Unit 3 Assignment

Sri Seshadri

11/5/2017

Chapter 6 Question 1

```
stress <- read.csv(file = "stress.csv")
knitr::kable(mosaic::fav_stats(stress$STRESS), caption = "Summary statistics")
```

Table 1: Summary statistics

min	Q1	median	Q3	max	mean	sd	n	missing
0	0	1	3	9	1.729647	1.849082	651	0

```
par(mfrow = c(1, 2))
hist(stress$STRESS, main = "Histogram of stress - Frequency",
    xlab = "Stress")

hist(stress$STRESS, prob = T, main = "Histogram of stress - Denisty",
    xlab = "Stress")
curve(dpois(x, lambda = 1.72), col = "darkblue", lwd = 2, add = TRUE,
    from = 0, to = 9, n = 10, type = "l")
```

a. Looking at the histogram and the summary statistic, where the mean and the standard deviaiton are alike, the most likely probability distribution in Poisson distribution.

```
qqnorm(y = stress$STRESS, main = "Normal Q-Q plot of stress")
qqline(y = stress$STRESS)
```

b. The variable "Stress" is a count variable and is discrete in nature. It does not align with a continuous distribution like normal. The evidence is seen in the clustering of points at different levels in the Q-Q plot.

Chapter 6 Question 2

```
options(width = 120)
# poisson regression
preg <- glm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH,
    family = poisson(link = "log"), data = stress)
preg.summary <- broom::tidy(preg)
knitr::kable(preg.summary, digits = 2, caption = "Poisson regression coefficients")
```

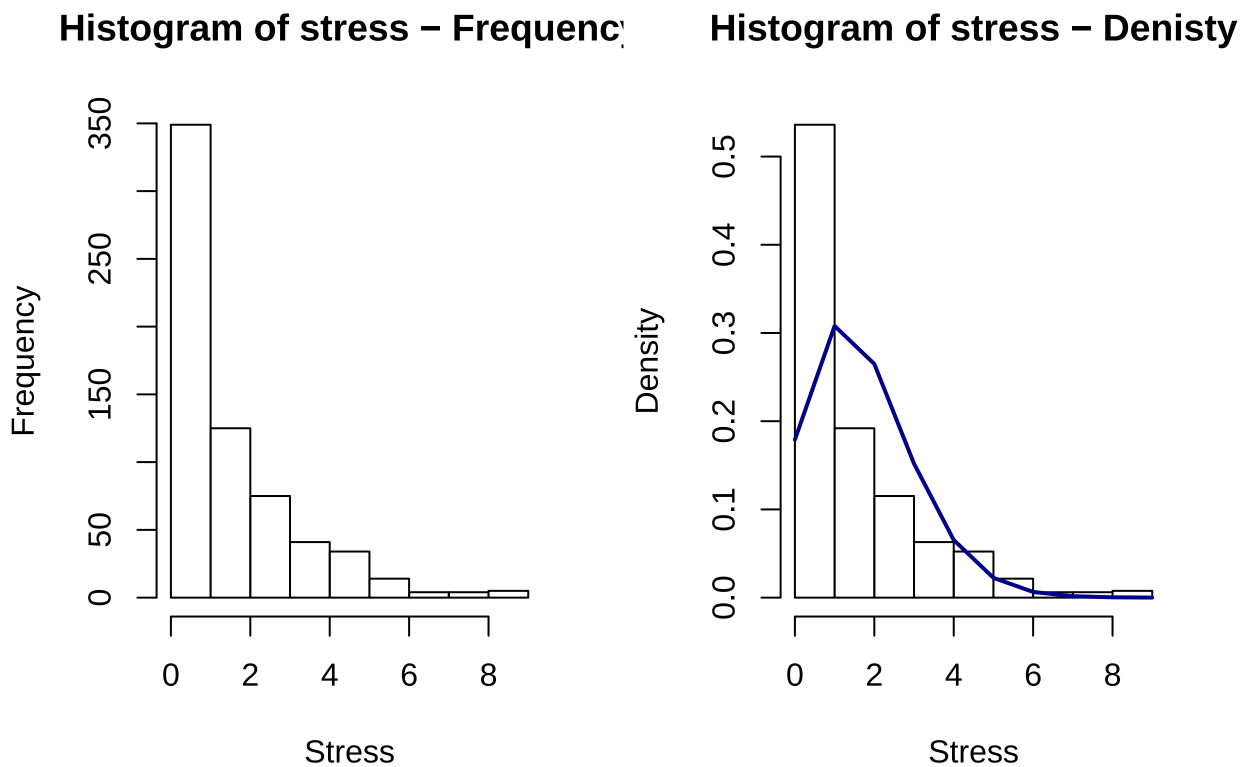

Figure 1: Histogram stress

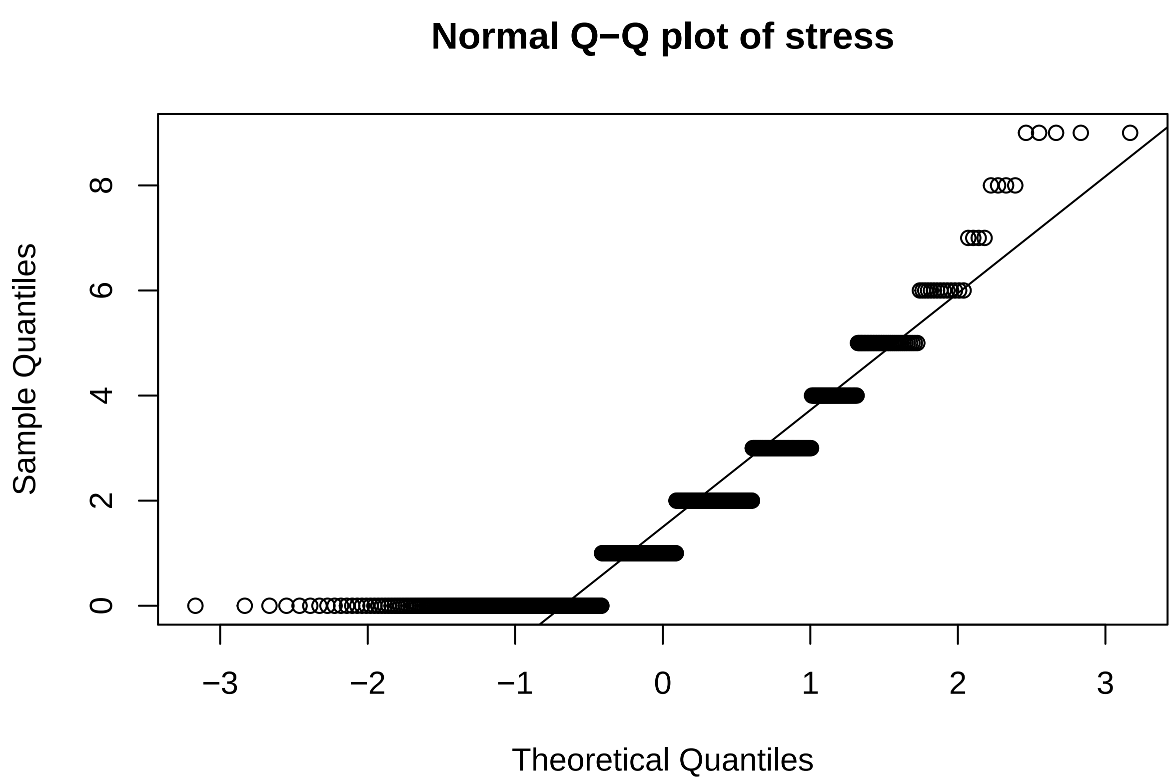

Figure 2: Normal Q-Q plot

Table 2: Poisson regression coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	2.73	0.23	11.68	0.00
COHES	-0.01	0.00	-4.47	0.00
ESTEEM	-0.02	0.01	-2.95	0.00
GRADES	-0.02	0.01	-2.38	0.02
SATTACH	-0.02	0.01	-2.85	0.00

```
# overdispersed poisson regression
overdisp.preg <- glm(formula = STRESS ~ COHES + ESTEEM + GRADES +
    SATTACH, family = quasipoisson(link = "log"), data = stress)
overdisp.preg.summary <- broom::tidy(overdisp.preg)
knitr::kable(overdisp.preg.summary, digits = 2, caption = "Overdispersed Poisson regression coefficient
```

Table 3: Overdispersed Poisson regression coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	2.73	0.31	8.76	0.00
COHES	-0.01	0.00	-3.35	0.00
ESTEEM	-0.02	0.01	-2.21	0.03
GRADES	-0.02	0.01	-1.78	0.07
SATTACH	-0.02	0.01	-2.14	0.03

```
# negative binomial
nbreg <- MASS::glm.nb(formula = STRESS ~ COHES + ESTEEM + GRADES +
    SATTACH, data = stress)
nbreg.summary <- broom::tidy(nbreg)
knitr::kable(nbreg.summary, digits = 2, caption = "Negative binomial regression coefficients")
```

Table 4: Negative binomial regression coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	2.76	0.34	8.08	0.00
COHES	-0.01	0.00	-3.24	0.00
ESTEEM	-0.02	0.01	-2.01	0.04
GRADES	-0.02	0.01	-1.74	0.08
SATTACH	-0.02	0.01	-2.02	0.04

a.

```
1. Poisson regression:

  An increase in cohes by 1 unit reduces stress by 1.28%.
  An increase in sattach by 1 unit reduces stress by 1.63%

2. Negative binomial

     An increase in cohes by 1 unit reduces stress by 1.33%.
```

```
An increase in sattach by 1 unit reduces stress by 1.66%
```

It is seen that the point estimates do not change much by the choice of distribution, because the mean and standard devation are not too far from eachother. (However the standard error does.)

b.

```r
result <- stress %>% mutate(COHES.LEVEL = ifelse(COHES < mean(COHES) -
    sd(COHES), 0, ifelse(COHES > mean(COHES) + sd(COHES), 2,
    1))) %>% mutate(poisson.predict = preg$fitted.values) %>%
    mutate(nb.predict = nbreg$fitted.values) %>% group_by(COHES.LEVEL) %>%
    summarise(mean.poisson.prediction = mean(poisson.predict),
        mean.nb.prediction = mean(nb.predict))

knitr::kable(result, digits = 2, caption = "mean prediction for low (0), medium(1) , high(2) COHES")
```

Table 5: mean prediction for low (0), medium(1) , high(2) COHES

COHES.LEVEL	mean.poisson.prediction	mean.nb.prediction
0	2.50	2.52
1	1.67	1.66
2	1.19	1.18

c.

Based on poisson regression the expected reduction in stress from high to low level of cohesion is : 52.52%

Based on negative binomial regression the expected reduction on stress from high to low lwvel of cohesion is : 53.33%

Chapter 6 Question 3

```r
library(broom)
glances <- rbind(glance(preg), glance(overdisp.preg), glance(nbreg))
glances$model <- c("Poisson", "Overdispersed Poisson", "Negative Binomial")


output <- glances %>% select(model, AIC, BIC, logLik)
# Notice no AICs for overdispersion in Output


# hand calculation of LogLik of overdispersed model
glances$logLik[2] <- sum(dpois(stress$STRESS, lambda = overdisp.preg$fitted.values,
    log = T))

# Notice the loglik for overdispersed model is same as for
# poisson. Since the AIC abd BICs are derived from the
# LogLik, the AIC BIC for overdispersed model would be
```

```r
# identical as poisson

output[2, ] <- output[1, ]

knitr::kable(output, digits = 2, caption = "AIC and BIC of models")
```

Table 6: AIC and BIC of models

model	AIC	BIC	logLik
Poisson	2417.22	2439.61	-1203.61
Poisson	2417.22	2439.61	-1203.61
Negative Binomial	2283.59	2310.46	-1135.79

c. Choosing Negative binomial in this case for lower AIC and BIC

Chapter 6 Question 4

```r
par(mfrow = c(1, 2))
plot(preg$fitted.values, residuals.glm(preg, type = "deviance"),
    xlab = "Fitted values", ylab = "deviance residuals")
plot(round(preg$fitted.values, 0), residuals.glm(preg, type = "deviance"))
```