# PSTAT127 Homework 5

Billy Dang

2024-11-12

## Problem 1

The dataset **wbca** (in the R library *faraway*) comes from a study of breast cancer in Wisconsin. The data represent 681 cases of potentially cancerous tumors of which 238 are actually malignant. At the time of this study, assessment of whether a tumor is malignant usually involved an invasive surgical procedure. The purpose of this study was to determine whether a new procedure (i.e., new at the time of this study) called fine needle aspiration (which draws only a small sample of tissue) could be effective in determining tumor status.

### 1.a

Fit a logistic regression with **CLASS** as the response and the other nine variables as predictors using glm in R. Report the residual deviance and associated degrees of freedom.

```
library(faraway)
data(wbca)

fit <- glm(Class ~., data = wbca, family = binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = wbca)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.16678    1.41491   7.892 2.97e-15 ***
## Adhes       -0.39681    0.13384  -2.965  0.00303 **
## BNucl       -0.41478    0.10230  -4.055 5.02e-05 ***
## Chrom       -0.56456    0.18728  -3.014  0.00257 **
## Epith       -0.06440    0.16595  -0.388  0.69795
## Mitos       -0.65713    0.36764  -1.787  0.07387 .
## NNucl       -0.28659    0.12620  -2.271  0.02315 *
## Thick       -0.62675    0.15890  -3.944 8.01e-05 ***
## UShap       -0.28011    0.25235  -1.110  0.26699
## USize        0.05718    0.23271   0.246  0.80589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.464  on 671  degrees of freedom
## AIC: 109.46
##
## Number of Fisher Scoring iterations: 8
```

– The residual deviance is 89.464 with 680 degrees of freedom.

### 1.b

Use AIC as the criterion to determine the best subset of variables if you only consider models obtained by dropping one explanatory variable out at a time. (Read the help file for the **step** function.)

```
## Start:  AIC=109.46
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap + USize
##
##         Df Deviance    AIC
## - USize  1   89.523 107.52
## - Epith  1   89.613 107.61
## - UShap  1   90.627 108.63
## <none>       89.464 109.46
## - Mitos  1   93.551 111.55
## - NNucl  1   95.204 113.20
## - Adhes  1   98.844 116.84
## - Chrom  1   99.841 117.84
## - BNucl  1  109.000 127.00
## - Thick  1  110.239 128.24
##
## Step:  AIC=107.52
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap
##
##         Df Deviance    AIC
## - Epith  1   89.662 105.66
## - UShap  1   91.355 107.36
## <none>       89.523 107.52
## - Mitos  1   93.552 109.55
## - NNucl  1   95.231 111.23
## - Adhes  1   99.042 115.04
## - Chrom  1  100.153 116.15
## - BNucl  1  109.064 125.06
## - Thick  1  110.465 126.47
##
## Step:  AIC=105.66
## Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap
##
##         Df Deviance    AIC
## <none>       89.662 105.66
## - UShap  1   91.884 105.88
## - Mitos  1   93.714 107.71
## - NNucl  1   95.853 109.85
## - Adhes  1  100.126 114.13
## - Chrom  1  100.844 114.84
## - BNucl  1  109.762 123.76
## - Thick  1  110.632 124.63
```

```
##
## Call:
## glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +
##     Thick + UShap, family = binomial, data = wbca)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.0333     1.3632   8.094 5.79e-16 ***
## Adhes        -0.3984     0.1294  -3.080  0.00207 **
## BNucl        -0.4192     0.1020  -4.111 3.93e-05 ***
## Chrom        -0.5679     0.1840  -3.085  0.00203 **
## Mitos        -0.6456     0.3634  -1.777  0.07561 .
## NNucl        -0.2915     0.1236  -2.358  0.01837 *
## Thick        -0.6216     0.1579  -3.937 8.27e-05 ***
## UShap        -0.2541     0.1785  -1.423  0.15461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.662  on 673  degrees of freedom
## AIC: 105.66
##
## Number of Fisher Scoring iterations: 8
```

```
## AIC of the best model using backwards elimination based on AIC criterion: 105.6618
```

### 1.c

Use the reduced model in step 1b to estimate the probability of malignancy for a new patient with predictor variables [1, 1, 3, 2, 1, 1, 4, 1, 1]. Hint: use function "predict.glm" with the appropriate choice of "type".

```r
new_patient <- data.frame(
  Adhes = 1,
  BNucl = 1,
  Chrom = 3,
  Epith = 2,
  Mitos = 1,
  NNucl = 1,
  Thick = 4,
  UShap = 1,
  USize = 1
)

predicted_probability <- predict.glm(mlr_backwards_elim, newdata = new_patient, type = "response")

cat("Estimated probability of malignancy for the new patient:", predicted_probability, "\n")
```

```
## Estimated probability of malignancy for the new patient: 0.9921115
```

### 1.d

Suppose that a cancer is classified as benign if p > .5 and malignant if p < .5. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

```
predicted_probabilities <- predict.glm(mlr_backwards_elim, newdata = wbca, type = "response")

predicted_classes <- ifelse(predicted_probabilities > 0.5, "benign", "malignant")

actual_classes <- ifelse(wbca$Class == 1, "malignant", "benign")

false_positives <- sum(predicted_classes == "malignant" & actual_classes == "benign")
false_negatives <- sum(predicted_classes == "benign" & actual_classes == "malignant")

cat("False Positives (Type I Errors):", false_positives, "\n")
```

```
## False Positives (Type I Errors): 227
```

```
cat("False Negatives (Type II Errors):", false_negatives, "\n")
```

```
## False Negatives (Type II Errors): 434
```

## Problem 2

A study was run to investigate whether a statistical model could be used to estimate the probability of a household purchasing a new car within a 12-month period, based both on the income of the household and on the age of the oldest car belonging to the household at the start of that 12 month period.

Data was collected from a random sample of $n$ households. Each household was asked the age of their oldest automobile (variable labelled "age" measured in years), and their income (variable labelled "income"). One year later, a follow-up visit asked if the household had brought a new car in that 12 month period (variable "purchase" - coded as "1" if they had brought a new car, and "0" otherwise).

Two models were fitted using R.

```
car.dat <- read.table("data/car.txt", header=T)
attach(car.dat)

fit1 <- glm(purchase ~ income + age, family=binomial(link = logit))
summary(fit1)
```

```
##
## Call:
## glm(formula = purchase ~ income + age, family = binomial(link = logit))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.73931    2.10195  -2.255   0.0242 *
## income       0.06773    0.02806   2.414   0.0158 *
## age          0.59863    0.39007   1.535   0.1249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 44.987  on 32  degrees of freedom
## Residual deviance: 36.690  on 30  degrees of freedom
## AIC: 42.69
##
## Number of Fisher Scoring iterations: 4
```

### 2.a

What is the value of n (i.e., how many households are in our sample)?

– There are 33 households in the sample (null deviance df = n - 1).

### 2.b

Write down model "fit1" and the assumptions of this model. Define all notation that you use.

– The logistic regression model "fit1" can be written as:

$$logit(P(purchase_i = 1)) = \beta_0 + \beta_1 * income_i + \beta_2 * age_i$$

where:

- $P(purhcase_i = 1)$ is the probability that household i purchased a new car within the 12 month period.
- $\beta_0$ is the intercept term
- $\beta_1$ is the coefficient for household income
- $\beta_2$ is the coefficient for the age of the oldest car in the household.
- $income_i$ represent the income of house hold i (where i is some value n = 1, 2, . . . ,33)
- $age_i$ represents the age of the oldest car for household i

And assumptions include linearity in log-odds, independence of observations, no perfect multicollinearity.

### 2.c

```r
fit2 <- update(fit1, .~.-age)
summary(fit2)
```

```
##
## Call:
## glm(formula = purchase ~ income, family = binomial(link = logit))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.98079    0.85720  -2.311   0.0208 *
## income       0.04342    0.02011   2.159   0.0308 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 44.987  on 32  degrees of freedom
## Residual deviance: 39.305  on 31  degrees of freedom
## AIC: 43.305
##
## Number of Fisher Scoring iterations: 4
```

```
round( vcov(fit2), digits = 5)
```

```
##             (Intercept)  income
## (Intercept)     0.73478 -0.0154
## income         -0.01540  0.0004
```

Load R library "MASS" and read the help file for "confint.glm". The last example in this help file has a glm for the bud worm data that you can run to practice to learn the R syntax.

```
library(MASS)
help("confint.glm")
```

```
## starting httpd help server ... done
```

Now use the "confint.glm" command to obtain a 99% confidence interval estimate for parameter $\beta_{income}$, assuming model "fit2" holds. Include your code and confidence interval results within your answer file.

```
confint(fit2, parm = "income", level = 0.99)
```

```
## Waiting for profiling to be done...

##        0.5 %       99.5 %
## -0.003272497  0.104611224
```

### 2.d(i)

Explain why I receive the warning message below when I run the following anova command in R.

```
anova(fit2, fit1, test = "F")
```

```
## Warning: using F test with a 'binomial' family is inappropriate

## Analysis of Deviance Table
##
## Model 1: purchase ~ income
## Model 2: purchase ~ income + age
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1        31     39.305
## 2        30     36.690  1   2.6149 2.6149 0.1059
```

– The following error message is a result of the logistic regression model, where the response variable is binary being combined with an F-test that is generally used in the context of linear models (i.e. OLS regression) where the residuals are assumed to be normally distributed.

### 2.d(ii)

How should I modify this anova command in order to obtain the p-value for the nested model hypothesis test that we studied in class?

– To test the hypothesis $H_0$ :model “fit2” (purchase $\sim$ income) versus $H_A$ : model “fit1” (purchase $\sim$ income + age), we need to perform a likelihood ratio test for the nested models. In R, we can use the anova() function with test = “Chisq” to obtain the Chi-squared test for nested model comparison. This is the appropriate test for comparing models in logistic regression (with a binomial family).

```r
# Perform the likelihood ratio test using anova with test = "Chisq"
anova_result <- anova(fit2, fit1, test = "Chisq")

# Display the results
print(anova_result)
```

```
## Analysis of Deviance Table
##
## Model 1: purchase ~ income
## Model 2: purchase ~ income + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        31     39.305
## 2        30     36.690  1   2.6149   0.1059
```

– Since the p-value is greater than $\alpha = 0.01$ with a value of .1059, we do not reject $H_0$ which suggests that adding age does not significantly improve the model fit at the 1% significance level.