HW4

Ben Tankus

4/24/2021

R Question

1. (5 points) In 1968, Dr. Benjamin Spock was tried in Boston on charges of conspiring to violate the Selective Service Act by encouraging young men to resist the draft for military service in Vietnam. The defense in the case challenged the method of jury selection claiming that women were underrepresented on jury panels by the process. The defense argued specifically that the judge in the Spock trial had a history of venires (panels of potential jurors) in which women were systematically underrepresented compared to the venires of six other Boston area district judges. These data can be found in case0502 in the Sleuth3 library.

Analyze the data by treating the number of women out of 30 people on a venire as a binomial response (that is, you'll change the percent women in the datasheet to a count by multiplying by 30 and rounding) and judge as an explanatory variable .

```
df <- case0502

df['Women'] <- round(df$Percent*30/100)


mod1 <- glm(cbind(Women, 30 - Women) ~ (Judge), data = df, family = binomial(link = "logit"))

summary(mod1)
```

```
##
## Call:
## glm(formula = cbind(Women, 30 - Women) ~ (Judge), family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.12843  -0.50536  -0.06147   0.52960   1.80026
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.66329    0.17236  -3.848 0.000119 ***
## JudgeB        0.01974    0.23305   0.085 0.932484
## JudgeC       -0.23749    0.21849  -1.087 0.277049
## JudgeD       -0.34831    0.33902  -1.027 0.304239
## JudgeE       -0.37691    0.24188  -1.558 0.119171
## JudgeF       -0.32945    0.22019  -1.496 0.134599
## JudgeSpock's -1.08591    0.24302  -4.468 7.88e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62.919  on 45  degrees of freedom
## Residual deviance: 31.119  on 39  degrees of freedom
## AIC: 208.53
##
## Number of Fisher Scoring iterations: 4
```

(a) Is there evidence of over dispersion in these data? Please explain (i.e., don't just answer "yes" or "no").

The dispersion parameter using quasibinomial fit is 0.784 which provides evidence of *under*-dispersion, but no *over*-dispersion.

(b) Do the odds of a female on a venire differ for the different judges? Please explain.

There is a clear difference in Spock's contribution to the number of women in the venire model. Spock's judge has the only significant p-vale at 7.88e-06, where the next closes is judge E at 0.12.

(c) Do judges A-F differ in their probabilities of selecting females to the venire? Please explain.

```
mod3 <- glm(cbind(Women, 30 - Women) ~ Judge, subset = (Judge !="Spock's"), data = df, family = binomial

summary(mod3)
```

```
##
## Call:
## glm(formula = cbind(Women, 30 - Women) ~ Judge, family = binomial(link = "logit"),
##     data = df, subset = (Judge != "Spock's"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.12843  -0.51805  -0.04574   0.52960   1.80026
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.66329    0.17236  -3.848 0.000119 ***
## JudgeB       0.01974    0.23305   0.085 0.932484
## JudgeC      -0.23749    0.21849  -1.087 0.277049
## JudgeD      -0.34831    0.33902  -1.027 0.304239
## JudgeE      -0.37691    0.24188  -1.558 0.119171
## JudgeF      -0.32945    0.22019  -1.496 0.134599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31.747  on 36  degrees of freedom
```

```
## Residual deviance: 26.168  on 31  degrees of freedom
## AIC: 173.19
##
## Number of Fisher Scoring iterations: 4
```

There is no significant difference between the remaining judges. Because the P-values remaining are all above 0.05 at a 95% confidence level.

(d) How different are the odds of having a woman on the Spock judge's venires from the odds of having a woman on the venires of other judges? Please explain. (Hint: In parts (b) through (d), think about what models you could compare using drop-in-deviance tests.)

```
mod4 <- glm(cbind(Women, 30 - Women) ~ factor(Judge), data = df, family = binomial(link = "logit"))
```

```
mod5 <- glm(cbind(Women, 30 - Women) ~ (Judge!="Spock's"), data = df, family = binomial(link = "logit"))
```

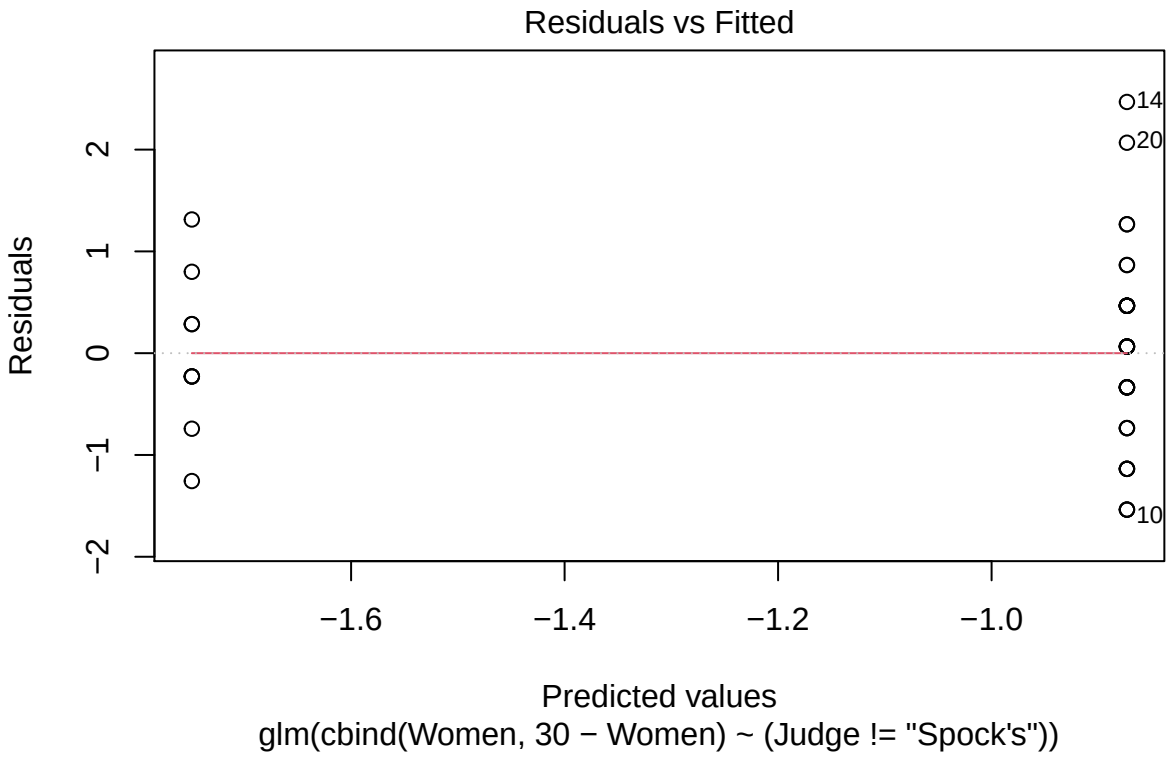
```
anova(mod4, mod5,test="Chisq")
```

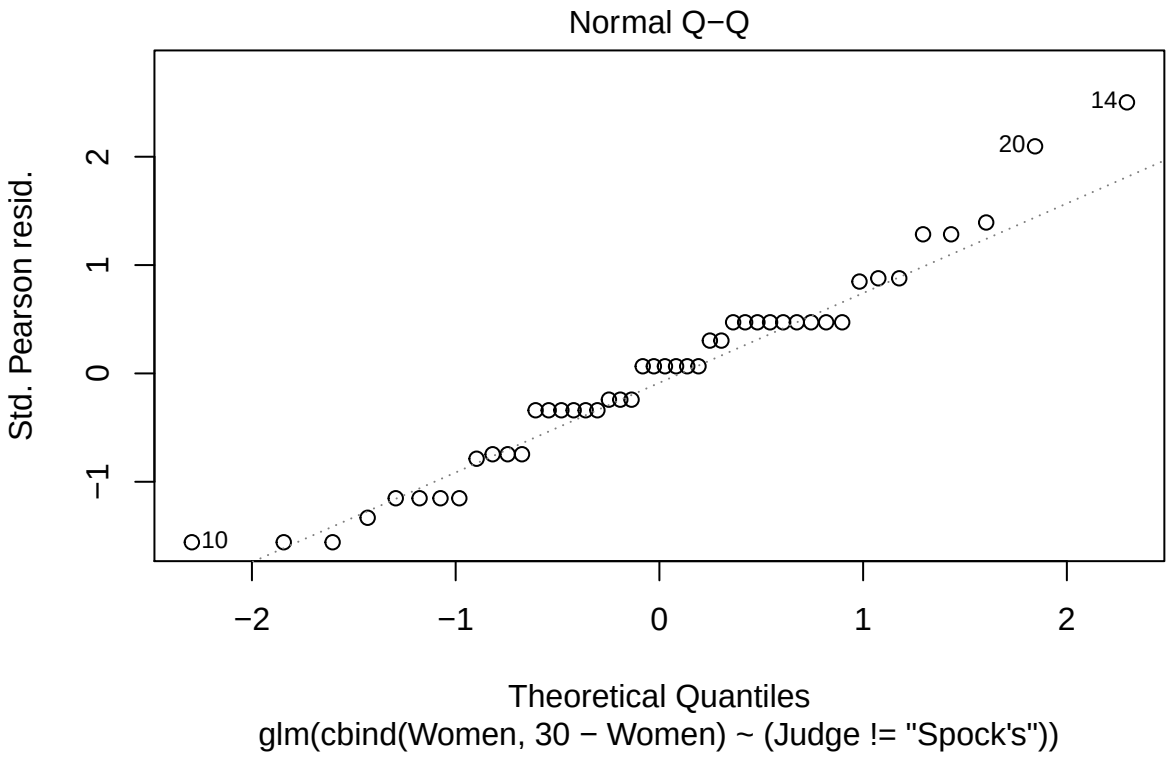
```
## Analysis of Deviance Table
##
## Model 1: cbind(Women, 30 - Women) ~ factor(Judge)
## Model 2: cbind(Women, 30 - Women) ~ (Judge != "Spock's")
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        39     31.119
## 2        44     36.698 -5  -5.5782   0.3494
```

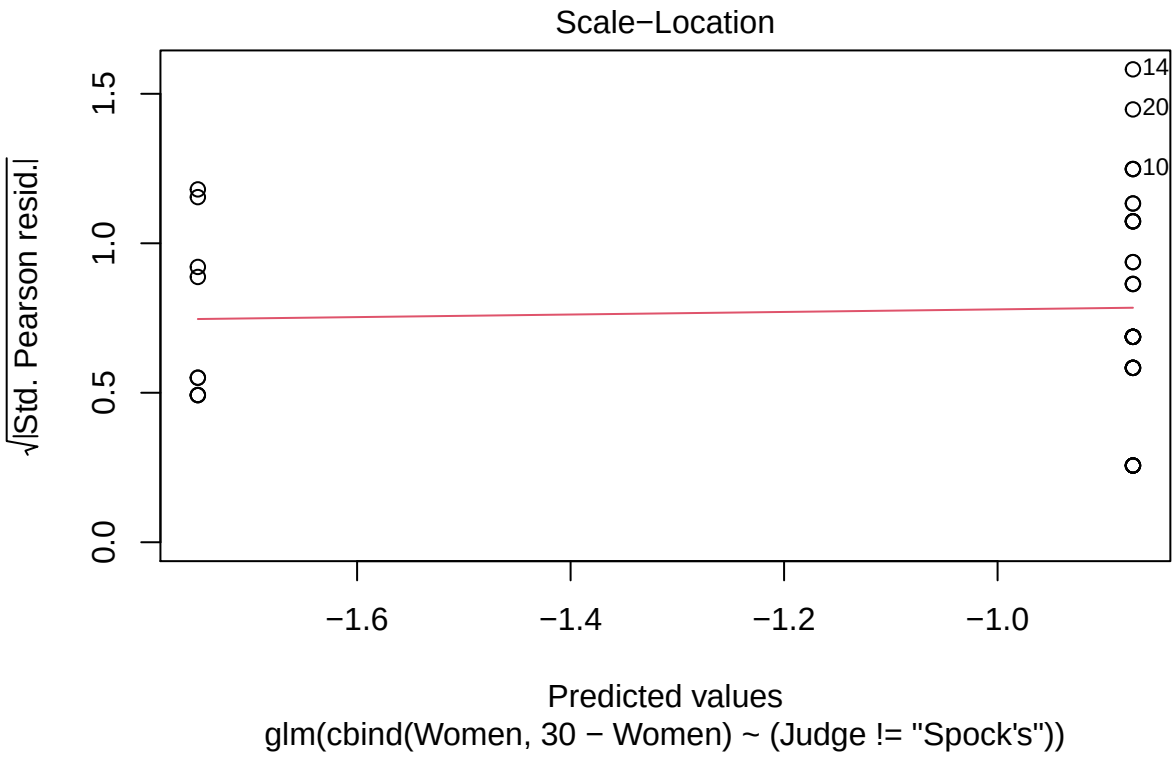
```
LRstats(mod4, mod5)
```

```
## Likelihood summary table:
##         AIC    BIC LR Chisq Df Pr(>Chisq)
## mod4 208.53 221.33   31.119 39     0.8116
## mod5 204.10 207.76   36.698 44     0.7746
```

```
plot(mod5)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
14
20
10
−1.6
−1.4
−1.2
−1.0
Predicted values
glm(cbind(Women, 30 − Women) ~ (Judge != "Spock's"))


Normal Q−Q
Std. Pearson resid.
−1
0
1
2
14
20
10
−2
−1
0
1
2
Theoretical Quantiles
glm(cbind(Women, 30 − Women) ~ (Judge != "Spock's"))


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−1.6
−1.4
−1.2
−1.0
14
20
10
Predicted values
glm(cbind(Women, 30 − Women) ~ (Judge != "Spock's"))

No significant difference between these too models. That means that the remaining judges are not significantly different in the proportion of women venirs they select.

Conceptual Questions

2. (2 points) Give three explanations for why you may see evidence of extra-binomial variation in a logistic regression.(M4L4)
3. Possible dependence among binary random variables; unaccounted for explanatory information.
4. Poor deviance goodness-of-fit after fitting a rich model
5. Outliers in the data
6. (2 points) In lecture, we saw that when we fit a logistic model when extra binomial variation is present, we get standard errors that are 'too small'. Explain why this gives misleading results.

When the standard errors are 'too small' this gives p-values that are 'too small' as well (mathematical computation of a p-value). This could contribute to incorrectly rejecting a true null hypothesis (type I error).

4. (1 point) Consider an experiment that studies the number of insects that survive a certain dose of an insecticide, using several batches of insects of size n each. The insects may be sensitive to factors that vary among batches during the experiment but these factors (such as temperature) were unmeasured. Explain why the distribution of the number of insects per batch surviving the experiment might show over dispersion relative to a binomial(n, p) distribution.

There may be dependent data in this study as temperature was unmeasured. As temperature fluctuates over the course of a year, it's possible that batches studied in July may behave differently than batches studied in January, thus introducing dependent data. This dependence could introduce over-dispersion in the study.