MATH3424 HW5

Chow, Hau Cheung Jasper (hcjchow / 20589533)

November 29, 2021

Q1a

```
setwd("/Users/jchow/Downloads/MATH3424 R")
cancer_data <- read.csv(file="BreastCancer.txt",header=TRUE)
#as.numeric(cancer_data$Class) # 1 = benign, 2 = malignant
cancer_data$Class <- relevel(cancer_data$Class, ref = "benign")
cancer_model_1 <- glm(Class ~ ., family="binomial", data=cancer_data)
summary(cancer_model_1)
```

```
##
## Call:
## glm(formula = Class ~ ., family = "binomial", data = cancer_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4841  -0.1153  -0.0619   0.0222   2.4698
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -10.10394    1.17488  -8.600  < 2e-16 ***
## Cl.thickness      0.53501    0.14202   3.767 0.000165 ***
## Cell.size        -0.00628    0.20908  -0.030 0.976039
## Cell.shape        0.32271    0.23060   1.399 0.161688
## Marg.adhesion     0.33064    0.12345   2.678 0.007400 **
## Epith.c.size      0.09663    0.15659   0.617 0.537159
## Bare.nuclei       0.38303    0.09384   4.082 4.47e-05 ***
## Bl.cromatin       0.44719    0.17138   2.609 0.009073 **
## Normal.nucleoli   0.21303    0.11287   1.887 0.059115 .
## Mitoses           0.53484    0.32877   1.627 0.103788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 102.89  on 673  degrees of freedom
## AIC: 122.89
##
## Number of Fisher Scoring iterations: 8
```

We notice that not all of the regression coefficients are significant at the 0.05 level; only intercept, Cl.thickness, Marg.adhesion, Bare.nuclei and Bl.cromatin are.

Instead of the F-statistic, we use the G-statistic $G = D_0 - D$ where if the null model is correct, $G \sim \chi^2(9)$ since the sample size $n = 683$ is large.

Since $G = D_0 - D = 884.35 - 102.89 = 781.46$ and as we can see below, the p-value is very close to 0, we reject the null hypothesis and the model is indeed significant.

```
pchisq(781.46, df=9, lower.tail=FALSE)
```

```
## [1] 2.079405e-162
```

```
rsq_1 <- 1-(102.89/884.35)
rsq_1
```

```
## [1] 0.8836547
```

Q1b

$\hat{\beta}_1$ is the estimated coefficient for CL.thickness. So the 95% CI is $[0.25666, 0.8133682]$.

```
beta_1_hat <- cancer_model_1$coefficients["Cl.thickness"]
se_beta_1_hat <- 0.14202
alpha <- 0.05
z_alpha_2 <- qnorm(alpha/2, mean=0, sd=1, lower.tail=FALSE)

beta_1_hat - z_alpha_2*se_beta_1_hat
```

```
## Cl.thickness
##      0.25666
```

```
beta_1_hat + z_alpha_2*se_beta_1_hat
```

```
## Cl.thickness
##    0.8133682
```

```
# https://stats.idre.ucla.edu/r/dae/logit-regression/
# this gets the CIs based on profiled log-likelihood function
#confint(cancer_model_1, level=0.95, parm="Cl.thickness")

# this gets them with the standard errors (which we want)
confint.default(cancer_model_1, level=0.95, parm="Cl.thickness")
```

```
##                  2.5 %    97.5 %
## Cl.thickness 0.256665 0.8133631
```

To test $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$, from the regression output in 1a, we notice that the p-value is 0.161688 (the z-statistic is 1.399) which is larger than our chosen significance of $\alpha = 0.1$. Therefore we fail to reject the null hypothesis that $H_0 : \beta_3 = 0$.

Q1c

```
cancer_model_2 <- glm(Class ~ Cl.thickness + Cell.shape + Marg.adhesion + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
summary(cancer_model_2)
```

```
##
## Call:
## glm(formula = Class ~ Cl.thickness + Cell.shape + Marg.adhesion +
##     Bare.nuclei + Bl.cromatin, family = "binomial", data = cancer_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2982  -0.1242  -0.0624   0.0234   2.3713
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.74114    1.04989  -9.278  < 2e-16 ***
## Cl.thickness   0.62576    0.13373   4.679 2.88e-06 ***
## Cell.shape     0.48994    0.15379   3.186 0.001444 **
## Marg.adhesion  0.33918    0.11221   3.023 0.002505 **
## Bare.nuclei    0.37330    0.09381   3.979 6.91e-05 ***
## Bl.cromatin    0.55731    0.16341   3.411 0.000648 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 112.57  on 677  degrees of freedom
## AIC: 124.57
##
## Number of Fisher Scoring iterations: 8
```

```
# p-value for G-test with new model
pchisq(771.78, df=5, lower.tail=FALSE)
```

```
## [1] 1.471799e-164
```

```
# R^2 for new model
rsq_2 <- 1-(112.57/884.35)
rsq_2
```

```
## [1] 0.8727088
```

This time we notice all regression coefficients are significant at the 0.05 level and the AIC of the model has increased slightly from 122.89 to 124.57, but not by a significant amount, so the reduced and full model are equally effective for regression. We also observe that $G = 884.35 - 112.57 = 771.78$ is slightly lower but the p-value is still very low and close to 0, so the model is also significant. The R^2 has decreased slightly from 0.88 to 0.87 but is still very high.

```
#2*(logLik(cancer_model_2)-logLik(cancer_model_1))
residual_dev_diff <- 112.57-102.89
residual_dev_diff
```

```
## [1] 9.68
```

```
qchisq(0.99, df=4, lower.tail=TRUE)
```

```
## [1] 13.2767
```

If we use the hypothesis that H_0 : $reducedsuitable$, H_1 : $fullsuitable$, then by examining the statistic $2[L(p+q) - L(p)] \sim \chi^2(q)$ we observe with $p = 5, p+q = 9$ that the residual deviance of reduced minus residual deviance of full is 9.68. We notice that at significance level $\alpha = 0.01$, the test statistic of 9.68 is smaller than the critical value of 13.2767, so we fail to reject H_0.

Q1d

```
new <- data.frame(Cl.thickness=6, Cell.shape=3, Marg.adhesion=8, Bare.nuclei=2, Bl.cromatin=5)
# can show log-odds = beta_0_hat + innerprod(new, beta_i_hat) = 1.72983
# where: log-odds = ln(pi/(1-pi)) where pi=probability tumor is malignant
# then to get pi need to do: pi = exp(log-odds)/(1+exp(log-odds))
pi <- predict(cancer_model_2, new, type="response")
1-pi
```

```
##         1
## 0.1506149
```

So the probability of this patient's tumour being benign is $1 - \pi = 0.1506149$.

Q1e

4 variables

```
cancer_model_3 <- glm(Class ~ Cell.shape + Marg.adhesion + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 151.3174
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Marg.adhesion + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 135.7746
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 132.7431
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape + Marg.adhesion + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 141.6494
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape + Marg.adhesion + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 135.7378
```

3 variables

```
cancer_model_3 <- glm(Class ~ Marg.adhesion + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 205.0236
```

```
cancer_model_3 <- glm(Class ~  Cell.shape + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 156.4433
```

```
cancer_model_3 <- glm(Class ~ Cell.shape + Marg.adhesion + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 182.4583
```

```
cancer_model_3 <- glm(Class ~ Cell.shape + Marg.adhesion + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 167.7037
```

```
# ========
cancer_model_3 <- glm(Class ~ Cl.thickness + Bare.nuclei + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 150.9232
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Marg.adhesion + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 174.8883
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Marg.adhesion + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 163.4411
```

```
# ========
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 160.0564
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 148.2462
```

```
# ========
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape + Marg.adhesion,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 168.2263
```

2 variables

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Cell.shape,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 206.3944
```

```
cancer_model_3 <- glm(Class ~  Cl.thickness + Marg.adhesion,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 260.1006
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 204.2192
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 231.4534
```

```
# ========
cancer_model_3 <- glm(Class ~ Cell.shape + Marg.adhesion,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 222.4743
```

```
cancer_model_3 <- glm(Class ~ Cell.shape + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 177.3853
```

```
cancer_model_3 <- glm(Class ~ Cell.shape + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 199.0765
```

```
# ========
cancer_model_3 <- glm(Class ~ Marg.adhesion + Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 257.0478
```

```
cancer_model_3 <- glm(Class ~ Marg.adhesion + Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 306.6519
```

```
# ========
cancer_model_3 <- glm(Class ~ Bare.nuclei + Bl.cromatin,
```

```
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 229.2795
```

1 or 0 variables

```
cancer_model_3 <- glm(Class ~ Cell.shape,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 271.5863
```

```
cancer_model_3 <- glm(Class ~ Marg.adhesion,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 467.3434
```

```
cancer_model_3 <- glm(Class ~ Bare.nuclei,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 344.6277
```

```
cancer_model_3 <- glm(Class ~ Bl.cromatin,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 392.2168
```

```
cancer_model_3 <- glm(Class ~ Cl.thickness,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 462.483
```

```
# ========
cancer_model_3 <- glm(Class ~ 1,
                      family="binomial", data=cancer_data)
AIC(cancer_model_3)
```

```
## [1] 886.3502
```

As we can see, the 4-variable model with the lowest AIC (132.7431) was with variables "Cl.thickness + Cell.shape + Bare.nuclei + Bl.cromatin." The 3-variable model with lowest AIC (148.2462) was with variables "Cl.thickness + Cell.shape + Bare.nuclei." The 2-variable model with lowest AIC (177.3853) was with variables "Cell.shape + Bare.nuclei." The 1-variable model with lowest AIC (271.5863) was with variable "Cell.shape". As we can see, the more variables we drop, the higher the AIC. In fact, out of all subset models of the variables "Cl.thickness + Cell.shape + Marg.adhesion + Bare.nuclei + Bl.cromatin," the one with the lowest AIC (124.57) is the model with all 5 of those variables, i.e. the model in part (c).