increase
> exp(cbind(OR = coef(m), confint.default(m)))
                  OR      2.5 %   97.5 %
(Intercept) 0.02410882 0.000776183 0.748838
data$chol   1.02387495 1.000858995 1.047420

> #OR per 10 unit increase
> exp(m$coefficients[2]*10)
data$chol
 1.266103

> exp((m$coefficients[2]-qnorm(0.975)*summary(m)$coefficients[2,2])*10)
data$chol
 1.008623

> exp((m$coefficients[2]+qnorm(0.975)*summary(m)$coefficients[2,2])*10)
data$chol
 1.589313
```

Inference

Formal inference in the simple logistic regression framework involves considering β_1 as an unknown population parameter and determining what we can or can't say about the unknown population parameters given the data we observed from our sample (using estimates of this parameter from our sample, $\widehat{\beta_1}$).

H_0: $\beta_1=0$ (H_0 : there is no association between x and odds of the outcome)
H_1: $\beta_1\neq0$ (H_1 : there is an association between x and odds of the outcome)

Note that $\beta_1=0$ is equivalent to that the regression line had a slope of 0 (it would be a horizontal line), $\beta_1=0$ means OR=$e^{\beta_1}=1$. That is, the null hypothesis $\beta_1=0$ is equivalent to the test of the odds ratio for a 1-unit increase in x being equal to 1 (H_0 :OR=1).

H_0: $\beta_1=0$ or OR=1 is rejected if $\widehat{\beta_1}$ is sufficiently far from 0. That is, we reject the claim that the population parameter β_1 is equal to 0 if $\widehat{\beta_1}$, the sample statistic, is far from 0.

An example: logistic regression inference

Formally test whether or not cholesterol is associated with risk of a coronary event at the α=0.05 level.

1. Set up the hypotheses and select the alpha level
H_0: $\beta_1=0$ or OR=1 (there is no association between cholesterol levels and risk for a coronary event)
H_1 :$\beta_1 \neq 0$ or OR ≠ 1 (there is an association between cholesterol levels and risk for a coronary event)
α=0.05

2. Select the appropriate test statistic $z = \frac{\beta_1}{SE_{\widehat{\beta_1}}}$

3. State the decision rule
- Determine the appropriate value from the standard normal distribution associated with a right hand tail probability of α/2=0.05/2=0.025
- Using the table, $z_{\frac{\alpha}{2}}$=1.960
- Decision Rule: Reject H_0 if $|z| \geq 1.960$ or Reject H_0 if $p \leq \alpha$
- Otherwise, do not reject H_0

An example: logistic regression inference

4. Compute the test statistic

$z = \frac{\beta_1}{SE_{\widehat{\beta_1}}} = \frac{0.024}{0.0116} = 2.069$

5. Conclusion

Reject H_0 since $z \geq 1.960$ or since $\leq \alpha$. We have significant evidence at the $\alpha = 0.05$ level that $\beta_1 \neq 0$. That is, there is evidence of an association between cholesterol level and risk of a coronary event.

The odds ratio for a coronary event is $e^{\beta_1} = 1.02$ for every 1-unit increase in cholesterol. (Or we could say that the odds ratio is 1.27 for every 10-unit increase in cholesterol as this may be a more reasonable and clinically relevant scale to report the results). We are 95% confident that the true odds ratio is between 1.00 and 1.047. (We could also report the 95% confidence interval for the 10-unit increase instead if we had chosen to present the odds ratio in the previous sentence based on this unit of increase).

R commands: Predict Method for GLM Fits

```
# predicted risk for each patient
risk <- predict(m, type=c("response"))
risk

 1         2         3         4         5         6         7
0.22720323 0.43615030 0.34653363 0.31520904 0.23137263 0.48299515 0.59393290
       8         9        10        11        12        13        14
0.44196119 0.49478598 0.47122323 0.70593666 0.61088437 0.41309669 0.18133869
      15        16        17        18        19        20        21
0.09429165 0.48299515 0.20715784 0.62757192 0.36273154 0.47710601 0.44778802
...
```

The parameter **“type”** indicates the type of prediction required. The default is on the scale of the linear predictors; the alternative **"response"** is on the scale of the response variable. Thus for a default binomial model the default predictions are of log-odds (probabilities on logit scale) and type = "response" gives the predicted probabilities.

```
# predicted risk for patient with cholesterol of 190
risk[41]
 41
0.6808668

> exp(m$coefficients[1]+m$coefficients[2]*190)/(1+exp(m$coefficients[1]+m$coefficients[2]*190))
(Intercept)
  0.6808668
```

Multiple Logistic Regression

The multiple logistic regression model is based on a linear relationship between the natural logarithm (ln) of the odds of an event and a linear combination of explanatory variable.

The equation for the full logistic regression line is given by

$$L = ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where
L is the log odds of the event
p is the probability of a success
$\frac{p}{1-p}$ are the odds of the event
$x_1, x_2, \ldots, x_k$ are the explanatory variables
β_0 is the intercept
$\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients
ϵ is the random error

If we solve the above regression equation for p, we find $p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\epsilon}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\epsilon}} = \frac{e^{\mathrm{L}}}{1+e^{\mathrm{L}}}$

In both the setting here as well as the simple logistic regression setting, the explanatory variable(s) may be dichotomous or continuous.

In the multiple logistic regression setting, there is a global test which should be conducted first (before evaluating tests based on each explanatory variable separately).

An example: Multiple Logistic Regression

Our explanatory variables are cholesterol level, age and gender and our outcome is whether or not the subject had another coronary event. The p-value for the global test was 0.0058.

A summary of the beta estimates from the model are shown below.

Test the global null hypothesis at the $\alpha=0.05$ level. If significant, then summarize the results from the tests of each of the regression coefficients.

Parameter	Estimate	Standard Error	p-value
beta_0	-8.536	2.684	0.001
beta_age	0.042	0.025	0.096
beta_chol	**0.029**	**0.013**	**0.024**
beta_M_vs._F	2.521	0.803	0.002

An example: logistic regression inference

The test for the global null hypothesis tests:
H_0 : $\beta_1=\beta_2=\cdots=\beta_k=0$
H_1 : at least one $\beta_i \neq 0$

The p-value for the global test was 0.0058. Since ≤α, we reject the null hypothesis and conclude that there is at least one $\beta_i \neq 0$.

Tests of the individual parameters follow the 5-step procedure discussed above for the simple logistic regression setting.

Age:

H_0 : $\beta_{age}=0$ or $OR_{age}=1$ (there is no association between age and risk for a coronary event, after controlling for cholesterol level and gender)

H_1 : $\beta_{age} \neq 0$ $OR_{age} \neq 1$ (there is an association between age and risk for a coronary event, after controlling for cholesterol level and gender)

We fail to reject the null hypothesis that or after adjusting for cholesterol level and gender since >α. We do not have significant evidence at the α=0.05 level that βage≠0 (p=0.096). That is, there is not evidence of an association between age and risk of a coronary event after adjusting for cholesterol level and gender. The odds ratio for a coronary event is $e^{\widehat{\beta_1}}$=1.04 for every 1-year increase in age.

An example: logistic regression inference

Cholesterol Level:

Reject H_0 : $\beta_{chol}=0$ or $OR_{chol}=1$ after adjusting for age and gender since $p \leq \alpha$.

We have significant evidence at the $\alpha=0.05$ level that $\beta_{chol} \neq 0$. That is, there is evidence of an association between cholesterol level and risk of a coronary event after adjusting for age and gender. The odds ratio for a coronary event is $e^{\widehat{\beta_{chol}}}=1.029$ for every 1-unit increase in cholesterol.

Gender:

Reject H_0 : $\beta_{gender}=0$ or $OR_{gender}=1$ after adjusting for age and cholesterol level since $p \leq \alpha$.

There is evidence of an association between gender and risk of a coronary event after adjusting for age and cholesterol level. The odds ratio for a coronary event is is $e^{\widehat{\beta_{gender}}}=12.44$ for males versus females.

An example: multiple logistic regression

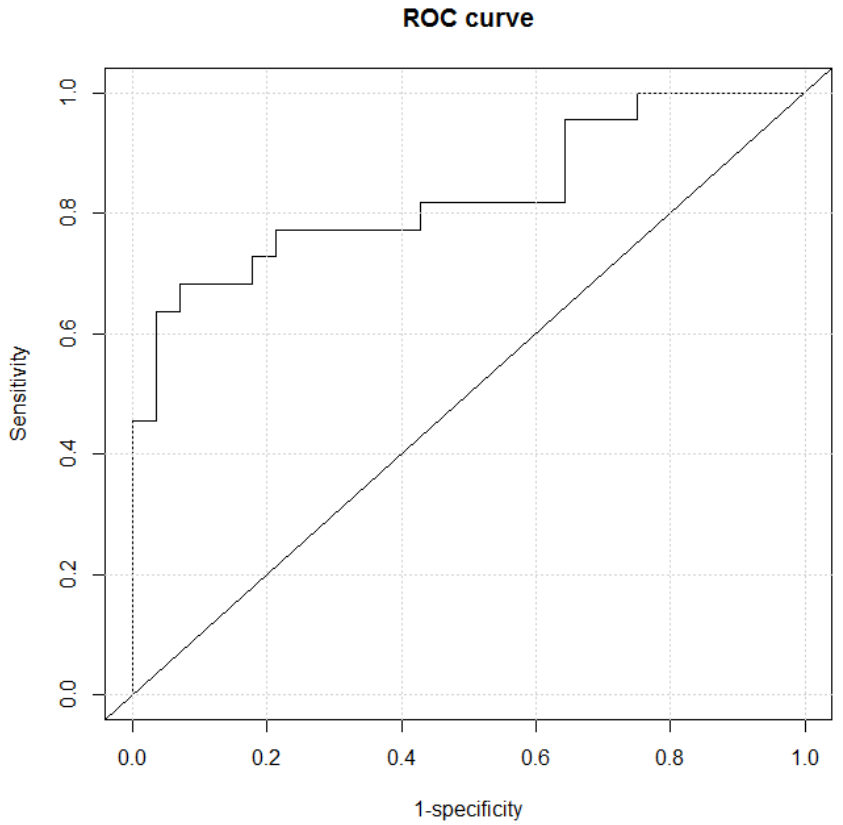
Use the regression model to predict the risk of a coronary event for a 60-year-old female with a cholesterol level of 150.

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\epsilon}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\epsilon}} = \frac{e^{-8.53+0.029\cdot 150+2.52\cdot 0+0.042\cdot 60}}{1+e^{-8.53+0.029\cdot 150+2.52\cdot 0+0.042\cdot 60}} = \frac{e^{-1.66}}{1+e^{-1.66}} = 15.97\%$$

The risk of a coronary event for a 60-year old woman with a cholesterol level of 150 is 15.97%.

Note in the above equation $x_{\text{M versus F}}=0$ since this is the dummy variable for males (which is =0 for women).

Receiver Operating Characteristic (ROC) Curve



Area under the ROC curve

In the logistic regression setting, the observed response takes on one of two values (0 = no event and 1= event).

The **predicted probabilities take on values between 0 and 1**. If the predicted probabilities for someone who had the event are close to or equal to 1, then we would say that the model was successful in predicting that individual's risk.

If the logistic model only contains a few categorical or dummy variables, then the possible values that the predicted probabilities may take on is limited. As such, it is possible the only predicted probabilities from a dataset one of two values (in the case of a logistic regression model with only one explanatory variable which is dichotomous).

For example, let's say that all **predicted probabilities are 0.32 or 0.72.**
What would we say about the fit of the model if all the events had a predicted probability of 0.72 and all the non-events had a predicted probability of 0.32? We'd like be inclined to think that the model was successful in discriminating between those with and without the event.

What if most of those with a predicted probability of 0.32 did not have the event and most of those with a predicted probability of 0.72 did. We'd still be inclined to say that the model fit was good. How we define "most" may vary and it becomes a little more difficult to assess without having a measure to quantify the fit.

Area under the ROC curve

We are interested in the association between gender and risk of having a coronary event in a high-risk patient population (who have had an event in the past). We conduct a simple logistic regression model with gender (specifically a dummy variable for male versus female) as the only explanatory variable. As before, our outcome is whether or not the subject had another coronary event. The regression summary is:

Parameter	Estimate	Standard Error	p-value
β_0	-1.153	0.468	0.014
β_1	1.728	0.627	0.006

Parameter	Estimate	Standard Error	p-value
β_0	-1.153	0.468	0.014
β_1	1.728	0.627	0.006

Here's the predicted risk for a male calculated from the regression equation:

$$\hat{p} = \frac{e^{\widehat{\beta_0}+\beta\widehat{gender}^x}}{1+e^{\widehat{\beta_0}+\beta\widehat{gender}^x}} = \frac{e^{-1.153\ +1.728*1}}{1+e^{1.153\ +1.728*1}} = \frac{e^{0.575}}{1+e^{0.575}} = 64\%$$

Here's the predicted risk for a female calculated from the regression equation:

$$\hat{p} = \frac{e^{\widehat{\beta_0}+\beta\widehat{gender}^x}}{1+e^{\widehat{\beta_0}+\beta\widehat{gender}^x}} = \frac{e^{-1.153\ +1.728*0}}{1+e^{1.153\ +1.728*0}} = \frac{e^{-1.153}}{1+e^{-1.153}} = 24\%$$

If we predicted that those with a predicted probabilities of >24% will have the event and assign them predicted outcomes of 1 ($\hat{Y}$=1) and we predicted that those with a predicted probabilities of ≤ 24% will not have the event and assign them predicted outcomes of 0 ($\hat{Y}$=0), then we could assess the fit of the predicted values versus the observed values.

Area under the ROC curve

In this example, all males have a predicted probability of 64% and as such as assigned via the rule suggested above as having the event of interest. All females have a predicted probability of 24% and as such as assigned via the rule suggested above as not having the event of interest. Here's a table summarizing actual versus predicted risk:

	Actual Outcome	
	$Y=0$ (Did not have the event)	$Y=1$ (Had the event)
Predicted Probability/Predicted Outcome	19	6
	9	16

16 of the 22 subjects who had a coronary event were male and had a predicted probability of 64%. The other 6 were female and had a predicted probability of 24%.
Per the rule suggested above, (where we suggested predicting those with predicted probabilities >24% as events) then we correctly predicted the event status of 16/22 (72.72%) subjects who did have the event. This is called the true positive rate or the sensitivity.

19 of the 28 subjects who did not have the coronary event were female and have predicted probabilities of 24%. The other 9 were male and had a predicted probability of 64%. Per the rule suggested above, (where we suggested predicting those with predicted probabilities ≤24% as non-events) then we correctly predicted the event status of 19/28 (67.86%) subjects who did not have the event. This is called the true negative rate or the specificity.

Area under the ROC curve

If we had instead used a cut off of 0% (if we predicted that those with a predicted probabilities of >0% will have the event and assign them predicted outcomes of 1 ($\hat{Y}=1$) and we predicted that those with a predicted probabilities of 0% will not have the event and assign them predicted outcomes of 0 ($\hat{Y}=0$, then we would have had:

		Actual Outcome	
		Y=0 (Did not have the event)	Y=1 (Had the event)
Predicted Probability/Predicted Outcome	Predicted Probability/Predicted Outcome: $\hat{p}$=24%/$\hat{Y}$=0 (Females)	0	0
	Predicted Probability/Predicted Outcome: $\hat{p}$=64%/$\hat{Y}$=1 (Males)	28	22
Predicted Probability/Predicted Outcome	Actual Outcome	Y=0 (Did not have the event)	Y=1 (Had the event)
	Predicted Probability/Predicted Outcome: $\hat{p}$=24%/$\hat{Y}$=0 (Females)	0	0
	Predicted Probability/Predicted Outcome: $\hat{p}$=64%/$\hat{Y}$=1 (Males)	28	22
Predicted Probability/Predicted Outcome		0	0
Predicted Probability/Predicted Outcome		28	22

The sensitivity (the proportion of true events that were classified correctly) would have been 22/22=100% and the specificity (the proportion of true non-events that were classified correctly) would have been 0/28=0%.

Area under the ROC curve

If we had instead used a cut off for our rule of 100% (if we predicted that those with a predicted probabilities of 100% will have the event and assign them predicted outcomes of 1 ($\hat{Y}=1$) and we predicted that those with a predicted probabilities of <100% will not have the event and assign them predicted outcomes of 0 ($\hat{Y}=0$), then we would have had:

		Actual Outcome	
		Y=0 (Did not have the event)	Y=1 (Had the event)
Predicted Probability/Predicted Outcome	Predicted Probability/Predicted Outcome: $\hat{p}$=24%/$\hat{Y}$=0 (Females)	28	22
	Predicted Probability/Predicted Outcome: $\hat{p}$=64%/$\hat{Y}$=1 (Males)	0	0
Predicted Probability/Predicted Outcome	Actual Outcome: Y=0 (Did not have the event)	28	22
	Actual Outcome: Y=1 (Had the event)	0	0

Then, the sensitivity (the proportion of true events that were classified correctly) would have been 0/22=0% and the specificity (the proportion of true non-events that were classified correctly) would have been 28/28=100%.

Sensitivity & specificity

Sensitivity and specificity are important measure in assessing the fit of a logistic regression model.

Simple logistic regression models with a single dichotomous factor as the explanatory variable are not very interesting as the range of predicted probabilities is limited.

Models with one or more continuous explanatory or independent variables have more possible values for the predicted probabilities and therefore **there are often many cutoffs that produce distinct values of sensitivity and specificity.**

The area under the **ROC (receiver operating characteristic) curve** (also known as the **c-statistic**) is a measure of the sensitivity and specificity across the range of all possible cutoffs.

It is often used to **measure the goodness of fit of a logistic regression model**.

The ROC curve is a plot of corresponding pairs of sensitivity (y-axis) and 1 minus the specificity (x-axis) for each possible cutoff point.

It ranges between **0.5 and 1.0 with larger values indicating better fit**.

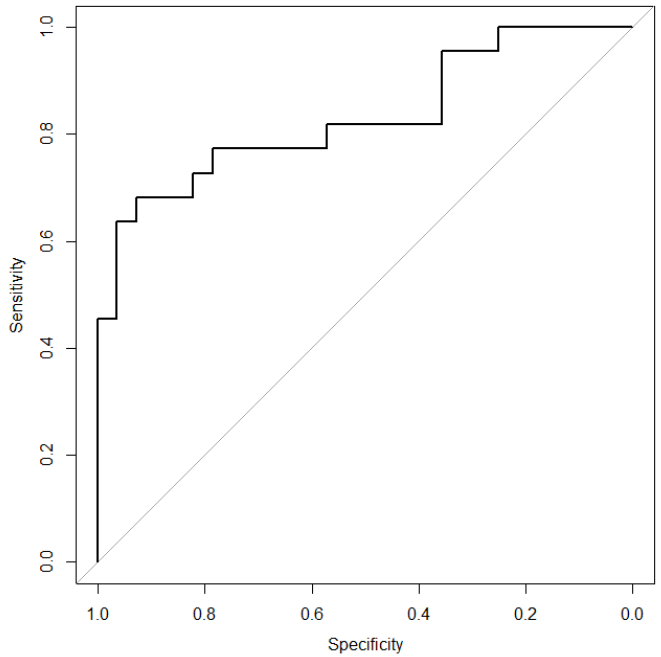
When the area under the curve is equal to 0.50, it is said that the model does no better at classifying events than at random or by chance.

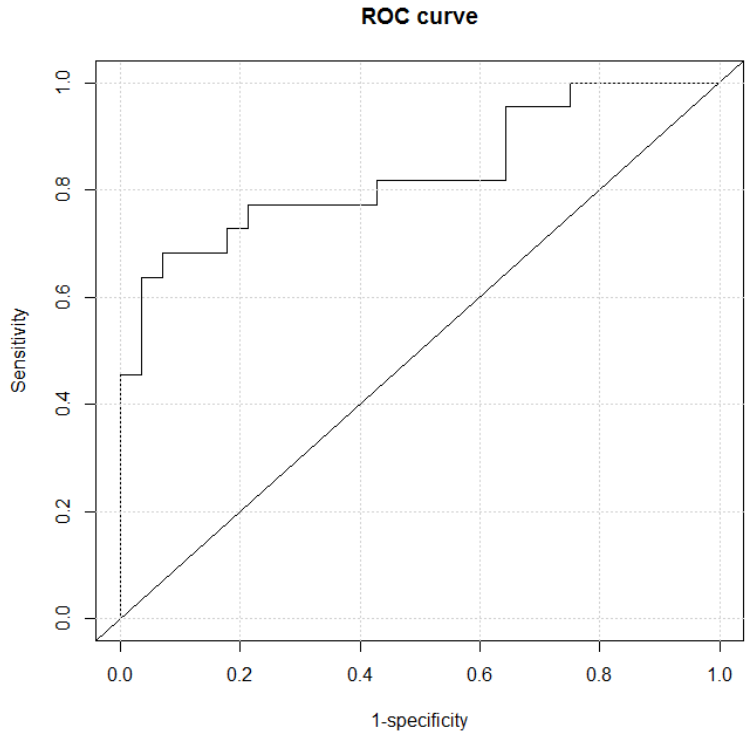
Area under the ROC curve: R commands

```
> #ROC curve
> library(pROC) # for visualizing, smoothing and comparing receiver operating characteristic (ROC curves).

> data$prob <- predict(m, type=c("response"))
> g <- roc(data$event ~ data$prob)
> plot(g)

> plot(1-g$specificities, g$sensitivities, type="l", xlab="1-specificity", ylab="Sensitivity", main="ROC  curve")
> abline(a=0, b=1)
> grid()
```





Multiple Logistic Regression

```
# multiple logistic regression
data$male <- ifelse(data$sex =="M", 1, 0)
m2<-glm(data$event ~ data$chol + data$male + data$age, family=binomial)
summary(m2)

# overall test
# install.package("aod")
library(aod)
wald.test(b=coef(m2), Sigma = vcov(m2), Terms = 2:4)
# Terms: An optional integer vector specifying which coefficients should be jointly tested
# Terms defines to compare which regression coefficients, here we want to compare the 2 to 4 (first is the
intercept)
# It gives as a result Chi-Squared test results, and p-value of it
# if p is smaller than 0.05 you can reject the null hypothesis

# ORs per 1 unit increase
exp(cbind(OR = coef(m2), confint.default(m2)))

# ROC curve  # install.package("pROC")
library(pROC)
# using model with chol and sex and age
data$prob <-predict(m2, type=c("response"))
g <- roc(data$event ~ data$prob)
plot(g)
```

Some Useful Statistical Web Apps

Try Different Distributions, Regressions and more

https://www2.stat.duke.edu/~mc301/shinyed/

These Web Apps are build with Shiny

http://shiny.rstudio.com/