STAT 426 HW 8

Nate Beebe

2025-05-02

Question 1 **Part a)**

```
library(tidyverse)
library(VGAM)

auto = read.csv("C:/Users/STP/Downloads/AutoAccidentInjuries.csv")

auto_long <- auto %>%
  pivot_longer(y1:y5, names_to = "injury", values_to = "freq") %>%
  mutate(injury = factor(injury, levels = paste0("y", 1:5), ordered = TRUE),
         Gender = factor(Gender, levels = c("Male","Female")),
         Location = factor(Location, levels = c("Urban","Rural")),
         Seatbelt = factor(Seatbelt, levels = c("No","Yes"))
)

model1a = vglm(injury ~ Gender + Location + Seatbelt + Location:Seatbelt, family = propodds(reverse = F
summary(model1a)
```

```
##
## Call:
## vglm(formula = injury ~ Gender + Location + Seatbelt + Location:Seatbelt,
##     family = propodds(reverse = FALSE), data = auto_long, weights = freq)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept):1               2.54727    0.02877  88.541   <2e-16 ***
## (Intercept):2               2.72169    0.02928  92.968   <2e-16 ***
## (Intercept):3               4.58923    0.04221 108.716   <2e-16 ***
## (Intercept):4               6.49618    0.08908  72.929   <2e-16 ***
## GenderFemale               -0.54625    0.02721 -20.075   <2e-16 ***
## LocationRural              -0.82326    0.03476 -23.687   <2e-16 ***
## SeatbeltYes                 0.76016    0.03938  19.302   <2e-16 ***
## LocationRural:SeatbeltYes   0.12442    0.05477   2.272   0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4])
##
## Residual deviance: 50874.33 on 152 degrees of freedom
##
```

```
## Log-likelihood: -25437.16 on 152 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):1', '(Intercept):3', '(Intercept):4'
##
##
## Exponentiated coefficients:
##              GenderFemale             LocationRural               SeatbeltYes
##                 0.5791160                 0.4389966                 2.1386111
## LocationRural:SeatbeltYes
##                 1.1324880
```

Part b)

There are 5 response categories so there needs to be 5-1 (4) cumulative logits, giving us 4 intercepts.

```
newdata = data.frame(
  Gender = factor("Male", levels = c("Male","Female")),
  Location = factor("Urban", levels = c("Urban","Rural")),
  Seatbelt = factor("Yes", levels = c("No","Yes")))

prediction1b <- predict(model1a, newdata = newdata, type = "response")
prediction1b
```

```
##          y1          y2         y3          y4           y5
## 1 0.9646826 0.005484223 0.02510459 0.004023422 0.0007051921
```

The probability of no injury is extremely high (~96%) for this particular combination.

Part c)

```
coef(model1a)["GenderFemale"]
```

```
## GenderFemale
##   -0.5462525
```

```
exp(coef(model1a)["GenderFemale"])
```

```
## GenderFemale
##     0.579116
```

For any level of injury, the cumulative odds for a female are exp(-0.54625) = 0.579116 higher than for a male with the same location and seatbelt usage.

Part d)

```
partd = summary(model1a)
beta_hat = coef(partd)["GenderFemale","Estimate"]
se_hat = coef(partd)["GenderFemale","Std. Error"]
z = qnorm(0.975)
ci_log = beta_hat + c(-1,1)*z* se_hat
ci_or = exp(ci_log)
ci_log
```

```
## [1] -0.5995848 -0.4929201
```

```
ci_or
```

```
## [1] 0.5490395 0.6108400
```

Assuming seatbelt usage and location are held constant, the cumulative odds that a female will have a less-severe injury (as defined by the dataset) is likely to be between 0.549 and 0.611. This is a statistically significant result because the entire interval is below the value of 1 (meaning no difference).

Part e)

```
coefs = coef(summary(model1a))
beta_sb = coefs["SeatbeltYes","Estimate"]
beta_int = coefs["LocationRural:SeatbeltYes","Estimate"]
OR_Urban = exp(beta_sb)
OR_Rural = exp(beta_sb + beta_int)
data.frame(
Location = c("Urban","Rural"),
Cumulative_OR = c(OR_Urban, OR_Rural)
)
```

```
##   Location Cumulative_OR
## 1    Urban      2.138611
## 2    Rural      2.421951
```

In an urban crash the odds multiply by 2.139 for wearing a seatbelt. In a rural crash the odds multiply by 2.422 for wearing a seatbelt.

Question 2 **Part a)**

```
data <- data.frame(
  SeatBelt = factor(c("Yes","Yes","No","No"), levels = c("Yes","No")),
  Fatal = factor(c("Yes","No","Yes","No"), levels = c("Yes","No")),
  Freq = c(25, 51790, 54, 10325))

model2a = glm(Freq ~ SeatBelt + Fatal, family = poisson, data = data)
summary(model2a)
```

```
##
## Call:
## glm(formula = Freq ~ SeatBelt + Fatal, family = poisson, data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.18687    0.11252   37.21   <2e-16 ***
## SeatBeltNo  -1.60790    0.01075 -149.52   <2e-16 ***
## FatalNo      6.66729    0.11257   59.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 115243.62  on 3  degrees of freedom
## Residual deviance:    104.07  on 1  degrees of freedom
## AIC: 144.74
##
## Number of Fisher Scoring iterations: 5
```

```
1-pchisq(deviance(model2a), df.residual(model2a))
```

```
## [1] 0
```

The p-value is 0. This means that we reject the null hyopthesis and conclude that this model does not fit the data.

Part b) The first lambda hat Y will be 0 due to the reference category, and the second is 6.66729. The difference of the lambda hat Y's is -6.66729.

```
exp(-6.66729)
```

```
## [1] 0.001271841
```

Fatal crashes are about 0.13% as frequent as non-fatal crashes after averaging over seatbelt use.

Part c)

```
model2c <- glm(Freq ~ SeatBelt * Fatal, family = poisson, data = data)
coef(summary(model2c))
```

```
##                       Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)          3.2188758  0.2000000 16.094379 2.793598e-58
## SeatBeltNo           0.7701082  0.2419060  3.183502 1.455052e-03
## FatalNo              7.6360765  0.2000483 38.171171 0.000000e+00
## SeatBeltNo:FatalNo -2.3827372  0.2421460 -9.840085 7.564672e-23
```

```
(25/51790) / (54/10325)
```

```
## [1] 0.0922976
```

```
exp(coef(summary(model2c))["SeatBeltNo:FatalNo","Estimate"])
```

```
## [1] 0.0922976
```

Due to the reference cell constraints, the only coefficient needed is the interaction term. From that, you simply need to exponentiate it.

Part d)

```
logOR = log((25*10325)/(51790*54))
SE = sqrt(1/25 + 1/51790 + 1/54 + 1/10325)
CI_log = logOR + c(-1,1)*1.96*SE
CI_OR = exp(CI_log)
CI_OR
```

```
## [1] 0.05742111 0.14835740
```

***Question 3* Part a)**

```
data3 = as.data.frame(UCBAdmissions)
model3a <- glm(Freq ~ Admit + Gender + Dept, family = poisson, data = data3)
summary(model3a)
```

```
##
## Call:
## glm(formula = Freq ~ Admit + Gender + Dept, family = poisson,
##     data = data3)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.37111    0.03964 135.498  < 2e-16 ***
## AdmitRejected  0.45674    0.03051  14.972  < 2e-16 ***
## GenderFemale  -0.38287    0.03027 -12.647  < 2e-16 ***
## DeptB         -0.46679    0.05274  -8.852  < 2e-16 ***
## DeptC         -0.01621    0.04649  -0.349 0.727355
## DeptD         -0.16384    0.04832  -3.391 0.000696 ***
## DeptE         -0.46850    0.05276  -8.879  < 2e-16 ***
## DeptF         -0.26752    0.04972  -5.380 7.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2650.1  on 23  degrees of freedom
## Residual deviance: 2097.7  on 16  degrees of freedom
## AIC: 2272.7
##
## Number of Fisher Scoring iterations: 5
```

```
1-pchisq(deviance(model3a), df.residual(model3a))
```

```
## [1] 0
```

The p-value comes out to be zero meaning that this model shows lack of fit. There is evidence to suggest the variables are not independent.

Part b)

```
data3 %>%
  mutate(pearson = rstandard(model3a, type = "pearson")) %>%
  arrange(desc(abs(pearson)))
```

```
##       Admit Gender Dept Freq     pearson
## 1  Admitted   Male    A  512  24.8802909
## 2  Admitted   Male    B  353  22.4162147
## 3  Rejected Female    A   19 -17.4033457
## 4  Rejected Female    E  299  15.4633866
```

```
## 5  Rejected Female    B    8 -13.7636600
## 6  Rejected Female    C  391  13.4448838
## 7  Admitted   Male    F   22 -13.4083131
## 8  Rejected Female    F  317  12.8305813
## 9  Rejected   Male    C  205  -9.6255354
## 10 Admitted Female    F   24  -9.5092774
## 11 Admitted Female    B   17  -8.8463067
## 12 Admitted   Male    E   53  -8.3962737
## 13 Admitted   Male    C  120  -7.7321646
## 14 Rejected   Male    F  351   7.5474457
## 15 Rejected   Male    E  138  -6.7507677
## 16 Admitted Female    C  202   5.5601331
## 17 Admitted Female    A   89  -5.5209782
## 18 Rejected Female    D  244   4.1600383
## 19 Admitted   Male    D  138  -4.0072036
## 20 Rejected   Male    A  313  -1.9711664
## 21 Rejected   Male    D  279  -0.7371577
## 22 Admitted Female    D  131   0.6674523
## 23 Rejected   Male    B  207  -0.5380980
## 24 Admitted Female    E   94   0.2584495
```

Many of these residuals are much higher than what would be expected if there was truly independence. This supports the findings in part a.

Part c)

```
model3c = glm(Freq ~ Admit + Gender*Dept, data = data3, family = poisson)
1 - pchisq(deviance(model3c), df.residual(model3c))
```

```
## [1] 0
```

The p-value here is also zero meaning this model does not fit. There is likely some dependence with Admissions and at least one of the other categories (or an interaction).

Part d)

```
model3d = glm(Freq ~ Admit*Dept + Gender*Dept, data = data3, family = poisson)
1 - pchisq(deviance(model3d), df.residual(model3d))
```

```
## [1] 0.001351993
```

This value is also lower that 0.05 meaning that we conclude it does not fit either. It is worth noting the massive change in deviance and that this model does a much better job than the previous one. This does suggest that there is another dependent relationship between Gender and Admissions.

Part e)

```
data3 %>%
  mutate(pearson = rstandard(model3d, type = "pearson")) %>%
  arrange(desc(abs(pearson)))
```

```
##       Admit Gender Dept Freq    pearson
## 1  Rejected Female    A   19 -4.1530742
```

```
## 2  Admitted Female    A   89  4.1530739
## 3  Rejected   Male    A  313  4.1530624
## 4  Admitted   Male    A  512 -4.1530621
## 5  Admitted   Male    E   53  1.0005342
## 6  Admitted Female    E   94 -1.0005342
## 7  Rejected   Male    E  138 -1.0005342
## 8  Rejected Female    E  299  1.0005342
## 9  Rejected   Male    C  205 -0.8680662
## 10 Admitted Female    C  202 -0.8680662
## 11 Admitted   Male    C  120  0.8680662
## 12 Rejected Female    C  391  0.8680662
## 13 Admitted   Male    F   22 -0.6197526
## 14 Admitted Female    F   24  0.6197526
## 15 Rejected Female    F  317 -0.6197526
## 16 Rejected   Male    F  351  0.6197526
## 17 Rejected Female    D  244 -0.5458732
## 18 Admitted   Male    D  138 -0.5458732
## 19 Admitted Female    D  131  0.5458732
## 20 Rejected   Male    D  279  0.5458732
## 21 Rejected Female    B    8 -0.5037077
## 22 Admitted Female    B   17  0.5037077
## 23 Admitted   Male    B  353 -0.5037077
## 24 Rejected   Male    B  207  0.5037077
```

These residuals are much closer to normal than the ones looked at previously. There is still some issues with the first four having absolute values too high. All of these residuals come from department A. Other than department A, the model fits pretty well.

Part f)

```r
model3f = glm(Freq ~ Admit*Gender + Admit*Dept + Gender*Dept, data = data3, family = poisson)
summary(model3f)
```

```
##
## Call:
## glm(formula = Freq ~ Admit * Gender + Admit * Dept + Gender *
##     Dept, family = poisson, data = data3)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  6.27150    0.04271 146.855  < 2e-16 ***
## AdmitRejected               -0.58205    0.06899  -8.436  < 2e-16 ***
## GenderFemale                -1.99859    0.10593 -18.866  < 2e-16 ***
## DeptB                       -0.40322    0.06784  -5.944 2.78e-09 ***
## DeptC                       -1.57790    0.08949 -17.632  < 2e-16 ***
## DeptD                       -1.35000    0.08526 -15.834  < 2e-16 ***
## DeptE                       -2.44982    0.11755 -20.840  < 2e-16 ***
## DeptF                       -3.13787    0.16174 -19.401  < 2e-16 ***
## AdmitRejected:GenderFemale -0.09987    0.08085  -1.235    0.217
## AdmitRejected:DeptB         0.04340    0.10984   0.395    0.693
## AdmitRejected:DeptC         1.26260    0.10663  11.841  < 2e-16 ***
## AdmitRejected:DeptD         1.29461    0.10582  12.234  < 2e-16 ***
## AdmitRejected:DeptE         1.73931    0.12611  13.792  < 2e-16 ***
## AdmitRejected:DeptF         3.30648    0.16998  19.452  < 2e-16 ***
```

```
## GenderFemale:DeptB         -1.07482    0.22861  -4.701 2.58e-06 ***
## GenderFemale:DeptC          2.66513    0.12609  21.137  < 2e-16 ***
## GenderFemale:DeptD          1.95832    0.12734  15.379  < 2e-16 ***
## GenderFemale:DeptE          2.79519    0.13925  20.073  < 2e-16 ***
## GenderFemale:DeptF          2.00232    0.13571  14.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2650.095  on 23  degrees of freedom
## Residual deviance:   20.204  on  5  degrees of freedom
## AIC: 217.26
##
## Number of Fisher Scoring iterations: 4
```

```r
#Estimate
exp(coef(model3f)["AdmitRejected:GenderFemale"])
```

```
## AdmitRejected:GenderFemale
##                   0.904955
```