Advanced Data Analysis HW5

Ao Liu, al3472

1.

For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the data Shuttle.csv shows the temperature in Fahrenheit at the time of the flight and whether at least one primary O-ring suffered thermal distress.

(a)

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

$$logit(\pi(TD|Temperature)) = \beta_0 + \beta_1 Temperature$$

$$\pi(TD|Temperature) = P(ThermalDistress = 1|Temperature)$$

Answer:

```
data = read.csv("Shuttle.csv", header = TRUE)
glm(ThermalDistress~Temperature, data = data, family = binomial("logit"))
```

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

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
```

(b)

Estimate β_1, the effect of temperature on the probability of thermal distress. Interpret your result.

Answer:

According to the result that we got in (a), our estimation of β_1, the effect of temperature on the probability of thermal distress is

$$-0.2322$$

This implies that when we increase the temperature by 1 degree, the odds of having Thermal Distress changes by a multiplicative factor of $e^{-0.2322}$

(c)

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 $e\beta_1$). Interpret your result

Answer:

```
confint(glm(ThermalDistress~Temperature, data = data, family = binomial("logit")))
```

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

According to the results in R, the 95% confidence interval for β_1 is

$$(-0.515718, -0.06082076)$$

so the the 95% confidence interval for e^{β_1} is

$$(0.597071743167396, 0.940991888047314)$$

This indicates that we are 95% confident that when we increase the temperature by 1 degree, the odds of having Thermal Distress changes by a multiplicative factor between 0.597071743167396 and 0.940991888047314.

(d)

Predict the probability of thermal distress at 31 degree, the temperature at the time of the Challenger flight.

Answer:

According to the estimation of the parameters, we have the following prediction function:

$$\hat{\pi}(TD|Temperature) = \frac{e^{15.0429-0.2322Temperature}}{1+e^{15.0429-0.2322Temperature}}$$

So when Temperature is 31 degree, we use the function above and get a prediction of the probability of Thermal Distress: 0.999608330327805.

(e)

At what temperature does the predicted probability equal 0.5?

Answer:

If the predicted probability equals to 0.5, then according to the function in (d), we have

$$2e^{15.0429-0.2322Temperature} = 1 + e^{15.0429-0.2322Temperature}$$

After solving this equation, the Temperature is 64.7842377260982.

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

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

(a)

Estimate β_1 and β_2 and interpret your result

Answer:

```
data = read.csv("adolescent.csv")
logit = glm(cbind(Yes,No)~factor(Gender)+factor(Race), data = data, family = binomial
  )
logit
```

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

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
```

The estimation for β_1 is 0.6478, the estimation for β_2 is -1.3135.
If we hold the gender fixed, then we estimate the odds that 15 or 16 year-old white adolescents having sexual intercourse is $e^{-1.3135}$ times the odds that 15 or 16 year-old black adolescents having sexual intercourse.
If we hold the race fixed, then we estimate the odds that 15 or 16 year-old male adolescents having sexual intercourse is $e^{0.6478}$ times the odds that 15 or 16 year-old female adolescents having sexual intercourse.

(b)

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 $e\beta_1$), Interpret your result

Answer:

```
confint = confint(glm(cbind(Yes,No)~factor(Gender)+factor(Race), data = data, family
  = binomial))
exp(confint)
```

```
                          2.5%      97.5%
(Intercept)           0.4077396  0.9764278
factor(Gender)Male    1.2343904  2.9872843
factor(Race)White     0.1682294  0.4279908
```

The 95% confidence interval to describe the effect of gender on the odds of Intercourse controlling for race is:

$$(1.2343904, 2.9872843)$$

We are 95% confident that, if we hold the race fixed, then we estimate the odds that 15 or 16 year-old male adolescents having sexual intercourse is between 1.2343904 and 2.9872843 times the odds that 15 or 16 year-old female adolescents having sexual intercourse.

(c)

Construct a 95% confidence interval to describe the effect of race on the odds of Intercourse controlling for gender (i.e. con- struct a 95% interval for $e\beta_2$), Interpret your result

Answer:

The 95% confidence interval to describe the effect of race on the odds of Intercourse controlling for race is:

$$(0.1682294, 0.4279908)$$

We are 95% confident that if we hold the gender fixed, then we estimate the odds that 15 or 16 year-old white adolescents having sexual intercourse is between 0.1682294 and 0.4279908 times the odds that 15 or 16 year-old black adolescents having sexual intercourse.

(d)

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

Answer:

```
summary(logit)
```

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

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 test statistics = 37.516984 - 0.058349 = 37.458635. Since p=2 we reject H_0 since $37.458635 > \chi^2_2(0.05) = 5.99$.

(e)

Test$H_0 : \beta_1 = 0$ **against** $H_a : \beta_1 \neq 0$**. Use** $\alpha = 0.05$**.**

Answer:

From the result of (d), we see that the p-value for testing that $H_0 : \beta_1 = 0$ against $H_a : \beta_1 = 0$ is 3.32e-08¡0.05. Therefore, we reject H_0.