Lung Cancer: A Multinomial Regression Data Analysis

Rong Duan

Abstract

Lung cancer is one of the most common cancers in the world, and the impact of lung cancer varies with the different levels of cancer. Compared with low cancer levels, medium and high levels have a higher risk and require much attention to longer patient's lives. In this study, a baseline odds model was implemented, and the low level was set as the baseline level. Alcohol use, air pollution, snoring, smoking, chronic lung disease, and swallowing difficulty are important characteristics related to cancer levels. The interaction term of alcohol use and swallowing difficulty is added through model diagnosis and model selection. After moving the influential outliers, the model with the final format is fitted. The result shows that with one unit increase in air pollution or chronic lung disease, and patients are more likely to fall into the high level of cancer. The effects of alcohol use or swallowing difficulty may be different since the interaction term is included in the model, the specific change varies with different base values. Among these predictors, smoking is highly correlated with alcohol consumption and chronic lung diseases, thus its influence is offset in the fitted model.

1 Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death in the world [1]. One in 16 people in the US will be diagnosed with lung cancer in their lifetime, but only 21% of all people with diagnosed lung cancer will survive 5 years or more [2]. It has been found that many factors, like smoking and air pollution, are the causes for lung cancer [3]. Some behaviors like drinking alcohol and snoring are found highly related to lung cancer in some studies [3]. In this study, 1000 lung cancer patients are involved, and their cancer levels('low', 'medium', and 'high') and some correlated features are recorded. This project aims to found correlated features for cancer levels, especially medium and high levels, and interpret how do these features correlated with cancer levels.

2 Method

2.1 Basic covariates exploration

A total of 1000 lung cancer patients are involved and 20 features are recorded. Two of them are basic biological features: Age and gender. 59.8% among them are male, and 40.1% among them are female. Among high cancer level patients, 69% of them are male, and 31% are female. The rest recorded features are the highly suspicious ones that may relate to lung cancer. All of them are recorded as ordered levels, and for simplicity, these data are treated as continuous variables for data analysis. Appendix: Table 2 is the summary mean value for the three cancer levels. Some features like alcohol use, the mean value is very different in each cancer levels; while some features, like air pollution, the mean value in low and medium levels are close, the mean value in high cancer level is much higher. Smoking is discovered that explains almost 90% of lung cancer risk in men and 70 to 80% in women [4]. The predictor 'Smoking' and 'Passive Smoker' are quite close with respect to the mean value in each group, thus a high correlation might exist between the two predictors.

2.2 Model building and data analysis

The data consists of 3 cancer levels, low (30%), medium(33%), and high(37%). In this data analysis, cancer level is the response variable, and other covariates are explored to found the relationship with cancer level. The baseline odds model and proportional odds model are considered for multinomial regression analysis. Since the proportional odds model assumes the all the level share the coefficients of the same covariates, baseline odds model are utilized to explore the possibility of the same coefficients in all levels.

The main interest is to find significant features related to medium and high cancer level since these two levels are riskier, thus low cancer level is set to be the baseline. The baseline odds model is implemented to explore the significant differences of the other two cancer levels compared with the low level. Since the numerous covariates and some of them are highly correlated, model selection with AIC criteria is implemented. After model selection, some significant covariates are kept in the model. To check the goodness-of-fit, Pearson residuals are plotted for the two compared groups. A smooth line is fitted to check whether there is any systematic pattern left in the residuals. If some pattern shows, consider quadratic terms and interaction terms. Repeat the procedure until no systematic pattern in the residual plots. Leverage plots and cook's distance plots are used to check whether influential outliers exist, and delete the detected influential outliers. Refit the model, and interpreted the results with estimated parameters.

3 Results

3.1 Model building

The baseline odds model is utilized to do a multinomial regression, low cancer level is regarded as the baseline level. Since smoking and air pollution are found can cause lung cancer in much scientific research, these two predictors are kept in the model while proceeding with the model selection. Six predictors are selected, they are alcohol use, air pollution, snoring, smoking, chronic lung disease, and swallowing difficulty. The p-value of each estimated parameter is significant. However, Pearson's residual plot shows a quadratic pattern in both two residual plots, which is an indication that the current model is lack fit. Thus, quadratic order and interaction terms are considered to be added in the model. After the model selection with AIC criteria, an interaction term (Alcohol use * Swallowing Difficulty) is added since the significant effect.

3.2 Model diagnostic

For the selected model, the Pearson residual plot is used to check the goodness of fit. In Figure 2, the Pearson residual plots don't show any pattern, which indicates the newly selected model fits quite well [Figure 1]. The leverage plot and cook's distance plot is checked for the influential outlier. For the medium-low comparison group, several points are especially influential compared with others. By checking the cook's distance plot, 524_{th} observation is an influential outlier [Figure 2]. Similarly, for the high-low comparison group, 25_{th} observation is detected as an influential outlier as well [Appendix.Figure 1]. These two outliers are deleted from the data and refit the model.

3.3 Model Explanation

The fitted model with 16 parameters are as follows.

$$\frac{p_{i2}}{p_{i1}} = \exp(\eta_{i2}), \quad \frac{p_{i3}}{p_{i1}} = \exp(\eta_{i3}) \quad i = 1, ..., 998$$

Where p_{i1} is the probability of baseline level, p_{i2} is the probability of medium cancer level, and p_{i3} is the probability of high cancer level. η_{ij} is the combination of linear predictors. To be more specific, $\eta_{ij} = \beta_{0j} + \beta_{1j}*\text{Alcohol-use}_i + \beta_{2j}*\text{Air Pollution}_i + \beta_{3j}*\text{Snoring}_i + \beta_{4j}*\text{Smoking}_i + \beta_{5j}*$ Chronic LungDisease$_i$ $+ \beta_{6j}*\text{Swallowing Difficulty}_i + \beta_{7j}*(\text{Alcohol use :Swallowing Difficulty})_i$, and $j = 2, 3$, which represent the medium and high level respectively. The estimated value is in Table 1.

When comparing medium and high level with Low level, some interesting results are found. Air pollute is found more related to high-level cancer. With one unit increase in air pollution, the odd

term	(Intercept)	Alcohol use	Air Pollution	Snoring
M-L:estimate(p-value)	-21.100(0.000)	3.883(0.000)	0.163(0.356)	1.203(0.000)
H-L:estimate(p-value)	-13.2(0.000)	1.296(0.000)	1.351(0.000)	0.766(0.000)
term	Smoking	chronic Lung Disease	Swallowing Difficulty	Alcohol:Swallowing Diff
M-L:estimate(p-value)	-0.680(0.000)	0.153(0.272)	3.348(0.000)	-0.663(0.000)
H-L:estimate(p-value)	-0.529(0.000)	0.294(0.128)	-0.057(0.888)	0.163(0.081)

Table 1: Estimated parameters

that the patient falls into high cancer level enlarge 3.86 times versus fall into the low level; while one unit increase in air pollution enlarge 1.17 times of the odds that patients fall into medium level compared with low level. Since high collinearity exists, smoking seems to lower the odds that the patient goes into a higher cancer level compared with a low cancer level. With one unit increase in chronic lung disease, the odds that patients fall into medium and high level enlarges 1.16 and 1.35 times compared with fall into a low level. With one unit increasing of snoring, it will enlarge 2.15 times the odds that a patient falls into the high cancer level instead of low cancer level.

If the value of swallowing difficulty is 0, the explanation of the Alcohol use change is kind of simple. With a one-unit increase in the Alcohol use, the odds of falling into medium level versus low level enlarge 48.6 times, holding the other terms constant. While one unit increases the "Alcohol use" only enlarges 3.65 times odds it falls into high cancer level. Combined with the value in Table 1, the mean value difference between high level and low level is around twice compared with the difference between medium level and low level. Thus, with a unit increase in Alcohol use, the odds of patient fall into high level is lower than medium level.

Similarly, if the value of alcohol use is 0, with one unit increasing of swallowing difficulty, it will enlarges 28.44 times the odds that the patients fall into the medium level instead of the low level. However, in reality the zero value of either alcohol use or swallowing difficulty is rare, thus a more specific explanation is in 4.1.

4 Discussion

4.1 Interaction term of baseline odds model

The interpretation of the increase of "Alcohol use" and "Swallowing difficulty" will change when considering the interaction term. For one unit increase of Alcohol use, the log odds between medium and low level will increase $\exp(3.883 - 0.663 * SwallowingDifficulty)$, which means the change of

log odds is determined by the swallowing difficulty. Similarly for Swallowing difficulty, if the patient increases one unit of Swallowing difficulty, the change of log odds is determined by his Alcohol use condition.

4.2 Influential effects for lung cancer levels

Based on the above-fitted model, air pollution and chronic lung disease" is an influential feature for high levels. In table 2, the mean value in medium and low levels aren't different much, but at a high level, the mean value of smoking is much larger. But in the regression model, the estimated coefficient of smoking is opposite, which shows one unit increasing in smoking will lower the odds that a patient will fall into the medium or high level rather than the low level. The reason is that smoking is highly correlated with alcohol use and chronic lung disease, thus the effect is offset by these features.

4.3 Proportional odds model

Based on the estimated parameters in the baseline odds model, although some estimated parameters have similar values in the two compared groups, still some parameters have a very different estimated coefficient value. The proportional odds model might not be appropriate in this case. And only the baseline odds model is implemented.

5 Conclusion

In this data analysis report, among many symptoms features, alcohol use, air pollution, snoring, smoking, chronic lung disease, and swallowing difficulty are detected as significant features related to cancer levels. These features can influence the probability that a lung cancer patient goes into a more severe cancer level, thus need to pay more attention when these symptoms go worse.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018), Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68: 394-424. https://doi.org/10.3322/caac.21492

[2] Howlader N, Noone AM, Krapcho M, Miller D, Bresi A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2017, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975$_2$017, $basedonNovember2019SEERdatasubmission, postedtotheSEERwebsi$ $AlbergAJ, BrockMV, SametJM(2016)."Chapter52: Epidemiologyoflungcancer".MurrayNadel'sTextbook$ $1-4557-3383-5$.

[3] Anne Sofie Christensen, BSc, Alice Clark, MSc, Paula Salo, PhD, Peter Nymann, PhD, Peter Lange, DMSc, Eva Prescott, DMSc, Naja Hulvej Rod, PhD, Symptoms of Sleep Disordered Breathing and Risk of Cancer: A Prospective Cohort Study, Sleep, Volume 36, Issue 10, 1 October 2013, Pages 1429–1435.

[4] Walser T, Cui X, Yanagawa J, Lee JM, Heinrich E, Lee G, Sharma S, Dubinett SM. Smoking and lung cancer: the role of inflammation. Proc Am Thorac Soc. 2008 Dec 1;5(8):811-5. doi: 10.1513/pats.200809-100TH. PMID: 19017734; PMCID: PMC4080902.

Appendix A: Table and Figure

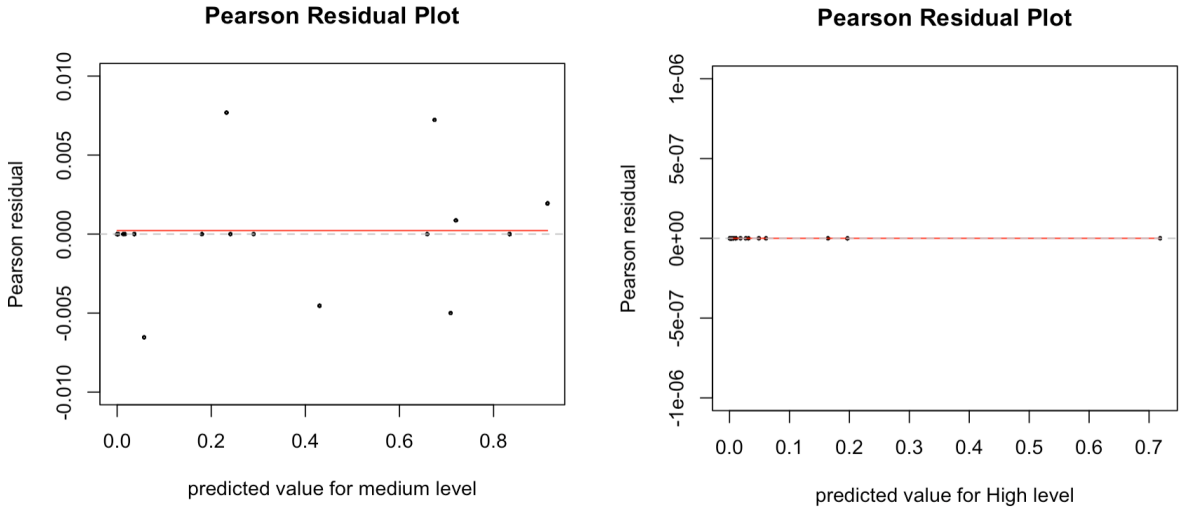

Figure 1: Pearson Residual plot for Medium level (Left); for High level (Right)

Table 2: Feature mean values within each cancer level.

Covariates	Low (Mean)	Medium (Mean)	High (Mean)
Age	35.00	38.62	37.32
Air Pollution	2.60	2.934	5.70
Alcohol use	2.23	4.20	6.83
Dust Allergy	3.11	5.44	6.62
chronic Lung Disease	3.09	3.96	5.83
Balanced Diet	3.00	3.52	6.62
Obesity	2.41	3.90	6.68
Smoking	3.02	2.46	6.07
Passive Smoker	2.63	3.01	6.53
Chest Pain	2.84	3.76	6.40
Coughing of Blood	2.86	3.85	7.43
Fatigue	2.17	3.49	5.59
Weight Loss	2.52	4.42	4.47
Shortness of Breath	2.50	4.63	5.33
Wheezing	2.57	4.76	3.89
Swallowing Difficulty	2.76	4.16	4.19
Clubbing of Finger Nails	2.47	4.94	4.21
Frequent Cold	2.37	3.68	4.38
Dry Cough	2.91	3.70	4.78
Snoring	2.14	3.31	3.23

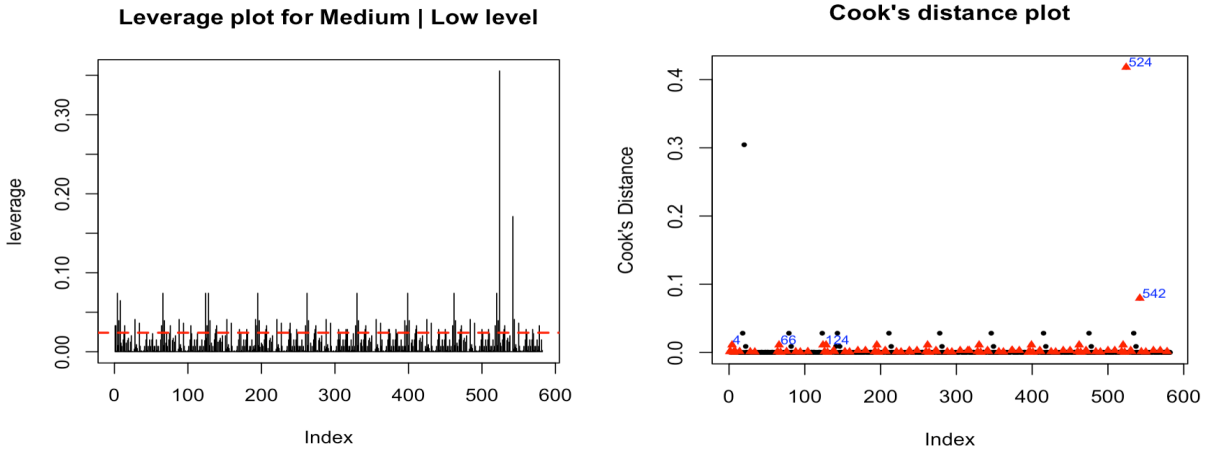

Figure 2: Leverage plot and Cook's distance plot for Medium | Low level

Appendix B: R code

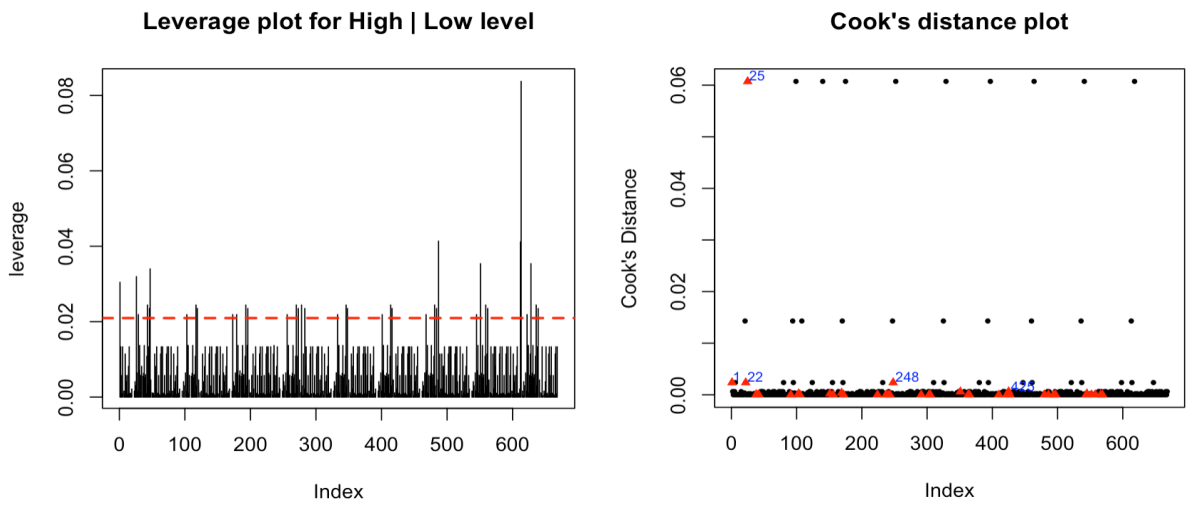

Figure 3: Leverage plot and Cook's distance plot for High | Low level

```
library("readxl")
library(MASS)
mydata <- read_excel("/Users/rduan/Desktop/cancer.xlsx")
mydata=mydata[,-1]
data1=mydata
data1$Level[data1$Level=="Low"]=0
data1$Level[data1$Level=="Medium"]=1
data1$Level[data1$Level=="High"]=2
data1[,c(2,24)] <- lapply(data1[,c(2,24)] , factor)
str(data1)
## Proportional odds model
plr<-polr(Level~`Dust Allergy`+`chronic Lung Disease`+`Obesity`,data=
   data2)
summary(plr)
## Baseline odds model
#install.packages("broom")
library("broom")
library(nnet)
library("dplyr")
library("knitr")
library("MASS")
my.model <- multinom(Level ~., data=data1)
tidy(my.model, exponentiate = FALSE)
```

```r
stepAIC(my.model)
# Model after selection
data2=data1[,c(24,4,5,6,8,10,11,19,23)]
#m2 <- multinom(Level ~`Alcohol use`+`Dust Allergy`+`chronic Lung
   Disease`+`Obesity`+`Smoking`+`Swallowing Difficulty`, data=data2)
m2<-multinom(Level~`Alcohol use`+`Air Pollution`+`Snoring`+`Smoking
   `+`chronic Lung Disease`+`Swallowing Difficulty`+`Alcohol use`:`
   Swallowing Difficulty`,data=data1)

# Tidy the result
tidy(m2, exponentiate = FALSE, conf.int = TRUE) %>%
  kable(digits = 3, format = "markdown")

# Predicted value

pred.probs <- fitted(m2,data2)
pred_probs <- as_tibble(fitted(m2)) %>% mutate(obs_num = 1:n())

# Merge with residuals and fitted
mydata=data2 %>% mutate(obs_num = 1:n())
res_p <- inner_join(mydata, pred_probs)

housing.bo <-  multinom(Level ~., data=data2)

# z values
zval.bo <- coef(housing.bo) / summary(housing.bo)$standard.errors
# two-sided p-values
pval.bo <- 2 * pnorm(abs(zval.bo), lower.tail=FALSE)
# Predicted value
prd_prob_bo = fitted(housing.bo)

#------- Pearson residuals ----------------
obslabel <- t(sapply(as.numeric(data2[[1]]), function(x) {
  res <- numeric(3)
  res[x] <- 1
  res
}))
```

```
# a list of (M-1) elements, each element contains the Pearson
   residuals for one submodel
resP.bo <- sapply(2:ncol(obslabel), function(m) {
  # baseline is column 1 here
  # otherwise you should replace "1" with the corresponding index and
    adjust the range of "m" accordingly
  obs_m <- obslabel[rowSums(obslabel[,c(1,m)]) > 0, m]
  fit_m <- prd_prob_bo[rowSums(obslabel[,c(1,m)]) > 0,c(1,m)]
  fit_m <- fit_m[,2] / rowSums(fit_m)
  a= 1:1000
  matrix(c((obs_m - fit_m) / sqrt(fit_m * (1 - fit_m)),a[rowSums(
   obslabel[,c(1,m)]) > 0]),ncol=2)
})

# MERGE TABLE & Plot residuals
resP2<-as_tibble(resP.bo[[1]])
names(resP2)<-c("r2","obs_num")
res2 <- inner_join(res_p,resP2)
res2<- res2[complete.cases(res2),]
plot(res2$`1`,res2$r2,type = "p",cex=0.3,xlab="predicted value for
   medium level",ylab="Pearson residual",ylim=c(-0.01,0.01), main="
   Pearson Residual Plot")
lines(smooth.spline(res2$'1', res2$r2, spar=2.4), col=2)
abline(h=0, lty=2, col="grey")
#
resP3<-as_tibble(resP.bo[[2]])
names(resP3)<-c("r3","obs_num")
res3 <- inner_join(res_p,resP3)
res3<- res3[complete.cases(res3),]
plot(res3$`2`,res3$r3,type = "p",cex=0.3,xlab="predicted value for
   High level",ylab="Pearson residual", ylim=c(-0.000001,0.000001),
   main="Pearson Residual Plot")
lines(smooth.spline(res3$'2', res3$r3,spar=1.6), col=2)
abline(h=0, lty=2, col="grey")

# Model after selection
names(data1)
```

```
# Covariate matrix
cor(model.matrix(m2)[,-1])
# Summary within each cancer level
data3=data1[data1$Level==2,-c(6,7)]
summary(data1)
# 0 & 1 leverage plots
data4=data1[data1$Level!=2,-c(6,7)]

bwtfit <- glm(formula = Level~`Alcohol use`+`Snoring`+`Smoking`+`
   chronic Lung Disease`+`Swallowing Difficulty`+`Alcohol use`:`
   Swallowing Difficulty`, family = binomial(), data = data4)
summary(bwtfit)
leverage = hatvalues(bwtfit)


W = diag(bwtfit$weights)
X = cbind(rep(1,nrow(data4)), data4[['Alcohol use']], data4[['Snoring
   ']],
          data4[['Smoking']], data4[['chronic Lung Disease']], data4
   [['Swallowing Difficulty']],data4[['Alcohol use']]*data4[['
   Swallowing Difficulty']])
Hat = sqrt(W) %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% sqrt(W)
all(abs(leverage - diag(Hat)) < 1e-15)

plot(names(leverage), leverage, xlab="Index", type="h",main='Leverage
   plot for Medium | Low level' )
#points(names(leverage), leverage, pch=16, cex=0.6)
p <- length(coef(bwtfit))
n <- nrow(data4)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)
# Cook's distance
cooks = cooks.distance(bwtfit)

plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6,main="Cook's
   distance plot")
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2)
```

```
susPts <- as.numeric(names(sort(cooks[infPts], decreasing=TRUE)[1:5])
    )
text(susPts, cooks[susPts], susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)

dispersion <- 1
# all(abs(cooks - (res.P/(1 - leverage))^2 * leverage/(dispersion * p
    ) < 1e-15))
# 0 & 1 leverage plots
data5=data1[data1$Level!=1,-c
    (1,2,3,5,6,7,9,10,12,13,14,15,16,17,18,20,21,22)]
bwtfit <- glm(formula = Level~`Alcohol use`+`Snoring`+`Smoking`+`
    chronic Lung Disease`+`Swallowing Difficulty`+`Alcohol use`:`
    Swallowing Difficulty`, family = binomial(),
                data = data5)

leverage=data_frame(leverage)
leverage = hatvalues(bwtfit)

# Outlier 25

W = diag(bwtfit$weights)
X = cbind(rep(1,nrow(data5)), data5[['Alcohol␣use']], data5[['Snoring
    ']],
          data5[['Smoking']], data5[['chronic␣Lung␣Disease']], data5
    [['Swallowing␣Difficulty']],data5[['Alcohol␣use']]*data5[['
    Swallowing␣Difficulty']])
Hat = sqrt(W) %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% sqrt(W)
all(abs(leverage - diag(Hat)) < 1e-15)

plot(names(leverage), leverage, xlab="Index", type="h",main='Leverage
    ␣plot␣for␣High␣|␣Low␣level')
#points(names(leverage), leverage, pch=16, cex=0.6)
p <- length(coef(bwtfit))
n <- 668
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)
# Cook's distance
cooks = cooks.distance(bwtfit)
```

```
plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6,main="Cook's
   distance plot")
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2)
susPts <- as.numeric(names(sort(cooks[infPts], decreasing=TRUE)[1:5])
   )
text(susPts, cooks[susPts], susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)
```