Discussion9

Jing Lyu

3/8/2023

Logistic regression

To model a binary outcome, (X_i, y_i), $y_i \in \{0, 1\}$, $X_i \in \mathbb{R}^p$, we use logistic regression

$$\text{logit}(\pi_i) = X_i^T \beta$$

where $\pi_i = p(y_i = 1|X_i)$ and $\text{logit}(a) = \log(a/(1-a))$.

For this section, we use `birthwt` dataset from the `MASS` package. The `birthwt` data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986. We want to study the risk factors associated with low baby birth weight.

(1) `low`: indicator of birth weight less than 2.5 kg.

(2) `lwt`: mother's weight in pounds at last menstrual period.

(3) `race`: mother's race (1 = white, 2 = black, 3 = other).

(4) `smoke`: smoking status during pregnancy.

(5) `ptd`: indiator of previous premature labours.

```
library(MASS)
example(birthwt) # run R codes from the Examples section of R's online help topic
```

```
##
## brthwt> bwt <- with(birthwt, {
## brthwt+ race <- factor(race, labels = c("white", "black", "other"))
## brthwt+ ptd <- factor(ptl > 0)
## brthwt+ ftv <- factor(ftv)
## brthwt+ levels(ftv)[-(1:2)] <- "2+"
## brthwt+ data.frame(low = factor(low), age, lwt, race, smoke = (smoke > 0),
## brthwt+            ptd, ht = (ht > 0), ui = (ui > 0), ftv)
## brthwt+ })
##
## brthwt> options(contrasts = c("contr.treatment", "contr.poly"))
##
## brthwt> glm(low ~ ., binomial, bwt)
##
## Call:  glm(formula = low ~ ., family = binomial, data = bwt)
##
## Coefficients:
## (Intercept)          age          lwt    raceblack    raceother    smokeTRUE
##     0.82302     -0.03723     -0.01565      1.19241      0.74068      0.75553
##     ptdTRUE       htTRUE       uiTRUE         ftv1         ftv2+
##     1.34376      1.91317      0.68020     -0.43638      0.17901
##
```

```
## Degrees of Freedom: 188 Total (i.e. Null);  178 Residual
## Null Deviance:       234.7
## Residual Deviance: 195.5     AIC: 217.5
```

Likelihood Ratio Test LRT can be applied to study whether a reduced model is preferred. Note that this test can only be used for nested models, where the null model is the smaller model (a special case of the larger model), and the alternative is the larger model.

$$\text{LR} = -2\big[\log \mathcal{L}(\text{reduced}) - \log \mathcal{L}(\text{full})\big]$$

We have that $\text{LR} \sim \chi^2_{K-k}$, under H_0 and large n. K is the number of parameters in full model and k is the number of parameters in reduced model.

Suppose we want to test

$$H_0 : \beta_{ptd} = 0 \quad H_a : \beta_{ptd} \neq 0$$

```
bwtfit = glm(low ~ lwt + race + smoke + ptd,
                   family = binomial(), data = bwt)
h0.fit = glm(low ~ lwt + race + smoke, family = binomial(), data = bwt)
anova(h0.fit, bwtfit, test = 'Chi')
```

```
## Analysis of Deviance Table
##
## Model 1: low ~ lwt + race + smoke
## Model 2: low ~ lwt + race + smoke + ptd
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       184     215.01
## 2       183     207.04  1   7.9752 0.004742 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The small p-value indicates the null is rejected. Therefore, the larger model is more appropriate.

Deviance Table Devaiance table is a sequential variable selection method, and is **sensitive to the order of parameters**.

In the model specification, the model selection is starting with an intercept only model, and sequentially testing additional terms to enter the model based on likelihood ratio test.

```
anova(glm(low ~ smoke + ptd + lwt, family = binomial(), data = bwt), test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: low
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    188     234.67
## smoke  1   4.8674       187     229.81  0.02737 *
## ptd    1  10.4757       186     219.33  0.00121 **
## lwt    1   4.1060       185     215.22  0.04273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(glm(low ~ lwt + ptd + smoke, family = binomial(), data = bwt), test = "Chi") # change the order
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: low
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    188     234.67
## lwt    1   5.9813       187     228.69 0.0144581 *
## ptd    1  11.1939       186     217.50 0.0008207 ***
## smoke  1   2.2739       185     215.22 0.1315675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have inconsistent results from the models with different orders of the predictors. We need to compare all the deviance tables corresponding to different orders of predictors. Deviance tables are used for model selection only when the number of the predictors is small.

```
summary(bwtfit)
```

Interpretation

```
##
## Call:
## glm(formula = low ~ lwt + race + smoke + ptd, family = binomial(),
##     data = bwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8628  -0.8209  -0.5589   0.9826   2.0806
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.380164   0.915144  -0.415   0.6778
## lwt         -0.012046   0.006463  -1.864   0.0623 .
## raceblack    1.278212   0.520349   2.456   0.0140 *
## raceother    0.898946   0.423585   2.122   0.0338 *
## smokeTRUE    0.877065   0.391881   2.238   0.0252 *
## ptdTRUE      1.223296   0.436828   2.800   0.0051 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 207.04  on 183  degrees of freedom
## AIC: 219.04
##
## Number of Fisher Scoring iterations: 4
```

The fitted model is :

$$logit\{P(having\ low\ birth\ weight)\} = -0.38 - 0.012X_{lwt} + 1.28X_{race=Black} + 0.90X_{race=Other} + 0.88X_{smoke} + 1.22X_{ptd}$$

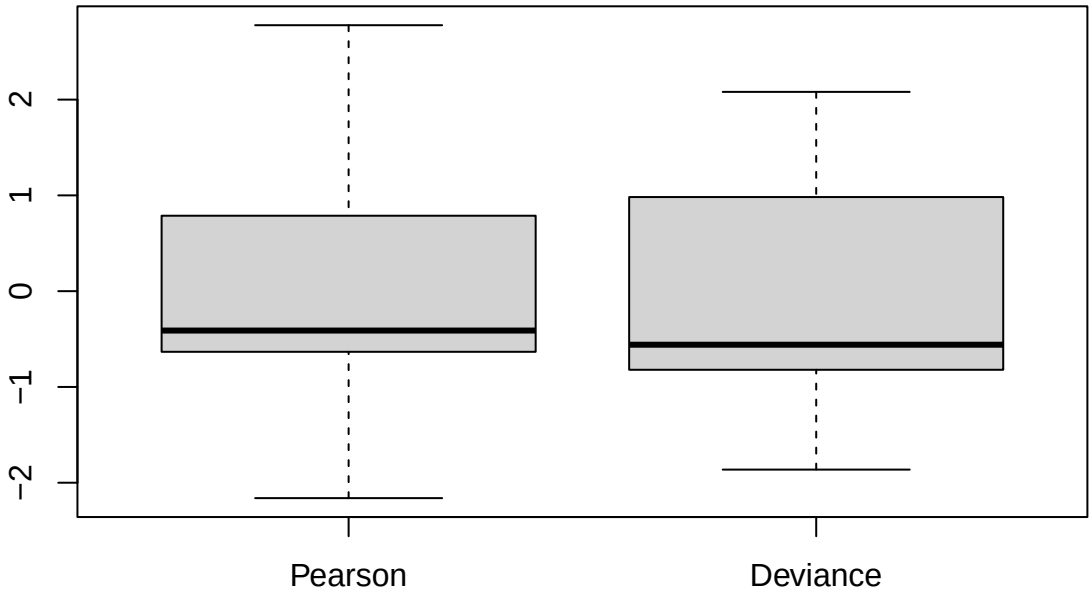
Effect of smoking status during pregnancy: The odds of having the baby with low birth weight for mothers smoking during pregnancy is $e^{0.88} = 2.41$ times that of having the baby with low birth weight for mothers not smoking during pregnancy.

Model Diagnostics

1. Pearson residuals and deviance residuals

If the two kinds of residuals are not quite similar to each other, the model may suffer from potential lack-of-fit.

```
res.P = residuals(bwtfit, type = "pearson")
res.D = residuals(bwtfit, type = "deviance")
boxplot(cbind(res.P, res.D), names = c("Pearson", "Deviance"))
```



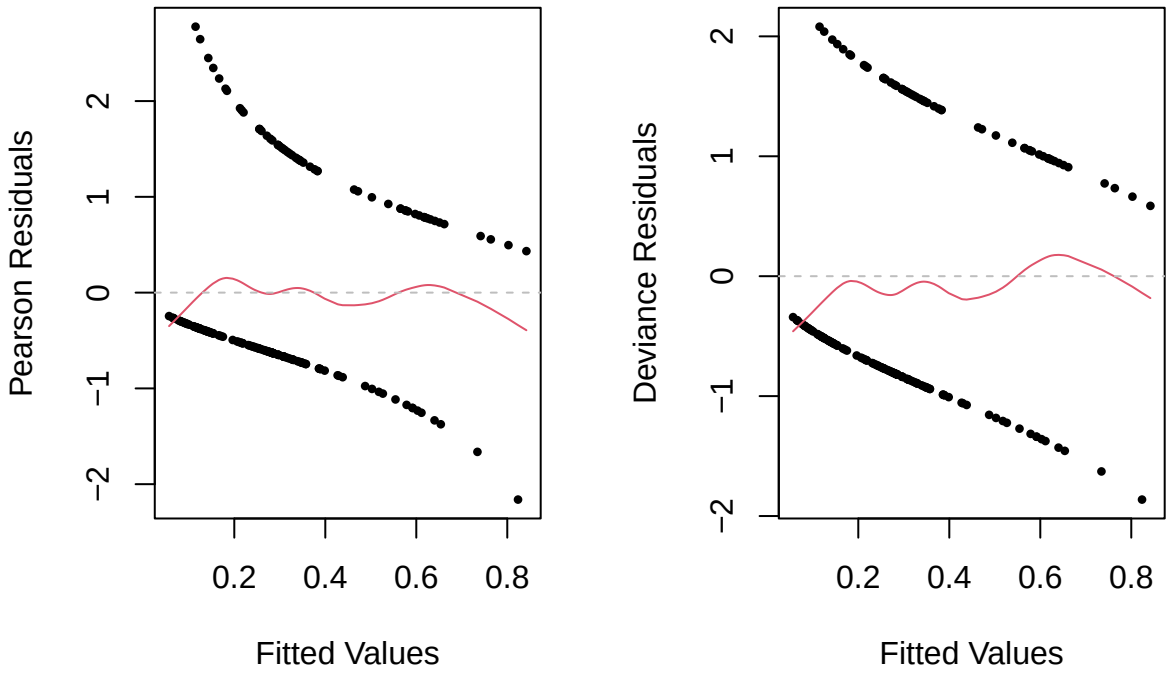
The boxplots show similar distributions of the two types of residuals, no lack-of-fit is provided.

2. Residual plots

The purpose is to check if there are any systematic patterns left in the residuals.

The scatter plot itself does not provide much information due to the special type of binary response type in logistic regression. It is useful to complement the residual plot with an overlaying smoothing splines fit, shown as red curve. The red curves are quite close to 0 in the two plots, but may have a slight quadratic pattern. Higher order terms or interaction terms can be added to see if the pattern exists.

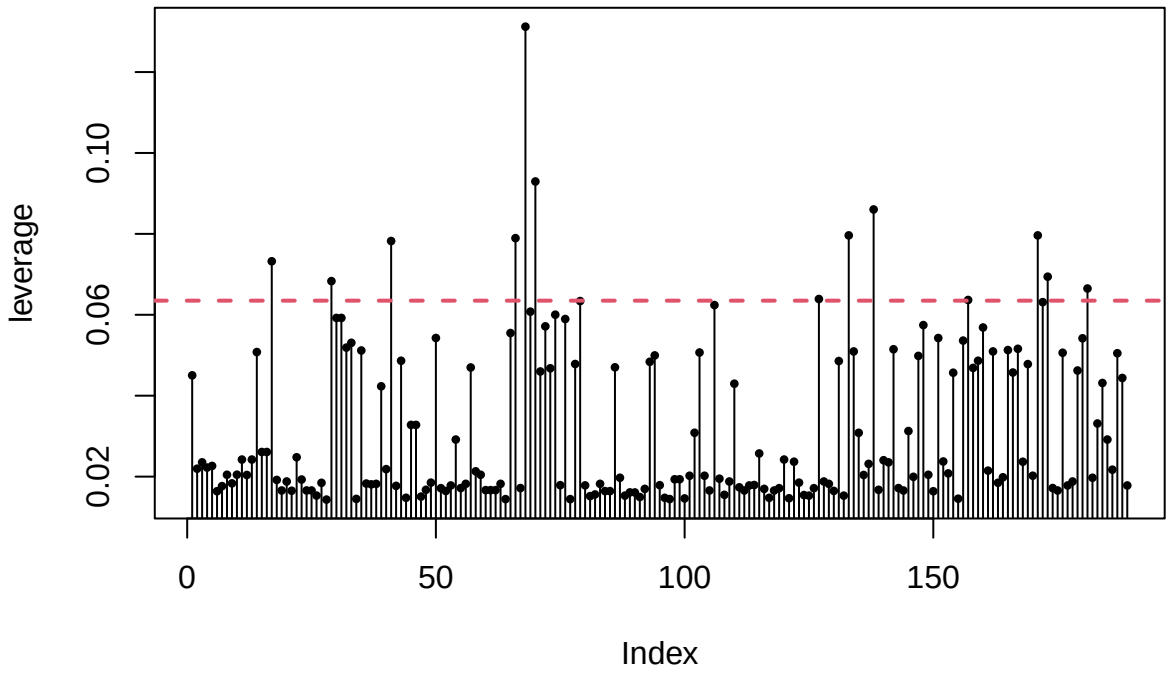
```
par(mfrow=c(1,2))
plot(bwtfit$fitted.values, res.P, pch=16, cex=0.6, ylab='Pearson Residuals', xlab='Fitted Values')
lines(smooth.spline(bwtfit$fitted.values, res.P, spar=0.9), col=2)
abline(h=0, lty=2, col='grey')
plot(bwtfit$fitted.values, res.D, pch=16, cex=0.6, ylab='Deviance Residuals', xlab='Fitted Values')
lines(smooth.spline(bwtfit$fitted.values, res.D, spar=0.9), col=2)
abline(h=0, lty=2, col='grey')
```



3. Leverage points

To identify influential data points, we plot the leverage h_{ii} (diagonal of hat matrix) against the index of the points. An observation is suspected as a leverage point if $h_{ii} > 2p/n$ where p is the number of coefficients and n is sample size.

```
par(mfrow=c(1,1))
leverage = hatvalues(bwtfit)
plot(names(leverage), leverage, xlab="Index", type="h")
points(names(leverage), leverage, pch=16, cex=0.6)
p = length(coef(bwtfit))
n = nrow(bwt)
abline(h=2*p/n,col=2,lwd=2,lty=2)
```

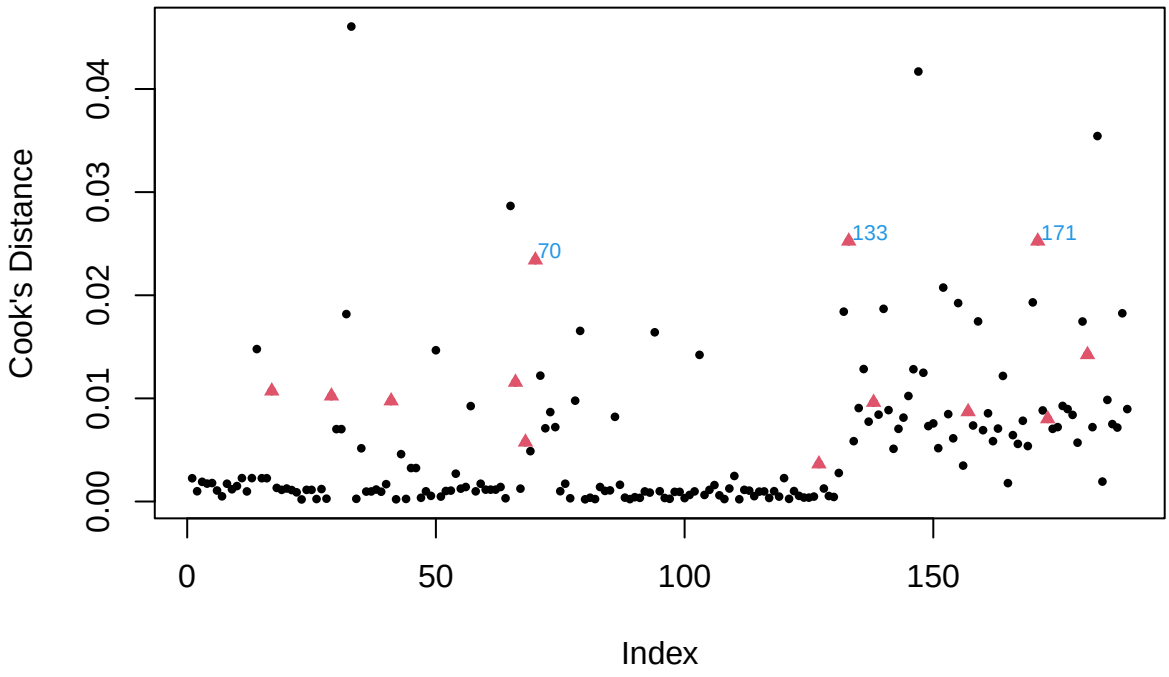



```
infPts = which(leverage>2*p/n)
```

4. Cook's distance

To detect outliers/influential observations, we can use Cook's distance.

```
cooks = cooks.distance(bwtfit)
plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6)
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2) # influential points
susPts = as.numeric(names(sort(cooks[infPts], decreasing=TRUE)[1:3]))
text(susPts, cooks[susPts], susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)
```



Differences between influential points and outliers:

Outlier: a point with a large residual.

Influential point: a point that has a large impact on the regression.

They are not the same thing. A point can be an outlier without being influential. A point can be influential without being an outlier. A point can be both or neither.

Prediction, Sensitivity and Specificity Now, we split the dataset into training set (70%) and test set (30%). Then, we use the model trained with training set to predict the birth weight indicator in the test set.

```
# Splitting dataset
library(caTools)
set.seed(123)
split = sample.split(bwt$low, SplitRatio = 0.7) # use 70% of dataset as training set and 30% as test se
bwt.train = subset(bwt, split == "TRUE")
bwt.test = subset(bwt, split == "FALSE")
bwtfit.train = glm(low ~ lwt + race + smoke + ptd,
                   family = binomial(), data = bwt.train)
threshold = 0.5
predicted_values = ifelse(predict(bwtfit.train, newdata = bwt.test)>threshold,1,0)
actual_values = bwt.test$low
conf_matrix = table(predicted_values, actual_values)
conf_matrix
```

```
##                 actual_values
## predicted_values  0  1
##                0 34 12
##                1  5  6
```

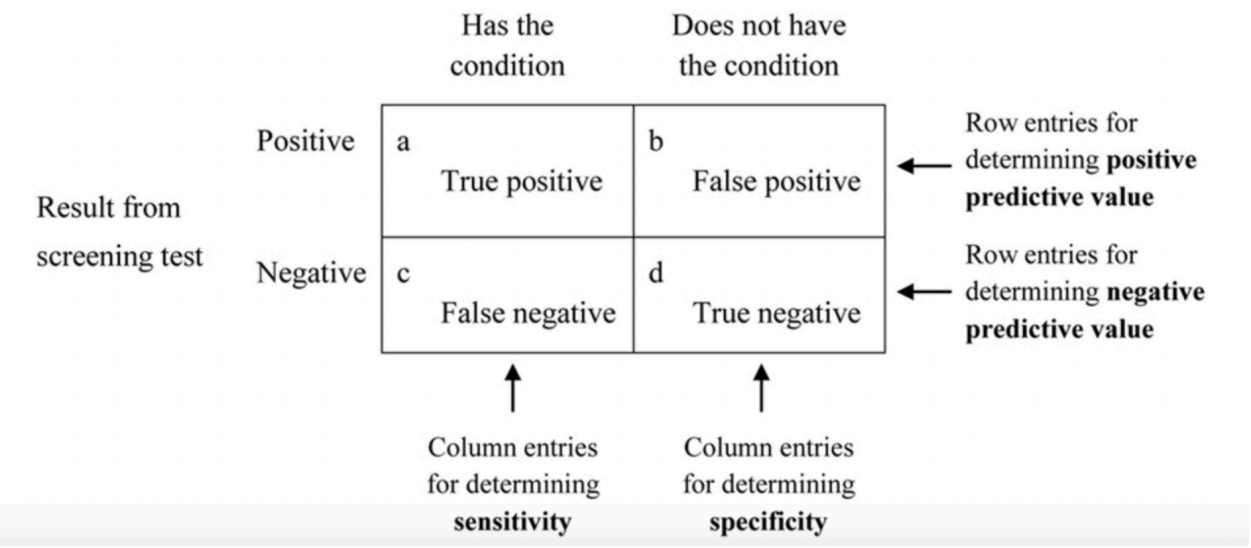
- Sensitivity (True positive rate): the probability of a positive test result, conditioned on the individual truly being positive. Formula: $\frac{TP}{TP+FN}$
- Specificity (True negative rate): the probability of a negative test result, conditioned on the individual truly being negative. Formula: $\frac{TN}{TN+FP}$

Based on the confusion matrix, we have

$$Sensitivity = 6/18 \approx 0.33, Specificity = 34/39 \approx 0.87$$

Low sensitivity can result in false negatives, incorrectly identifying low birth weight as normal birth weight.

Sensitivity and specificity are inversely related.


Has the condition
Does not have the condition
Result from screening test
Positive
Negative
a
True positive
b
False positive
c
False negative
d
True negative
Row entries for determining positive predictive value
Row entries for determining negative predictive value
Column entries for determining sensitivity
Column entries for determining specificity