Assignment 3

STAT34700 Generalized Linear Models

Seung Chul Lee

2/7/2022

Problem 1 (Agresti, 6.4)

The given model is

$$\pi_{ij} = \frac{e^{\beta_{j0}+\beta_j x_i}}{1 + e^{\beta_{10}+\beta_1 x_i} + e^{\beta_{20}+\beta_2 x_i}}$$

Note that $j = 3$ is the baseline category. The partial differentiation of π_{i3} w.r.t x_i is

$$\begin{aligned}
\frac{\partial \pi_{i3}}{\partial x_i} &= \frac{\beta_3 e^{\beta_{30}+\beta_3 x_i}(1 + e^{\beta_{10}+\beta_1 x_i} + e^{\beta_{20}+\beta_2 x_i}) - e^{\beta_{30}+\beta_3 x_i}(\beta_1 e^{\beta_{10}+\beta_1 x_i} + \beta_2 e^{\beta_{30}+\beta_3 x_i})}{(1 + e^{\beta_{10}+\beta_1 x_i} + e^{\beta_{20}+\beta_2 x_i})^2} && (1)\\
&= \pi_{i3}\left(\beta_3 - \frac{\beta_1 e^{\beta_{10}+\beta_1 x_i} + \beta_2 e^{\beta_{30}+\beta_3 x_i}}{1 + e^{\beta_{10}+\beta_1 x_i} + e^{\beta_{20}+\beta_2 x_i}}\right) && (2)\\
&= \pi_{i3}(\beta_3 - \beta_1 \pi_{i1} - \beta_2 \pi_{i2}) && (3)\\
&= \pi_{i3}(-\beta_1 \pi_{i1} - \beta_2 \pi_{i2}) && (4)
\end{aligned}$$

The last equality holds due to $\beta_3 = 0$, since $j = 3$ is the baseline category. Note that we have $\pi_{ij} > 0$.

a) If $\beta_1 > 0$ and $\beta_2 > 0$, then $\frac{\partial \pi_{i3}}{\partial x_i} < 0$. Hence, π_{i3} is decreasing in x_i.
b) If $\beta_1 < 0$ and $\beta_2 < 0$, then $\frac{\partial \pi_{i3}}{\partial x_i} > 0$. Hence, π_{i3} is increasing in x_i.
c) If β_1 and β_2 differ in signs, then we cannot determine the sign of the partial derivative and thus depends on the observation i. Hence, it is nonmonotone.

Problem 2 (Agresti, 6.10)

Our textbook suggests that it may not be sensible to force a cumulative logit model of proportional odds, and the default model to consider is the baseline-category logit model. This is because the cumulative logit model assumes a common β across different levels, which does not allow crossovers across individuals. Such restrictions may make intuitive sense for an ordinal variable, since there is a clear order in the response. However, there is little justification for forcing the same increment in log odds with a nominal response. There is also the problem of assigning an arbitrary order to a nominal variable to designate cutoffs that need to be defined for a cumulative logit model. The model will be sensitive to the choice of the arbitrary order that we select.

The cumulative logit model is not a special case of the baseline-category logit model. Under the baseline-category logit model, $\mathbb{P}(y_i = k) = p_{ik}$ simplifies to

$$p_{ik} = \frac{e^{X_i^T \beta_k}}{1 + \sum_h^{c-1} e^{X_i^T \beta_h}}$$

This gives us the desirable property that

$$\log \frac{p_{ik}}{p_{ij}} = X_i^T(\beta_k - \beta_j)$$

However, note that $\mathbb{P}(y_i = k) = p_{ik}$ under the cumulative logit model is

$$p_{ik} = \mathbb{P}(y_i \leq k) - \mathbb{P}(y_i \leq k-1) \tag{5}$$

$$= \frac{e^{\alpha_k + X_i^T\tilde{\beta}}}{1 + e^{\alpha_k + X_i^T\tilde{\beta}}} - \frac{e^{\alpha_{k-1} + X_i^T\tilde{\beta}}}{1 + e^{\alpha_{k-1} + X_i^T\tilde{\beta}}} \tag{6}$$

which cannot be transformed to the form specified in the baseline-category model.

Problem 3 (Agresti, 6.13)

(a)

With the given model,

$$\text{logit}[\mathbb{P}(y_i \leq j)] = \alpha_j + \beta_j x_i, \ \forall x_i \in \mathbb{R}$$

Then,

$$\text{logit}[\mathbb{P}(y_i \leq j+1)] - \text{logit}[\mathbb{P}(y_i \leq j)] = (\alpha_{j+1} - \alpha_j) + (\beta_{j+1} - \beta_j)x_i$$

Note that we should have $\text{logit}[\mathbb{P}(y_i \leq j+1)] - \text{logit}[\mathbb{P}(y_i \leq j)] \geq 0$, since $\mathbb{P}(y_i \leq j+1) \geq \mathbb{P}(y_i \leq j)$. However, it is possible to have values $x_i > \frac{\alpha_{j+1} - \alpha_j}{\beta_{j+1} - \beta_j}$ if $\beta_{j+1} \neq \beta_j$. Hence, it is possible to have a misordered cumulative probabilities under a complex cumulative logit model.

(b)

Suppose $x_i \in \{0, 1\}$. Then, the difference in logit of cumulative probabilities above is

$$\text{logit}[\mathbb{P}(y_i \leq j+1)] - \text{logit}[\mathbb{P}(y_i \leq j)] = \begin{cases} \alpha_{j+1} - \alpha_j \\ \alpha_{j+1} - \alpha_j + \beta_{j+1} - \beta_j \end{cases}$$

Or, more generally,

$$\text{logit}[\mathbb{P}(y_i \leq k)] - \text{logit}[\mathbb{P}(y_i \leq j)] = \begin{cases} \alpha_k - \alpha_j \\ \alpha_k - \alpha_j + \beta_k - \beta_j \end{cases}$$

for $k > j$. This model is free from the misorder problem in (a) with some constraints. We need $\text{logit}[\mathbb{P}(y_i \leq k)] - \text{logit}[\mathbb{P}(y_i \leq j)] \geq 0$ for both cases. The first cases is already satisfied since $\alpha_k > \alpha_j$ by the usual ordering constraint for the cutoffs. The second case also needs to be satisfied, which can be achieved by having $(\alpha_k + \beta_k) \geq (\alpha_j + \beta_j)$, $\forall k > j$. That is, the sequence $\{\alpha_j + \beta_j\}_{j=0}^c$ needs to be monotone increasing.

Fitting the above model will require an estimation of $2(c-1)$ many parameters, which is equal to the number of parameters required in the saturated model. Hence, this is equivalent to the saturated model.

Problem 4 (Agresti, 6.17)

Let y_i be defined as

$$y_i = \begin{cases} 1, \text{ if response is "strongly disagree"} \\ 2, \text{ if response is "mildly disagree"} \\ 3, \text{ if response is "mildly agree"} \\ 4, \text{ if response is "strongly agree"} \\ z, \text{ if response is "do not know"} \end{cases}$$

Then, $p_{iz} := \mathbb{P}(y_i = z)$ with a logit model will be

$$\text{logit}(p_{iz}) = X_i^T \beta_z \Rightarrow p_{iz} = \frac{e^{X_i^T \beta_z}}{1 + e^{X_i^T \beta_z}}$$

On the other hand, for the other levels, the cumulative logit model will be

$$\text{logit}(\mathbb{P}(y_i \leq k | y_i \neq z)) = \alpha_k + \tilde{X}_i^T \tilde{\beta} \Rightarrow \mathbb{P}(y_i \leq k | y_i \neq z) = \frac{e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}}$$

Then,

$$p_{ik|y_i \neq z} := \mathbb{P}(y_i \leq k | y_i \neq z) - \mathbb{P}(y_i \leq k-1 | y_i \neq z) = \frac{e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}} - \frac{e^{\alpha_{k-1} + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_{k-1} + \tilde{X}_i^T \tilde{\beta}}}$$

By definition of conditional probability,

$$p_{ik|y_i \neq z} = \frac{\mathbb{P}(y_i = k, y_i \neq z)}{\mathbb{P}(y_i \neq z)} = \frac{\mathbb{P}(y_i = k)}{\mathbb{P}(y_i \neq z)} = \frac{p_{ik}}{1 - p_{iz}}$$

$$\Rightarrow p_{ik} = p_{ik|y_i \neq z} \cdot (1 - p_{iz}) = \left(\frac{e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}} - \frac{e^{\alpha_{k-1} + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_{k-1} + \tilde{X}_i^T \tilde{\beta}}} \right) \frac{1}{1 + e^{X_i^T \beta_z}}$$

Hence, the likelihood function for fitting the two models simultaneously will be

$$L_i(y_i) = p_{iz}^{y_{iz}} \prod_{k=1}^{4} p_{ik}^{y_{ik}} \tag{7}$$

$$= \left(\frac{e^{X_i^T \beta_z}}{1 + e^{X_i^T \beta_z}} \right)^{y_{iz}} \prod_{k=1}^{4} \left\{ \left(\frac{e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_k + \tilde{X}_i^T \tilde{\beta}}} - \frac{e^{\alpha_{k-1} + \tilde{X}_i^T \tilde{\beta}}}{1 + e^{\alpha_{k-1} + \tilde{X}_i^T \tilde{\beta}}} \right) \frac{1}{1 + e^{X_i^T \beta_z}} \right\}^{y_{ik}} \tag{8}$$

Problem 5 (Agresti, 6.23)

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
# Create the table as a dataframe
Housing = c("Tower", "Apartments", "Atrium", "Terraced")
Influence = c("Low", "Medium", "High")
Contact = c("Low", "High")
sat = data.frame(expand.grid(Housing, Influence, Contact))
low = c(21, 61, 13, 18,
        34, 43, 8, 15,
        10, 26, 6, 7,
        14, 78, 20, 57,
        17, 48, 10, 31,
        3, 15, 7, 5)
med = c(21, 23, 9, 6,
```

```
        22, 35, 8, 13,
        11, 18, 7, 5,
        19, 46, 23, 23,
        23, 45, 22, 21,
        5, 25, 10, 6)
high = c(28, 17, 10, 7,
         36, 40, 12, 13,
         36, 54, 9, 11,
         37, 43, 20, 13,
         40, 86, 24, 13,
         23, 62, 21, 13)
satis = cbind(sat, low, med, high)
colnames(satis) = c("Housing", "Influence", "Contact", "Low", "Med", "High")
head(satis)
```

```
##      Housing Influence Contact Low Med High
## 1      Tower       Low     Low  21  21   28
## 2 Apartments       Low     Low  61  23   17
## 3     Atrium       Low     Low  13   9   10
## 4   Terraced       Low     Low  18   6    7
## 5      Tower    Medium     Low  34  22   36
## 6 Apartments    Medium     Low  43  35   40
```

```
mod1 = vglm(cbind(low, med, high) ~ factor(Housing) + factor(Influence) + factor(Contact), family = cumu
summary(mod1)
```

```
##
## Call:
## vglm(formula = cbind(low, med, high) ~ factor(Housing) + factor(Influence) +
##     factor(Contact), family = cumulative(parallel = T), data = satis)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept):1              -0.49614    0.12454  -3.984 6.78e-05 ***
## (Intercept):2               0.69071    0.12521   5.516 3.46e-08 ***
## factor(Housing)Apartments   0.57235    0.11875   4.820 1.44e-06 ***
## factor(Housing)Atrium       0.36619    0.15677   2.336 0.019498 *
## factor(Housing)Terraced     1.09101    0.15151   7.201 5.99e-13 ***
## factor(Influence)Medium    -0.56639    0.10496  -5.396 6.81e-08 ***
## factor(Influence)High      -1.28882    0.12670 -10.172  < 2e-16 ***
## factor(Contact)High        -0.36028    0.09536  -3.778 0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 47.7276 on 40 degrees of freedom
##
## Log-likelihood: -123.432 on 40 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
```

```
##
##
## Exponentiated coefficients:
## factor(Housing)Apartments     factor(Housing)Atrium   factor(Housing)Terraced
##                 1.7724273                 1.4422234                 2.9772934
##   factor(Influence)Medium     factor(Influence)High       factor(Contact)High
##                 0.5675685                 0.2755961                 0.6974781
```

I first fit an additive model that fits a constant β. It is necessary to keep in mind that the sign of coefficients should be interpreted in the opposite direction. That is, the negatively significant coefficient on `factor(influence):High` should be regarded as having a positive effect on satisfaction (i.e., more likely to have a higher level).

```
mod2 = vglm(cbind(low, med, high) ~ factor(Housing) + factor(Influence) + factor(Contact), family = cumu
anova(mod1, mod2, type = "I")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(low, med, high) ~ factor(Housing) + factor(Influence) +
##     factor(Contact)
## Model 2: cbind(low, med, high) ~ factor(Housing) + factor(Influence) +
##     factor(Contact)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        40     47.728
## 2        34     39.157  6   8.5706   0.1992
```

I then fit a model with a nonconstant $\{\beta_k\}$. As the textbook suggests, forcing a common value of β may result in a poor fit. The textbook also mentions that fitting a nonconstant $\{\beta_k\}$ is likely to have a better statistical significance but should be cautious since having a simple model is desirable when practical difference is little. Note that the deviance analysis tells us the difference between the two models is not even statistically significant. Hence, I refrain from carrying on with the more complex model.

```
mod3 = vglm(cbind(low, med, high) ~ factor(Housing) * factor(Influence) * factor(Contact), family = cumu
anova(mod1, mod3, type = "I")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(low, med, high) ~ factor(Housing) + factor(Influence) +
##     factor(Contact)
## Model 2: cbind(low, med, high) ~ factor(Housing) * factor(Influence) *
##     factor(Contact)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        40     47.728
## 2        23     15.037 17   32.691  0.01233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I then consider interactions. I first fit up to three-way interactions. Deviance analysis suggests that adding interactions has a significant effect.

```
anova(mod3, test = "LRT")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Model: 'cumulative', 'VGAMordinal', 'VGAMcategorical'
##
## Links: 'logitlink'
##
## Response: cbind(low, med, high)
##
##                                                Df Deviance Resid. Df
## factor(Housing)                                 3   55.537        41
## factor(Influence)                               2  106.489        39
## factor(Contact)                                 1   15.120        35
## factor(Housing):factor(Influence)               6   21.717        35
## factor(Housing):factor(Contact)                 3    7.925        32
## factor(Influence):factor(Contact)               2    0.112        31
## factor(Housing):factor(Influence):factor(Contact)  6    2.125        29
##                                                Resid. Dev  Pr(>Chi)
## factor(Housing)                                   103.056 5.274e-12 ***
## factor(Influence)                                 145.550 < 2.2e-16 ***
## factor(Contact)                                    40.339 0.0001009 ***
## factor(Housing):factor(Influence)                  38.879 0.0013622 **
## factor(Housing):factor(Contact)                    25.086 0.0475965 *
## factor(Influence):factor(Contact)                  17.273 0.9457255
## factor(Housing):factor(Influence):factor(Contact)     17.162 0.9078529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I test for significance of the interaction terms to simplify the model. The three-way interaction term appears to be insignificant. Also, the `Influence:Contact` interaction term, which happens to be the next from the last, is also insignificant. Thus, I remove the two from the model.

```
mod4 = vglm(cbind(low, med, high) ~ factor(Housing) + factor(Influence) + factor(Contact) + factor(Housi
anova(mod4, test = "LRT")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Model: 'cumulative', 'VGAMordinal', 'VGAMcategorical'
##
## Links: 'logitlink'
##
## Response: cbind(low, med, high)
##
##                                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## factor(Housing)                   3   55.910        43    103.638 4.391e-12
## factor(Influence)                 2  106.489        39    145.550 < 2.2e-16
## factor(Contact)                   1   15.120        35     40.339 0.0001009
## factor(Housing):factor(Influence)  6   21.788        37     39.061 0.0013226
## factor(Housing):factor(Contact)   3    7.945        34     25.218 0.0471616
##
## factor(Housing)                   ***
```

```
## factor(Influence)                 ***
## factor(Contact)                   ***
## factor(Housing):factor(Influence) **
## factor(Housing):factor(Contact)   *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod5 = vglm(cbind(low, med, high) ~ factor(Housing) + factor(Influence) + factor(Contact) + factor(Housi
anova(mod5, test = "LRT")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Model: 'cumulative', 'VGAMordinal', 'VGAMcategorical'
##
## Links: 'logitlink'
##
## Response: cbind(low, med, high)
##
##                                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## factor(Housing)                  3   55.910        43    103.638 4.391e-12
## factor(Influence)                2  106.489        39    145.550 < 2.2e-16
## factor(Contact)                  1   15.120        35     40.339 0.0001009
## factor(Housing):factor(Contact)  3    7.945        34     25.218 0.0471616
## factor(Housing):factor(Influence) 6   21.788        37     39.061 0.0013226
##
## factor(Housing)                 ***
## factor(Influence)               ***
## factor(Contact)                 ***
## factor(Housing):factor(Contact)   *
## factor(Housing):factor(Influence) **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I then check for the significance of the remaining two interaction terms. Both two-way interactions are significant at the 5% significance level, so I stop my backward selection process here.

Problem 6 (Agresti, 7.7)

Let $y_{ij} \overset{ind}{\sim} Poi(\mu_i)$. Then, by property of the Poisson distribution, $y_{i+} \overset{ind}{\sim} Poi(n_i\mu_i)$, where the underscore $+$ denotes a sum with regards to that index. Also, let $N = \sum_i \sum_j y_{ij}$, i.e. the total count. Then, conditioning on N gives

$$(y_{1+}, \cdots, y_{c+}) \Big| \left(\sum_i y_i = N \right) \sim Multinomial(N, \vec{p})$$

where $\vec{p} = (p_1, \cdots, p_c)$ and $p_i = \frac{n_i\mu_i}{\sum_i n_i\mu_i}$.

(a)

Suppose we have $n_1 = \cdots = n_c = n_0$, and our null hypothesis is $H_0 : \mu_1 = \cdots = \mu_c = \mu_0$. Then, under the null, $p_i = \frac{n_0\mu_0}{\sum_i n_0\mu_0} = \frac{1}{c}$.

Note that the Pearson chi-squared test statistic is defined as

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\text{var}[y_i]}$$

In the case of a Poisson response, the statistic can be simplified to

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

since $\text{var}[X] = E[X]$ for a Poisson random variable. The fitted values are replaced with expected values (under the null) to test the validity of a given null hypothesis.

Then,

$$X^2 = \sum_i \frac{(y_{i+} - \hat{\mu}_i)^2}{\hat{\mu}_i} = \sum_i \frac{(y_{i+} - N/c)^2}{N/c} \sim \chi^2_{c-1}$$

for large N, since $\hat{\mu}_i = \mathbb{E}[y_{i+}] = Np_i = \frac{N}{c}$.

If the null hypothesis is true, the y_{i+}s should not deviate too much from the expected value of N/c and will not be a large number. On the other hand, if the null is false, then the numerator will be large and thus yield a larger statistic overall. Hence, we can reject the null hypothesis if $X^2 > \chi^2_{1-\alpha,c-1}$, where α is an appropriate significance level.

(b)

Now suppose $n = \sum_i n_i$ and no constraint is placed on n_is. Then, under the null, $p_i = \frac{n_i \mu_0}{\sum_i n_i \mu_0} = \frac{n_i}{n}$. This will yield $\mathbb{E}[y_{i+}] = N\frac{n_i}{n}$. We can construct a test statistic by the same token as in part (a) as follows

$$X^2 = \frac{(y_{i+} - N\frac{n_i}{n})^2}{N\frac{n_i}{n}} \sim \chi^2_{c-1}$$

The rejection criterion based on this statistic follows the same logic as in part (a). That is, reject the null if $X^2 > \chi^2_{1-\alpha,c-1}$.

Problem 7 (Agresti, 7.15)

Suppose $A = 0, 1$. Given $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$, we can write the joint probability of A, B and C as follows:

$$\mathbb{P}(A = i,\ B = j,\ C = k) = \mathbb{P}(A = i)\ \mathbb{P}(B = j,\ C = k)$$

The corresponding loglinear model is

$$\log \mu_{ijk} = \beta_0 + \beta^A_i + \beta^B_j + \beta^C_k + \gamma^{BC}_{jk}$$

The logit model for the conditional distribution of A given n_{+jk} is

$$\begin{aligned}
\text{logit}\ \mathbb{P}(A = 1|B = j, C = k) &= \log \frac{\mathbb{P}(A = 1|B = j, C = k)}{\mathbb{P}(A = 0|B = j, C = k)} && (9)\\
&= \log \mu_{1jk} - \log \mu_{0jk} && (10)\\
&= (\beta_0 + \beta^A_1 + \beta^B_j + \beta^C_k + \gamma^{BC}_{jk}) - (\beta_0 + \beta^A_0 + \beta^B_j + \beta^C_k + \gamma^{BC}_{jk}) && (11)\\
&= \beta^A_1 - \beta^A_0 && (12)
\end{aligned}$$

Let $\delta = \beta^A_1 - \beta^A_0$. Then, the model is simply

$$\text{logit}\ \mathbb{P}(A = 1|B = j, C = k) = \delta$$

This makes intuitive sense, since B and C should play no role in explaining the odds of A.

Problem 8 (Agresti, 7.16)

The homogeneous association loglinear model is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}$$

Treating A as a response variable will yield a multinomial response with $i = 1, \cdots, r$ response levels or categories. Let $i = 1$ be the baseline category. Then, treating the other variables as predictors gives the following baseline-category logit model.

$$\begin{aligned}
\log \frac{\mathbb{P}(A = i | B = j, C = k)}{\mathbb{P}(A = 1 | B = j, C = k)} &= \log \mu_{ijk} - \log \mu_{1jk} && (13) \\
&= (\beta_i^A - \beta_1^A) + (\gamma_{ij}^{AB} - \gamma_{1j}^{AB}) + (\gamma_{ik}^{AC} - \gamma_{1k}^{AC}) && (14) \\
&= \delta_0 + \delta_j^B + \delta_k^C && (15)
\end{aligned}$$

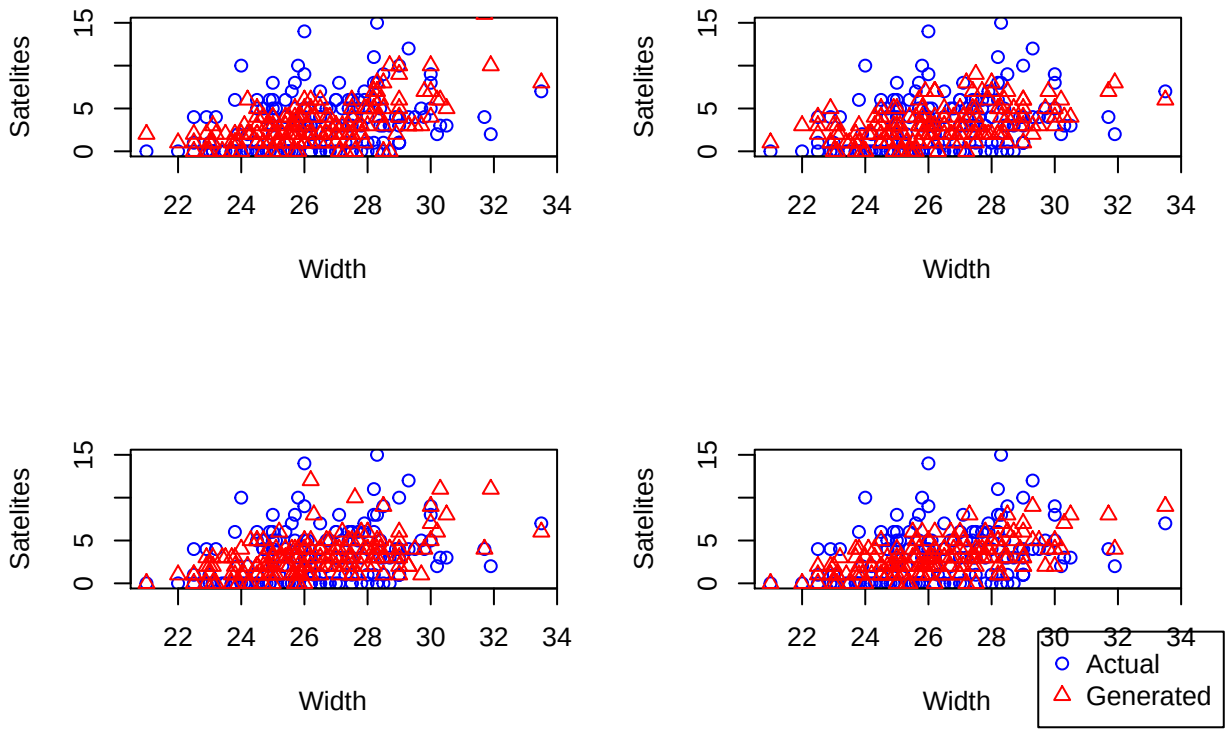
Note that the model is a simple additive model with an intercept and main effects of B and C.

Problem 9 (Agresti, 7.20)

Overdispersion relative to the Poisson distribution will occur when there are unobserved (or, rather, unused) variables that actually explain the response. The problem does not mention any predictors in particular. Hence, any factor that may contribute to the number of accidents over time will cause overdispersion. For instance, the amount of snowfall is likely to affect (increase) the average number of accidents. Moreover, the presence of police cars on the highway will also affect (decrease) the average number of accidents. The response will follow a Poisson distribution only when conditioning on each combination of variables like the ones suggested. If such covariates are omitted, the model will suffer from heterogeneity and thus have a larger variation than Poisson.

Problem 10 (Agresti, 7.28)

```
crabs = read.table("Crabs.dat", header = T)
glm_test = glm(y ~ width, family = poisson(link = "log"), data = crabs)
set.seed(19950328)
n = dim(crabs)[1]
legend = c("Actual", "Generated")
col = c("blue", "red")
pch = c(1, 2)
par(mfrow = c(2, 2))
for(i in 1:4){
  plot(crabs$width, crabs$y, xlab = "Width   ", ylab = "Satelites", col = "blue")
  points(crabs$width, rpois(n, fitted(glm_test)), pch = 2, col = "red")
}
legend("bottomright", inset = c(-0.1, -1), xpd = T, legend = legend, col = col, pch = pch)
```



The plot shows that the actual points have a larger variability than the randomly generated points. There are more actual points with unusually large satelite count values than there are with generated points. Also, the number of zeros seem to be more frequent in the actual data. I suspect there exists both an overdispersion and a zero-inflation problem with the given model.

Problem 11 (Agresti, 7.30)

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     crabs
```

```
year = c(2001:2013)
attacks = c(33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23)
shark = data.frame(cbind(year, attacks))
mod_poi = glm(attacks ~ 1, family = poisson, data = shark)
mod_nb = glm.nb(attacks ~ 1, data = shark)
# Poisson model
summary(mod_poi)
```

```
##
## Call:
## glm(formula = attacks ~ 1, family = poisson, data = shark)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.70106  -1.22008   0.09689   1.30274   2.05954
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.11522    0.05842   53.32   <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: 31.392  on 12  degrees of freedom
## Residual deviance: 31.392  on 12  degrees of freedom
## AIC: 97.129
##
## Number of Fisher Scoring iterations: 4
```

```
# Poisson log-likelihood
(l_poi = logLik(mod_poi))
```

```
## 'log Lik.' -47.56432 (df=1)
```

```
# Negative binomial model
summary(mod_nb)
```

```
##
## Call:
## glm.nb(formula = attacks ~ 1, data = shark, init.theta = 15.49441181,
##     link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.83716  -0.79986   0.06172   0.81039   1.26361
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.11522    0.09153   34.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15.4944) family taken to be 1)
##
##     Null deviance: 13.363  on 12  degrees of freedom
## Residual deviance: 13.363  on 12  degrees of freedom
## AIC: 92.608
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15.5
##           Std. Err.:  10.5
##
##  2 x log-likelihood:  -88.608
```

Note that the standard error in the negative binomial model is much larger than the Poisson model.

```
# Negative binomial log-likelihood
(l_nb = logLik(mod_nb))
```

```
## 'log Lik.' -44.30402 (df=2)
```

```
# LRT
cat("-2 log LR = ", -2 * as.numeric(l_poi - l_nb), " > ", qchisq(0.95, 1))
```

```
## -2 log LR =  6.520599  >  3.841459
```

The log-likelihood is also larger for the negative binomial model than the Poisson model. Using the degrees of freedom output of the `logLik` function, the likelihood ratio test yields that the Poisson null can be rejected at the 5% significance level. It seems that a negative binomial GLM is more appropriate for this data.

```
mod_nb_alt = glm.nb(attacks ~ year, data = shark)
summary(mod_nb_alt)
```

```
##
## Call:
## glm.nb(formula = attacks ~ year, data = shark, init.theta = 20.15274625,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0006  -1.0233   0.3683   0.7313   1.1484
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 64.26589   45.75078   1.405    0.160
## year        -0.03047    0.02280  -1.337    0.181
##
## (Dispersion parameter for Negative Binomial(20.1527) family taken to be 1)
##
##     Null deviance: 15.391  on 12  degrees of freedom
## Residual deviance: 13.522  on 11  degrees of freedom
## AIC: 92.879
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  20.2
##           Std. Err.:  15.7
##
##  2 x log-likelihood:  -86.879
```

Using the negative binomial model, I try adding `year` as a predictor to see whether there is an increase in the number of attacks as time passes. The coefficient is not significant and even negative in value. This does not support the claim that there is an increase in shark attacks in recent years.

```r
shark$ind = ifelse(shark$year %in% c(2012, 2013), 1, 0)
mod_nb_alt2 = glm.nb(attacks ~ ind, data = shark)
summary(mod_nb_alt2)
```

```
##
## Call:
## glm.nb(formula = attacks ~ ind, data = shark, init.theta = 15.84139055,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8069  -0.7596  -0.1643   0.8680   1.3265
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.09927    0.09918   31.25   <2e-16 ***
## ind          0.09940    0.24861    0.40    0.689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15.8414) family taken to be 1)
##
##     Null deviance: 13.532  on 12  degrees of freedom
## Residual deviance: 13.371  on 11  degrees of freedom
## AIC: 94.448
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15.8
##           Std. Err.:  10.9
##
##  2 x log-likelihood:  -88.448
```

I also try creating an indicator variable equal to 1 if the `year` is the two most recent years (i.e., 2012 and 2013) and 0 otherwise. Again, the coefficient estimate is not statistically significant. Hence, there is little reason to believe that there is a spike of shark attacks in two most recent years.