Binary Data

Samiran Sinha
Texas A&M University
sinha@stat.tamu.edu

September 3, 2021

Binary data

- In a binary dataset, the variable of interest can be categorized as 0 and 1, and they are usually referred as failure or success.
- Suppose that Y is a binary variable. To characterize this variable we talk about its success probability, i.e., $\pi = \text{pr}(Y = 1)$.
- For a binary variable Y with success probability π, the mean is

  $$\mu = E(Y) = 0 \times \text{pr}(Y = 0) + 1 \times \text{pr}(Y = 1) = 0 + \pi = \pi,$$

  and the variance is $\text{Var}(Y) = E(Y^2) - E^2(Y)$. Note that $E(Y^2) = E(Y) = \pi$. So,

  $$\text{Var}(Y) = \pi - \pi^2 = \pi(1 - \pi).$$

Inference on proportion (estimation)

- One of our goals is to make inference about the success probability, π based on a sample data.
- Suppose that the interest is in identifying the prevalence of influenza C virus (ICV) infection among among cattle with respiratory disease. The paper[1] makes use of information from an observational study to estimate the prevalence.

[1] Influenza C Virus in Cattle with Respiratory Disease, United States, 2016–2018

Inference on proportion

- Briefly, the study collected clinical samples from 1,525 infected animals, and sent the samples for laboratory testing.
- The result showed 64 were positive for ICV, hence the estimate of the prevalence was $64/1525 = 0.042$.

Inference on proportion

- Suppose that we have observed data on n independent units (think each unit is an animal in the context of our data example), $Y_1, \ldots, Y_n$.
- Note $Y_1 = 1$ if the 1st unit is a success and 0 otherwise. In our data example, $Y_1 = 1$ if the 1st animal is positive for the virus and 0 otherwise. This way we can define $Y_2, \ldots, Y_n$ for unit $2, \ldots, n$, respectively.
- The estimator of π is $\widehat{\pi} = \sum_{i=1}^{n} Y_i/n =$ sum of ones$/n$.
- For a large sample size n, we use $\widehat{\pi} \pm Z_{\alpha/2}\sqrt{\{\widehat{\pi}(1-\widehat{\pi})/n\}}$ as the $(1-\alpha)100\%$ confidence interval, where $Z_{\alpha/2}$ denotes the upper $\alpha/2$ percentile point of $N(0,1)$ distribution. This is known as Wald's CI.
- This large sample confidence interval is usually used when $n\widehat{\pi} \geq 15$ and $n(1-\widehat{\pi}) \geq 15$.

Inference on proportion

- However, there are issues (the actual coverage probability is different from the nominal coverage probability) about this confidence interval specially for a) small sample size (like when both conditions do not hold together) and b) when the true π is close to 0 or 1.
- When the true π is close to zero, then for a small sample size $\widehat{\pi}$ could be zero that leads to zero-length confidence interval.
- Also, for a small sample size the CI may have small/larger confidence level than the nominal level.
- Several alternatives have been proposed, like Wilson, Agresti-Coull, Jeffreys, Clopper–Pearson interval etc.

Inference on proportion

- Wilson interval:

$$\frac{\widehat{\pi} + z^{*2}/(2n)}{1 + z^{*2}/n} \pm \frac{z^*}{1 + z^{*2}/n}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n} + \frac{z^{*2}}{4n^2}}$$

  This interval has good properties even for a small number of trials and for π close to zero or one. Here $z^* = Z_{\alpha/2}$.

- Jeffrey's interval is a Bayesian credible interval based on the posterior distribution of π using the Jeffrey's prior. It is $\alpha/2$ and $1-\alpha/2$ percentile points of the Beta$(\sum_{i=1}^n Y_i + 0.5, n - \sum_{i=1}^n Y_i + 0.5)$ distribution. To avoid low coverage probability, if $\sum_{i=1}^n Y_i = 0$, the lower limit of the confidence interval is replaced by 0 and when $\sum_{i=1}^n Y_i = n$ the upper limit of the interval is replaced by 1.

- Another simple CI is Agresti-Coull CI, $\tilde{p} \pm z^*\sqrt{\tilde{p}(1-\tilde{p})/n}$, where $\tilde{p} = (\sum_{i=1}^n Y_i + z^{*2}/2)/(n + z^{*2})$.

- Once a $(1-\alpha)100\%$ two-sided CI is calculated, we can use that interval to conduct two sided hypothesis test.

Different types of CI for π

Code

```
library(DescTools)
BinomCI(2, 6, conf.level = 0.95,
        method = c("agresti-coull", "jeffreys",
   "wilson", "clopper-pearson"))
                      est     lwr.ci    upr.ci
agresti-coull   0.3983890 0.09252396 0.7042541
jeffreys        0.3333333 0.07677014 0.7135770
wilson          0.3333333 0.09677141 0.7000067
clopper-pearson 0.3333333 0.04327187 0.7772219
# This function can produce many other types of CI for
# proportion
```

Inference on proportion (testing)

- Suppose that the interest is in testing $H_0 : \pi = \pi_0$ against an alternative hypothesis H_a.
- One can use $T = \sqrt{n}(\widehat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)}$ as the test statistic, and depending on H_a calculate the p-value using the following table.
- Note that for a large n, T approximately follows $N(0,1)$ distribution (Z distribution).

H_a	p-value
$H_a : \pi > \pi_0$	$\text{pr}_{H_0}(T > T_{\rm obs}) = \text{pr}(Z > T_{\rm obs})$
$H_a : \pi < \pi_0$	$\text{pr}_{H_0}(T < T_{\rm obs}) = \text{pr}(Z < T_{\rm obs})$
$H_a : \pi \neq \pi_0$	$2\text{pr}_{H_0}(T > \lvert T_{\rm obs}\rvert) = 2\text{pr}(Z > \lvert T_{\rm obs}\rvert)$

- $T_{\rm obs}$: the observed value of T
- This test is valid for a large sample and it is usually recommended when $n\pi_0 \geq 10$ and $n(1-\pi_0) \geq 10$.
- If at least one of the above conditions does not hold, then the exact test is recommended.

Inference on proportion (exact test)

- Let $Y = \sum_{i=1}^{n} Y_i$, then $Y \sim \mathrm{Binomial}(n, \pi)$. Let Y_o be the observed value of Y.
- Then the p-value is the probability of observing the random variable Y that is at least as *extreme* as Y_o under H_0. The meaning of *extreme* changes with the alternative hypothesis.
- The following table contains the exact p-value formula. Define $p_0(u) = \binom{n}{u}\pi_0^u(1-\pi_0)^{n-u}$.

H_a	p-value
$H_a : \pi > \pi_0$	$\sum_{u \geq Y_o} p_0(u)$
$H_a : \pi < \pi_0$	$\sum_{u \leq Y_o} p_0(u)$
$H_a : \pi \neq \pi_0$	$\sum_{u: p_0(u) \geq p_0(Y_o)} p_0(u)$

Large sample test

Consider the previous example where $Y_o = 2$ and $n = 6$, and we want to test $H_0 : \pi = 0.5$ versus $H_a : \pi < 0.5$.

Code

```
prop.test(2, 6, correct=F, alternative="less")

1-sample proportions test without continuity correction

data:  2 out of 6, null probability 0.5
X-squared = 0.66667, df = 1, p-value = 0.2071
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.6529852
sample estimates:
        p
0.3333333
```

Since the test is one sided, the above CI is a one sided CI.

Exact test

Code

```
binom.test(2, 6, p=0.5, alternative="less")

Exact binomial test

data:  2 and 6
number of successes = 2, number of trials = 6, p-value = 0.3437
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.7286616
sample estimates:
probability of success
             0.3333333
```

Since the test is one sided, the above CI is a one sided CI.

Two sided test

Code

```
binom.test(2, 10,p=0.18, alternative="two.sided")
Exact binomial test
data: 2 and 10 number of successes = 2, number of trials = 10, p-value =
0.6983
alternative hypothesis: true probability of success is not equal to 0.18

95 percent confidence interval: 0.02521073 0.55609546
sample estimates: probability of success 0.2
```

Association between two categorical variables

- Suppose that there are two variables, each of them is a categorical variable having two categories. The observed data can be categorized in the following table:

	B		Total
A	No	Yes	
No	n_{00}	n_{01}	n_{0+}
Yes	n_{10}	n_{11}	n_{1+}
Total	n_{+0}	n_{+1}	n

- n_{ij}: observed frequency corresponding to the ith category of the row variable (like, A) and the jth category of the column variable (like, B).
- The goal is to test if there is any association between the two. The null hypothesis H_0 is that there is no association between the variables. The alternative hypothesis H_a is either *Yes* of A tend to occur more with *Yes* of B or *No* of B.

- This is tested via the test statistic

$$T = \frac{\{n_{00} - E(n_{00})\}^2}{E(n_{00})} + \cdots + \frac{\{n_{11} - E(n_{11})\}^2}{E(n_{11})}.$$

- $E(n_{ij})$: expected value of the count n_{ij} under the null hypothesis, that there is no such association, and it is $n(n_{i+}/n) \times (n_{+j}/n) = n_{i+}n_{+j}/n$, and n is the total number of observations
- Under H_0, T follows the χ^2 distribution with 1 degree of freedom. The p-value is $\text{pr}(\chi_1^2 > t_{obs})$, where t_{obs} is the observed value of T. This is Pearson Chi-square test.

- Now assume that A and B have I and J categories, respectively. The categories of A (B) are now denoted by $1, \ldots, I$ $(1, \ldots, J)$.

- The test statistics is

$$T = \frac{\{n_{11} - E(n_{11})\}^2}{E(n_{11})} + \cdots + \frac{\{n_{IJ} - E(n_{IJ})\}^2}{E(n_{IJ})}.$$

- $E(n_{ij})$: expected value of the count n_{ij} under the null hypothesis, that there is no such association, and it is $n(n_{i+}/n) \times (n_{+j}/n) = n_{i+}n_{+j}/n$, and n is the total number of observations

- Under H_0, T follows the χ^2 distribution with $(I-1) \times (J-1)$ degrees of freedom. The p-value is $\text{pr}(\chi^2_{(I-1)\times(J-1)} > t_{obs})$.

An example

Let us look at the survey dataset available in the R package MASS. We shall study the association between smoking and exercise of people.

Code

```
library(MASS)       # load the MASS package
tbl = table(survey$Smoke, survey$Exer)
        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
 chisq.test(tbl)

Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
# Indeed, because some cells have counts less than 5
```

- The above mentioned Chi-square test are asymptotic test. If the sample sizes are large then the performance of the test is satisfactory (neither conservative nor liberal). To check if the chi-square test is okay, we need to make sure that the expected cell frequencies are greater than 5 for all cells of the contingency table. Otherwise, we adopt Yate's continuity correction in the test statistic. However, this correction method is also not a foolproof solution.

- Continuity corrected test statistic is

$$T = \frac{\{|n_{11} - E(n_{11})| - 0.5\}^2}{E(n_{11})} + \cdots + \frac{\{|n_{IJ} - E(n_{IJ})| - 0.5\}^2}{E(n_{IJ})}.$$

- Under H_0, T approximately follows the χ^2 distribution with $(I-1)\times(J-1)$ degrees of freedom. The p-value is pr$(\chi^2_{(I-1)\times(J-1)} > t_{obs})$.

- For a large sample size, corrected or uncorrected methods will produce similar results.

An example

Revisit the survey data example of `library(MASS)`.

Code

```
library(MASS)       # load the MASS package
tbl = table(survey$Smoke, survey$Exer)
        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
 chisq.test(tbl, correct=T)

Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl, correct = T) :
Chi-squared approximation may be incorrect
# Indeed, because some cells have counts less than 5
```

Fisher's exact test

The best solution in the small sample scenario (actually in all samples) is to use the Fisher exact test (FET). FET is time consuming for a large sample data, however, due to the availability of superior computing power nowadays we can easily use FET for small to moderate sample size data. The asymptotic test that is quick to compute, can be used for a large sample data. I will not not show you the formula of FET that involves with the Hypergeometric distribution, rather show you how to conduct the test in `R`.

Fisher's exact test

Code

```
library(MASS)       # load the MASS package
tbl = table(survey$Smoke, survey$Exer)
        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7

 fisher.test(tbl)

Fisher's Exact Test for Count Data

data:  tbl
p-value = 0.4138
alternative hypothesis: two.sided
```

Odds ratio

- Odds ratio is a commonly used measure to check association between two variables. It is frequently used for checking association between two binary variables.
- Revisit the 2 × 2 scenario.

	B		Total
A	N	Y	
N	n_{00}	n_{01}	n_{0+}
Y	n_{10}	n_{11}	n_{1+}
Total	n_{+0}	n_{+1}	n

- Odd ratio is defined as ratio of the odds of A being Y when B is Y to the odds of A being Y when B is N. Note that the odds of A being Y when B is Y$=\text{pr}(A=Y|B=Y)/\text{pr}(A=N|B=Y)$ and the odds of A being Y when B is N $=\text{pr}(A=Y|B=N)/\text{pr}(A=N|B=N)$. Hence, the odds ratio is

$$\begin{aligned} OR &= \frac{\text{pr}(A=Y|B=Y)/\text{pr}(A=N|B=Y)}{\text{pr}(A=Y|B=N)/\text{pr}(A=N|B=N)} \\ &= \frac{\text{pr}(A=Y|B=Y)\text{pr}(A=N|B=N)}{\text{pr}(A=N|B=Y)\text{pr}(A=Y|B=N)}. \end{aligned}$$

Odds ratio

It is defined as the odds of disease in exposed individuals relative to the odds of disease in the unexposed. It can also be defined as the odds of the exposure.

Odds ratio

- If there is no association between A and B, then $\mathrm{pr}(A=\mathrm{Y}|B=\mathrm{Y})=\mathrm{pr}(A=\mathrm{Y}|B=\mathrm{N})$. Consequently, $\mathrm{pr}(A=\mathrm{N}|B=\mathrm{Y})=\mathrm{pr}(A=\mathrm{N}|B=\mathrm{N})$, and OR becomes 1. Thus, OR (log-OR) equals 1 (0) implies no association between A and B.
- One can estimate $\mathrm{pr}(A=\mathrm{Y}|B=\mathrm{Y})$ by n_{11}/n_{+1}, $\mathrm{pr}(A=\mathrm{Y}|B=\mathrm{N})$ by n_{10}/n_{+0}, $\mathrm{pr}(A=\mathrm{N}|B=\mathrm{Y})$ by n_{01}/n_{+1} and $\mathrm{pr}(A=\mathrm{N}|B=\mathrm{N})$ by n_{00}/n_{+0}.
- So the estimated odds ratio is $\widehat{OR}=n_{11}n_{00}/n_{01}n_{10}$. A larger or a smaller value of $\widehat{OR}$ ($\log(\widehat{OR})$) compared to 1 (0) indicates a possible association.
- For a large sample size, the variance of $\log(\widehat{OR})$ can be estimated by $\tau^2=(1/n_{00}+1/n_{10}+1/n_{01}+1/n_{11})$.
- For a large sample size, $\{\log(\widehat{OR})-\log(OR)\}/\tau$ approximately follows the Normal$(0,1)$ distribution.
- Approximate $(1-\alpha)100\%$ CI for $\log(OR)$: $\log(\widehat{OR})\pm Z_{\alpha/2}\tau$

Odds ratio: Example

In a simple unmatched case-control study, 1327 women aged 50-81 years with hip fractures, who lived in a largely urban area in Sweden, were compared with 3262 controls with the same age range randomly selected from the national registry. The objective was to determine whether women currently taking postmenopausal hormone replacement therapy (HRT) were less likely to have hip fractures than those who are not under HRT. The results were given in a 2x2 table format show the number of women who were current users of HRT and those who had never used or formerly used HRT[2].

Hip	HRT		Total
fracture	Never/former user	Current user	
No (controls)	3023	239	3262
Yes (cases)	1287	40	1327
Total			4589

[2] Measures derived from a 2 × 2 table for an accuracy of a diagnostic test

Odds ratio: Example

- $\widehat{OR} = 40 \times 3023/(239 \times 1287) = 0.393$. So, $\log(\widehat{OR}) = -0.93$
- Standard error of $\log(\widehat{OR})$ is 0.174.
- 95% CI: $(-1.27, -0.59)$
- Suppose that we are interested in testing H_0: there is no association between the two variables, hip fracture and current use of HRT against H_a : there is an association. In other words, $H_0 : \log(OR) = 0$ versus $H_a : \log(OR) \neq 0$.
- Based on the two-sided 95% CI, we reject H_0 at the 5% level (the CI does not include 0, the null value of $\log(OR)$), and conclude that the data provide a strong evidence that the two variables are associated.

Odds ratio: Example

- For the stated objective in the question "The objective was to determine whether women currently taking postmenopausal hormone replacement therapy (HRT) were less likely to have hip fractures than those who are not under HRT", we have to test $H_0 : OR = 1$ versus $H_a : OR < 1$.
- It is important to understand why the objective of the question is reflected in $OR < 1$.
- Under the objective $\text{pr}(Hip = Y|HRT = Y) < \text{pr}(Hip = Y|HRT = N)$ which implies $\text{pr}(Hip = Y|HRT = Y)/\text{pr}(Hip = Y|HRT = N) < 1$. Also, consequently, $\text{pr}(Hip = N|HRT = Y) > \text{pr}(Hip = N|HRT = N)$ which implies $1 > \text{pr}(Hip = N|HRT = N)/\text{pr}(Hip = N|HRT = Y)$.
- So, $OR < 1$, or $\log(OR) < 0$.
- To test this hypothesis at $100\alpha\%$ level one <u>may use</u> the one sided CI for $\log(OR)$, $[\log(\widehat{OR}) - Z_\alpha \tau, \infty]$.

Odds ratio

Code

```
library(epitools)       # load the MASS package
mydata=array(c(3023, 239, 1287, 40), dim=c(2, 2),
dimnames=list( HRT=c("never or former", "current"), HF=c("no", "yes")))
> mydata
                 HF
HRT                  no  yes
  never or former 3023 1287
  current          239   40
> oddsratio.wald(mydata)
$data
                 HF
HRT                  no  yes Total
  never or former 3023 1287  4310
  current          239   40   279
  Total           3262 1327  4589
$measure
                 odds ratio with 95% C.I.
HRT                estimate     lower     upper
  never or former 1.0000000        NA        NA
  current         0.3931169 0.2794948 0.5529293
$p.value
                 two-sided
HRT                 midp.exact fisher.exact   chi.square
  never or former           NA           NA           NA
  current         4.101568e-09 4.975764e-09 2.978215e-08
$correction
[1] FALSE
attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
```

Risk ratio

- Risk ratio is a ratio of two risks. For the two variables example the rsik ratio can be defined as $RR = \text{pr}(A = \mathrm{Y}|B = \mathrm{Y})/\text{pr}(A = \mathrm{Y}|B = \mathrm{N})$.
- If there is no association between A and B, then $\text{pr}(A = \mathrm{Y}|B = \mathrm{Y}) = \text{pr}(A = \mathrm{Y}|B = \mathrm{N})$. Consequently, RR becomes 1. Thus, RR (log-RR) equals 1 (0) implies no association between A and B.
- Assuming that we have random sample from the poulation, we can estimate $\text{pr}(A = \mathrm{Y}|B = \mathrm{Y})$ by n_{11}/n_{+1}, $\text{pr}(A = \mathrm{Y}|B = \mathrm{N})$ by n_{10}/n_{+0}.
- So the estimated risk ratio is $\widehat{RR} = n_{11}n_{+0}/n_{+1}n_{10}$. A larger or a smaller value of $\widehat{RR}$ ($\log(\widehat{RR})$) compared to 1 (0) indicates a possible association.
- For a large sample size, the variance of $\log(\widehat{RR})$ can be estimated by $\tau^2 = (n_{+1} - n_{11})/n_{11}n_{+1} + (n_{+0} - n_{10})/n_{10}n_{+0} = n_{01}/n_{11}n_{+1} + n_{00}/n_{10}n_{+0}$.
- For a large sample size, $\{\log(\widehat{RR}) - \log(RR)\}/\tau$ approximately follows the Normal$(0, 1)$ distribution.
- Approximate $(1 - \alpha)100\%$ CI for $\log(RR)$: $\log(\widehat{RR}) \pm Z_{\alpha/2}\tau$

Risk ratio

- Risk ratio and relative risk are the same.
- For a random sample of data, we can also estimate the risk ratio if it is defined the other way, $RR = \text{pr}(B = \mathrm{Y}|A = \mathrm{Y})/\text{pr}(B = \mathrm{Y}|A = \mathrm{N})$.
- However, for a case-control data, if A denotes the case-control status, then we can only estimate the latter risk ratio $RR = \text{pr}(B = \mathrm{Y}|A = \mathrm{Y})/\text{pr}(B = \mathrm{Y}|A = \mathrm{N})$ but not the former $RR = \text{pr}(A = \mathrm{Y}|B = \mathrm{Y})/\text{pr}(A = \mathrm{Y}|B = \mathrm{N})$.

risk ratio

Code

```
  mydata
                 HF
HRT                 no  yes
  never or former 3023 1287
  current          239   40
> riskratio.wald(mydata)
$data
                 HF
HRT                 no  yes Total
  never or former 3023 1287  4310
  current          239   40   279
  Total           3262 1327  4589
$measure
                 risk ratio with 95% C.I.
HRT                estimate     lower     upper
  never or former 1.0000000        NA        NA
  current         0.4801252 0.3590985 0.6419415
$p.value
                 two-sided
HRT                 midp.exact fisher.exact   chi.square
  never or former           NA           NA           NA
  current         4.101568e-09 4.975764e-09 2.978215e-08
```

Confounder

- Confounder is a third variable other than the disease and main exposure that 1) has some direct effect on the disease and 2) is also associated with the exposure.
- As a measure of the degree of confounding one may compare ψ_p and ψ, where ψ_p denotes the the pooled odds ratio (or relative risk) of the disease from exposed to unexposed group, and ψ is the common odds ratio across the tables defined for various levels of the confounding variable.
- Let us consider the following data from the book *Case-Control Studies* by *James Schlesselman*. It is report on recent oral contraceptive (OC) use among 234 cases of myocardial infraction (MI) and 1742 controls. Age is a confounding factor. The pooled data are

Recent OC use	MI	Controls
Yes	29	135
No	205	1607

- The estimated odds ratio is approximately 1.7. Thus, the odds of MI among OC users is 1.7 times that among non users.

Confounder

Age	Recent OC use	MI	Controls	Estimated OR
25-29	Yes	4	62	7.2
	No	2	224	
30-34	Yes	9	33	8.9
	No	12	390	
35-39	Yes	4	26	1.5
	No	33	330	
40-44	Yes	6	9	3.7
	No	65	362	
45-49	Yes	6	5	3.9
	No	93	301	

Based on this stratified data, all age-specific odds ratios exceed 1.7 except in one instance, so the common odds ratio must be significantly higher than 1.7. This difference between $\widehat{\psi}_p$ and $\widehat{\psi}$ indicates a profound confounding effect of age.

Common odds ratio

Let the data from the *i*th table (*i*th level of confounding) is

Confounding	E	Disease	No disease	
ith group	Yes	a_i	b_i	
	No	c_i	d_i	n_i

then the estimate of the common odds ratio after adjusting for the confounding variables can be obtained via the Mantel-Haenszel's method,

$$\widehat{\psi} = \frac{\sum_i a_i d_i / n_i}{\sum_i b_i c_i / n_i}.$$

Code

```
mydata <-array(c(4, 2, 62, 224,
        9, 12, 33, 390,
        4, 33, 26, 330,
        6, 65, 9, 362,
        6, 93, 5, 301),
       dim = c(2, 2, 5),
       dimnames = list(OCuse = c("Yes", "No"),
Response = c("MI", "Controls"),
Age = c("25-29", "30-34", "35-39", "40-44", "45-49")))
```

Mantel-Haenszel estimate

Code

```
 mantelhaen.test(mydata)

Mantel-Haenszel chi-squared test with continuity
correction

data:  mydata
Mantel-Haenszel X-squared = 32.793, df = 1, p-value =
1.025e-08
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 2.426983 6.493688
sample estimates:
common odds ratio
         3.969895
```

Real data example[3]

- Sixty-four women with chronic, treatment-resistant PTSD were randomly assigned to either trauma-informed yoga or supportive women's health education, each as a weekly 1-hour class for 10 weeks. Assessments were conducted at pretreatment, midtreatment, and posttreatment and included measures of DSM-IV PTSD, affect regulation, and depression. The study ran from 2008 through 2011.
- The primary outcome measure was the Clinician-Administered PTSD Scale (CAPS). At the end of the study, 16 of 31 participants (52%) in the yoga group no longer met criteria for PTSD compared to 6 of 29 (21%) in the control group ($n = 60$, $\chi^2 = 6.17$, p-value$=$ 0.013).

[3] Yoga as an adjunctive treatment for posttraumatic stress disorder: a randomized controlled tria, *J Clin Psychiatry*, 2014

- In order to check if yoga has any effect on PTSD, create two variables.
    - The first one is the treatment with two categories yoga and control.
    - The second one is if the PTSD criteria are met (call it eligible PTSD), this has two groups yes and no.
- Then we shall test dependence between these two variables using the Chi-square test. Do it on your own.

Sample size calculation for inference on proportion

- Sample size needs to be calculated before conducting a clinical trials so that necessary plan/resources can be allocated for collecting data.
- Suppose that our end goal is to test $H_0 : \pi = \pi_0$ versus $H_a : \pi \neq \pi_0$ at the $100\alpha\%$ level of significance. Also, we desire to reject H_0 with probability $(1-\beta)$ when in fact $\pi = \pi_1$. Then the required sample size is

$$n = \left(\frac{Z_{1-\alpha/2}\sqrt{\pi_0(1-\pi_0)} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)}}{(\pi_1 - \pi_0)} \right)^2$$

- For one sided alternative $H_a : \pi > \pi_0$ the required sample size is

$$n = \left(\frac{Z_{1-\alpha}\sqrt{\pi_0(1-\pi_0)} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)}}{(\pi_1 - \pi_0)} \right)^2$$

- The only difference between these two formulas is the coefficient associated with the first term.

Sample size calculation for inference on proportion

- Often we want to estimate the proportion with a prespecified margin of error.
- For instance we want to estimate π with the margin of error no more than $m = 0.01$ for the 95% CI.
- Then the required sample size would be

$$n = \frac{Z^2_{\alpha/2}\pi^*(1-\pi^*)}{m^2},$$

  where π^* denotes the potential success rate of the trial (this can be obtained from a pilot study).
- If π^* is difficult to obtain, then replace it by 0.5 that will yield a sample size that is more than what is required.
- The margin of error often referred to as the precision, and expressed as a percentage of the actual proportion, like m is 10% of π^*.

Binary regression

- Supposed that we have observed data on X and Y, where X is a continuous/binary explanatory and Y is a binary response.
- We want to see how the mean of Y depends on X.
- Note that $E(Y|X) = \text{pr}(Y = 1|X)$, and suppose that the success probability depends on X through the following model

$$\text{pr}(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

- Note that this model (logistic model) ensures that this probability lies between 0 and 1.

Interpretation of β_1

- If $\beta_1 > 0$, then $\text{pr}(Y = 1|X)$ increases with X and for $\beta_1 < 0$ the probability decreases with X.
- If $\beta_1 = 0$ then there is no association between X and Y.
- Odds ratio interpretation, the odds ratio of the disease for $X_0 + 1$ to X_0 is $\exp(\beta_1)$

$$\begin{aligned}\frac{\text{odds of pr}(Y = 1|X = X_0 + 1)}{\text{odds of pr}(Y = 1|X = X_0)} &= \frac{\text{pr}(Y = 1|X = X_0 + 1)/\text{pr}(Y = 0|X = X_0 + 1)}{\text{pr}(Y = 1|X = X_0)/\text{pr}(Y = 0|X = X_0)} \\ &= \frac{\exp\{\beta_0 + \beta_1(X_0 + 1)\}}{\exp\{\beta_0 + \beta_1(X_0)\}} \\ &= \exp(\beta_1).\end{aligned}$$

- Logarithm of the odds ratio is β_1.

Data example copied from a website

This dataset has a binary response (outcome, dependent) variable called admit. There are three predictor variables: gre, gpa and rank. We will treat the variables gre and gpa as continuous The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. We can get basic descriptives for the entire data set by using summary command.

We want to do a model fitting to the data, like fit a logistic model to the data, and estimate the model parameters and other quantities.

Fitting the logistic model to a dataset

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
head(mydata)
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2

nrow(mydata) # to check the number of observations

mydata$rank <- factor(mydata$rank)
mylogit <- glm(admit ˜ gre + gpa + rank, data = mydata, family = "binomial")
```

Code

```
summary(mylogit)
Call:
glm(formula = admit ~ gre + gpa + rank, family = "binomial",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
gre          0.002264   0.001094   2.070 0.038465 *
gpa          0.804038   0.331819   2.423 0.015388 *
rank2       -0.675443   0.316490  -2.134 0.032829 *
rank3       -1.340204   0.345306  -3.881 0.000104 ***
rank4       -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

Interpretation

- Interpretation of coefficient estimates (MLE)
- Inference on the coefficents (hypothesis test and confidence interval)
- Residual analysis for model checking (deviance residual) *plot(mylogit$deviance.resid)*
- Leverage points and influential observations *plot(hatvalues(mylogit))* and *plot(cooks.distance(mylogit))*
- Estimation of probability for a given set of covariates
- Cautionary note: do not use residual deviance to check goodness-of-fit for binary data

Test

- Interest is in checking in higher GPA is associated with the higher chance of admission
- $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 > 0$
- Test statistic $TS = \widehat{\beta}_2/se(\widehat{\beta}_2)$; for a large sample TS follows the Z distribution (standard normal),
- The observed value of TS is $0.8/0.33 = 2.42$.
- The p-value is $\text{pr}(Z > 2.42) = 0.0079$.
- We reject H_0, and we conclude that the data provide sufficient evidence for the H_a that higher GPA is associated with the higher chance of admission at the 1% level of significance.

Confidence interval

- 99% CI (two-sided) for β_2: $\widehat{\beta}_2 \pm Z_{0.005} se(\widehat{\beta}_2)$
- Based on this 99% CI, we can test $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$. Check if the CI contains the null value, 0 in this particular example. If the CI contains the null value, then we fail to reject H_0 otherwise we reject H_0 at the 1% level.

CI

Code

```
confint(mylogit)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept) -6.2716202334 -1.792547080
gre          0.0001375921  0.004435874
gpa          0.1602959439  1.464142727
rank2       -1.3008888002 -0.056745722
rank3       -2.0276713127 -0.670372346
rank4       -2.4000265384 -0.753542605
```

Test of multiple parameters

- Interest is in checking if the variable rank is associated with the chance of admission.
- That means $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ versus H_a : at least one of $\beta_3, \beta_4, \beta_5$ is non-zero.
- We do likelihood ratio test (given in the next slide).
- Since the p-value came out to be so small, we decide to reject H_0 and conclude that at least one of $\beta_3, \beta_4, \beta_5$ is non-zero, or conclude that the chance of admission depends on the rank of the school.

Test of multiple parameters

Code

```
mylogit.ha <- glm(admit ˜ gre + gpa + rank, data = mydata, family = "binomial")


mylogit.h0 <- glm(admit ˜ gre + gpa, data = mydata, family = "binomial")
anova(mylogit.h0, mylogit.ha, test="LRT")
#Analysis of Deviance Table

#Model 1: admit ˜ gre + gpa
#Model 2: admit ˜ gre + gpa + rank
#  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
#1       397     480.34
#2       394     458.52  3   21.826 7.088e-05 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Wald test

- Interest is in $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ versus H_a : at least one of $\beta_3, \beta_4, \beta_5$ is non-zero.
- We may use Wald's procedure to test this hypothesis.
- We need the estimate of the parameters and the variance-covariance matrix of the estimator.
- Basic idea has already been discussed in the linear regression context. If the interest is in testing $H_0 : A\boldsymbol{\beta} = b$ versus $H_a : A\boldsymbol{\beta} \neq b$, then the test statistic is $(A\widehat{\boldsymbol{\beta}} - b)^T \{A\text{Var}(\widehat{\boldsymbol{\beta}})A^T\}^{-1}(A\widehat{\boldsymbol{\beta}} - b)$ and this test statistic follows χ^2 distribution with degrees of freedom $\mathrm{rank}(A)$.
- We can write $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ as $A\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ and

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Code

```
library(aod)
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:6)
Wald test:
----------

Chi-squared test:
X2 = 20.9, df = 3, P(> X2) = 0.00011
```

Wald test

- Next suppose that interest is in testing $H_0 : \beta_3 = \beta_4 = \beta_5$ versus H_a : at least one of $\beta_3, \beta_4, \beta_5$ is different from the other two.
- We can write H_0 as $H_0 : \beta_3 - \beta_4 = 0$ and $\beta_3 - \beta_5 = 0$, so that we can write H_0 as $A\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ and

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}.$$

- Here $\text{rank}(A) = 2$.

Wald test

- Note that one may wonder why we did not include $\beta_4 - \beta_5 = 0$ in H_0
- Suppose that we had included it in H_0, the resulting A would have been

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

- Note that the 3rd row of A can be obtained as $(-1)\times$ 1st row of $A+$ $(1)\times$ 2nd row of A. Hence, the 3rd row is dependent on the first two. In other words, $\beta_4 - \beta_5 = 0$ does not carry anything extra in the presence of the other two statements $\beta_3 - \beta_4 = 0$ and $\beta_3 - \beta_5 = 0$, consequently it is omitted.

Code

```
library(aod)
myl=cbind( c(0, 0), c(0, 0), c(0, 0), c(1, 1), c(-1, 0), c(0, -1))
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), L=myl)
Wald test:
----------

Chi-squared test:
X2 = 8.7, df = 2, P(> X2) = 0.013
```

Prediction

- Suppose that the interest is in estimating the probability of success when the explanatory variables are $X_1 = X_1^*, \ldots, X_p = X_p^*$.
- First, based on the observed data we obtain estimate and standard error for β-parameters.
- Define $\eta = \beta_0 + \beta_1 X_1^* + \cdots + \beta_p X_p^*$, then the success probability, when $X_1 = X_1^*, \ldots, X_p = X_p^*$, is

$$p^* \overset{def}{=} \text{pr}(Y = 1 | X_1 = X_1^*, \ldots, X_p = X_p^*) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Prediction

- The estimator of p^* is

$$\frac{\exp(\widehat{\eta})}{1+\exp(\widehat{\eta})} \text{ or } \frac{1}{1+\exp(-\widehat{\eta})},$$

where $\widehat{\eta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1^* + \cdots + \widehat{\beta}_p X_p^*$.

- A 95% CI (large sample) for p^* is

$$\left[\frac{1}{1+\exp[-\{\widehat{\eta} - 1.96 se(\widehat{\eta})\}]}, \frac{1}{1+\exp[-\{\widehat{\eta} + 1.96 se(\widehat{\eta})\}]}\right],$$

where $se(\widehat{\eta})$ standard error of $\widehat{\eta}$.

Prediction

Code

```
myout=predict.glm(mylogit, newdata=data.frame(gre=600, gpa=3.8, rank=as.factor(1)), se.fit=TRUE)
 1/(1+exp(-as.numeric(myout[1]))) # estimate of the probability
#[1] 0.6044446
ll=1/(1+exp(-(as.numeric(myout[1])-1.96*as.numeric(myout[2]))))
ul=1/(1+exp(-(as.numeric(myout[1])+1.96*as.numeric(myout[2]))))
print(c(ll, ul)) # 95 percent confidence interval
#[1] 0.4634509 0.7299750
```

Another data example, response is a Binomial count

This is a survey data on 1607 currently married and fecund women interviewed in the Fiji Fertility Survey. Age, education, desire for more children and current use of contraception are recorded.

Code

```
mydata2 <- read.table("http://data.princeton.edu/wws509/datasets/cuse.dat", header=TRUE)
head(mydata2)
    age education wantsMore notUsing using
1   <25       low       yes       53     6
2   <25       low        no       10     4
3   <25      high       yes      212    52
4   <25      high        no       50    10
5 25-29       low       yes       60    14
6 25-29       low        no       19    10

out=glm( cbind(using, notUsing) ~ age + education + wantsMore, data=mydata2, family=binomial)
summary(out)
```

Goodness-of-fit

- The purpose of a goodness-of-fit test is to check if a model fits the data well.
- For the linear model, we usually do the goodness-of-fit via residual analysis.
- For the binary data context, we cannot use residual $e_i = Y_i - \widehat{\pi}_i$ to check model goodness-of-fit because e_i does not behave like the linear model residual even when the model fits the data well.
- We need a different method – Hosmer Lemeshow test.

Goodness-of-fit

- Based on the MLE of β, we first obtain the estimate of the success probability

$$\widehat{\pi}_i = \frac{\exp(Z_i^T \widehat{\beta})}{1 + \exp(Z_i^T \widehat{\beta})},$$

  where $Z_i = (1, X_{i,1}, \ldots, X_{i,p})^T$.

- Then observations are grouped into g mutually exclusive groups based on their estimated probabilities. Usually g is taken as 10 (the default value), then the groups are $(mn, p_1], (p_1, p_2], \ldots, (p_9, mx]$, where p_j denotes the $j \times 10$th percentile point of $(\widehat{\pi}_1, \ldots, \widehat{\pi}_n)$, and mn and mx denote the minimum and maximum of the estimated probabilities.

Goodness-of-fit

- Now, for each group, we count the number of successes based on the observations that fall in that group. Let the $O_{k,1}$ ($O_{k,0}$) be the number of observed successes (failures) in the kth group.
- The expected number of successes (failures) $e_{k,1}$ ($e_{k,0}$) in the kth group will be $n(p_k - p_{k-1})$ $(O_{k,0} + O_{k,1} - np_k + np_{k-1})$, where n is the total sample size.
- If the model holds true, we expect $e_{k,1}$ and $O_{k,1}$ will be close, and motivated by this fact Hosmer-Lemeshow considered the Pearson Chi-square test statistic

$$T = \sum_{k=1}^{g} \left\{ \frac{(O_{k,0} - e_{k,0})^2}{e_{k,0}} + \frac{(O_{k,1} - e_{k,1})^2}{e_{k,1}} \right\}.$$

Goodness-of-fit

- Under the null hypothesis that the model fits the data well T follows approximate χ^2 distribution with $g-2$ degrees of freedom.
- The major limitation is that the test result varies with the number of groups g. A large g is good for detecting model misspecification, but that may result in some groups to have 0 observed frequency. On the other hand a small g will not be able to detect a model misspecification.

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
mylogit <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")

library(generalhoslem)
logitgof(mydata$admit, fitted(mylogit))

# Hosmer and Lemeshow test (binary model)
#
#data:  mydata$admit, fitted(mylogit)
#X-squared = 3.2185, df = 8, p-value = 0.9199
```

Goodness-of-fit

- In linear model R^2 measures the percentage of variability in the response explained by the regression model.
- However, in the binary data regression R^2 cannot be defined.
- There is a concept of pseudo-R^2 (pR^2) defined as

$$pR^2 = 1 - \frac{\log(L_f)}{\log(L_n)},$$

  where L_f and L_n denote the likelihood of the full model and the likelihood of the null model containing only the intercept term.
- pR^2 close to zero indicates that the model does not have much predictive power, whereas a higher value of pR^2 indicates a better fit.

Goodness-of-fit

- Since deviance$= -2\times$ log-likelihood, pR^2 can be expressed as

$$pR^2 = 1 - \frac{\text{Residual deviance of the full model}}{\text{Residual deviance of the null model}}.$$

- Note that this pR^2 is known as McFadden's pR^2. There are other pseudo-R^2 statistics too, and none of them is the best in all situations.
- Some cautionary note: it is generally not presented to common people who are used to use R^2 for linear regression because pR^2 does not have the same interpretation as of the R^2 value. For the logistic regression usually it is small, and even for a very accurate model fitting (with high prediction) pR^2 may be close to 0.5.

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ˜ gre + gpa + rank, data = mydata, family = "binomial")

out0 <- glm(admit ˜ 1, data = mydata, family = "binomial")

pr2= 1- as.numeric(logLik(outf))/as.numeric(logLik(out0))
print(pr2)
# 0.08
```

Prediction accuracy measures

- Other than finding association between the success probability and explanatory variables, sometimes we are interested in finding predictive power of the model.
- To check the predictive power, using the parameter estimate the response is predicted in a *test* dataset.
- Note that the *test* dataset is not used in model fitting.
- Then we compare the predicted value (responses) and observed value (responses) in the test data.

Prediction accuracy measures

- To find the predicted responses, first we estimate the success probabilities for every subject of the test data. Then based on a chosen cutoff value, we declare the predicted response to be one (success) or zero (failure). If the estimated success probability is greater than the cutoff we call it a success otherwise failure.
- Suppose that we obtain the following table comparing the predicted and observed values of the response on a test data

Predicted	Observed response	
response	Fail	Success
Fail	m_{00}	m_{01}
Success	m_{10}	m_{11}

- The above matrix is called a confusion matrix.

Prediction accuracy measures

- One measure is false positive prediction, that means, predicted to be success while actually it is a failure.
- Usually we want the number of false positives to be low.
- In the above table, it is the $(2, 1)$th entry, m_{10}.
- Another thing that we may pay attention to is false negative prediction, that means, predicted to be failure while actually it is a success.
- Usually we want the number of false negatives to be low.
- In the above table, it is the $(1, 2)$th entry, m_{01}.

Prediction accuracy measure

- Now we talk about sensitivity and specificity that takes into account the concepts of false positive and false negative.
- Sensitivity (Sens) is the proportion of observed positives that were predicted to be positives.
- In the above table it is

$$\frac{(2,2)\text{th entry}}{(1,2)\text{th entry} + (2,2)\text{th entry}} = \frac{m_{11}}{m_{01} + m_{11}}.$$

- Specificity (Spec) is the proportion of observed negatives that were predicted to be negatives.
- In the above example it is

$$\frac{(1,1)\text{th entry}}{(1,1)\text{th entry} + (2,1)\text{th entry}} = \frac{m_{00}}{m_{00} + m_{10}}.$$

Prediction accuracy measures

- One measure of goodness of prediction is higher value of sensitivity plus specificity

$$\text{proportion of predicted successes among the observed successes}$$
$$+$$
$$\text{proportion of predicted failures among the observed failures}$$

Revisit the admission data

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")
 testd=data.frame(admit=c(0, 0, 0, 1, 1, 1),
 gre=c(300, 350, 380, 700, 800, 790),
 gpa=c(3.5, 3.2, 3.2, 3.88, 3.9, 3.95),
 rank=c(4, 4, 4, 1, 1, 1))

out10=predict(outf, newdata=testd) #
# the above only calculates X-traspose x beta-hat
 out10=predict(outf, newdata=testd, type="response")
# to calculate the pr(admit=1|given X)
mypred.admit=rep(0, length(testd$admit))
### Here I have used 0.6 as the cutoff point, so if the estimated probability is
### greater than 0.6, we declare the predicted response to be 1 otherwise 0
mypred.admit[out10>0.6]<-1
mypred.admit[out10<=0.6]<-0
  mytab=table(mypred.admit, testd$admit)
accuracy= sum(diag(mytab))/sum(mytab)
print(accuracy)
#1#
```

Revisit the admission data

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ˜ gre + gpa + rank, data = mydata, family = "binomial")
 testd=data.frame(admit=c(0, 0, 0, 1, 1, 1),
 gre=c(300, 350, 380, 700, 800, 790),
 gpa=c(3.5, 3.2, 3.2, 3.88, 3.9, 3.95),
 rank=c(4, 4, 4, 1, 1, 1))

out10=predict(outf, newdata=testd) #
 library(e1071)
 library(caret)
confusionMatrix(data = as.factor(as.numeric(out10>0.6)),
reference = as.factor(testd$admit))
```

Revisit the admission data

Code

```
Confusion Matrix and Statistics

          Reference
Prediction 0 1
         0 3 0
         1 0 3

               Accuracy : 1
                 95% CI : (0.5407, 1)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.01563

                  Kappa : 1

 Mcnemar’s Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.5
         Detection Rate : 0.5
   Detection Prevalence : 0.5
      Balanced Accuracy : 1.0

       ’Positive’ Class : 0
```

Revisit the admission data, continues

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ˜ gre + gpa + rank, data = mydata, family = "binomial")
 testd=data.frame(admit=c(0, 0, 0, 1, 1, 1),
 gre=c(300, 350, 380, 600, 750, 790),
 gpa=c(3.5, 3.5, 3.2, 3.88, 3.7, 3.95),
 rank=c(4, 2, 4, 2, 1, 1))

out10=predict(outf, newdata=testd) #
# the above only calculates X-traspose x beta-hat
 out10=predict(outf, newdata=testd, type="response")
# to calculate the pr(admit=1|given X)
mypred.admit=rep(0, length(testd$admit))
mypred.admit[out10>0.6]<-1
mypred.admit[out10<=0.6]<-0
  mytab=table(mypred.admit, testd$admit)
accuracy= sum(diag(mytab))/sum(mytab)
print(accuracy)
#0.833
```

Spec and Sens change with the cutoff

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")
 testd=data.frame(admit=c(0, 0, 0, 1, 1, 1),
 gre=c(300, 350, 380, 600, 750, 790),
 gpa=c(3.5, 3.5, 3.2, 3.88, 3.7, 3.95),
 rank=c(4, 2, 4, 2, 1, 1))

out10=predict(outf, newdata=testd) #
# the above only calculates X-traspose x beta-hat
 out10=predict(outf, newdata=testd, type="response")
# to calculate the pr(admit=1|given X)
mypred.admit=rep(0, length(testd$admit))
mypred.admit[out10>0.6]<-1
mypred.admit[out10<=0.6]<-0
  mytab=table(mypred.admit, testd$admit)
mysp= mytab[1, 1]/sum(mytab[, 1])
mysn= mytab[2, 2]/sum(mytab[, 2])

print(c(mysp, mysn))
#1  0.666
```

Spec and Sens change with the cutoff

Code

```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ˜ gre + gpa + rank, data = mydata, family = "binomial")
 testd=data.frame(admit=c(0, 0, 0, 1, 1, 1),
 gre=c(300, 350, 380, 600, 750, 790),
 gpa=c(3.5, 3.5, 3.2, 3.88, 3.7, 3.95),
 rank=c(4, 2, 4, 2, 1, 1))

out10=predict(outf, newdata=testd) #
# the above only calculates X-traspose x beta-hat
 out10=predict(outf, newdata=testd, type="response")
# to calculate the pr(admit=1|given X)
mypred.admit=rep(0, length(testd$admit))
mypred.admit[out10>0.7]<-1
mypred.admit[out10<=0.7]<-0
  mytab=table(mypred.admit, testd$admit)
mysp= mytab[1, 1]/sum(mytab[, 1])
mysn= mytab[2, 2]/sum(mytab[, 2])

print(c(mysp, mysn))
#1  0.33
```

ROC curve

- ROC curve is a plot of the values of sensitivity against one minus specificity, as the value of the cutoff that increases from 0 to 1.
- A higher area under the ROC curve indicates a better discriminatory power of the model.
- If the ROC curve closely follows the 45^o line, that is not a good indication of the discriminatory power of the model.

ROC curve

Code

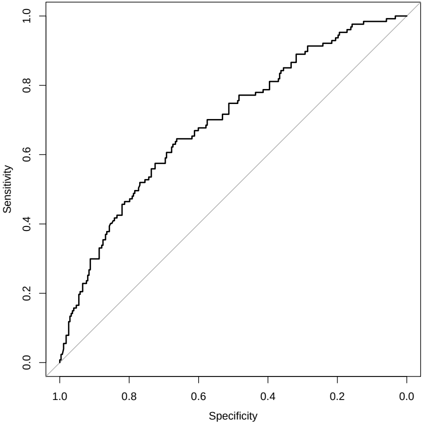
```
library(MASS)
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
outf <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")


out10=predict(outf, type="response")
library(pROC)
roccurve <- roc(mydata$admit ~ out10)
roccurve

Call:
roc.formula(formula = mydata$admit ~ out10)

Data: out10 in 273 controls (mydata$admit 0) < 127 cases (mydata$admit 1).
Area under the curve: 0.6921
plot(roccurve)
```

ROC curve


Sensitivity
Specificity