ADA2: Chapter 11 Logistic Regression

April, 2019

THE UNIVERSITY OF
NEW MEXICO

Generalized Linear Model (GLM)

- Generalization of ordinary linear regression model that allows for response variables that have other than a normal distribution (such as binary response with disease vs no disease).
- The linear model is related to the response variable via a link function.
- Logistic regression is a special case of GLM when the link function is logit link.

Admission data

```
#Binary response: admit, 1 admit, 0 no admission.
#Three predictor variables: gre, gpa and rank.
#variables gre and gpa are continuous.
#The variable rank is categorical
> head(ex.data)
  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
```

what happen if we fit a simple linear regression model by using "admit" as response variable, and "gpa" as predictor variable?

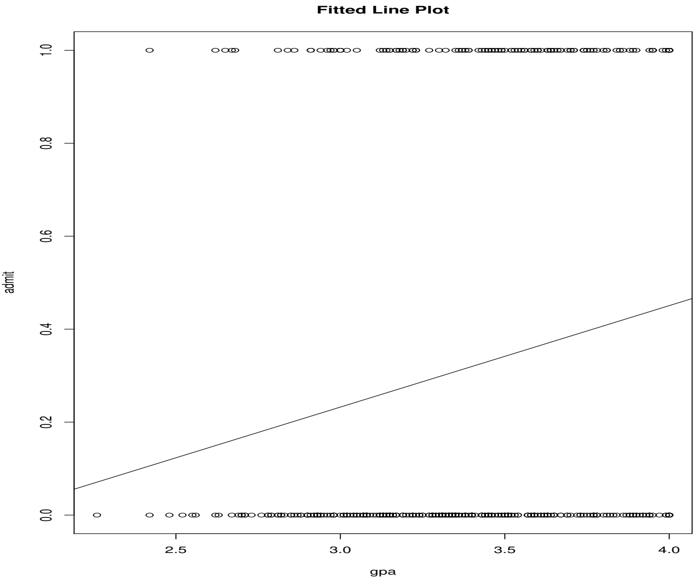


Figure 1: Fitted line plot

Odds ratio

- Let's say that the probability of success is 0.8, thus

$$p = 0.8, q = 1 - p = 0.2$$

- The odds of success are defined as
  $\mathrm{odds(success)} = p/q = 0.8/0.2 = 4$,
  —- that is, the odds of success are 4 to 1.
  Odds(success)

$$\begin{cases} > 1, \text{ or } p > 0.5 & \text{a success is more likely than a failure} \\ = 1, \text{ or } p = 0.5 & \text{same likelihood of success and failure} \\ < 1, \text{ or } p < 0.5 & \text{a success is less likely than a failure} \end{cases}$$

- The odds of failure would be
  $\mathrm{odds(failure)} = q/p = 0.2/0.8 = 0.25$,
  —-that is, the odds of failure are 1 to 4.

- Odds ratio 1
  $\mathrm{OR1} = \mathrm{odds(success)/odds(failure)} = 4/0.25 = 16$
  the odds of success are 16 times greater than for failure.
- Odds ratio 2
  $\mathrm{OR2} = \mathrm{odds(failure)/odds(success)} = 0.25/4 = 0.0625$
  the odds of failure are one-sixteenth the odds of success.

In medical examples, we often interpret the relative risk and odds ratio. Suppose individuals can be classified according to whether they have been exposed to a risk factor and ultimately whether they developed a specific disease.

$$y_i = \begin{cases} 1 & \text{if developing disease} \\ 0 & \text{if not} \end{cases}$$

$$E_i = \begin{cases} 1 & \text{if exposed} \\ 0 & \text{if not} \end{cases}$$

Let $P(y_i = 1|E_i = 1) = p_1$ and $P(y_i = 1|E_i = 0) = p_2$

Outcome	Exposed population	non-exposed population
Diseased	p_1	p_2
Non-diseased	$1 - p_1$	$1 - p_2$

Relative risk and odds ratio

Outcome	Exposed population	non-exposed population
Diseased	p_1	p_2
Non-diseased	$1 - p_1$	$1 - p_2$

- Relative ratio
$$RR = p_1/p_2$$
is the probability of disease in the exposed population divided by the probability in the non-exposed population.
- The odds of having the disease for the exposed population is $p_1/(1 - p_1)$.
- The odds of having the disease for the non-exposed population is $p_2/(1 - p_2)$.
- The odds ratio is
$$OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}$$

The odds ratio is

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

- $OR > 1 \rightarrow$ more likely to develop diseases given exposed versus not exposed
- $OR < 1 \rightarrow$ less likely to develop diseases given exposed versus not exposed
- $OR = 1 \rightarrow$ as likely to develop diseases given exposed versus not exposed

Regression models for probability of pay bill on time

Example: credit-scoring $g\left\{P(\text{a subject pays a bill on time})\right\} \sim$ size of the bill + annual income +occupation + mortage and debt obligations +percentage of bills paid on time in the past $+ \cdots$

Question: How do we relate the outcome, y (binary, pays a bill on time or not one time) , to an exposure, x?

$$g(E(y_i|x_i]) = g(\mu_i) = \beta_0 + \beta_1 x_i$$

$$E(y_i|x_i] = \mu_i = g^{-1}(\beta_0 + \beta_1 x_i)$$

$g()$ is called a link function, when $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$, we call the link function a logit function, and the regression is called logistic regression.

Logistic Regression

- y: a binary outcome
  x: explanatory variable
  $y_i \overset{indep}{\sim} Bernoulli(\mu_i)$
  $\mu_i = P(y_i = 1|X = \mathbf{x}) = 1 - P(y_i = 0|X = \mathbf{x})$
  $$\text{logit}(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i(p-1)}$$
  or
  $$\mu_i = \frac{\exp(\beta_0 + \beta_1 x_i \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i(p-1)}}{1 + \exp(\beta_0 + \beta_1 x_i \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i(p-1)}}$$
- $\ln\left(\frac{\mu}{1-\mu}\right)$ is called a logit link function, logit transformed probability
- the logit transformed probability is linearly related to $\mathbf{x}$ with intercept β_0 and slopes $\beta_1, \cdots, \beta_{p-1}$

Consider the simple logistic regression model for the disease case,

$$y_i = \begin{cases} 1 & \text{if developing disease} \\ 0 & \text{if not} \end{cases}, \; E_i = \begin{cases} 1 & \text{if exposed} \\ 0 & \text{if not} \end{cases}$$

- $y_i \overset{indep}{\sim} Bernoulli(\mu_i)$ where $\mu_i = P(y_i = 1|E_i) =$ p(develop disease given exposure status), μ_i can take on values of p_1 and p_2
- $\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 E_i$

—— $E_i = 1, \mu_i = p_1$

$$\ln\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 = \text{ log odds of disease given exposed}$$

—— $E_i = 0, \mu_i = p_2$

$$\ln\left(\frac{p_2}{1-p_2}\right) = \beta_0 = \text{ log odds of disease given not exposed}$$

$$\begin{aligned}\beta_1 &= \text{log odds of disease given exposed} - \text{log odds of disease given not exposed} \\ &= \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) \\ &= \ln\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)\end{aligned}$$

$$e^{\beta_1} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = OR$$

this is an unadjusted OR—measures association between exposure and disease without consideration of other factors.

More complicated model

Suppose x_i is continuous, E_i is binary as before

$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 E_i + \beta_2 x_i$

——$E_i = 1, \mu_i = p_1$

$$\ln\left(\frac{p_1}{1-p_1}\right) = (\beta_0 + \beta_1) + \beta_2 x_i$$

——$E_i = 0, \mu_i = p_2$

$$\ln\left(\frac{p_2}{1-p_2}\right) = \beta_0 + \beta_2 x_i$$

- β_1 measure the change in intercepts between exposed $(E = 1)$ and non-exposed individuals $(E = 0)$, called adjusted $log(OR)$.
- β_0: intercept for non-exposed individuals $(E = 0)$–the "baseline group" to which other groups are compared

- OR—measures association between exposure and disease without consideration of other factors
- Adjusted OR—are ORs obtained from multi variable models, which adjust effects relative to other factors included in the model. We need to always specify what other effects are included in model.

Fix E and vary $x \to x+1$

$$\begin{aligned} \ln\left(\frac{\mu}{1-\mu}\right) &= \beta_0 + \beta_1 E + \beta_2(x+1) \\ &= \beta_0 + \beta_1 E + \beta_2 x + \beta_2 \end{aligned}$$

- $\beta_0 + \beta_1 E + \beta_2 x$: log odds when $X = x$
- β_2: increase in log odds of developing the disease when $X = x \to X = x+1$ holding E fixed. This is adjusted $log(OR)$ for exposure, and e^{β_2} is the corresponding adjusted OR.

Another model

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 E_i + \beta_2 x_i + \beta_3 (E_i * x_i)$$

a model where each exposure group has its own intercept and slope.

The logistic family of distributions

The logistic family of distributions has density (for any real x):

$$f(x|\mu,\sigma) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma\left(1+e^{-\frac{x-\mu}{\sigma}}\right)^2}$$

and cdf

$$F(x) = \frac{1}{1+e^{-\frac{x-\mu}{\sigma}}} = \frac{e^{\frac{x-\mu}{\sigma}}}{1+e^{\frac{x-\mu}{\sigma}}}$$

The logistic family of distributions

If we plug in $\mu = 0$ and $\sigma = 1$, we get

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{1 + e^{x}}$$

Part of the motivation for logistic regression is we imagine that there is some threshold t, and if $T \leq t$, then the event occurs, so $Y = 1$. Thus, $P(Y = 1) = P(T \leq t)$ where T has this logistic distribution, so the CDF of T is used to model this probability.

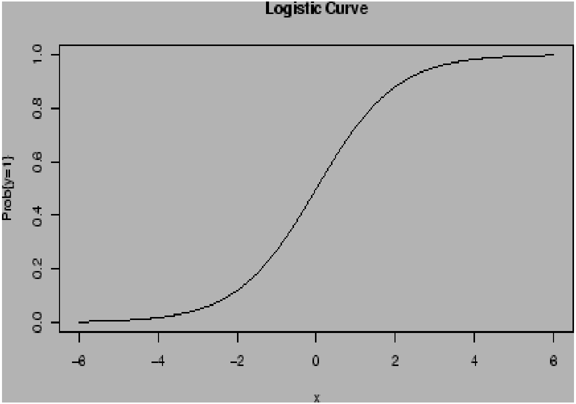


Figure 2: Shape of the logistic curve

The shape suggests that for some values of the predictor(s), the probability remains low. Then, there is some threshhold value of the predictor(s) at which the estimated probability of event begins to increase.

The logistic distribution

The logistic distribution looks very different from the normal distribution but has similar (but not identical) shape and cdf when plotted. For $\mu = 0$ and $\sigma = 1$, the logistic distribution has mean 0 but variance $\pi^3/3$ so we will compare the logistic distribution with mean 0 and $\sigma = 1$ to a $N(0, \pi^2/3)$.

The two distributions have the same first, second, and third moment, but have different fourth moments, with the logistic distribution being slightly more peaked. The two densities disagree more in the tails also, with the logistic distribution having larger tails (probabilities of extreme events are larger).

The logistic distribution

In R, you can get the density, cdf, etc. for the logistic distribution using

```
> dlogis()
> plogis()
> rlogis()
> qlogis()
```

As an example

```
> plogis(-8)
[1] 0.0003353501
> pnorm(-8,0,pi/sqrt(3))
[1] 5.153488e-06
```

Logistic versus normal

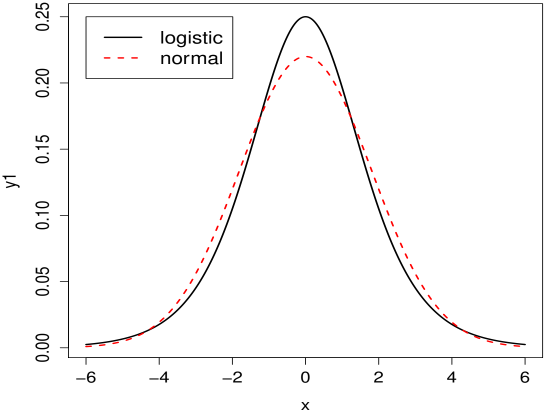


Figure 3: Pdfs of logistic versus normal distributions with the same mean and variance

Logistic versus normal

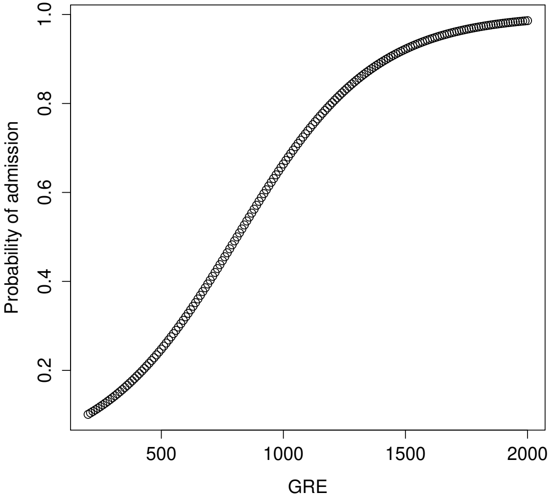


Figure 4: Cdfs of logistic versus normal distributions with the same mean and variance

Example continued: admission data

```
#Binary response: admit, 1 admit, 0 no admission.
#Three predictor variables: gre, gpa and rank.
#variables gre and gpa are continuous, rank is categorical
> head(ex.data)
  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
```

Interest: whether gpa of the student was related to the probability that the student got admitted.

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 gpa_i$$

where $\mu_i = P(ith \text{ student got admitted } |\text{gpa}_i)$

```
> nrow(ex.data)
[1] 400

> tapply(ex.data$gpa,ex.data$rank,mean)
       1        2        3        4
3.453115 3.361656 3.432893 3.318358
> tapply(ex.data$gre,ex.data$rank,mean)
       1        2        3        4
611.8033 596.0265 574.8760 570.1493
> xtabs(~admit + rank, data = ex.data)
     rank
admit  1  2  3  4
    0 28 97 93 55
    1 33 54 28 12
```

Fitting glm in R, we have the following results

```
myfit_gpa <- glm(admit ~ gpa, data = ex.data,
family = "binomial")
summary(myfit_gpa)
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.3576     1.0353  -4.209 2.57e-05 ***
gpa           1.0511     0.2989   3.517 0.000437 ***
```

- The fitted model is

$$\text{logit}(\mu_i) = -4.3576 + 1.0511 * gpa_i$$

- The column labelled “z value” is the Wald test statistic. $3.517 = 1.0511/0.2989$, since p-value $<< 0$, reject $H_0 : \beta_1 = 0$, conclude that GPA has an significant effect on log odds of admission.

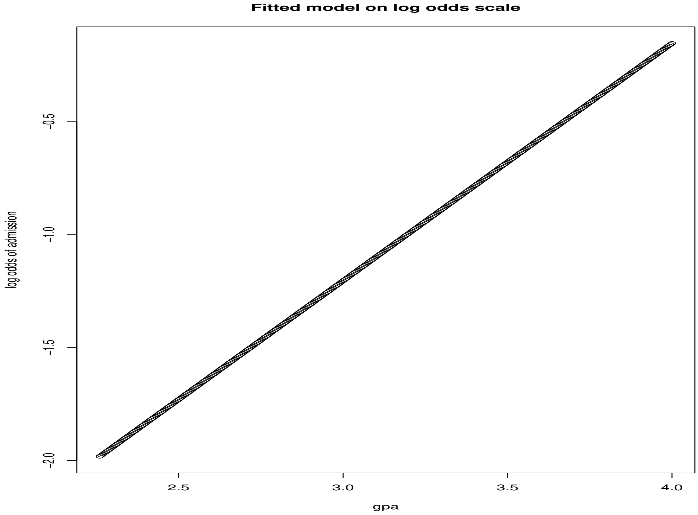


Figure 5: Fitted model on log-odds scale

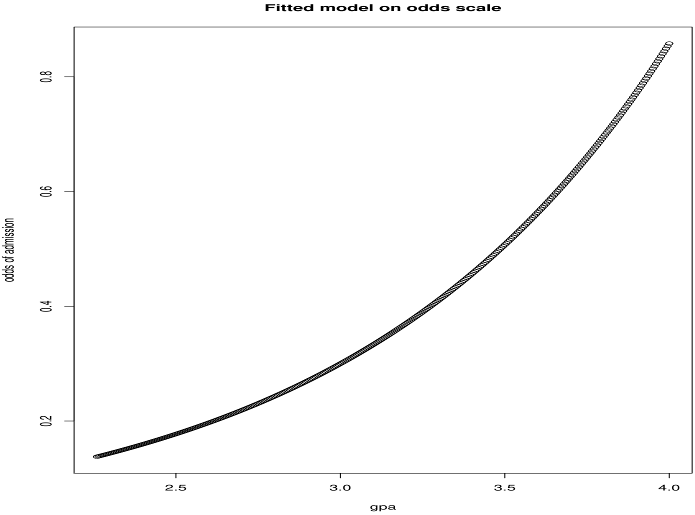


Figure 6: Fitted model on odds scale

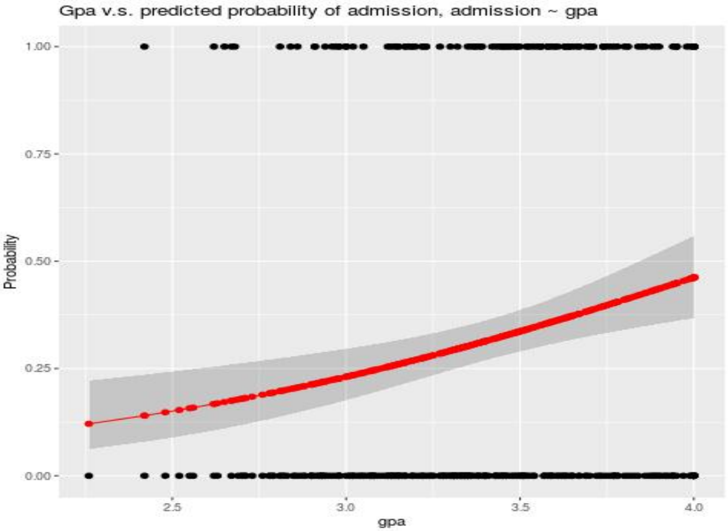


Figure 7: Fitted model on probability scale

Confidence intervals for the coefficients and the odds ratios

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} = \mathbf{x}_i' \boldsymbol{\beta}$$

- A $(1-\alpha) \times 100\%$ confidence interval for $\beta_j, j = 0, 1, \cdots, p-1$ can be calculated as

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \hat{se}(\hat{\beta}_j)$$

- The $(1-\alpha) \times 100\%$ confidence interval for the odds ratio over a one unit change in x_j is

$$\left[\exp(\hat{\beta}_j - Z_{1-\alpha/2} \hat{se}(\hat{\beta}_j)), \exp(\hat{\beta}_j + Z_{1-\alpha/2} \hat{se}(\hat{\beta}_j))\right]$$

Example

Fit admission status with gre, gpa and rank

```
###fit data with all variables
myfit <- glm(admit ~ gre + gpa + rank, data = ex.data,
 family = "binomial")
summary(myfit)

 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 ***
```

Example

- All predictors are significant, with gpa being a slightly stronger predictor than GRE score.
- The log-odds of being accepted increases by .804 for every unit increase in GPA when other variables held constant. ——- Of course a unit increase in GPA (from 3.0 to 4.0) is huge.
- The log-odds of being admitted to grad school is $-3.99 + .002gre + .804gpa - .675rank2 - 1.34rank3 - 1.55rank4$, so the probability of being admitted to grad school p is

$$p = \frac{e^{(-3.99+.002gre+.804gpa-.675rank2-1.34rank3-1.55rank4)}}{1 + e^{(-3.99+.002gre+.804gpa-.675rank2-1.34rank3-1.55rank4)}}$$

Note that the default is that the school has rank1.

Example

- Fitted probability

  The first observation is

  ```
  > ex.data[1,]
    admit gre  gpa rank
  1     0 380 3.61    3
  ```

  For this individual, the predicted probability of admission is

  $$p = \frac{e^{-3.99+.002(380)+.804(3.61)-1.34}}{1+e^{-3.99+.002(380)+.804(3.61)-1.34}} = 0.1726$$

  (If you only use as many decimals as I did here, you'll get 0.159 due to round off error).

You can get the predicted probabilities for this individual by

```
> myfit$fitted.values[1]
        1
0.1726265
```

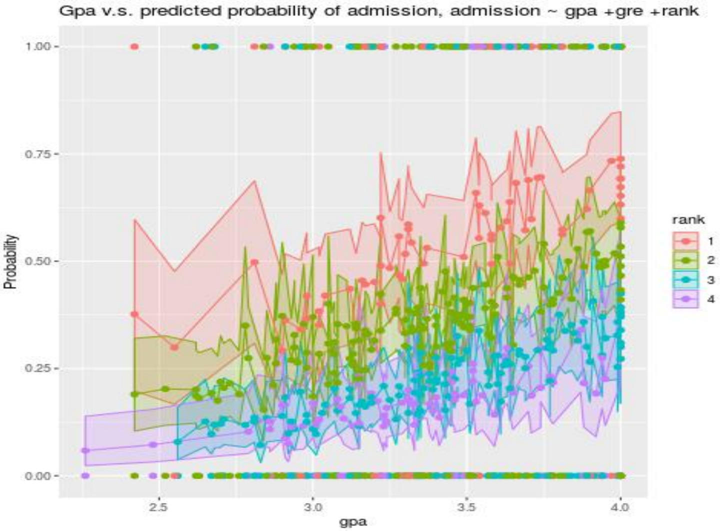


Figure 8: Fitted model on probability scale

Example

```
> names(myfit)
 [1] "coefficients"      "residuals"         "fitted.values"
 [4] "effects"           "R"                 "rank"
 [7] "qr"                "family"            "linear.predictors"
[10] "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"
[16] "df.residual"       "df.null"           "y"
[19] "converged"         "boundary"          "model"
[22] "call"              "formula"           "terms"
[25] "data"              "offset"            "control"
[28] "method"            "contrasts"         "xlevels"
>
```

- odds ratio with one unit change in gpa when all other variables are held constant is

$$\exp(0.804038) = 2.2345448$$

- 95% CI of odds ratio for one unit change in gpa is $[\exp(0.8040 - 1.96 * 0.3318), \exp(0.8040 + 1.96 * 0.3318)] = [e^{0.1537}, e^{1.4543}] = [1.1661, 4.2816]$

```
exp(cbind(OR = coef(myfit), confint(myfit)))
## Waiting for profiling to be done...
##                    OR       2.5 %    97.5 %
## (Intercept) 0.0185001 0.001889165 0.1665354
## gre         1.0022670 1.000137602 1.0044457
## gpa         2.2345448 1.173858216 4.3238349
## rank2       0.5089310 0.272289674 0.9448343
## rank3       0.2617923 0.131641717 0.5115181
## rank4       0.2119375 0.090715546 0.4706961
```

Model selection

```
myfit0<-glm(admit ~ 1, data = ex.data, family = "binomial")
upper<-formula(~gre+gpa+rank,data=ex.data)
model.aic = step(myfit0, scope=list(lower= ~., upper= upper
## Start:  AIC=501.98
## admit ~ 1
##
##          Df Deviance    AIC
## + rank   3   474.97 482.97
## + gre    1   486.06 490.06
## + gpa    1   486.97 490.97
## <none>       499.98 501.98
```

The **Akaike information criterion (AIC)** is an estimator of the relative quality of statistical models for a given set of data.

- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.
- AIC provides a means for model selection.

```
## Step:  AIC=472.88
## admit ~ rank + gpa
##
##        Df Deviance    AIC
## + gre   1   458.52 470.52
## <none>      462.88 472.88
## - gpa   1   474.97 482.97
## - rank  3   486.97 490.97
##
## Step:  AIC=470.52
## admit ~ rank + gpa + gre
##
##        Df Deviance    AIC
## <none>      458.52 470.52
## - gre   1   462.88 472.88
## - gpa   1   464.53 474.53
## - rank  3   480.34 486.34
```

- The smallest AIC $=$ 470.52, with variables rank, gpa and gre
- The second smallest one with AIC $=$472.88, with variables rank and gpa
- By model comparison for these two models, we would like to choose the full model with rank, gpa and gre.

```
myfit <- glm(admit ~ gre + gpa + rank, data = ex.data,
family = "binomial")
myfit3<-glm(admit ~ gpa+rank, data = ex.data,
 family = "binomial")
anova(myfit3,myfit)
qchisq(0.95,1)
pchisq(4.3578,1,lower.tail = FALSE)
> anova(myfit3,myfit)
Analysis of Deviance Table

Model 1: admit ~ gpa + rank
Model 2: admit ~ gre + gpa + rank
  Resid. Df Resid. Dev Df Deviance
1       395     462.88
2       394     458.52  1   4.3578
> qchisq(0.95,1)
[1] 3.841459
> pchisq(4.3578,1,lower.tail = FALSE)
[1] 0.03683985
```

Wald test

```
#test that the coefficient for rank=2 is equal to the
 coefficient for rank=3
 coef(myfit)
 (Intercept)          gre          gpa        rank2
-3.989979073  0.002264426  0.804037549 -0.675442928
 rank3        rank4
 -1.340203916 -1.551463677
l <- cbind(0, 0, 0, 1, -1, 0)
wald.test(b = coef(myfit), Sigma = vcov(myfit), L = l)
## Wald test:
## Chi-squared test:
## X2 = 5.5, df = 1, P(> X2) = 0.019
```

Since p-value for the test is 0.019, conclude that the coefficient for rank=2 is not equal to the coefficient for rank=3, or there is a significant difference between the effect on log odds of admission from rank 2 and rank 3 university applicants.

Assessment of model fit

- Model selection
- Residuals: can be useful for identifying potential outliers (observations not well fit by the model) or misspecified models. Residuals not very useful in logistic regression.
  —-Raw residual
  —Deviance residuals
  —-Pearson residuals
- Influence
  ——Cook's distance: measures the influent of case *i* on all of the fitted g_is
  ——Leverage
- Prediction

Example: logistic regression

$$\log\frac{\mu_i}{1-\mu_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

- $\hat{\mu}_i$: fitted probabilities
- raw residual: $y_i - \hat{\mu}_i$
- Pearson residuals: $\Gamma_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{\mu}_i)}}$

  —this is based on the idea of subtracting off the mean and dividing by the standard deviation

  —-if we replace $\hat{\mu}_i$ by μ_i, then Γ_i has mean 0 and variance 1.
- Deviance residuals: based on the contribution of each point to the likelihood

  —For logistic regression, $l = \sum_{i=1}^{n}\left\{y_i\log\hat{\mu}_i + (1-y_i)\log\hat{(}1-\mu_i)\right\}$

  —-

$$d_j = \text{sign}(y_j - \hat{\mu}_j)\sqrt{-2\left\{y_i\log\hat{\mu}_i + (1-y_i)\log\hat{(}1-\mu_i)\right\}}$$

if $y_i = 1$, $\text{sign}(y_j - \hat{\mu}_j) = 1$

—-if $y_i = 0$, $\text{sign}(y_j - \hat{\mu}_j) = -1$

- Each of these type of residuals can be squared and added together to create an (residual sum of squares) RSS-like statistic
—-Deviance: $D = \sum_{i=1}^{n} d_i^2$
—-Pearson statistic: $X^2 = \sum_{i=1}^{n} \Gamma_i^2$

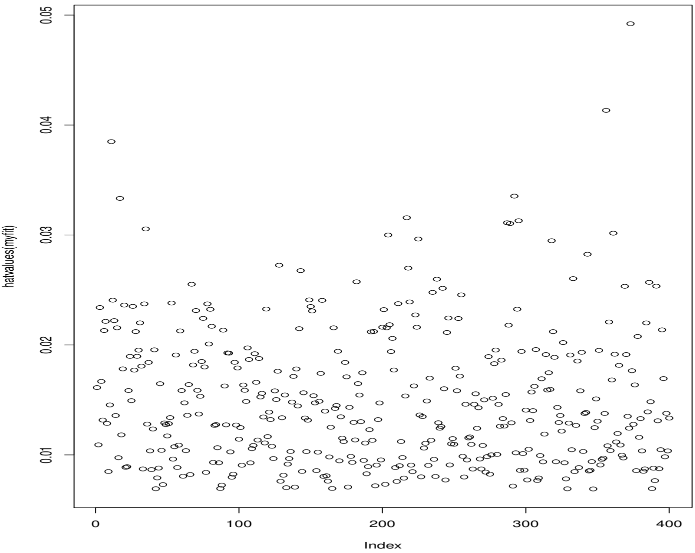
- Influential data, if removing the observation substantially changes the estimate of coefficients or fitted probabilities
- An observation with an extreme value on a predictor variable is called a point with high leverage.

  —— Leverage is a measure of how far an independent variable deviates from its mean. In fact, the leverage indicates the geometric extremeness of an observation in the multi-dimensional covariate space.

  —-These leverage points can have an unusually large effect on the estimate of logistic regression coefficients

  ——Leverages greater than $2\bar{h}$ or $3\bar{h}$ cause concerns, where $\bar{h} = p/n$

plot(hatvalues(myfit))



```
> highleverage <- which(hatvalues(myfit) > .045)
#0.45 = 3*p/n = 3*6/400
> hatvalues(myfit)[highleverage]
       373
0.04921401
> ex.data[373,]
    admit gre  gpa rank
373     1 680 2.42    1
> myfit$fit[373]
      373
0.3765075
> mgre
       1        2        3        4
611.8033 596.0265 574.8760 570.1493
> mgpa
       1        2        3        4
3.453115 3.361656 3.432893 3.318358
```

- Cook's distance
  If $\hat{\beta}$ is the MLE of β under the model

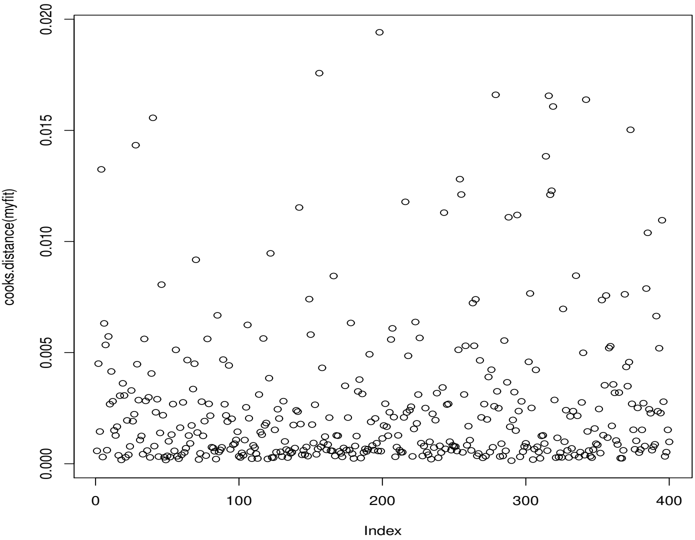
$$g(\mu_i) = \mathbf{x}_i' \beta$$

  and $\hat{\beta}_{(-j)}$ is the MLE based on the data but holding out the jth observation, then cooks distance for case j is

$$\begin{aligned} c_k &= \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(-j)})'[\widehat{Var}(\hat{\beta})]^{-1}(\hat{\beta} - \hat{\beta}_{(-j)}) \\ &= \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(-j)})'\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}(\hat{\beta} - \hat{\beta}_{(-j)}) \end{aligned}$$

  Some package doesn't scale c_j by p.

plot(cooks.distance(myfit))



```
> max(cooks.distance(myfit))
[1] 0.01941192
> highcook <- which((cooks.distance(myfit)) > .05)
#0.05 is simply a very small critical number in $F$
distribution
> cooks.distance(myfit)[highcook]
named numeric(0)
```

Comments:

- In a binomial setup where all n_i are big the standardized deviance residuals should be closed to Gaussian. The normal probability plot can be used to check this.
- In a binomial setup where x_i (number of successes) are very small in some of the groups numerical problems sometimes occur in the estimation. This is often seen in very large standard errors of the parameter estimates.

- Residuals are less informative for logistic regression than they are for linear regression:
  ——yes/no (1 or 0) outcomes contain less information than continuous ones
  —— the fact that the adjusted response depends on the fit hampers our ability to use residuals as external checks on the model
- We are making fewer distributional assumptions in logistic regression, so there is no need to inspect residuals for, say, skewness or non constant variance
- Issues of outliers and influential observations are just as relevant for logistic regression and GLM models as they are for linear regression
- If influential observations are present, it may or may not be appropriate to change the model, but you should at least understand why some observations are so influential

Prediction

Fitted probabilities:

```
###prediction, fitted probabilities
myfit$fit[1:20]     #fitted probabilities
##          1      2       3      4         5
## 0.17262654 0.29217496 0.73840825 0.17838461 0.11835391
6      7       8       9       10
0.36996994  0.41924616 0.21700328 0.20073518   0.51786820
##     11       12     13      14         15
##0.37431440   0.40020025   0.72053858 0.35345462 0.6923798
##    16      17     18      19         20
## 0.18582508    0.33993917 0.07895335 0.54022772 0.573511
```

Predicted probabilities:

```
mgre<-tapply(ex.data$gre, ex.data$rank, mean)
# mean of gre by rank
mgpa<-tapply(ex.data$gpa, ex.data$rank, mean)
# mean of gpa by rank
newdata1 <- with(ex.data, data.frame(gre = mgre,
gpa = mgpa, rank = factor(1:4)))
newdata1
##        gre      gpa rank
## 1 611.8033 3.453115    1
## 2 596.0265 3.361656    2
## 3 574.8760 3.432893    3
## 4 570.1493 3.318358    4
```

```
newdata1$rankP <- predict(myfit, newdata = newdata1,
type = "response")
newdata1
##        gre      gpa rank     rankP
## 1 611.8033 3.453115    1 0.5428541
## 2 596.0265 3.361656    2 0.3514055
## 3 574.8760 3.432893    3 0.2195579
## 4 570.1493 3.318358    4 0.1704703
```

- The predicted probability of being accepted into a graduate program is 0.5429 for students from the highest prestige undergraduate institutions (rank$=$ 1), with gre $=$ 611.8 and gpa$=$3.45 .

Translate the estimated probabilities into a predicted outcome

1. Use 0.5 as a cutoff.
   ——if $\hat{\mu}_i$ for a new observation is greater than 0.5, its predicted outcome is $y = 1$.
   —- if $\hat{\mu}_i$ for a new observation is less than or equal to 0.5, its predicted outcome is $y = 0$.

- This approach is reasonable when
  (a) it is equally likely in the population of interest that the outcomes 0 and 1 will occur, and
  (b) the costs of incorrectly predicting 0 and 1 are approximately the same.

2. Find the best cutoff for the data set on which the logistic regression model is based.
   ——we evaluate different cutoff values and for each cutoff value, calculate the proportion of observations that are incorrectly predicted.
   ——select the cutoff value that minimizes the proportion of incorrectly predicted outcomes.

- This approach is reasonable when
  (a) the data set is a random sample from the population of interest, and
  (b) the costs of incorrectly predicting 0 and 1 are the same.

Example:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 gre_i + \beta_2 gpa_i + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i}$$

if we use the cutoff of 0.5, we get the following results

```
> table(ex.data$admit,fitted(myfit)>.5)

    FALSE TRUE
  0   254   19
  1    97   30
> t1<-table(ex.data$admit,fitted(myfit)>.5)
> (t1[1,2]+t1[2,1])/sum(t1)
[1] 0.29
```

Recall that 1 means admission, 0 no admission. We misclassify people (97+19)/400=29% of the time.

Instead, let's try finding a classification rule that minimizes misclassification in our data set.

```
for(p in seq(.15,.9,.05))
{t1<-table(ex.data$admit,fitted(myfit)>p)
cat(p,(t1[1,2]+t1[2,1])/sum(t1),"\n")}
0.35 0.325
0.4 0.3
0.45 0.3075
0.5 0.29
0.55 0.29
0.6 0.3025
0.65 0.3075
0.7 0.315
Error in t1[2, 1] : subscript out of bounds
> max(fitted(myfit))  [1] 0.7384082
```

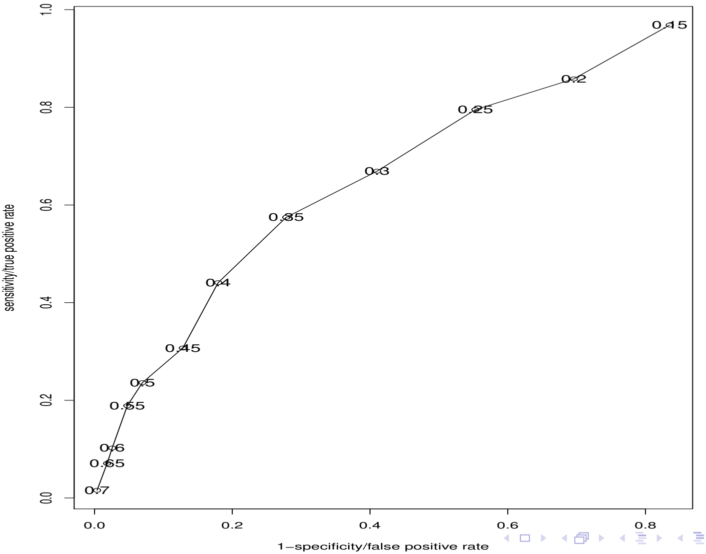
It looks like we can't do much better than 29%.

Receiver operating characteristic (ROC) curve

ROC curve is a plot of 1-specificity against sensitivity.

- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The true-positive rate is also known as sensitivity. The false-positive rate is also known as the fall-out or probability of false alarm, and can be calculated as $(1 - \text{specificity})$.
- The ROC curve is the sensitivity as a function of fall-out.

```
#Roc curve
p1<-matrix(0,nrow=12,ncol=3)
i=1
for(p in seq(0.15,.7,.05)){
t1<-table(ex.data$admit,fitted(myfit)>p)
p1[i,]=c(p,1-(t1[1,1])/sum(t1[1,]),(t1[2,2])/sum(t1[2,]))
i=i+1
}
plot(p1[,2],p1[,3],type = "o",
xlab="1-specificity/false positive rate",
ylab="sensitivity/true positive rate")
text(p1[,2],p1[,3],p1[,1],cex=1.2)
#p1[,2] false positive rate (type I error)
#p1[,3] true postive rate (power)
```


0.0
0.2
0.4
0.6
0.8
1.0
sensitivity/true positive rate
0.0
0.2
0.4
0.6
0.8
1−specificity/false positive rate
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7

```
dp1<-data.frame(p1)
names(dp1)<-c("cutt off prob","type I error","power")
print(dp1)
> print(dp1)
   cutt off prob type I error      power
1           0.15  0.835164835 0.96850394
2           0.20  0.695970696 0.85826772
3           0.25  0.553113553 0.79527559
4           0.30  0.410256410 0.66929134
5           0.35  0.278388278 0.57480315
6           0.40  0.179487179 0.44094488
7           0.45  0.128205128 0.30708661
8           0.50  0.069597070 0.23622047
9           0.55  0.047619048 0.18897638
10          0.60  0.025641026 0.10236220
11          0.65  0.018315018 0.07086614
12          0.70  0.003663004 0.01574803
```

Comments:

- The area under the ROC curve can give us insight into the predictive ability of the model.
- If it is equal to 0.5 (an ROC curve with slope $= 1$), the model can be thought of as predicting at random.
- Values close to 1 indicate that the model has good predictive ability.
- It can also be thought of as a plot of the Power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities).