Exploring the Relationship between Gender, Political Orientation, and Education Level

Greg Johnson

Introduction

The following data represent the responses of students with regard to their education level (Degree), political affiliation (Party) and gender (Gender). Our data may be grouped and represented as a three-way contingency table (2 by 5 by 7).

```
politic <- array(data = c(32, 67, 12, 23, 16, 20, 85, 14, 21, 9, 18, 63, 6,
    29, 12, 29, 68, 9, 20, 13, 11, 48, 13, 19, 7, 12, 65, 17, 32, 14, 9, 44,
    6, 20, 13, 31, 118, 20, 33, 38, 25, 98, 16, 23, 20, 16, 69, 13, 28, 8, 58,
    88, 13, 11, 13, 8, 30, 7, 16, 3, 8, 82, 16, 44, 13, 16, 54, 7, 23, 9), dim = c(5,
    7, 2), dimnames = list(Degree = c("LT highschool", "High school", "Junior college",
    "Bachelor", "Graduate"), Party = 1:7, Gender = c("Male", "Female")))
politic
```

```
## , , Gender = Male
##
##                 Party
## Degree            1  2  3  4  5  6  7
##   LT highschool  32 20 18 29 11 12  9
##   High school    67 85 63 68 48 65 44
##   Junior college 12 14  6  9 13 17  6
##   Bachelor       23 21 29 20 19 32 20
##   Graduate       16  9 12 13  7 14 13
##
## , , Gender = Female
##
##                 Party
## Degree            1  2  3  4  5  6  7
##   LT highschool   31 25 16 58  8  8 16
##   High school    118 98 69 88 30 82 54
##   Junior college  20 16 13 13  7 16  7
##   Bachelor        33 23 28 11 16 44 23
##   Graduate        38 20  8 13  3 13  9
```

Association of Party and Gender

Let's take a look at the association between Party and Gender, ignoring Degree. We can test the statistical hypothesis that Gender and Party are statistically independent: $H_0 : \pi_{ij} = \pi_{i+} \cdot \pi_{+j}$ for $\forall i, j$. One appropriate test statistic for this exact situation is the chi-squared test of independence.

Some quick comments before running the model: the test statistic is asymptotically distributed as chi-squared. Our large sample size ensures this convergence in distribution. Also, the chi-squared test treats both variables as nominal when in fact Party is ordinal. Thus a more powerful test would exploit the ordering of Party. Regardless, we proceed with the chi-squared test using an (arbitrary) alpha level of: $\alpha = 0.05$.

```
tab1 <- t(apply(politic, MARGIN = c(2, 3), sum))
tab1
```

```
##         Party
## Gender    1    2    3    4   5    6    7
##   Male    150 149 128 139 98 140   92
##   Female 240 182 134 183 64 163 109
```

```r
chi1 <- chisq.test(tab1)  #chi-squared model
chi1
```

```
##
##   Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 24.474, df = 6, p-value = 0.0004271
```

Since our p-value is less than our alpha level, we reject our null hypothesis and can say that, ignoring Degree, Party and Gender are associated i.e. statistically dependent. One drawback of the chi-squared test is that it only tells us about the presence of a relationship (which it has in this case). To investigate the association we can look at the standardized residuals:

```r
chi1[["stdres"]]
```

```
##          Party
## Gender            1          2          3          4          5          6
##   Male    -3.098763 -0.1778679  1.185479 -0.9028193  4.011472  0.2832832
##   Female   3.098763  0.1778679 -1.185479  0.9028193 -4.011472 -0.2832832
##          Party
## Gender            7
##   Male     0.09373869
##   Female  -0.09373869
```

Since the standardized residuals are null distributed as standard normal, any whose absolute value is greater than 2 indicate a great lack of fit for that category. According to the standardized residuals, the model of independence predicts much less strong-democratic males than there actually are; the model also predicts more independent (nearly Republican) males than there actually are.

An alternative to the parameteric test of independence is to use a permutation test. Permutation tests are a type of resampling technique (Givens & Hoeting, 2013). Say there are two categorical variables, baldness and heart disease. Under the null hypothesis, the two variables are independent and their joint distribution is the product of their marginal distributions i.e. P(Baldness = $i \cap$ HeartDisease = j) = P(Baldness = i) $\cdot P$(HeartDisease = j) for all values of i and j. In other words, for any one realization of heart disease, it doesn't matter what group of baldness it is associated with. For the whole dataset, the baldness "labels" can be randomly shuffled (permuted) without changing the joint null distribution (which is just the product of the marginals). Each of the M number of permutations we make is an equally likely random sample (under the null hypothesis) as the original sample of ours. If we compute our test statistic for each of the permuted samples, we can create an empirical null sampling distribution against which we can compare our computed test statistic. Since the number of permutations for our sample size is way too large to reasonably compute, we settle for just a very large $M = 10000$.

```r
test.obs = chisq.test(tab1)$statistic

Party <- c(rep("1", 390), rep("2", 331), rep("3", 262), rep("4", 322), rep("5",
    162), rep("6", 303), rep("7", 201))
Gender <- c(rep("Male", 896), rep("Female", 1075))

table(Party)
```

```
## Party
##   1   2   3   4   5   6   7
```

```
## 390 331 262 322 162 303 201
```

```r
table(Gender)
```

```
## Gender
## Female   Male
##   1075    896
```

```r
y <- Party
x <- Gender
teststat = numeric(10000)
for (i in 1:10000) {
    y.s = sample(y)
    table.s = table(x, y.s)
    teststat[i] = chisq.test(table.s)$statistic
}
(p = sum(teststat >= test.obs)/10000)
```
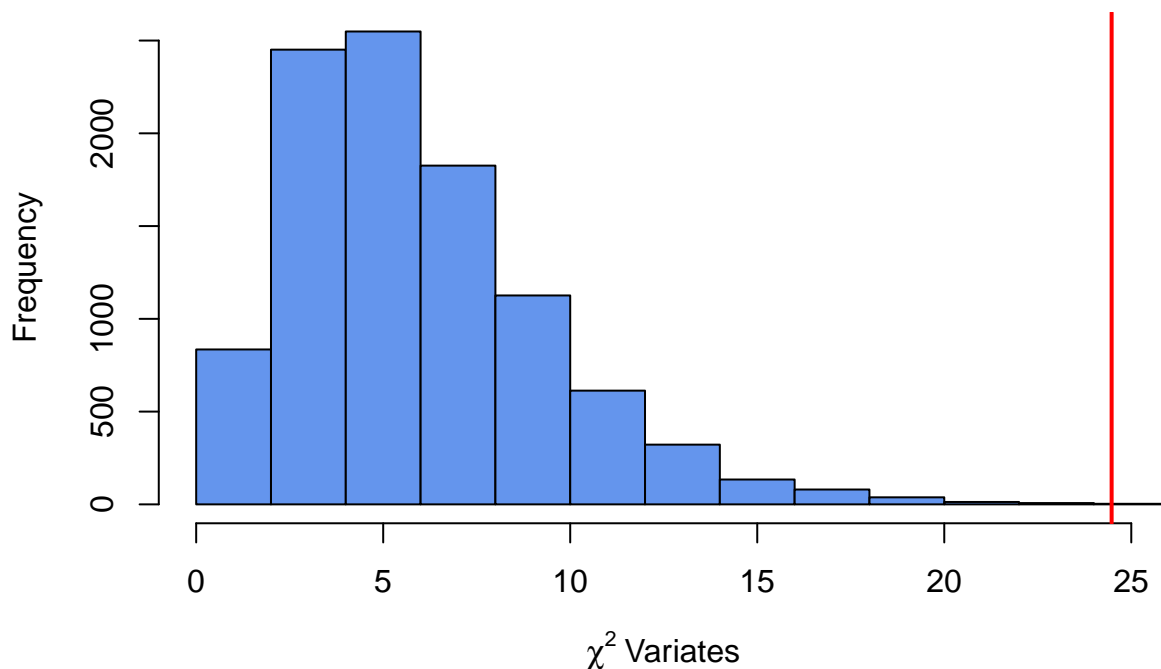
```
## [1] 6e-04
```

```r
(ci = quantile(teststat, prob = c(0.025, 0.975)))
```

```
##      2.5%     97.5%
##  1.223017 14.364170
```

```r
hist(teststat, col = "cornflowerblue", xlab = expression(chi^2 ~ "Variates"),
    xlim = c(0, 25), main = expression("Figure 1. Histogram of" ~ chi^2 ~ "Variates under" ~
        H[0]))
abline(v = test.obs, lwd = 2, col = "red")
```

## Figure 1. Histogram of χ^2 Variates under H_0

The p-value obtained from using the theoretical sampling distribution of the chi-squared test statistic and the p-value obtained from the chi-squared resampling method are virtually the same - both are less than our

$\alpha = 0.05$.

Visualizing the Three-Way Contingency Table

```r
# rename dimensions to prevent cluttering in the plots
dimnames(politic) <- list(Degree = c("LT HS", "HS", "JC", "B", "G"), Party = 1:7,
    Gender = c("M", "F"))
mosaic(politic, main = "Figure 2. Mosaic Plot of Politic Data")
```

Figure 2. Mosaic Plot of Politic Data
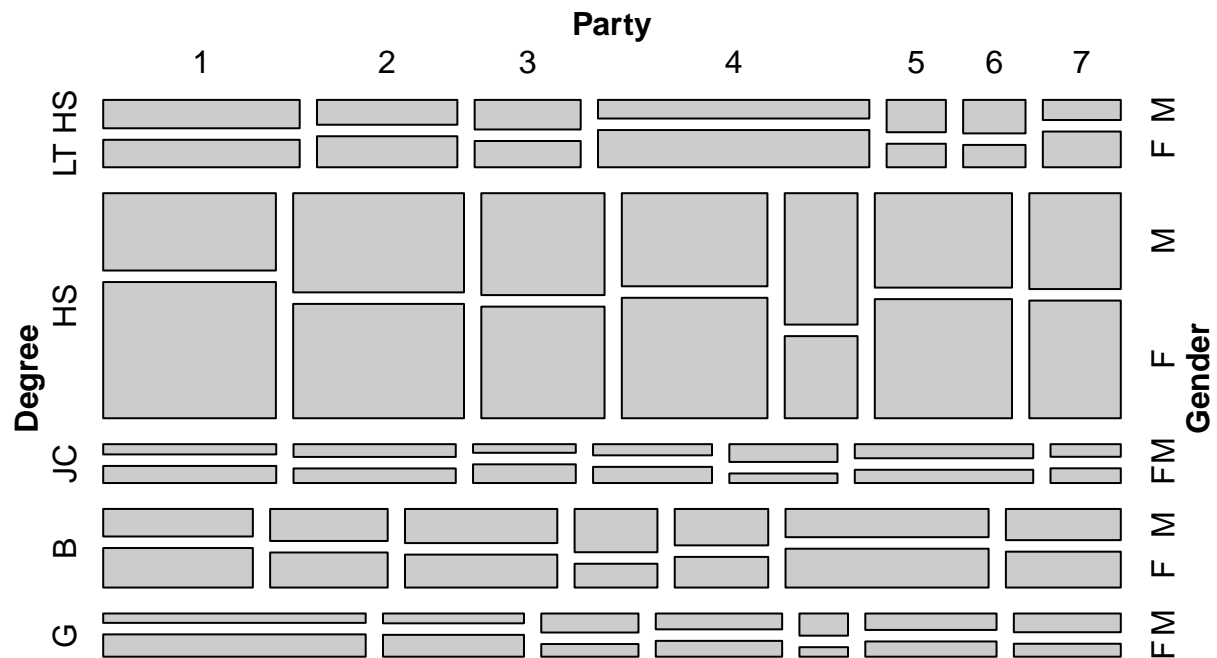
A Mosaic plot visually depicts the expected cell counts (frequencies) for each group formed by cross-classifying by the variables under consideration. Each group is assigned an area which is proportional to the expected cell counts. When applied to our data, we can get a feel for how large groups should be relative to each under independence. For example, under independence, we should expect to see much more Strong Democratic females in high school than Strong Republican females in high school however there isn't as much of a difference for the same two groups in Bachelors.

```r
sieve(politic, main = "Figure 3. Sieve Display of Politic Data")
```

Figure 3. Sieve Display of Politic Data
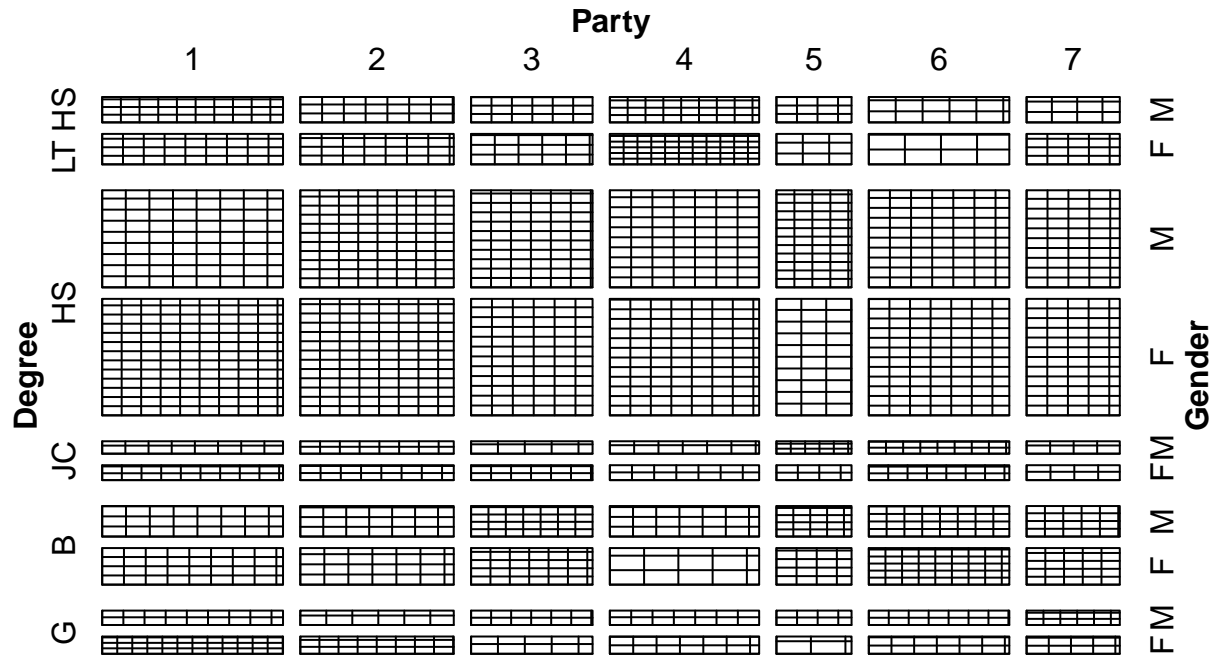
The sieve plot goes a step further than the mosaic plot and represents the observed cell counts in addition to the expected cell counts. It adds in the observed cell counts as the number of squares in the area. The deviations apparent in the plot can be used to evaluate log-linear models which includes evaluating independence in this case. A positive deviation means the expected cell count was larger than expected under independence and this is shown by partial squares in the area (it implies there are more squares, they just don't fit in the proportioned area). A negative deviation means the cell count was smaller than expected under independence and this is shown by black strips (it implies the squares don't fill up the whole proportioned area). The sieve plot appears to show many deviations, implying that a model of independence between Gender and Party does not fit the data well. There is some sort of dependence between the two variables that warrants exploration.

Using the Binomial GLM to Predict Education

Instructions: Dichotomize the Degree variable into a new variable called Education. Then perform a logistic regression analysis in which Party and Gender predict Education. Interpret, analyse residuals, and perform model selection.

Let's fit the logistic regression model: $log\text{Odds} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ where X_1 is gender and X_2 is political party.

```
# dichotomize degree
pd1 <- as.data.frame(as.table(politic))
pd1[["Education"]] <- as.numeric(pd1[["Degree"]] %in% c("JC", "B", "G"))

# reshape data into a dataframe with a frequency variable
pd2 <- as.data.frame(as.table(tapply(pd1[["Freq"]], list(pd1[["Party"]], pd1[["Gender"]]),
    pd1[["Education"]]), sum)))
names(pd2) <- c("Party", "Gender", "Education", "Freq")
pd2[["Party"]] <- as.numeric(pd2[["Party"]])
```

```r
glm2 <- glm(Education ~ Party + Gender, weights = Freq, data = pd2, family = "binomial")
summary(glm2)[["coefficients"]]
```

```
##               Estimate Std. Error     z value      Pr(>|z|)
## (Intercept) -0.76555469 0.11297897 -6.7760812 1.234793e-11
## Party        0.05342662 0.02336966  2.2861530 2.224531e-02
## GenderF     -0.05329264 0.09474158 -0.5625053 5.737718e-01
```

```r
glm4 <- glm(Education ~ 1, weights = Freq, data = pd2, family = "binomial")  #null
# (intercept-only) model LRT, test all coefficients are zero
LRT <- -2 * (as.numeric(logLik(glm4)) - as.numeric(logLik(glm2)))
pchisq(LRT, df = 2, lower.tail = FALSE)
```

```
## [1] 0.05794517
```

```r
# wald test coefficients
B <- summary(glm2)[["coefficients"]][, 1]
B.SE <- summary(glm2)[["coefficients"]][, 2]
pchisq((B/B.SE)^2, df = 1, lower.tail = FALSE)
```

```
##  (Intercept)        Party       GenderF
## 1.234793e-11 2.224531e-02 5.737718e-01
```

```r
# 95% confidence intervals for odds ratios
exp(confint(glm2))
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.3722153 0.579702
## Party       1.0076603 1.104363
## GenderF     0.7874838 1.141740
```

The result of our LRT test is $p = 0.058$ which is bad news for our model - it's poor evidence for any of our predictors being significant. However an $\alpha = 0.05$ is arbitrary and it would be silly to dismiss the entire model based on a binary decision. So we can follow up the LRT test by looking at Wald tests for the coefficients. The effect for party is statistically significant ($p < 0.05$) but not so for gender ($p > 0.05$). Let's interpret the effect of party then:

$$\frac{Odds_{party=x+1}}{Odds_{party=x}} = e^{0.053} = 1.054$$

$$Odds_{party=x+1} = 1.054 \cdot Odds_{party=x}$$

So conditional on gender, when party increases by one unit, the odds of being educated increase by 5%. A 95% confidence interval for the log odds ratio is $(1.008, 1.104)$. If we contend that gender is of practical significance and should stay in our model, we can also interpret its effect with the caveat that the observed effect is probably due to random chance.

$$\frac{Odds_{female}}{Odds_{male}} = e^{-0.053} = 0.948$$

$$Odds_{female} = 0.948 \cdot Odds_{male}$$

Conditional on party, males have 5% greater odds of being educated than females. Again, this effect is not statistically significant.

Since the predictors are categorical, we can use G^2 as a measure of goodness of fit as it will have an asymptotic chi-squared distribution. If there is good fit, we should expect G^2 to be close to its degrees of freedom (its

expectation). $G^2 = 2557.7$ on $n - p = 1968$ degrees of freedom so the model does not fit very well. We can further explore the lack-of-fit by analyzing the residuals of the model.
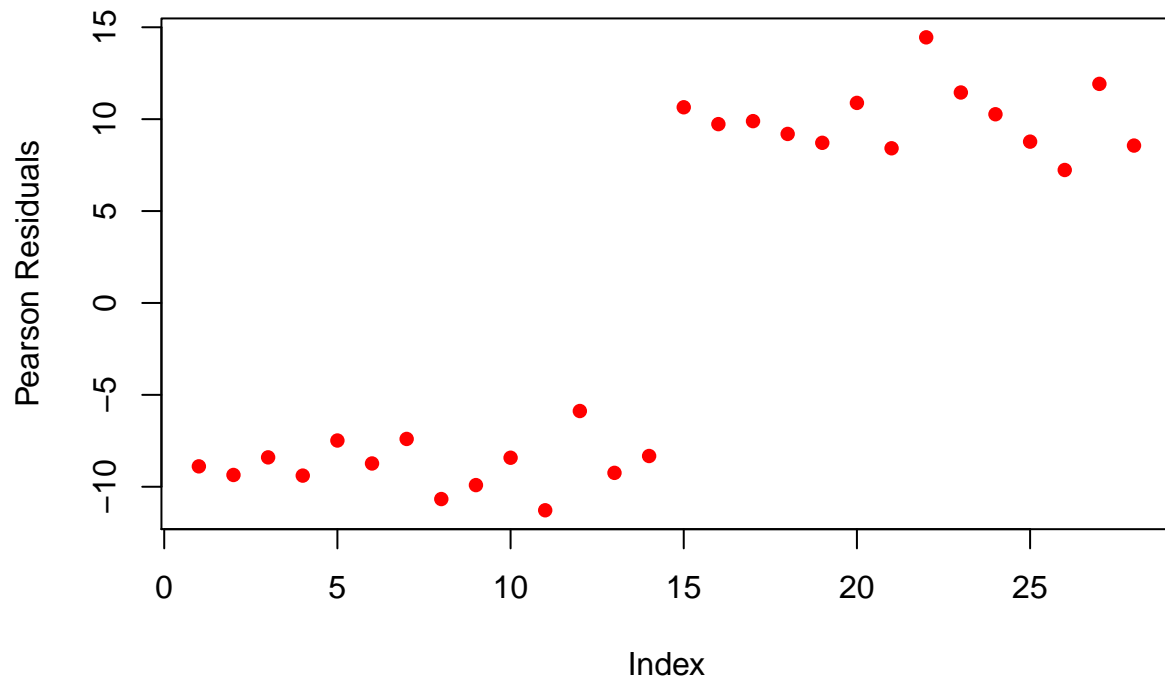
```
### Pearson Residuals v.s. observation
plot(residuals(glm2, type = "pearson"), main = "Figure 4. Pearson Residual Plot",
    col = "red", pch = 16, ylab = "Pearson Residuals")
```

Figure 4. Pearson Residual Plot
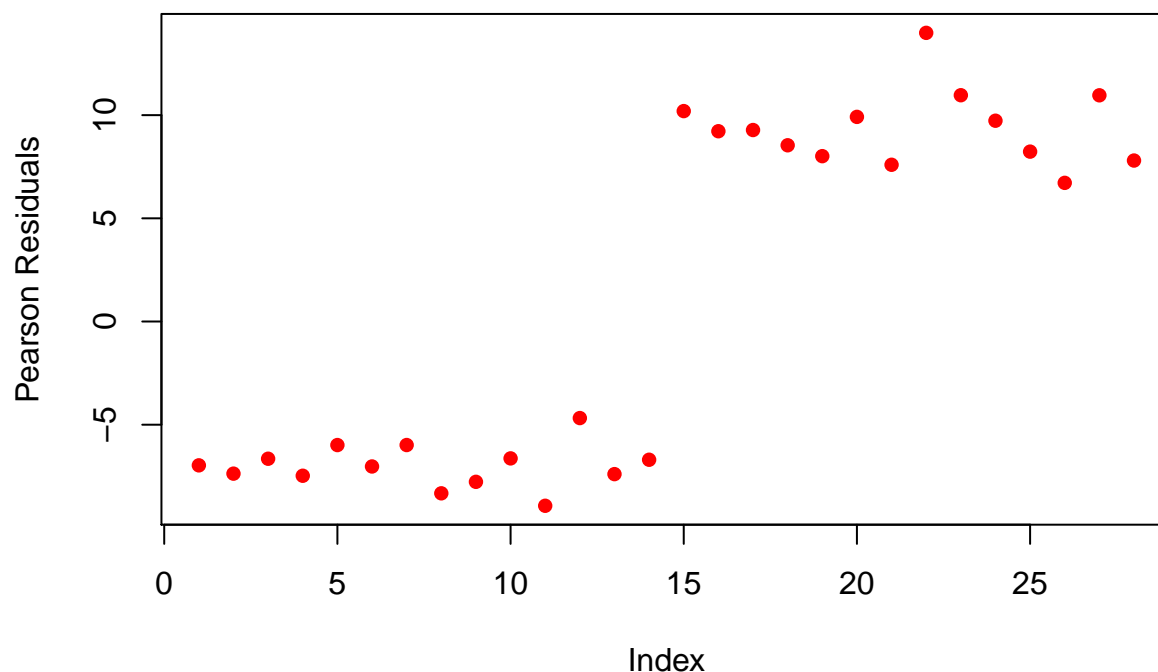

```
### Deviance Residuals v.s. observation
plot(residuals(glm2, type = "deviance"), main = "Figure 5. Deviance Residual Plot",
    col = "red", pch = 16, ylab = "Pearson Residuals")
```
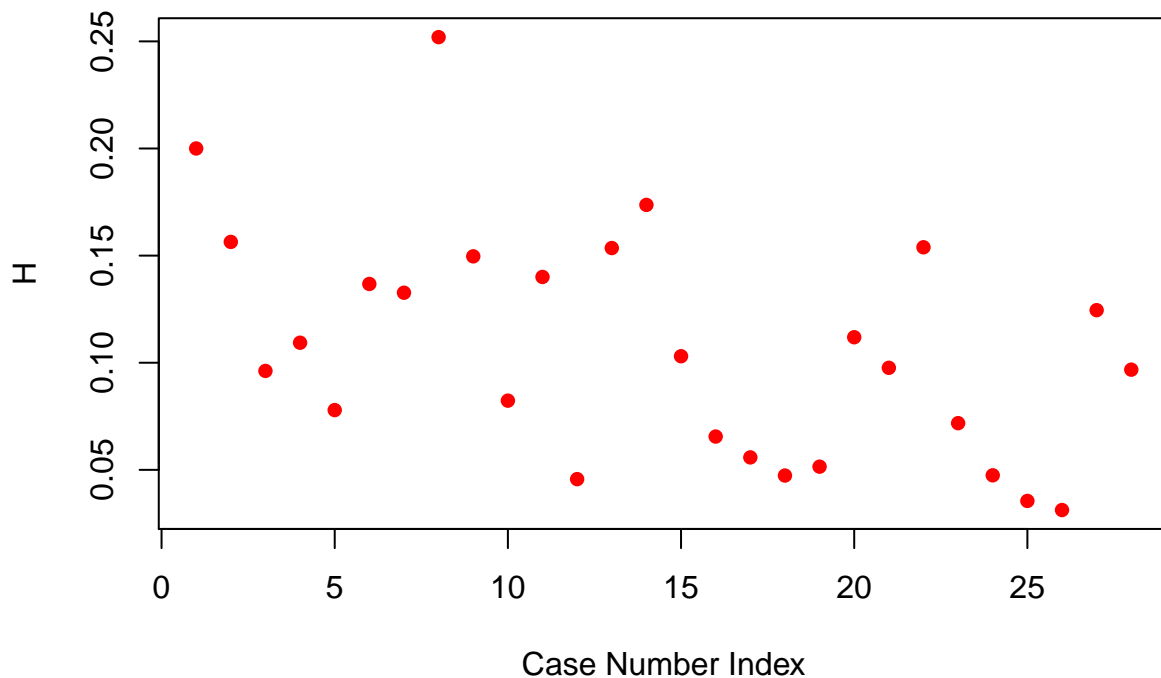
Figure 5. Deviance Residual Plot

The Pearson residual and Deviance residual plots are similar. No one case has an overwhelmingly larger residual than any other.

```
### Hat Diagonal Plot
plot(hatvalues(glm2), ylab = "H", xlab = "Case Number Index", col = "red", pch = 16,
    main = "Figure 6. Hat Diagonal Plot")
```
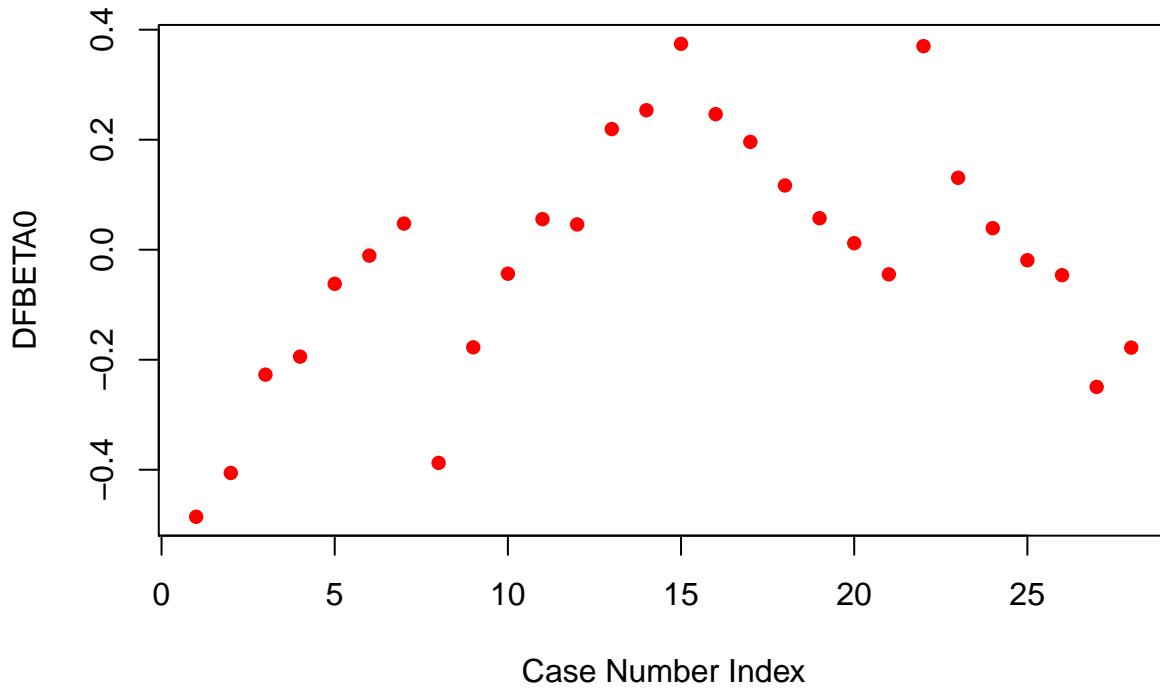
Figure 6. Hat Diagonal Plot

Our heuristic cutoff for leverage is $\frac{3p}{n} = \frac{9}{1971} = 0.005$ It apppears that many points are high leverage. This may be a result of the extremely large sample size.
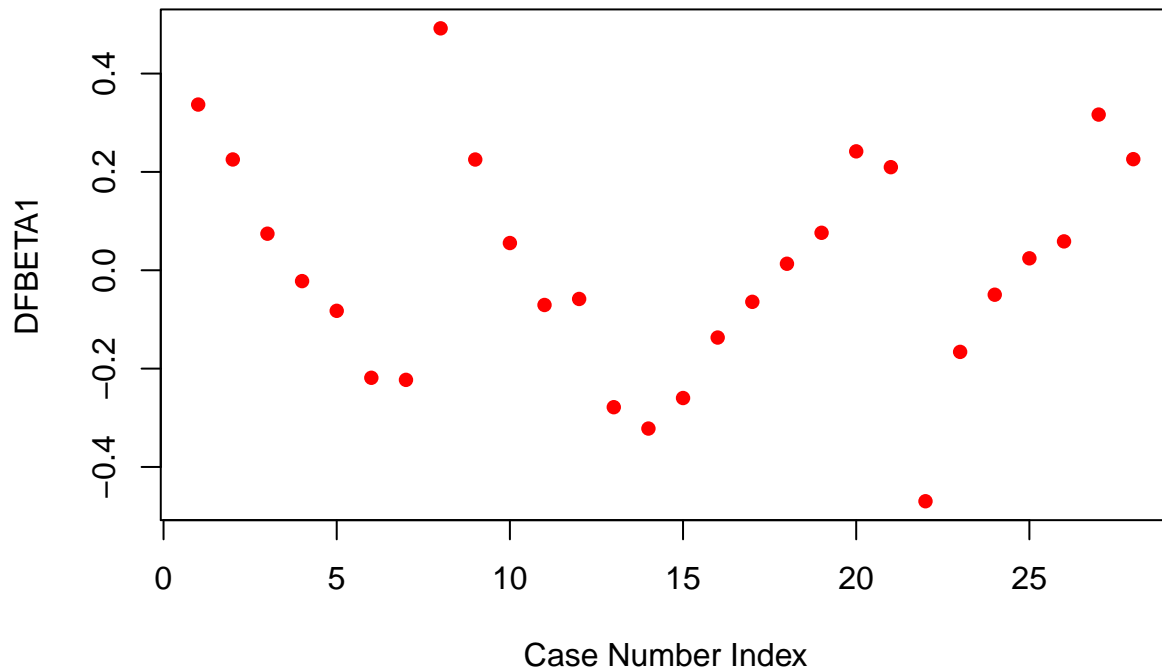
```
### Intercept DfBeta0 Plot
plot(dfbetas(glm2)[, 1], ylab = "DFBETA0", xlab = "Case Number Index", col = "red",
    pch = 16, main = "Figure 7. DFBeta0 Plot")
```

Figure 7. DFBeta0 Plot


```
### Intercept DfBeta1 Plot
plot(dfbetas(glm2)[, 2], ylab = "DFBETA1", xlab = "Case Number Index", col = "red",
    pch = 16, main = "Figure 8. DFBeta1 Plot")
```
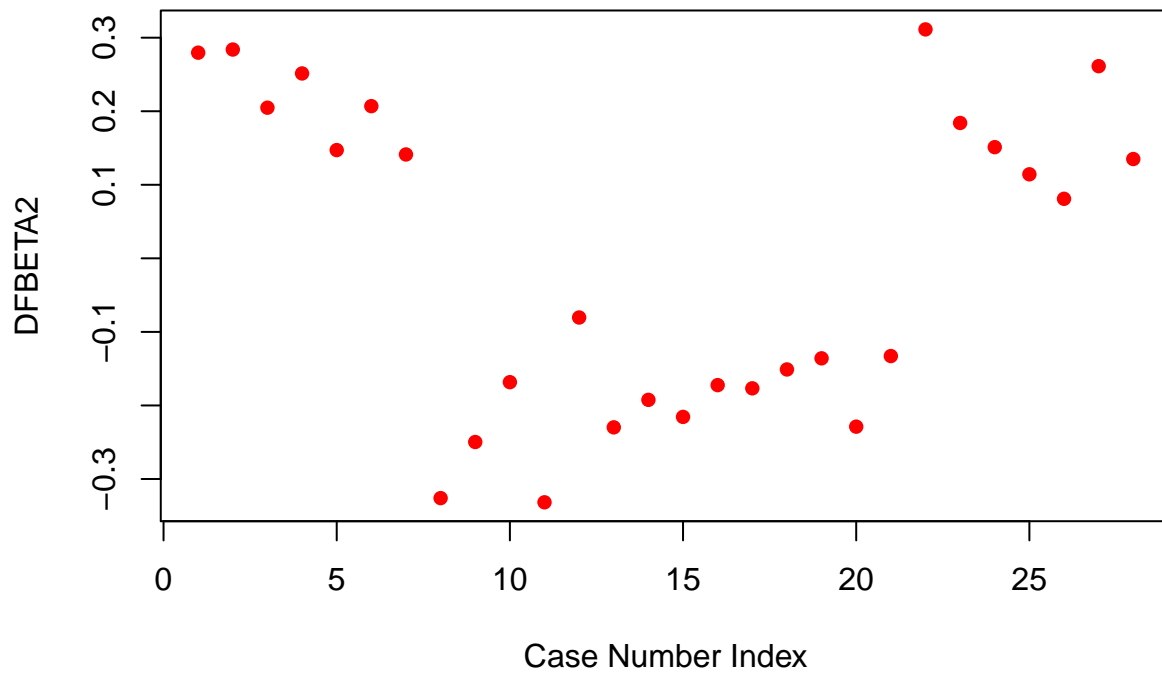
Figure 8. DFBeta1 Plot


```
### Intercept DfBeta2 Plot
plot(dfbetas(glm2)[, 3], ylab = "DFBETA2", xlab = "Case Number Index", col = "red",
    pch = 16, main = "Figure 9. DFBeta2 Plot")
```

Figure 9. DFBeta2 Plot

Our heuristic cutoff for the DFBETA statistic is $\frac{2}{\sqrt{n}} = \frac{2}{44.4} = 0.05$. It appears that no point has undue

influence on any of the three parameters.

Now that the inadequacies of our current model have been exposed, let's perform a model selection to find a model that fits the data well. Since our number of predictors is so small (2), the total number of possible models is very small (5) and we can consider all of them. We can build a table for important model diagnostics that can help us select the right model.

```
glm1 <- glm(Education ~ Party * Gender, weights = Freq, data = pd2, family = "binomial")
glm3a <- glm(Education ~ Party, weights = Freq, data = pd2, family = "binomial")
glm3b <- glm(Education ~ Gender, weights = Freq, data = pd2, family = "binomial")

fit.table <- matrix(NA, 5, 7, dimnames = list(Model = c("1", "2", "3a", "3b",
    "4"), c("Predictors", "G Squared", "df", "AIC", "Models Comp.", "Delta G Sq.",
    "p-val")))
fit.table[, 1] <- c("P+G+P*G", "P+G", "P", "G", "None")
k <- 1
for (model in c("glm1", "glm2", "glm3a", "glm3b", "glm4")) {
    fit.table[k, 2] <- round(get(model)[["deviance"]], 2)
    fit.table[k, 3] <- get(model)[["df.residual"]]
    fit.table[k, 4] <- round(get(model)[["aic"]], 2)
    k <- k + 1
}
fit.table[, 5] <- c("-", "(2)-(1)", "(3a)-(2)", "(3b)-(2)", "(4)-(3b)")
fit.table[1, 6] <- "-"
fit.table[2, 6] <- round(as.numeric(fit.table[2, 2]) - as.numeric(fit.table[1,
    2]), 2)
fit.table[3, 6] <- round(as.numeric(fit.table[3, 2]) - as.numeric(fit.table[2,
    2]), 2)
fit.table[4, 6] <- round(as.numeric(fit.table[4, 2]) - as.numeric(fit.table[2,
    2]), 2)
fit.table[5, 6] <- round(as.numeric(fit.table[5, 2]) - as.numeric(fit.table[4,
    2]), 2)
for (i in 2:5) {
    fit.table[i, 7] <- round(pchisq(as.numeric(fit.table[i, 6]), df = 1, lower.tail = FALSE),
        2)
}
fit.table
```

```
##
## Model Predictors G Squared df   AIC      Models Comp. Delta G Sq. p-val
##    1   "P+G+P*G"  "2555.52"  "24" "2563.52" "-"         "-"         NA
##    2   "P+G"      "2557.68"  "25" "2563.68" "(2)-(1)"   "2.16"      "0.14"
##    3a  "P"        "2558"     "26" "2562"    "(3a)-(2)"  "0.32"      "0.57"
##    3b  "G"        "2562.91"  "26" "2566.91" "(3b)-(2)"  "5.23"      "0.02"
##    4   "None"     "2563.38"  "27" "2565.38" "(4)-(3b)"  "0.47"      "0.49"
```

If we start with Model 1, we can see that it compares favorably with the others based on deviance. This isn't necessarily an interesting diagnostic since Model 1 has more parameters, it's bound to have less deviance. What is more telling is the AIC, which penalizes Model 1 for the interaction term - since it doesn't have the lowest AIC, it appears that the interaction term is not worth the reduction in deviance that it provides. In terms of AIC, the best model is one in which only Party is a predictor (Model 3a). This also makes sense in light of our investigation of Model 2 where the gender predictor was found to be insignificant. It does not have significantly higher deviance than Model 2 yet it's AIC is lower than both Model 1 and Model 2. Model 3a appears to have the best tradeoff between fitting the data and parsimony. It's also important to note that the deviance for Model 3a is still very large and is nowhere near its expected value (as a chi-squared variable, this would be its $df = n - p = 1979$), indicating poor fit as model.