HW5_1803

Yue Zhang

2024-11-11

#4.14

```
infection = read_table("E:/Biostat/Biostatistics/PH 1830/Infection.dat")

infection$Fail = infection$n - infection$y
infection = infection[, -5]
infection1 = infection %>%
    rowwise() %>%
    do(data.frame(Center = .$center, Treat = .$treat, y = c(rep(1,
        .$y), rep(0, .$Fail))))

infection_model = glm(y ~ Treat, family = binomial(link = "logit"),
    data = infection1)
summary(infection_model)
```

```
##
## Call:
## glm(formula = y ~ Treat, family = binomial(link = "logit"), data = infection1)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7142     0.1780  -4.012 6.03e-05 ***
## Treat         0.4040     0.2514   1.607    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 360.83  on 272  degrees of freedom
## Residual deviance: 358.23  on 271  degrees of freedom
## AIC: 362.23
##
## Number of Fisher Scoring iterations: 4
```

```
exp(0.404)
```

```
## [1] 1.497804
```

Null hypothesis: there's no treatment difference between drug and control group. Alternative hypothesis: there's treatment difference between drug and control group. Based on the result, we can see that the odds

of successes in drug group are 49.8% higher than the control group. However, the p-value for treatment effect is 0.108 which is larger than 0.05. Therefore, we fail to reject the null hypothesis, and there's no treatment difference between the two groups.

#4.17 ##(a) The equation is: $logit[\pi(x)/1-\pi(x)] = 0.1 * alcohol + 1.2 * smoking + 0.3 * race + 0.2 * (race * smoking) - 7$ For r=1, the equation is: $logit[\pi(x)/1-\pi(x)] = 0.1 * alcohol + 1.4 * smoking - 6.7$ For r=0, the equation is: $logit[\pi(x)/1-\pi(x)] = 0.1 * alcohol + 1.2 * smoking - 7$ For s=1, the equation is: $logit[\pi(x)/1-\pi(x)] = 0.1 * alcohol + 0.5 * race - 5.8$ For s-0, the equation is: $logit[\pi(x)/1-\pi(x)] = 0.1 * alcohol + 0.3 * race - 7$

```
# For r=1,
exp(1.4)
```

```
## [1] 4.0552
```

```
# For r=0,
exp(1.2)
```

```
## [1] 3.320117
```

```
# For s=1,
exp(0.5)
```

```
## [1] 1.648721
```

```
# For s=0,
exp(0.3)
```

```
## [1] 1.349859
```

The fitted conditional odds ratio for smoking effect: 4.0552(r=1), 3.3201(r=0). The fitted conditional odds ratio for race effect: 1.6487(s=1), 1.3499(s=0).

##(b) The coefficient of smoking is 1.2. This means that the log odds of the presence of squamous cell esophageal cancer when s=1 while holding other variables constant. Null hypothesis: smoking has no effect on the presence of squamous cell esophageal cancer ($\beta_{smoking} = 0$) Alternative hypothesis: smoking has an effect on the presence of squamous cell esophageal cancer ($\beta_{smoking} \neq 0$) Since the p-value < 0.01, we can state that smoking has a statistically significant effect on the presence of cancer and reject the null hypothesis.

The coefficient of race is 0.3. This means that the log odds of the presence of squamous cell esophageal cancer when r=1 while holding other variables constant. Null hypothesis: race has no effect on the presence of squamous cell esophageal cancer ($\beta_{race} = 0$) Alternative hypothesis: race has an effect on the presence of squamous cell esophageal cancer ($\beta_{race} \neq 0$) Since the p-value $= 0.02 < 0.05$, we can state that race has a statistically significant effect on the presence of cancer and reject the null hypothesis.

The coefficient of the cross-product is 0.2. This means that the log odds of the presence of squamous cell esophageal cancer when r=1 and s=1 while holding other variables constant. Null hypothesis: the cross-product of race and smoking has no effect on the presence of squamous cell esophageal cancer ($\beta_{race*smoking} = 0$) Alternative hypothesis: the cross-product of race and smoking has an effect on the presence of squamous cell esophageal cancer ($\beta_{race*smoking} \neq 0$) Since the p-value $= 0.04 < 0.05$, we can state that the cross product has a statistically significant effect on the presence of cancer and reject the null hypothesis.

#3

```r
burn = read.csv("E:/Biostat/Biostatistics/PH 1830/Data burn1000.csv")
burn = burn %>%
    mutate(death = ifelse(death == "Dead", 1, 0))

knots_4 = c(1.1, 19, 44.37, 78.87)
model_4 = glm(death ~ rcs(age, knots_4) + tbsa + race + inh_inj,
    family = binomial, data = burn)
summary(model_4)
```

```
##
## Call:
## glm(formula = death ~ rcs(age, knots_4) + tbsa + race + inh_inj,
##     family = binomial, data = burn)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.721231   0.757697  -7.551 4.33e-14 ***
## rcs(age, knots_4)age   -0.063392   0.060739  -1.044   0.2966
## rcs(age, knots_4)age'   0.507012   0.264028   1.920   0.0548 .
## rcs(age, knots_4)age'' -0.919094   0.519642  -1.769   0.0769 .
## tbsa                    0.091012   0.009177   9.917  < 2e-16 ***
## raceWhite              -0.562026   0.306528  -1.834   0.0667 .
## inh_injYes              1.516034   0.356538   4.252 2.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 845.42  on 999  degrees of freedom
## Residual deviance: 331.93  on 993  degrees of freedom
## AIC: 345.93
##
## Number of Fisher Scoring iterations: 7
```

```r
# Model with 3 knots at 10th, 50th, and 90th percentiles
quantiles_3 = quantile(burn$age, probs = c(0.1, 0.5, 0.9))
print(quantiles_3)
```

```
##   10%   50%   90%
##  1.50 31.95 67.11
```

```r
knots_3 = c(1.5, 31.95, 67.11)
model_3 = glm(death ~ rcs(age, knots_3) + tbsa + race + inh_inj,
    family = binomial, data = burn)
summary(model_3)
```

```
##
## Call:
## glm(formula = death ~ rcs(age, knots_3) + tbsa + race + inh_inj,
##     family = binomial, data = burn)
##
## Coefficients:
```

```
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -6.250213   0.729866  -8.564  < 2e-16 ***
## rcs(age, knots_3)age   0.024138   0.024886   0.970   0.3321
## rcs(age, knots_3)age'  0.057648   0.023634   2.439   0.0147 *
## tbsa                   0.089484   0.008988   9.955  < 2e-16 ***
## raceWhite             -0.614118   0.304880  -2.014   0.0440 *
## inh_injYes             1.473561   0.352619   4.179 2.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 845.42  on 999  degrees of freedom
## Residual deviance: 334.13  on 994  degrees of freedom
## AIC: 346.13
##
## Number of Fisher Scoring iterations: 7
```

```
# Model with 5 knots at 5th, 27.5th, 50th, 73.5th, and 95th
# percentiles
quantiles_5 = quantile(burn$age, probs = c(0.05, 0.275, 0.5,
    0.735, 0.95))
print(quantiles_5)
```

```
##     5%  27.5%    50%  73.5%    95%
##  1.100 13.100 31.950 50.200 78.235
```

```
knots_5 = c(1.1, 13.1, 31.95, 50.2, 78.235)
model_5 = glm(death ~ rcs(age, knots_5) + tbsa + race + inh_inj,
    family = binomial, data = burn)
summary(model_5)
```

```
##
## Call:
## glm(formula = death ~ rcs(age, knots_5) + tbsa + race + inh_inj,
##     family = binomial, data = burn)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -5.706947   0.842387  -6.775 1.25e-11 ***
## rcs(age, knots_5)age    -0.069039   0.109985  -0.628   0.5302
## rcs(age, knots_5)age'    0.656409   1.097112   0.598   0.5496
## rcs(age, knots_5)age''  -0.856560   2.173857  -0.394   0.6936
## rcs(age, knots_5)age''' -0.124725   1.627924  -0.077   0.9389
## tbsa                     0.091201   0.009195   9.919  < 2e-16 ***
## raceWhite               -0.564139   0.306283  -1.842   0.0655 .
## inh_injYes               1.516759   0.357027   4.248 2.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 845.42  on 999  degrees of freedom
```

```
## Residual deviance: 331.72  on 992  degrees of freedom
## AIC: 347.72
##
## Number of Fisher Scoring iterations: 7
```

```
anova(model_4, model_3, model_5)
```

```
## Analysis of Deviance Table
##
## Model 1: death ~ rcs(age, knots_4) + tbsa + race + inh_inj
## Model 2: death ~ rcs(age, knots_3) + tbsa + race + inh_inj
## Model 3: death ~ rcs(age, knots_5) + tbsa + race + inh_inj
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       993     331.93
## 2       994     334.13 -1  -2.2059   0.1375
## 3       992     331.72  2   2.4171   0.2986
```

When comparing model_4(has 4 knots) with model_3(has 3 knots), the p-value is 0.1375 which means that there isn't a statistically significant improvement in fit when moving from 4 to 3 knots. When comparing model_4(has 4 knots) with model_5(has 5 knots), the p-value is 0.2986 which suggests that there isn;t a statistically significant improvement in fit when moving from 4 to 5 knots. Also, the AIC value is the least in model_4. Therefore, the four knots model provides a better fit than others.