Stsci4110HW5

Nick Gembs

11/2/2022

df = read.csv("C:/Users/Nick/Downloads/bookspoll\_f2022.csv")

df

#a  
  
# Total Fit  
  
fit.total = glm(books~news+prelim+summerjob, data = df, family = "poisson")  
summary(fit.total)

##   
## Call:  
## glm(formula = books ~ news + prelim + summerjob, family = "poisson",   
## data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2335 -1.1946 0.0007 0.7888 3.0734   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.456863 0.471048 0.970 0.3321   
## newsYes 0.012671 0.087626 0.145 0.8850   
## prelim 0.012338 0.005322 2.318 0.0204 \*  
## summerjobYes -0.166093 0.106632 -1.558 0.1193   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 264.36 on 127 degrees of freedom  
## Residual deviance: 257.28 on 124 degrees of freedom  
## AIC: 643.88  
##   
## Number of Fisher Scoring iterations: 5

# Backward Selection 1  
  
fit.np = glm(books~news+prelim, data = df, family = "poisson")  
fit.ns = glm(books~news+summerjob, data = df, family = "poisson")  
fit.ps = glm(books~prelim+summerjob, data = df, family = "poisson")  
  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

#HO : summerjob is not significant in predicting book  
#H1 : summerjob is significant in predicting book  
lrtest(fit.total,fit.np)

## Likelihood ratio test  
##   
## Model 1: books ~ news + prelim + summerjob  
## Model 2: books ~ news + prelim  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 4 -317.94   
## 2 3 -319.12 -1 2.358 0.1246

#HO : prelim is not significant in predicting book  
#H1 : prelim is significant in predicting book  
lrtest(fit.total,fit.ns)

## Likelihood ratio test  
##   
## Model 1: books ~ news + prelim + summerjob  
## Model 2: books ~ news + summerjob  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 4 -317.94   
## 2 3 -320.85 -1 5.8213 0.01583 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#HO : news is not significant in predicting book  
#H1 : news is significant in predicting book  
lrtest(fit.total,fit.ps)

## Likelihood ratio test  
##   
## Model 1: books ~ news + prelim + summerjob  
## Model 2: books ~ prelim + summerjob  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 4 -317.94   
## 2 3 -317.95 -1 0.0209 0.885

# highest p-value was in the model testing news, meaning that news was the worst predictor of book. Remove news.  
  
# Backward Selection 2  
  
fit.p = glm(books~prelim, data = df, family = "poisson")  
fit.s = glm(books~summerjob, data = df, family = "poisson")  
  
#HO : summerjob is not significant in predicting book  
#H1 : summerjob is significant in predicting book  
lrtest(fit.ps,fit.p)

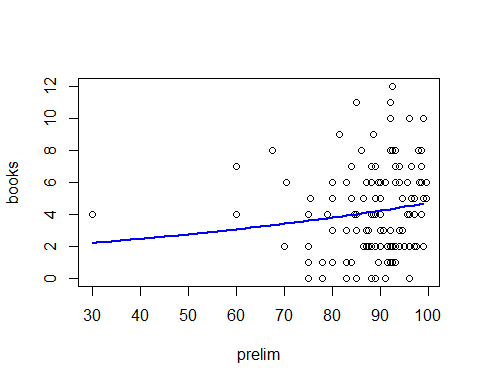
## Likelihood ratio test  
##   
## Model 1: books ~ prelim + summerjob  
## Model 2: books ~ prelim  
## #Df LogLik Df Chisq Pr(>Chisq)  
## 1 3 -317.95   
## 2 2 -319.12 -1 2.3395 0.1261

#HO : prelim is not significant in predicting book  
#H1 : prelim is significant in predicting book  
lrtest(fit.ps,fit.s)

## Likelihood ratio test  
##   
## Model 1: books ~ prelim + summerjob  
## Model 2: books ~ summerjob  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 3 -317.95   
## 2 2 -320.85 -1 5.8032 0.016 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# highest p-value was in the model testing summerjob, meaning that summerjob was the next worst predictor of book. Remove summerjob.  
  
finalmodel = glm(books~prelim, data = df, family = "poisson")

#b  
  
x = seq(min(df$prelim),max(df$prelim), 1)  
dataframe = data.frame(prelim = x)  
  
p.final = predict.glm(finalmodel, newdata = dataframe, type = "response")  
  
plot(df$prelim, df$books, xlab = "prelim", ylab = "books")  
  
lines(x,p.final, col = "blue", lwd = 2)



#c  
  
#summary(fit.total)  
#summary(finalmodel)  
  
ans.total = exp(fit.total$coefficients[1] + 0\*fit.total$coefficients[2] + 80\*fit.total$coefficients[3] + 0\*fit.total$coefficients[4])  
ans.final = exp(finalmodel$coefficients[1] + finalmodel$coefficients[2]\*80)  
  
cat("Using the full model, the expected number of books is: ", ans.total, "\n" , "Using the reduced selected model, the expected number of books is: ", ans.final)

## Using the full model, the expected number of books is: 4.237246   
## Using the reduced selected model, the expected number of books is: 3.791931

#d  
  
ans.total = exp(fit.total$coefficients[1] + 1\*fit.total$coefficients[2] + 95\*fit.total$coefficients[3] + 1\*fit.total$coefficients[4])  
ans.final = exp(finalmodel$coefficients[1] + finalmodel$coefficients[2]\*95)  
  
cat("Using the full model, the expected number of books is: ", ans.total, "\n" , "Using the reduced selected model, the expected number of books is: ", ans.final)

## Using the full model, the expected number of books is: 4.373497   
## Using the reduced selected model, the expected number of books is: 4.464408

#e  
  
# 𝐻0: ‘extra’ parameters (those in the saturated model, but not in our model are equal to zero.) ”Good Fit of Model M”  
# 𝐻𝐴: at least one of these ‘extra’ parameters is not zero “Lack-of-fit of Model M”  
  
summary(finalmodel)

##   
## Call:  
## glm(formula = books ~ prelim, family = "poisson", data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.00442 -1.24925 -0.06769 0.83097 3.01209   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.46215 0.46579 0.992 0.3211   
## prelim 0.01088 0.00520 2.093 0.0364 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 264.36 on 127 degrees of freedom  
## Residual deviance: 259.64 on 126 degrees of freedom  
## AIC: 642.24  
##   
## Number of Fisher Scoring iterations: 5

pvalue = 1- pchisq(259.64,df = 126)  
pvalue

## [1] 2.664635e-11

cat("The P-value is 0, reject the null. This model has a lack of fit, it could be suggested to be less restrictive in the backward selection, adding parameters to the model. The model needs to have an increased # of parameters or increased flexibility.")

## The P-value is 0, reject the null. This model has a lack of fit, it could be suggested to be less restrictive in the backward selection, adding parameters to the model. The model needs to have an increased # of parameters or increased flexibility.

df2 = read.csv("C:/Users/Nick/Downloads/birthdata2.csv")

df2

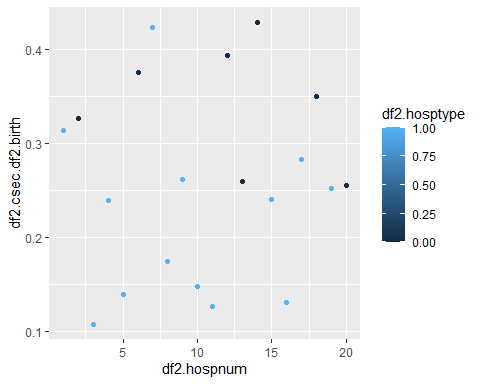
#a  
  
raw <- data.frame(df2$hospnum, df2$csec/df2$birth, df2$hosptype)  
raw

## df2.hospnum df2.csec.df2.birth df2.hosptype  
## 1 1 0.3135593 1  
## 2 2 0.3261164 0  
## 3 3 0.1072165 1  
## 4 4 0.2387178 1  
## 5 5 0.1391586 1  
## 6 6 0.3755523 0  
## 7 7 0.4230769 1  
## 8 8 0.1745283 1  
## 9 9 0.2612446 1  
## 10 10 0.1470588 1  
## 11 11 0.1260504 1  
## 12 12 0.3935185 0  
## 13 13 0.2590068 0  
## 14 14 0.4285714 0  
## 15 15 0.2401276 1  
## 16 16 0.1304348 1  
## 17 17 0.2828685 1  
## 18 18 0.3497668 0  
## 19 19 0.2516364 1  
## 20 20 0.2552083 0

print("It appears that private hospitals (hosptype = 0) may have higher c-section rates than public hospitals.")

## [1] "It appears that private hospitals (hosptype = 0) may have higher c-section rates than public hospitals."

#b  
  
library(ggplot2)  
  
ggplot(data = raw, aes(x = df2.hospnum, y = df2.csec.df2.birth, color = df2.hosptype)) +   
 geom\_point()



#c  
  
cat("The model is ln(mean\_csec) = ln(births) + B0 + B1\*hosptype + B2\*ses")

## The model is ln(mean\_csec) = ln(births) + B0 + B1\*hosptype + B2\*ses

#d  
  
c.fit = glm(df2$csec~ df2$hosptype + df2$ses + offset(log(df2$births)), family = "poisson")  
summary(c.fit)

##   
## Call:  
## glm(formula = df2$csec ~ df2$hosptype + df2$ses + offset(log(df2$births)),   
## family = "poisson")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0917 -0.6024 -0.1610 1.0360 2.0946   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.86113 0.05784 -32.177 <2e-16 \*\*\*  
## df2$hosptype -0.26653 0.03070 -8.683 <2e-16 \*\*\*  
## df2$ses 0.24234 0.01613 15.021 <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: 479.178 on 19 degrees of freedom  
## Residual deviance: 23.871 on 17 degrees of freedom  
## AIC: 165.81  
##   
## Number of Fisher Scoring iterations: 3

# e  
  
#HO : predictor is not significant in predicting mean(csec)  
#H1 : predictor is significant in predicting mean(csec)  
  
anova(c.fit)

## Analysis of Deviance Table  
##   
## Model: poisson, link: log  
##   
## Response: df2$csec  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 19 479.18  
## df2$hosptype 1 208.72 18 270.46  
## df2$ses 1 246.59 17 23.87

1-pchisq(208.72,1)

## [1] 0

1-pchisq(246.59,1)

## [1] 0

cat("P-values for both hosptype and ses are less than .05. They are both significant predictors of csec.")

## P-values for both hosptype and ses are less than .05. They are both significant predictors of csec.

ses\_rate = exp(c.fit$coefficients[1]+c.fit$coefficients[3]+c.fit$coefficients[2])/exp(c.fit$coefficients[1]+c.fit$coefficients[2])  
ses\_rate

## (Intercept)   
## 1.274222

hosptype\_rate = exp(c.fit$coefficients[1]+c.fit$coefficients[2]+c.fit$coefficients[3])/exp(c.fit$coefficients[1]+c.fit$coefficients[3])  
hosptype\_rate

## (Intercept)   
## 0.7660308

cat("An increase in ses by one unit results in a ", ses\_rate , " multiplicative change in mean csec\n")

## An increase in ses by one unit results in a 1.274222 multiplicative change in mean csec

cat("A change from private to public hospital results in a ", hosptype\_rate , " multiplicative change in mean csec")

## A change from private to public hospital results in a 0.7660308 multiplicative change in mean csec

#f  
  
  
# 𝐻0: ‘extra’ parameters (those in the saturated model, but not in our model are equal to zero.) ”Good Fit of Model M”  
# 𝐻𝐴: at least one of these ‘extra’ parameters is not zero “Lack-of-fit of Model M”  
  
summary(c.fit)

##   
## Call:  
## glm(formula = df2$csec ~ df2$hosptype + df2$ses + offset(log(df2$births)),   
## family = "poisson")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0917 -0.6024 -0.1610 1.0360 2.0946   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.86113 0.05784 -32.177 <2e-16 \*\*\*  
## df2$hosptype -0.26653 0.03070 -8.683 <2e-16 \*\*\*  
## df2$ses 0.24234 0.01613 15.021 <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: 479.178 on 19 degrees of freedom  
## Residual deviance: 23.871 on 17 degrees of freedom  
## AIC: 165.81  
##   
## Number of Fisher Scoring iterations: 3

pvalue = 1- pchisq(23.871,df = 17)  
pvalue

## [1] 0.1229826

cat("The P-value is .1229826 > .05, fail to reject the null. This model has a good fit")

## The P-value is .1229826 > .05, fail to reject the null. This model has a good fit

#g  
  
x = seq(min(df2$hospnum),max(df2$hospnum), 1)  
dataframe = data.frame(hospnum = x, actual = df2$csec)  
  
p.final = predict.glm(c.fit, newdata = dataframe, type = "response")  
  
dataframe = cbind(dataframe, predicted= p.final, hosptype = df2$hosptype, births = df2$births, ses = df2$ses)  
dataframe

## hospnum actual predicted hosptype births ses  
## 1 1 74 62.544334 1 236 3.3  
## 2 2 241 274.946197 0 739 3.6  
## 3 3 104 121.279547 1 970 0.2  
## 4 4 566 517.620369 1 2371 2.5  
## 5 5 43 41.547785 1 309 0.5  
## 6 6 255 258.819845 0 679 3.7  
## 7 7 11 7.968864 1 26 3.9  
## 8 8 222 234.366358 1 1272 1.8  
## 9 9 848 860.249609 1 3246 3.3  
## 10 10 280 288.987421 1 1904 1.0  
## 11 11 45 52.887823 1 357 0.9  
## 12 12 425 411.672213 0 1080 3.7  
## 13 13 266 247.020133 0 1027 1.8  
## 14 14 12 8.581538 0 28 2.8  
## 15 15 602 588.582980 1 2507 2.8  
## 16 16 18 19.010476 1 138 0.6  
## 17 17 142 157.634520 1 502 4.0  
## 18 18 525 532.028539 0 1501 3.4  
## 19 19 692 694.319913 1 2750 3.1  
## 20 20 49 39.931536 0 192 1.2

#h  
  
  
#plot(df2$ses,(c.fit$fitted.values/df2$births)\*1000)  
  
ggplot(data = df2, aes(x = df2$ses, y = (c.fit$fitted.values/df2$births)\*1000, color = df2$hosptype)) +   
 geom\_point() + xlab("ses") + ylab("Predicted csec per 1000 births")

## Warning: Use of `df2$ses` is discouraged. Use `ses` instead.

## Warning: Use of `df2$births` is discouraged. Use `births` instead.

## Warning: Use of `df2$hosptype` is discouraged. Use `hosptype` instead.

