DATA-581 Lab 2

Dhun Sheth

2023-11-29

Question 3 (Exercise 5 from Ch. 4)

Part A

```
y <- c(1,3,5,6,5,7)
x <- c(0,1,2,3,4,5)
X <- matrix(c(1,1,1,1,1,1,0,1,2,3,4,5), nrow=6)
print(X)
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    1    1
## [3,]    1    2
## [4,]    1    3
## [5,]    1    4
## [6,]    1    5
```

Part B

```
X.QR <- qr(X)
R <- qr.R(X.QR, complete=TRUE)
Q <- qr.Q(X.QR, complete=TRUE)
print(Q)
```

```
##            [,1]       [,2]        [,3]       [,4]       [,5]       [,6]
## [1,] -0.4082483 -0.5976143 -0.39966534 -0.3614756 -0.3232858 -0.2850961
## [2,] -0.4082483 -0.3585686 -0.06213786  0.1879790  0.4380958  0.6882127
## [3,] -0.4082483 -0.1195229  0.88170280 -0.1180607 -0.1178243 -0.1175878
## [4,] -0.4082483  0.1195229 -0.12913187  0.8228959 -0.2250764 -0.2730487
## [5,] -0.4082483  0.3585686 -0.13996653 -0.2361476  0.6676714 -0.4285096
## [6,] -0.4082483  0.5976143 -0.15080120 -0.2951910 -0.4395807  0.4160295
```

```
print(R)
```

```
##          [,1]      [,2]
## [1,] -2.44949 -6.123724
## [2,]  0.00000  4.183300
```

```
## [3,]  0.00000  0.000000
## [4,]  0.00000  0.000000
## [5,]  0.00000  0.000000
## [6,]  0.00000  0.000000
```

```r
print(Q%*%R)
```

```
##      [,1]          [,2]
## [1,]    1 -8.881784e-16
## [2,]    1  1.000000e+00
## [3,]    1  2.000000e+00
## [4,]    1  3.000000e+00
## [5,]    1  4.000000e+00
## [6,]    1  5.000000e+00
```

Part C

```r
U <- qr.R(X.QR)
U_inv <- solve(U)
print(U_inv)
```

```
##            [,1]       [,2]
## [1,] -0.4082483 -0.5976143
## [2,]  0.0000000  0.2390457
```

Part D

```r
p <- ncol(R)
n <- nrow(X)
Q1 <- Q[,1:p]

beta <- U_inv %*% t(Q1) %*% y
print(beta)
```

```
##          [,1]
## [1,] 1.857143
## [2,] 1.057143
```

Part E

```r
y_pred <- beta[1] + beta[2]*x
QR_RSS <- sum((y_pred-y)^2)
print(QR_RSS)
```

```
## [1] 3.942857
```

Part F

```
error_var <- QR_RSS/(n-p)
print(error_var)
```

```
## [1] 0.9857143
```

Part G

```
SE_beta <- sqrt(error_var)*sqrt(diag(solve(t(U)%*%U)))
t_statistic <- beta[2]/SE_beta[2]
p_value <- 2*(1-pt(t_statistic,n-p))
print(p_value)
```

```
## [1] 0.01120966
```

Part H

The test statistic follows a t-distribution on n-p degrees of freedom, 6-2 = 4 D.o.F. The p-value of the t-test is 0.0112097 and because it is below the 0.05 significant level, there is enough evidence to reject the null hypothesis.

Question 5 (Exercise 2 from Ch. 8)

Part A

```
glm_reg <- glm(y~x, data=p13.2, family=binomial())
summary(glm_reg)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(), data = p13.2)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.7395139  4.4394326  -1.969   0.0490 *
## x            0.0002009  0.0001006   1.998   0.0458 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.526  on 19  degrees of freedom
## Residual deviance: 22.435  on 18  degrees of freedom
## AIC: 26.435
##
## Number of Fisher Scoring iterations: 4
```

$log(\frac{p}{1-p}) = -8.7395139 + 2.0090564 \times 10^{-4} * x$

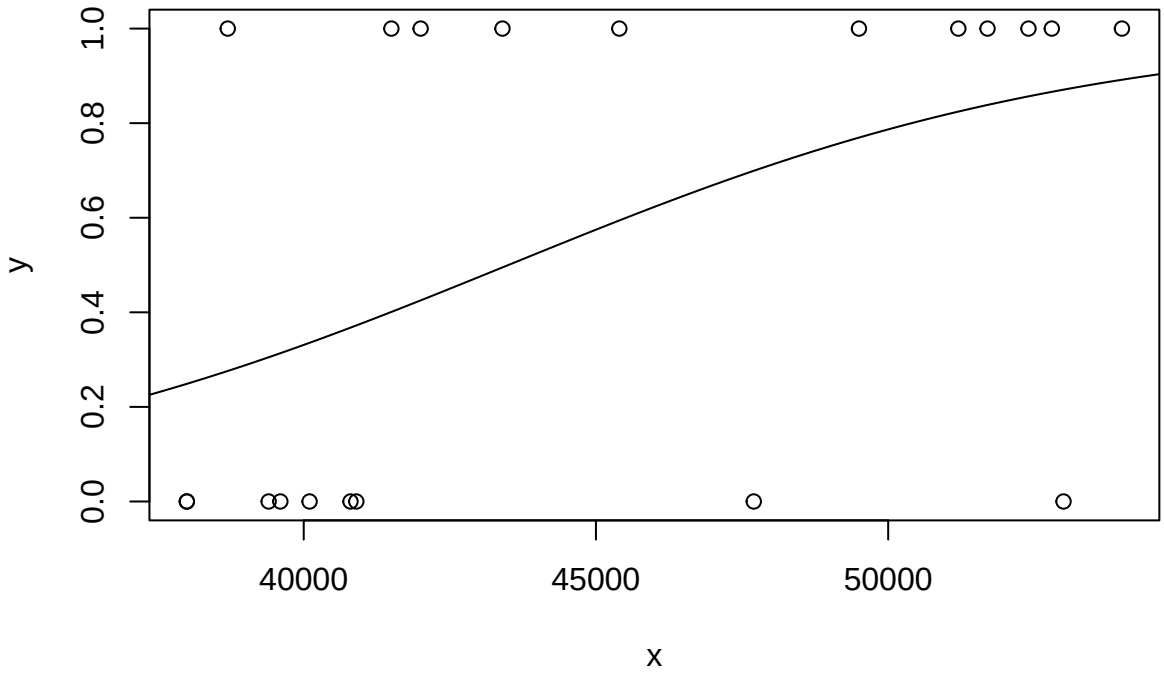
$p = \frac{e^{-8.7395139+2.0090564\times10^{-4}*x}}{1+e^{-8.7395139+2.0090564\times10^{-4}*x}}$

Part B

```
y.logit2 <- glm(y~x, data = p13.2, family = binomial(link = "logit"))
anova(y.logit2, glm_reg)
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ x
## Model 2: y ~ x
##   Resid. Df Resid. Dev Df Deviance
## 1        18     22.435
## 2        18     22.435  0        0
```

```
plot(y~x, data=p13.2)
x_value <- 35000:75000
lines(x_value, predict(glm_reg,
newdata=data.frame(x = x_value), type = "response"))
```



There is only 1 predictor variable and the residual deviance is not much larger than the degree's of freedom, hence, the model is acceptable.

In addition, looking at the graph and estimated curve, the model seems acceptable.

Part C

$p = 0.3310835 =>$ hence unlikely they own their home.

Question 6

Part A, B

```
epil_reg <- glm(y~age+trt, data=epil, family=poisson)
print(summary(epil_reg))
```

```
##
## Call:
## glm(formula = y ~ age + trt, family = poisson, data = epil)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.533031   0.110638  22.895  < 2e-16 ***
## age          -0.013331   0.003708  -3.596 0.000324 ***
## trtprogabide -0.092514   0.045596  -2.029 0.042460 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2517.8  on 235  degrees of freedom
## Residual deviance: 2502.0  on 233  degrees of freedom
## AIC: 3273.9
##
## Number of Fisher Scoring iterations: 6
```

Based on the p-values being below the 0.05 significance level, all coefficients (intercept, age, trt) are significant.

Part C

```
print(confint(epil_reg))
```

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

```
##                    2.5 %       97.5 %
## (Intercept)   2.31587948  2.749603939
## age          -0.02062141 -0.006086226
## trtprogabide -0.18188033 -0.003117740
```

Part D

The beta for treatment is -0.0925138. Because it is negative, this shows a negative relationship indicating a reduction in seizures with treatment. In addition, the treatment coefficient is significant.

Part E

```
new_data <- data.frame(age = 20, trt = "progabide")
pred_y <- predict(epil_reg, newdata = new_data, type = "response")
print(pred_y)
```

```
##        1
## 8.792404
```

Predicted number of seizures for a 20-year-old with treatment: ” 9