ADVANCED DATA ANALYSIS

HW5

Fan Yang
UNI: fy2232
02/28/2018

Problem 1

(10pt) For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the data Shuttle.csv shows the temperature in fahrenheit (^{o}F) at the time of the flight and whether at least one primary O-ring suffered thermal distress.

Load the data:

In [1]:
```
Shuttle <- read.csv("Shuttle.csv",header = T)
Shuttle <- cbind(Shuttle, TD = factor(Shuttle$ThermalDistress))
head(Shuttle,5)
```

Temperature	ThermalDistress	TD
66	0	0
70	1	1
69	0	0
68	0	0
67	0	0

(a)

(2pt) Use logistic regression to model the effect of the temperature on the probability of thermal distress. That is, fit the model

$$\text{logit}(\pi(\text{TD}|\text{Temperature})) = \beta_0 + \beta_1 \text{Temperature}$$
$$\pi(\text{TD}|\text{Temperature}) = P(\text{Termal Distress} = 1|\text{Temperature})$$

In [2]:
```
glm.1a <- glm(ThermalDistress ~ Temperature,
              family = binomial("logit"), data =Shuttle)
glm.1a
```

```
Call:  glm(formula = ThermalDistress ~ Temperature, family = binomial
("logit"),
    data = Shuttle)

Coefficients:
(Intercept)  Temperature
    15.0429      -0.2322

Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
Null Deviance:      28.27
Residual Deviance: 20.32        AIC: 24.32
```

$$\text{logit}(\pi(\text{TD}|\text{Temperature})) = 15.0429 - 0.2322 \times \text{Temperature}$$

where $\pi(\text{TD}|\text{Temperature}) = P(\text{Termal Distress} = 1|\text{Temperature})$.

(b)

(2pt) Estimate β_1, the effect of temperature on the probability of thermal distress. Interpret your result.

In [3]:
```
glm.1a$coefficients
```

(Intercept)	15.0429016476891
Temperature	-0.2321627442184

$$\hat{\beta}_1 = -0.2321627442184$$

.

(c)

(2pt) Construct a 95% confidence interval to describe the effect of the temperature on the odds of thermal distress (i.e. construct a 95% interval for e^{β_1}), Interpret your result

In [4]:
```
confint(glm.1a)
```

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

	2.5 %	97.5 %
(Intercept)	3.3305848	34.34215133
Temperature	-0.5154718	-0.06082076

A 95% interval for β_1 is

$$[-0.5154718, -0.06082076]$$

A 95% interval for e^{β_1} is

$$[e^{-0.5154718}, e^{-0.06082076}] = [0.5972188, 0.9409919]$$

(d)

(2pt) Predict the probability of thermal distress at 31^{o}F, the temperature at the time of the Challenger flight.

In [5]:
```
newdata.1d <- data.frame(Temperature = 31)
predict(glm.1a, newdata.1d,type = "response")
```

1: 0.999608782884929

$$\pi(\text{TD}|\text{Temperature}) = 0.999608782884929$$

$$\text{logit}(\pi(\text{TD}|\text{Temperature})) = 15.0429 - 0.2322 \times 31 = 7.8447$$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$
$$\pi = \frac{e^{\text{logit}}}{1+e^{\text{logit}}} = 0.999608782884929$$

Therefore, $\pi(\text{TD}|\text{Temperature}) = 0.999608782884929.$

(e)

(2pt) At what temperature does the predicted probability equal 0.5?

$$\begin{aligned} \text{logit}(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) \\ &= \log\left(\frac{0.5}{1-0.5}\right) \\ &= 0 \\ \text{logit}(\pi) &= 15.0429 - 0.2322 \times \text{Temperature} = 0 \\ \text{Therefore, Temperature} &= 64.7842377260982 \end{aligned}$$

At temperature 64.7842377260982 ^{o}F, the predicted probability equal 0.5

Problem 2

The data in the file adolescent.csv appeared in a national study of 15 and 16 year-old adolescents. The event of interest is ever having sexual intercourse. The goal is to study the effect if any of race and gender on having sexual intercourse (Yes, No). Consider the following model

$$\text{logit}(\pi(\text{Intercourse} = \text{Yes}|\text{Gender}, \text{Race})) = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Race}$$

Load the data:

In [6]:
```
adolescent <- read.table("adolescent.csv",header = T,sep=",")
adolescent
```

```
Warning message in read.table("adolescent.csv", header = T, sep = ","):
“incomplete final line found by readTableHeader on 'adolescent.csv'”
```

Race	Gender	Yes	No
White	Male	43	134
White	Female	26	149
Black	Male	29	23
Black	Female	22	36

(a)

(2pt) Estimate β_1 and β_2 and interpret your result

In [7]:
```
attach(adolescent)
glm.2a <- glm(cbind(Yes, No)~factor(Gender)+factor(Race), family=binomia
l)
glm.2a
```

```
Call:  glm(formula = cbind(Yes, No) ~ factor(Gender) + factor(Race),
    family = binomial)

Coefficients:
       (Intercept)  factor(Gender)Male   factor(Race)White
           -0.4555              0.6478             -1.3135

Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
Null Deviance:      37.52
Residual Deviance: 0.05835      AIC: 25.19
```

For model

$$\text{logit}(\pi(\text{Intercourse} = \text{Yes}|\text{Gender}, \text{Race})) = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Race}$$

we get

$$\hat{\beta}_1 = \hat{\beta}_{\text{Gender}} = 0.6478$$
$$\hat{\beta}_2 = \hat{\beta}_{\text{Race}} = -1.3135$$

Therefore

$$\text{logit}(\pi) = -0.4555 + 0.6478 \times \text{Gender} - 1.3135 \times \text{Race}$$

(b)

(2pt) Construct a 95% confidence interval to describe the effect of gender on the odds of Intercourse controlling for race (i.e. construct a 95% interval for e^{β_1}), Interpret your result

In [8]:
```
confint(glm.2a)
```

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

	2.5 %	97.5 %
(Intercept)	-0.8971266	-0.02385449
factor(Gender)Male	0.2105773	1.09436472
factor(Race)White	-1.7824267	-0.84865350

A 95% interval for β_1 is

$$[0.2105773, 1.09436472]$$

A 95% interval for e^{β_1} is

$$[e^{0.2105773}, e^{1.09436472}] = [1.2343904, 2.9872843]$$

(c)

(2pt) Construct a 95% confidence interval to describe the effect of gender on the odds of Intercourse controlling for race (i.e. con- struct a 95% interval for e^{β_2}), Interpret your result

In [9]:
```
confint(glm.2a)
```

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

	2.5 %	97.5 %
(Intercept)	-0.8971266	-0.02385449
factor(Gender)Male	0.2105773	1.09436472
factor(Race)White	-1.7824267	-0.84865350

A 95% interval for β_2 is

$$[-1.7824267, -0.84865350]$$

A 95% interval for e^{β_2} is

$$[e^{-1.7824267}, e^{-0.84865350}] = [0.1682294, 0.4279908]$$

(d)

(2pt)Test $H_0 : \beta_1 = \beta_2 = 0$ against H_a : at least one of them is not zero. Use $\alpha = 0.05$.

Using Likelihood Ratio Test:
Under general H_0

$$-2(\text{log of the likelihood ratio}) = -2[\log(L(R)) - \log(L(F))] \sim \chi_k^2$$

where k is the number of parameters set equal to zero to get the reduced model.

Reject H_0 if

$$-2\text{ (log of the likelihood ratio)} > \chi_k^2(1-\alpha)$$

In [10]:
```
summary(glm.2a)
```

```
Call:
glm(formula = cbind(Yes, No) ~ factor(Gender) + factor(Race),
    family = binomial)

Deviance Residuals:
       1         2         3         4
-0.08867   0.10840   0.14143  -0.13687

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.4555     0.2221  -2.050  0.04032 *
factor(Gender)Male   0.6478     0.2250   2.879  0.00399 **
factor(Race)White   -1.3135     0.2378  -5.524 3.32e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37.516984  on 3  degrees of freedom
Residual deviance:  0.058349  on 1  degrees of freedom
AIC: 25.186

Number of Fisher Scoring iterations: 3
```

The deviances are
Null deviance: 37.516984 on 3 degrees of freedom
Residual deviance: 0.058349 on 1 degrees of freedom

The test statistics = 37.516984 – 0.058349 = 37.458635. Since $p = 2$ we reject H_0 since $37.458635 > \chi_2^2(0.95) = 5.99$.

(e)

(2pt)Test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. Use $\alpha = 0.05$.

In [11]:
```
summary(glm.2a)
```

```
Call:
glm(formula = cbind(Yes, No) ~ factor(Gender) + factor(Race),
    family = binomial)

Deviance Residuals:
       1         2         3         4
-0.08867   0.10840   0.14143  -0.13687

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.4555     0.2221  -2.050  0.04032 *
factor(Gender)Male   0.6478     0.2250   2.879  0.00399 **
factor(Race)White   -1.3135     0.2378  -5.524 3.32e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37.516984  on 3  degrees of freedom
Residual deviance:  0.058349  on 1  degrees of freedom
AIC: 25.186

Number of Fisher Scoring iterations: 3
```

From this output we see that the p-value for testing that $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ is **0.00399**. Under $\alpha = 0.05$ we can reject H_0 and conclude that $\beta_1 \neq 0$.