Lab 23 R Script

Alexander Hernandez

11/17/2022

1) Clinical Trial Drop Outs

a)Import the data and determine its dimension

```
trials = read.table("https://media.pearsoncmg.com/aw/aw_sharpe_business_3/datasets/txt/Clinical%20Trials
                    sep="\t", header=TRUE)
dim(trials)
```

```
## [1] 428   3
```

b) The Missing values are left blank. Clean the data by removing them

```
trials_new = na.omit(trials)
attach(trials_new)
```

c) Fit a multiple logistic regression model using Age and HDRS as predictor variables

```
model1 = glm(DRP ~ AGE + HD2114,
            family = binomial(logit))
summary(model1)
```

```
##
## Call:
## glm(formula = DRP ~ AGE + HD2114, family = binomial(logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2995  -0.8156  -0.6617   1.2711   2.1195
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.44197    0.48827  -0.905 0.365370
## AGE         -0.03790    0.01151  -3.293 0.000992 ***
## HD2114       0.04682    0.01590   2.944 0.003241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 464.61  on 399  degrees of freedom
```

```
## Residual deviance: 445.80  on 397  degrees of freedom
## AIC: 451.8
##
## Number of Fisher Scoring iterations: 4
```

```
# DRP = [1 + exp( 0.44197 + 0.03790(AGE) -0.04682(HD2114) )]^-1
```

d) What is the predicted dropout probability of a 30 year old patient with HDRS score of 30?

```
predict(model1, data.frame(AGE=30, HD2114=30), type="resp")
```

```
##         1
## 0.4564631
```

```
# Dropout Probability of 30 year-old with HDRS score of 30: 0.4564631
```

2) Respiratory Fucnction and Smoking

a) Import the data and identify its dimension

```
resp = read.table("http://jse.amstat.org/datasets/fev.dat.txt")
colnames(resp) = c("age", "fev", "height", "sex", "smoke")
attach(resp)

dim(resp)
```

```
## [1] 654   5
```

b) Test whether smoking status differ by gender (2-sample t-test)

```
# H0: There is no significant difference of smoking status between sexes
# Ha: There is a significant different of smoking status between sexes
table(sex)
```

```
## sex
##   0   1
## 318 336
```

```
xtabs(~sex + smoke)
```

```
##    smoke
## sex   0   1
##   0 279  39
##   1 310  26
```

```
prop.test(c(39, 26), n=c(318, 336), correct=F)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(39, 26) out of c(318, 336)
## X-squared = 3.739, df = 1, p-value = 0.05316
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0007400171  0.0912611312
## sample estimates:
##     prop 1     prop 2
## 0.12264151 0.07738095
```

```
# With a p-value of 0.05316 and a standard confidence of 95%,
# there is not enough evidence to reject the null
# and claim there is no difference in smoking status between sexes.
```

c) Fit a multiple linear regression model to study fev, using age, height, sex, and smoking status as predictor variables

```
model2 = lm(fev ~ age + height + sex + smoke)
summary(model2)
```

```
##
## Call:
## lm(formula = fev ~ age + height + sex + smoke)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37656 -0.25033  0.00894  0.25588  1.92047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.456974   0.222839 -20.001  < 2e-16 ***
## age          0.065509   0.009489   6.904 1.21e-11 ***
## height       0.104199   0.004758  21.901  < 2e-16 ***
## sex          0.157103   0.033207   4.731 2.74e-06 ***
## smoke       -0.087246   0.059254  -1.472    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4122 on 649 degrees of freedom
## Multiple R-squared:  0.7754, Adjusted R-squared:  0.774
## F-statistic:   560 on 4 and 649 DF,  p-value: < 2.2e-16
```

```
# fev = -4.457 + 0.066(age) + 0.104(height) + 0.157(sex) - (0.087(smoke))
```

d) Use model to predict the FEV of a 50 inches tall, 12 year-old girl who is not a smoker. Construct a 95% confidence interval

```
predict(model2, data.frame(age=12, height=50, sex=0, smoke=0), interval="conf", level=0.95)
```

```
##        fit      lwr      upr
## 1 1.539109 1.399775 1.678443
```

```
# Predicted FEV is: 1.539109
# Confidence Interval: (1.399775, 1.678443)
```

3) Real Estate on 1115 Houses

```r
homes = read.csv("C:\\repos\\STAT 50001\\Lab 23\\home.csv")
attach(homes)

model3 = glm(Sold ~ .,
             family = binomial(logit),
             data = homes)
summary(model3)
```

```
##
## Call:
## glm(formula = Sold ~ ., family = binomial(logit), data = homes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9173  -0.7681  -0.5527   0.9337   2.3872
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.222e+00  3.826e-01  -8.422  < 2e-16 ***
## Living.Area -1.444e-03  2.518e-04  -5.734  9.8e-09 ***
## Age          4.900e-03  2.823e-03   1.736 0.082609 .
## Price        1.693e-05  1.444e-06  11.719  < 2e-16 ***
## Bedrooms     4.805e-01  1.366e-01   3.517 0.000436 ***
## Bathrooms   -1.813e-01  1.829e-01  -0.991 0.321493
## Fireplaces  -1.253e-01  1.633e-01  -0.767 0.442885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1401.2  on 1114  degrees of freedom
## Residual deviance: 1159.9  on 1108  degrees of freedom
## AIC: 1173.9
##
## Number of Fisher Scoring iterations: 4
```

```r
# Sold = [1 + exp(  3.222                 +
#                   0.001444(Living.Area)  -
#                   0.0049(Age)            -
#                   0.00001693(Price)      -
#                   0.4805(Bedrooms)       +
#                   0.1813(Bathrooms)      +
#                   0.1253(Fireplaces)      )]^-1
```

4) Health Clinic for Flu Shots

a) Fit a multiple logistic regression model and check for significance of each variable (X1, X2, and X3)

```
flu = read.table("C:\\repos\\STAT 50001\\Lab 23\\flu.txt", header=TRUE)
model4 = glm(y ~ .,
             family = binomial(logit),
             data = flu)
summary(model4)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(logit), data = flu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4037  -0.5637  -0.3352  -0.1542   2.9394
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## x1           0.07279    0.03038   2.396  0.01658 *
## x2          -0.09899    0.03348  -2.957  0.00311 **
## x3           0.43397    0.52179   0.832  0.40558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

```
# y = [1 + exp( 1.17716     -
#               0.07279(x1) +
#               0.09899(x2) -
#               0.43397(x3)   ]^-1
```

b) What is the estimate probability that a male client aged 55 with a health awareness index 60 will receive a flu shot?

```
predict(model4, data.frame(x1=55, x2=60, x3=1), interval="conf", level=0.95, type="resp")
```

```
##          1
## 0.06422197
```

```
# The chance that the male client will get the flu shot:
# 0.0642
```