PHP2514: Applied Generalized Linear Models

Homework 3

Antonella Basso

Question 1:

Use the data from the British Doctors' study (Table 9.1 on page 201 in your textbook) to answer the following questions:

a) Fit Model 9.9 (page 202).

b) Reproduce Table 9.2 (page 203).

c) Based on this model, what is the effect of smoking on deaths from coronary heart disease?

d) Calculate the expected number of deaths in each age and smoking status category based on this model and plot them against the observed frequencies.

e) Assess the overall fit of the model checking for outliers, influential points, and form of the model covariates.

f) Fit another model that includes age as a categorical covariate. Use as the reference group "Ages: 35 to 44" and include in the model the interaction of age with smoking status.

- i. Perform a hypothesis test to check whether the interaction between age and smoking is significant.
- ii. Compare model in part (a) with model in part (f). Which one best fits the data? Explain. Use both quantitative and visual approaches to justify your answer.

In [42]:

```r
#DATA WRANGLING
#installing packages
suppressMessages(install.packages("tidyverse"))
suppressMessages(library(tidyverse))
suppressMessages(library(nnet)) #multinomial regression
suppressMessages(install.packages("VGAM")) #ordinal regression
```

In [2]:

```r
suppressMessages(library(VGAM))
```

Data:

Table 9.1: *Deaths from coronary heart disease after 10 years among British male doctors categorized by age and smoking status in 1951.*

Smokers:

Age Group	Deaths	Person-Years
35-44	32	52407
45-54	104	43248
55-64	206	28612
65-74	186	12663
75-84	102	5317

Non-Smokers:

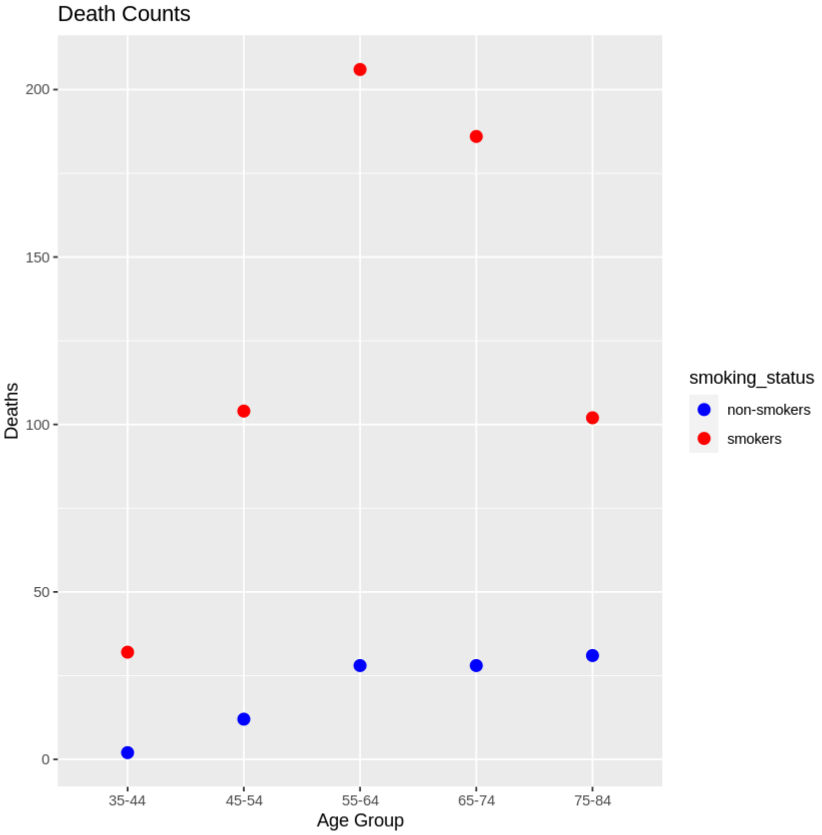
Age Group	Deaths	Person-Years
35-44	2	18790
45-54	12	10673
55-64	28	5710
65-74	28	2585
75-84	31	1462

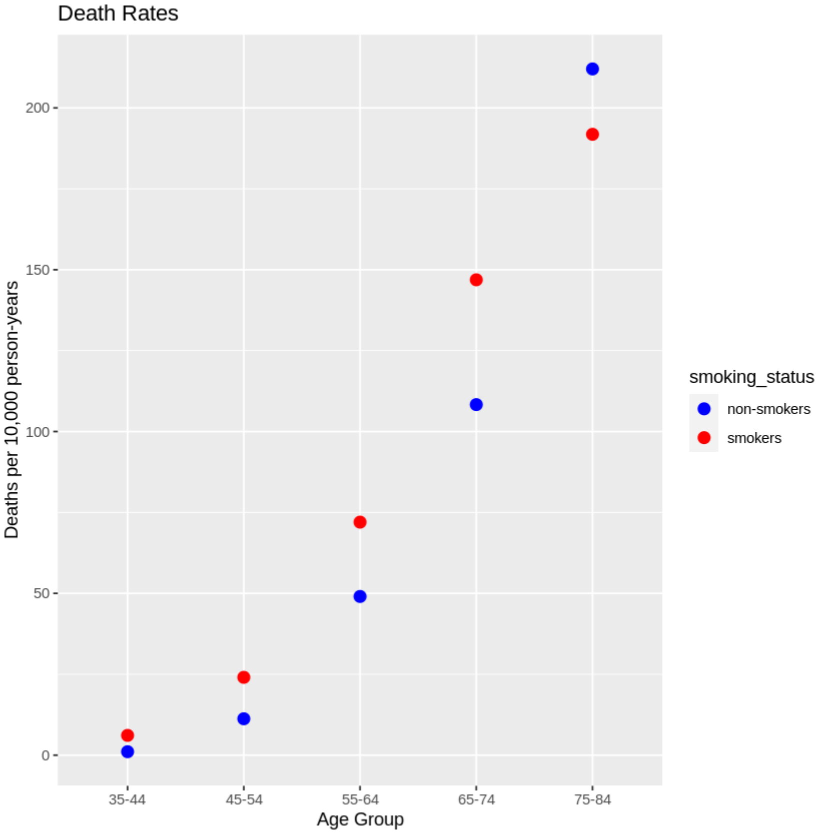
In [3]:

```
#Data: Table 9.1 (on page 201)

british_docs <- data.frame(age_group=rep(c(1, 2, 3, 4, 5), 2), #nominal age vari
                           age_group2=rep(c(1, 2, 3, 4, 5)^2, 2), #square of age
                           age=rep(c("35-44", "45-54", "55-64", "65-74", "75-84"
                           smoking_status=as.vector(cbind(rep(c("smokers"), 5),
                           deaths=c(32, 104, 206, 186, 102, 2, 12, 28, 28, 31),
                           person_years=c(52407, 43248, 28612, 12663, 5317, 1879
```

In [4]:

```
#Visualizing Death Counts and Death Rates

#counts
ggplot(british_docs, aes(x=age, y=deaths, color=smoking_status)) +
    geom_point(size=3) +
    scale_color_manual(values=c("blue", "red")) +
    labs(x="Age Group", y="Deaths", title="Death Counts")

#rates
ggplot(british_docs, aes(x=age, y=(deaths/person_years)*10000, color=smoking_sta
    geom_point(size=3) +
    scale_color_manual(values=c("blue", "red")) +
    labs(x="Age Group", y="Deaths per 10,000 person-years", title="Death Rates")
```





a) Model (9.9):

The outcome of interest, Y_i, indicates the number of deaths from coronary heart disease among male British doctors.

For $i = \{1, 2, ..., 10\}$, let i be the ith row in the `british_docs` data frame above, denoting the ith subgroup defined by age group and smoking status. Then, the Poisson regression model to model the differential death rate (for smokers and non-smokers) with respect to age is as follows:

$$ log(\text{deaths}_i) = log(\text{person_years}_i) + \beta_0 + \beta_1\text{smoking_status}_i + \beta_2\text{age_group}_i + \beta_3\text{age_group}2_i + \beta_4\text{age_group*smoking_status}_i $$
Where the non-smoker group is the reference group (denoted as 0), and $\text{age_group}2_i$ is the square of age_group_i and is used to account for the non-linearity of the rate of increase.

In [5]:

```
pr_glm <- glm(deaths ~ smoking_status + age_group + age_group2 + age_group*smoki
              family=poisson, data=british_docs)
#summary(pr_glm)
```

b) Table (9.2)

In [6]:

```
#Table 9.2

#confidence intervals
coeff_ci <- as.data.frame(round(exp(confint(pr_glm)), 2))
lower <- coeff_ci[, 1]
upper <- coeff_ci[, 2]

coeff_ci2 <- c()

for (i in 1:5){
    comb_ci <- paste(lower[i], upper[i], sep=", ")
    coeff_ci2 = c(coeff_ci2, comb_ci)
    }

#table
table92 <- as.data.frame(round(summary(pr_glm)$coefficients, 5)) %>%  #coefficie
               mutate(rate_ratio=round(exp(as.vector(coefficients(pr_glm))), 5)

names(table92) <- c("Coefficient", "Standard Error", "Wald Statistic", "p-value"

table92 <- as.data.frame(t(table92)) %>% select(!(starts_with("(Intercept)"))) #
table92
```

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

A data.frame: 6 × 4

	smoking_statussmokers	age_group	age_group2	smoking_statussmokers:age_group
	<chr>	<chr>	<chr>	<chr>
Coefficient	1.44097	2.37648	-0.19768	-0.30755
Standard Error	0.37220	0.20795	0.02737	0.09704
Wald Statistic	3.87151	11.42820	-7.22306	-3.16925

	smoking_statussmokers	age_group	age_group2	smoking_statussmokers:age_group
	<chr>	<chr>	<chr>	<chr>
p-value	0.00011	0.00000	0.00000	0.00153
Rate Ratio	4.22480	10.76692	0.82064	0.73525
95% Confidence Interval	2.09, 9.01	7.23, 16.34	0.78, 0.87	0.61, 0.89

In [7]:

```
#Table 9.3

fitted_values = fitted(pr_glm)
pearson_residuals <- residuals(pr_glm, type = "pearson")
deviance_residuals <- residuals(pr_glm, type = "deviance")

chisq <- sum(pearson_residuals^2)
residual_deviance <- deviance(pr_glm) #sum(deviance_residuals^2)

#data.frame(fitted_values, pearson_residuals, deviance_residuals)
cbind(british_docs[, c(3, 4, 5)], fitted_values, pearson_residuals, deviance_res

#chisq
#residual_deviance
```

A data.frame: 10 × 6

	age	smoking_status	deaths	fitted_values	pearson_residuals	deviance_residuals
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	35-44	smokers	32	29.584734	0.44404929	0.43820403
2	45-54	smokers	104	106.811960	-0.27208163	-0.27328873
3	55-64	smokers	206	208.198646	-0.15237591	-0.15264528
4	65-74	smokers	186	182.827893	0.23459923	0.23392570
5	75-84	smokers	102	102.576767	-0.05694769	-0.05700118
6	35-44	non-smokers	2	3.414801	-0.76561908	-0.83049031
7	45-54	non-smokers	12	11.541629	0.13492231	0.13404370
8	55-64	non-smokers	28	24.743377	0.65469354	0.64106682
9	65-74	non-smokers	28	30.229155	-0.40544060	-0.41058325
10	75-84	non-smokers	31	31.071038	-0.01274427	-0.01274913

c) Poisson Regression Model Interpretation

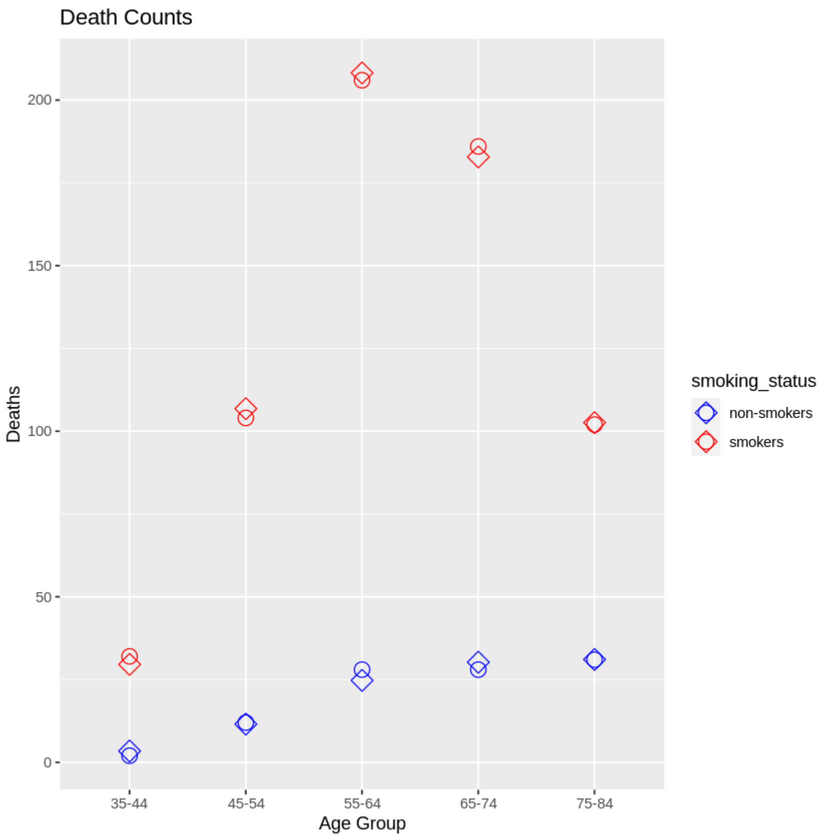
Based on the Poisson regression model for this data (`pr_glm`), which produced statistically significant estimates of coefficients, the risk of death from coronary heart disease is increased by a factor of 4 for smokers (with non-smokers as the reference group). That is, the exponentiated beta coefficient for smoking status, which yields a corresponding death rate ratio, tells us that the risk of a coronary heart disease-related death is approximately 4 times higher for those who smoke than for those who do not, irrespective of age. This risk is then presumably exacerbated with increase in age (given the interaction term).

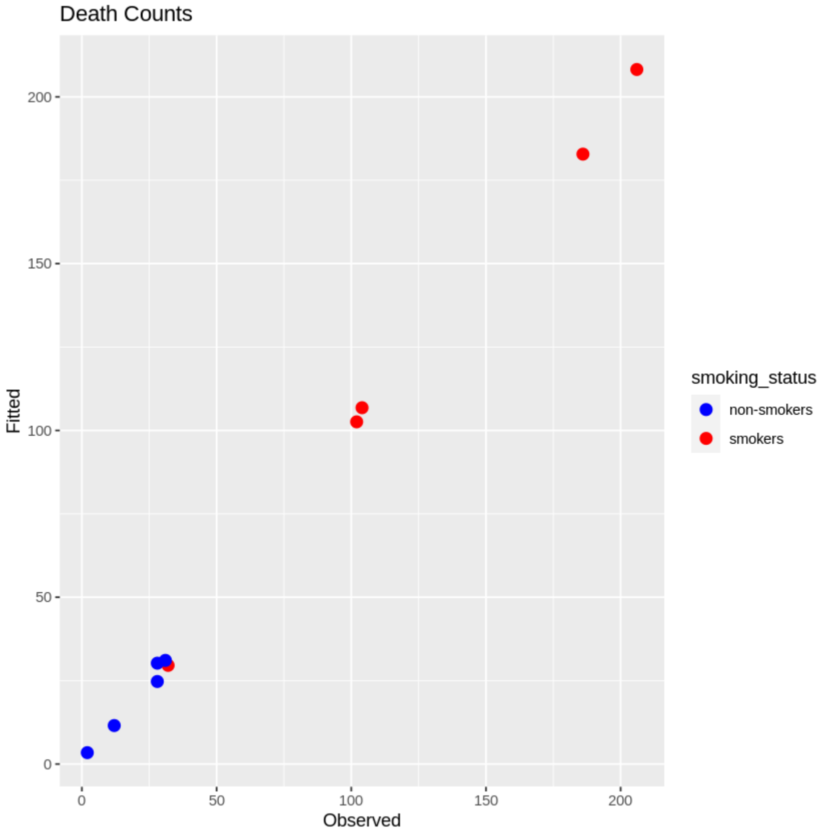
d) Fitted vs. Observed Number of Deaths

Plotting the expected number of deaths in each age and smoking status category based on the Pisson regression model (`pr_glm`) against their corresponding observed frequencies.

In [8]:

```
#Fitted vs. Observed Deaths Plot

#data frame
deaths <- cbind(british_docs[, c(3, 4, 5)], round(fitted_values, 2))
names(deaths) <- c("age", "smoking_status", "observed", "fitted")

#plot of fitted and observed against age group
ggplot(deaths) +
    geom_point(aes(x=age, y=observed, color=smoking_status), shape=1, size=4) +
    geom_point(aes(x=age, y=fitted, color=smoking_status), shape=5, size=4) +
    scale_color_manual(values=c("blue", "red")) +
    labs(x="Age Group", y="Deaths", title="Death Counts")

#plot of fitted against observed
ggplot(deaths, aes(x=observed, y=fitted, color=smoking_status)) +
    geom_point(size=3) +
    scale_color_manual(values=c("blue", "red")) +
    labs(x="Observed", y="Fitted", title="Death Counts")
```





e) Checking Model Fit & Model Assumptions:

Model Fit:

- **GoF/Deviance**: Given the χ^2-statistic of ≈ 1.55 and corresponding p-value of $\approx 0.91 > 0.05$, we may not reject the null hypothesis that assumes equality between our observed and fitted values. That is, under this assumption, we can be roughly 90% certain that there is no difference between the model's predictions and the observed outcomes. Thus, we conclude that this model fits the data well.

Model Assumptions:

- **Linearity** (functional form of model covariates): As residuals display no significant trend with respect to observed outcomes, (first plot below), we may assume that there is a linear relationship between covariates in the model and death rate.

- **Normality**: Since the Normal Q-Q Plot (second plot below) displays values roughly along the diagonal line (especially near the center), we deduce that residuals are approximately normally distributed.

- **Outliers**: As no standardized residuals exceed observations by 3 in absolute value (fourth plot below), there are no outliers (observations with a response far away from the regression plane) in the data.

- **Influential Points**: Given that the Cook's distance is less than 1 (fifth plot below), it follows that there are no influential/high-leverage points in the data (i.e. observations with a relatively large effect on estimates of model coefficients).

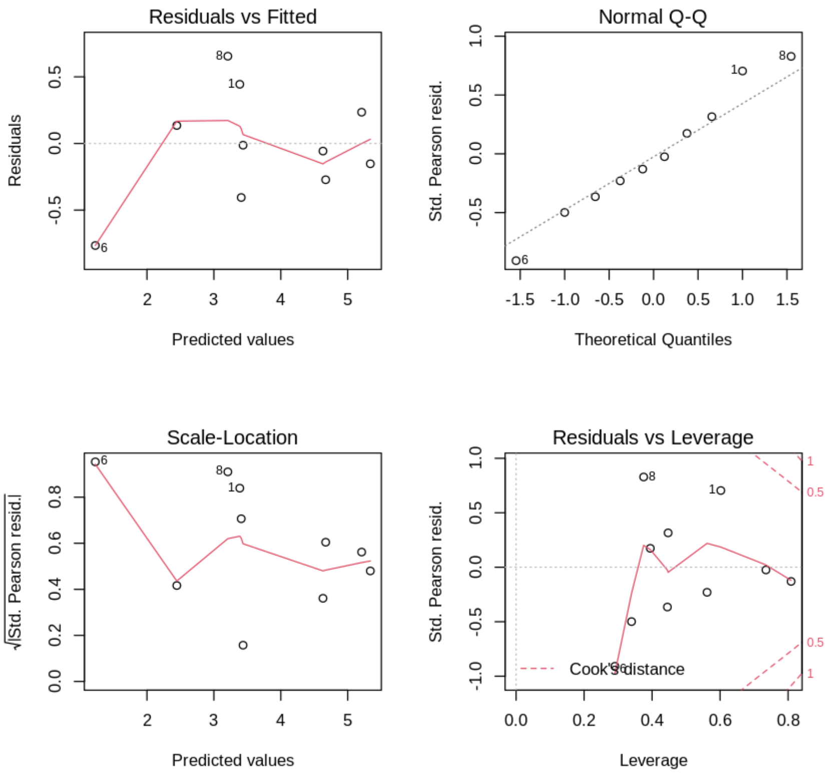
In [9]:

```
#MODEL FIT
#GoF: Chi Square Test
#p-value for chisq statistic with 5 degrees of freedom is very high
#we fail to reject the null hypothesis and assume that there is little differenc
#thus, the model is a good fit for the data

1 - pchisq(chisq, df.residual(pr_glm)) #same as pchisq(q=chisq, df=5, lower.tail
```

```
0.907199008588632
```

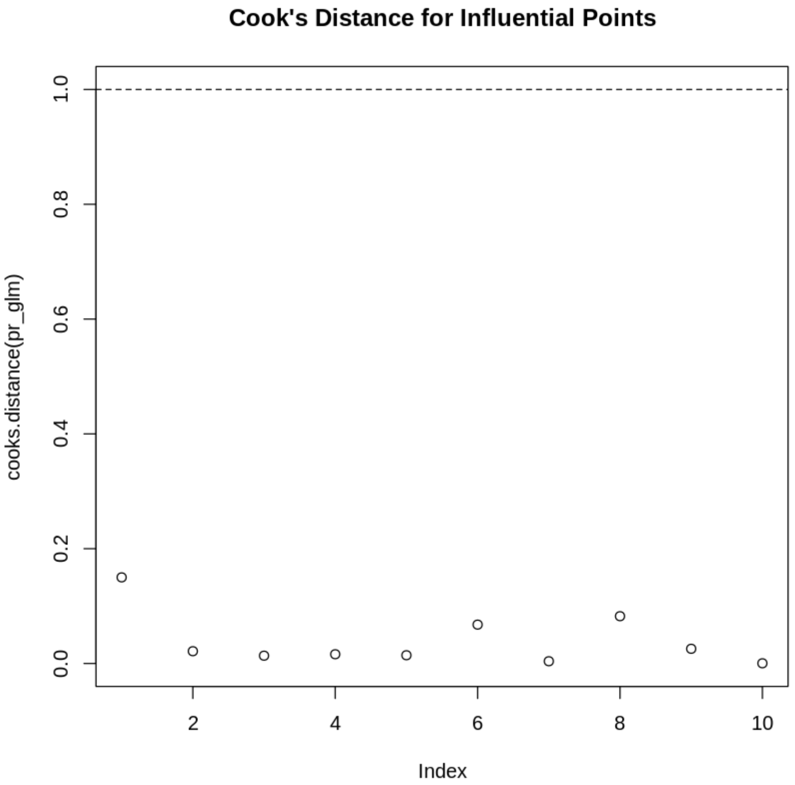
In [10]:

```
#MODEL ASSUMPTIONS:

#plots to assess linearity, normality, outliers
par(mfrow=c(2,2))
plot(pr_glm)
```



In [11]:

```
#cook's distance for influential points
plot(cooks.distance(pr_glm), ylim=c(0,1), main = "Cook's Distance for Influentia
abline(h = 1, lty = 2) #cutoff line at 1 (degress of freedom/number of observati
```



f) Model for Categorical Covariates

Fitting another model that includes age as a categorical covariate (with age 35-44 as the reference group) and the interaction of age with smoking status. Is the interaction between age and smoking significant?

- Referencing the ANOVA table below (between the models with and without the interaction term), it is evident that, given the corresponding p-value of $\approx 0.016 < 0.05$, the interaction term is statistically significant. This implies furthermore, that there exists a statistically significant association (although not very strong) between the two covariates. And, given the form of the model, this association is homogeneous. That is, all covariates in the model (in this case, smoking status and age group) are not independent. Also observing the coefficient estimate p-values for interactions above (from the `summary` table), we see that, although this relationship is not extremely significant, it is significant enough to infer a dependency. Specifically, since the most significant interaction occurs between the smoking and maximum age group, we may conjecture moreover that whether or not a person smokes is somewhat dependent on whether or not they belong to an older age group (and vise versa). Which model best fits the data?
- Not only is the AIC value for the model which regards age as a continuous explanatory variable (66.703) smaller than that for the model which treats it as a categorical explanatory variable (75.068), but when comparing their corresponding summary plots and looking at their densities (below), we see that the latter model (from part (f)) produces residuals that are far more spread out and less normally distributed than the model from part (a). Taking these visuals in tandem with the values produced (including the sums of their squared residuals), it is clear that the model which takes age as a continuous covariate fits the data

best. This could, in part, be due to the fact that the former model includes a squared term for age in addition to the terms in the latter model, as well as the fact that it accounts for the ordinal nature of this predictor, whereas the new model treats it merely as nominal (and hence, gives us less information about the data). However, we are unable to make definitive claims about the reason behind this difference in overall fits. What we can say with certainty is that the model from part (a) approximates observed values and hence fits the data far better than this new model from part (f).

In [12]:

```
#Models with Age as Categorical (age 35-44 = reference group)

#Log-linear model without offset for all categorical covariates
#not a significant difference from log-linear model (more of a terminology matte
ll_glm <- glm(deaths ~ smoking_status + age + age*smoking_status,
              family=poisson, data=british_docs)

#Poisson regression with offset (even with all categorical covariates), since n
#"saturated model" assumes homogeneous assosiation
pr_glm2 <- glm(deaths ~ smoking_status + age + age*smoking_status + offset(log(p
              family=poisson, data=british_docs) #(only interraction term needed
#summary(pr_glm2)

#Poisson regression without interaction term (only the main effects)
#assumes mutual independence of covariates
pr_glm3 <- glm(deaths ~ smoking_status + age + offset(log(person_years)),
              family=poisson, data=british_docs)
#summary(pr_glm3)
```

In [13]:

```
#Is there any association between smoking status and age group (categorical cova
#with a p-value of 0.016, there is a statistically significant assosiation
#thus, smoking status and age group are NOT independent and the interaction term

anova(pr_glm3, pr_glm2, test="LRT")
```

A anova: 2 × 5

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	4	1.213237e+01	NA	NA	NA
2	0	2.664535e-15	4	12.13237	0.01639363

In [14]:

```
#Model Comparisons:

#residual deviances - 1.64 vs. 12.13
anova(pr_glm, pr_glm3, test="LRT")

#pearson chi-square statistics - 1.56 vs. 11.16
deaths2 <- deaths %>% mutate(fitted2=fitted(pr_glm3),
                             pearson_residuals=residuals(pr_glm, type = "pearson
                             pearson_residuals2=residuals(pr_glm3, type = "pears

#deaths2
chisq <- sum(deaths2$pearson_residuals^2)
chisq2 <- sum(deaths2$pearson_residuals2^2)
```

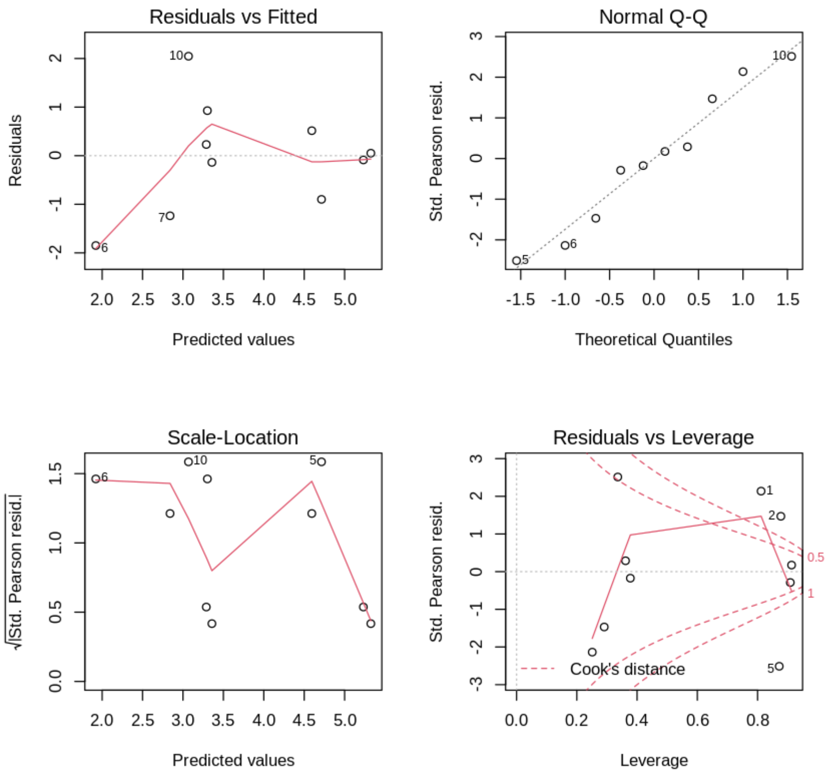
```
#chisq
#chisq2
```

A anova: 2 × 5

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	5	1.63537	NA	NA	NA
2	4	12.13237	1	-10.497	NA

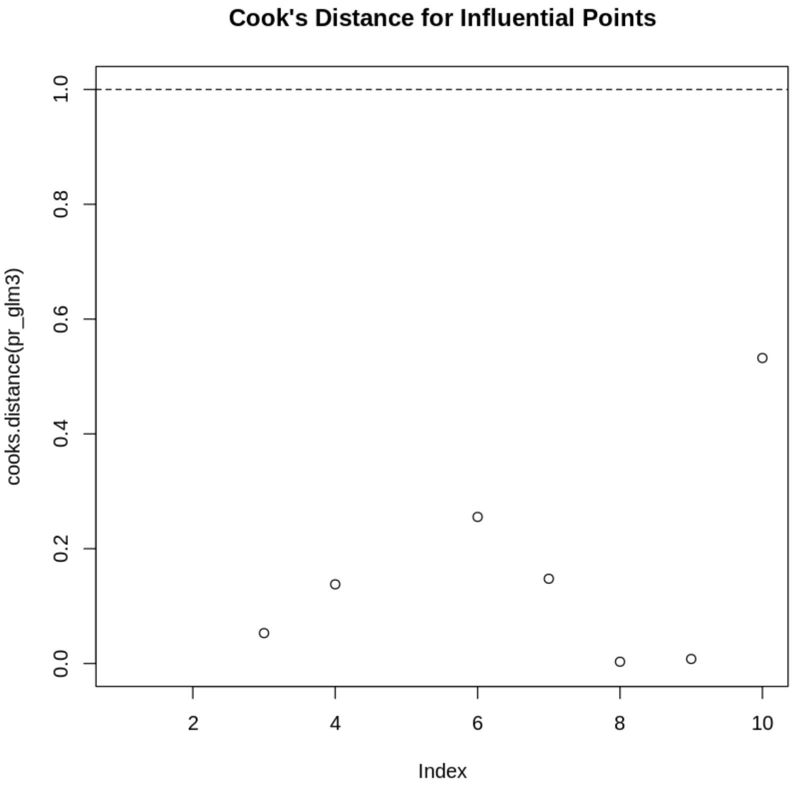
In [15]:

```
#Second Model - Assumptions:

#plots to assess linearity, normality, outliers
par(mfrow=c(2,2))
plot(pr_glm3)
```



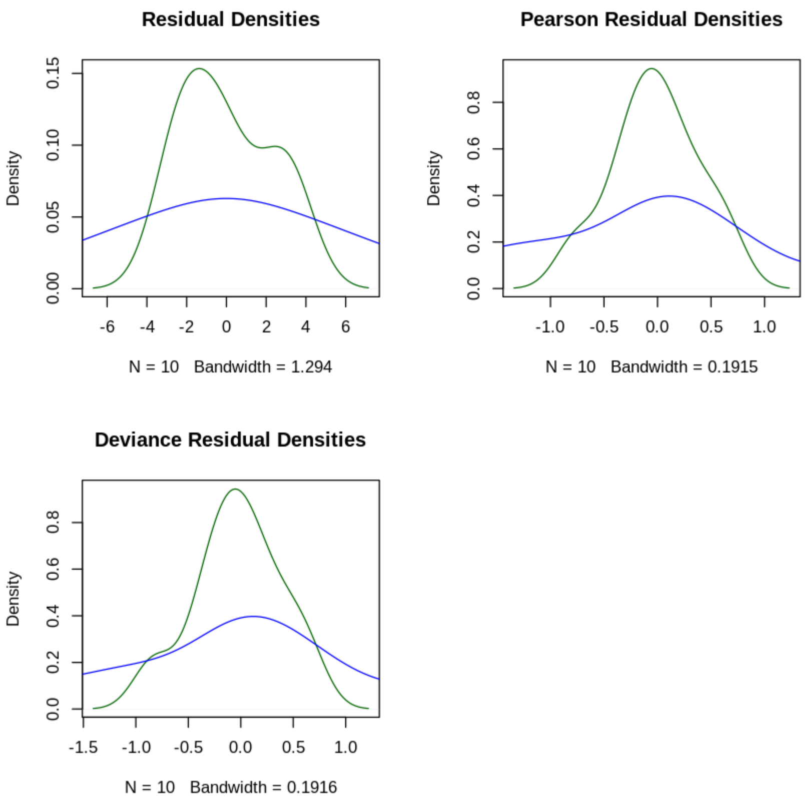
In [16]:

```
#cook's distance for influential points
plot(cooks.distance(pr_glm3), ylim=c(0,1), main = "Cook's Distance for Influenti
abline(h = 1, lty = 2) #cutoff line at 1 (degress of freedom/number of observati
```


Cook's Distance for Influential Points

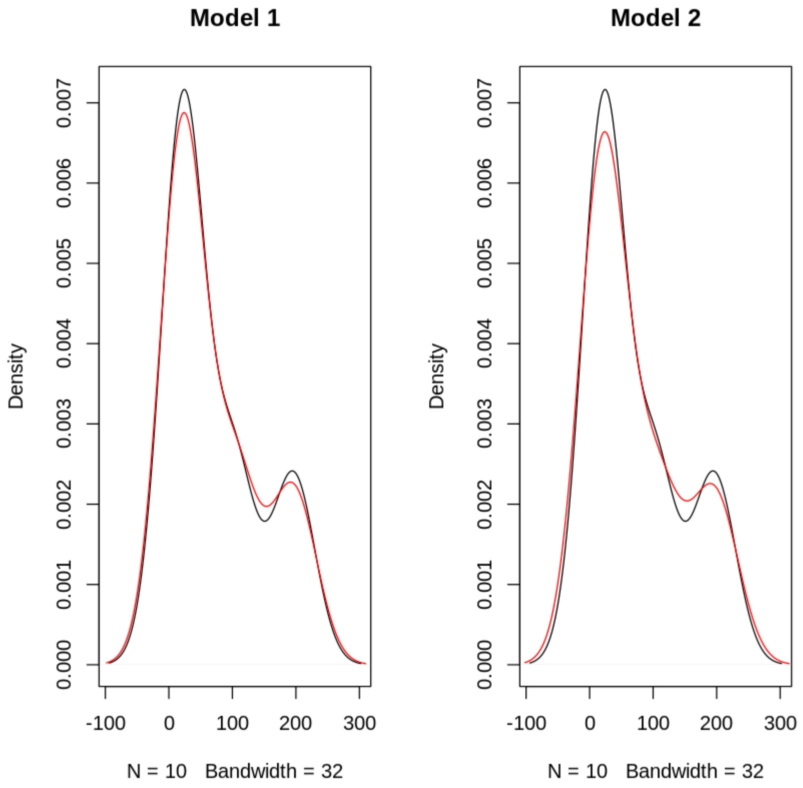
In [17]:

```
#Residual Densities

#model from part (a) in green
#model from part (f) in blue

par(mfrow=c(2,2))

plot(density(resid(pr_glm, type="response")), col="darkgreen", main="Residual De
    lines(density(resid(pr_glm3, type="response")), col="blue")

plot(density(resid(pr_glm, type="pearson")), col="darkgreen", main="Pearson Resi
    lines(density(resid(pr_glm3, type="pearson")), col="blue")

plot(density(resid(pr_glm, type="deviance")), col="darkgreen", main="Deviance Re
    lines(density(resid(pr_glm3, type="deviance")), col="blue")
```



In [18]:

```
#Observed vs. Fitted Densities

#observed in black
#fitted in red

par(mfrow=c(1,2))

plot(density(british_docs$deaths), main='Model 1')
lines(density(predict(pr_glm, type="response")), col='red')

plot(density(british_docs$deaths), main='Model 2')
lines(density(predict(pr_glm3, type='response')), col='red')
```



Question 2:

The dataset "vrt.csv" contains data collected from a hypothetical randomized control trial for evaluating an influenza vaccine. Patients were randomly allocated to two groups, one of which was given the new vaccine and the other a saline placebo. The number of patients with immune response after vaccination was recorded and the amount of immune response was classified as "small", "moderate", and "large".

a) Conduct a comprehensive Exploratory Data Analysis (EDA) to inspect, understand and describe the information collected in this dataset. Use appropriate summary statistics and plots to present your results from the EDA.

b) Suppose that you want to evaluate the association between immune response, vaccination status, and sex. Use the appropriate GLM to analyze the data and answer the following:

- i. Explain what GLM would you use to answer this research question and why.
- ii. Apply a model selection procedure (forward, backward or stepwise) to find the model that best fits the data. In this procedure also consider inclusion of interaction terms in the model.
- ii. State the form of the model that best fits the data. What does this model imply about the association among vaccine, immunity level, and sex?

c) You are also interested in the effect of vaccine type on the immunity level. Use an appropriate GLM to analyze the data and answer the following:

- i. What GLM would you use to answer this question?

- ii. Using the type of GLM you determined, implement a model selection procedure to find the model that best fits the data. In this procedure also consider inclusion of interaction terms in the model. Use as reference groups the "placebo", "female", and "small" immunity level.
- ii. State the form of the model that best fits the data. What is the effect of vaccines on the immunity level?

In [19]:

```r
#DATA WRANGLING

#importing "vrt" data
vrt <- read.csv("/home/jovyan/AGLM/HW3/vrt.csv")

#renaming values
names(vrt)[names(vrt) == "antibody.level"] <- "antibody_level"

vrt$vaccine[vrt$vaccine == "Placebo"] <- "P"
vrt$vaccine[vrt$vaccine == "Yes"] <- "V"

vrt$sex[vrt$sex == "Male"] <- "M"
vrt$sex[vrt$sex == "Female"] <- "F"

vrt$antibody_level[vrt$antibody_level == "small"] <- "S"
vrt$antibody_level[vrt$antibody_level == "moderate"] <- "M"
vrt$antibody_level[vrt$antibody_level == "large"] <- "L"

#vrt
```

a) Exploratory Data Analysis (EDA)

The variables in this dataset are as follows:

- Outcome (Y: number of subjects in each antibody level group (S, M, L))
- Covariate (X_1: vaccine type (placebo, vaccine), X_2: sex (M, F))

The primary outcome of interest is a categorical random variable with ordinal scale, while the predictor variables are categorical and binary with nominal scale.

This EDA consists of:

- Descriptive Statistics
- Boxplots

Descriptive Statistics:

- Count of subjects in the study.
- Count of subjects in each antibody level group.
- Summary of subject count in each antibody level group (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data, by vaccine type, and sex.

In [20]:

```r
#total number of subjects in study
#sum(vrt$freq)

#subject count in each antibody level group
```

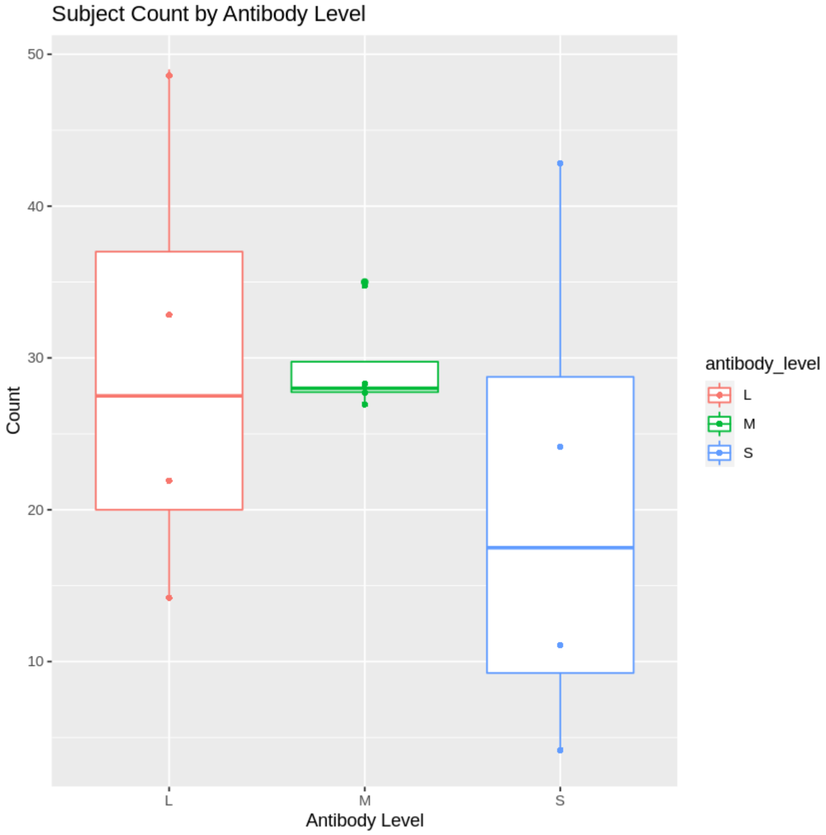
```
#by(vrt$freq, vrt$antibody_level, sum, na.rm=TRUE)

#summary of subject count in each antibody level group
#by(vrt$freq, vrt$antibody_level, summary, na.rm=TRUE)

#SD and of subject count in each antibody level group
#sd(vrt[vrt$antibody_level == "S",]$freq)
#sd(vrt[vrt$antibody_level == "M",]$freq)
#sd(vrt[vrt$antibody_level == "L",]$freq)

#summary of subject count in each antibody level group and vaccine group
#summary(vrt[vrt$antibody_level == "S" & vrt$vaccine == "V",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "M" & vrt$vaccine == "V",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "L" & vrt$vaccine == "V",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "S" & vrt$vaccine == "P",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "M" & vrt$vaccine == "P",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "L" & vrt$vaccine == "P",]$freq, na.rm=TRUE)

#summary of subject count in each antibody level group and sex
#summary(vrt[vrt$antibody_level == "S" & vrt$sex == "M",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "M" & vrt$sex == "M",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "L" & vrt$sex == "M",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "S" & vrt$sex == "F",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "M" & vrt$sex == "F",]$freq, na.rm=TRUE)
#summary(vrt[vrt$antibody_level == "L" & vrt$sex == "F",]$freq, na.rm=TRUE)
```

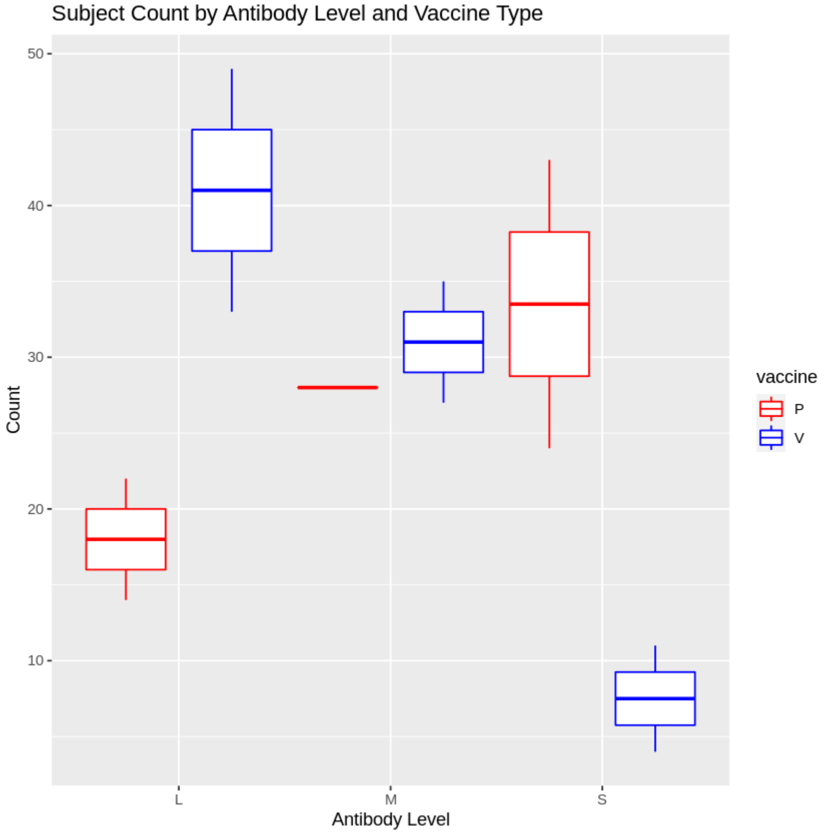
Boxplots:

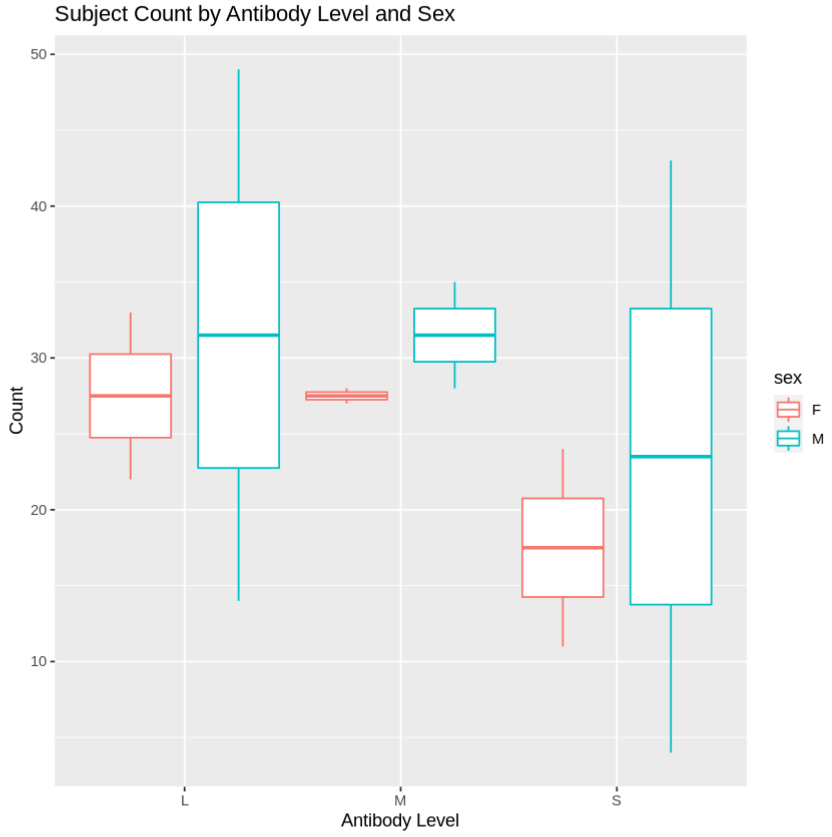
- The first boxplot shows a side by side comparison of the mean and spread of subject count for each immunity/antibody level group.
- The second boxplot shows side by side comparisons of the mean and spread of subject count for each immunity/antibody level group and vaccine type.
- The third boxplot shows side by side comparisons of the mean and spread of subject count for each immunity/antibody level group and sex.

In [21]:

```
#Boxplots

#spread of subject count by antibody level
ggplot(vrt, aes(x=antibody_level, y=freq, color=antibody_level)) +
    geom_boxplot() +
    geom_jitter(shape=16, position=position_jitter(0)) +
    labs(x = "Antibody Level", y = "Count", title = "Subject Count by Antibody L

#spread of subject count by antibody level and vaccine type
ggplot(vrt, aes(x=antibody_level, y=freq, color=vaccine)) +
    geom_boxplot(position=position_dodge(1)) +
    scale_color_manual(values=c("red", "blue")) +
    labs(x = "Antibody Level", y = "Count", title = "Subject Count by Antibody L

#spread of subject count by antibody level and sex
ggplot(vrt, aes(x=antibody_level, y=freq, color=sex)) +
    geom_boxplot(position=position_dodge(1)) +
    labs(x = "Antibody Level", y = "Count", title = "Subject Count by Antibody L
```


Subject Count by Antibody Level
Count
50
40
30
20
10
L
M
S
Antibody Level
antibody_level
L
M
S


Subject Count by Antibody Level and Vaccine Type
Count
50
40
30
20
10
L
M
S
Antibody Level
vaccine
P
V



b) GLM to Evaluate Variable Association

Research Question: Is there any association between immune response, vaccination status, and sex?

GLM: Log-Linear Model for Contingency Tables

- To answer this research question, it is best to employ a Log-Linear model to the data. Given the fact that we are solely interested in testing for association between variables, and not the effect of a subset of them on another, this type of model is preferable to a multinomial regression model.

Best Model: Joint Independence (Although, NO ASSUMPTION HOLDS)

While it is evident, based on the model selection procedure and model comparisons (using LR tests and AIC scores), that neither model fits the data well enough to assume any particular association between variables, the model which best describes the data (purely based on parsimony and closeness to the saturated model in fit) is that of the following form:

Let $X^A_{:i=1, 2, 3}$ be the antibody level random variable of three groups with $i=1, 2, 3$ representing the "small", "moderate", and "large" groups, respectively: $X^B_{:j=1, 2}$ be the vaccine type random variable of two groups with $j=1, 2$ representing the "placebo" and "vaccine" groups, respectively; and $X^C_{:k=1, 2}$ be the sex random variable with $k=1, 2$ representing "female" and "male", respectively. Moreover, let $i=3$, $j=2$, and $k=1$ be the reference groups for antibody level, vaccine type, and sex. Then,

$$ log(E[Y_{ij}]) = \beta_0 + \beta^A_i + \beta^B_j + \beta^C_k + \beta^{AB}_{ij} $$

This model not only implies that sex is jointly independent from immunity/antibody level and vaccine type, but that immunity/antibody level and vaccine are strongly associated.

In [22]:

```
#LOG-LINEAR MODEL
#Model Selection: Backward Elimination

#saturated model
ll_glm1 <- glm(freq ~ antibody_level + vaccine + sex + #main effects
               antibody_level*vaccine + antibody_level*sex + vaccine*sex + #two-
               antibody_level*vaccine*sex, #three-way interaction
               family=poisson, data=vrt)

#model with two-way interaction terms <- assumes homogeneous assosiation
ll_glm2 <- glm(freq ~ antibody_level + vaccine + sex + #main effects
               antibody_level*vaccine + antibody_level*sex + vaccine*sex, #two-w
               family=poisson, data=vrt)

#model without least significant two-way interaction term <- assumes conditional
ll_glm3 <- glm(freq ~ antibody_level + vaccine + sex + #main effects
               antibody_level*vaccine + vaccine*sex, #two-way interactions
               family=poisson, data=vrt)

#model with 1 two-way interaction term <- assumes joint independence (best model
#sex is jointly independent of vaccine and antibody level
#second smallest AIC score aside from ll_glm5 and saturated model
ll_glm4 <- glm(freq ~ antibody_level + vaccine + sex + #main effects
               antibody_level*vaccine, #two-way interaction
               family=poisson, data=vrt)

#model without least significant term (sex)
#best model (if we were not only focused on association between all three variab
#smallest AIC score aside from saturated model
ll_glm5 <- glm(freq ~ antibody_level + vaccine + antibody_level*vaccine, family=

#additive model (main effects) <- assumes mutual independence
ll_glm6 <- glm(freq ~ antibody_level + vaccine + sex, family=poisson, data=vrt)

#null model
ll_glm0 <- glm(freq ~ 1, family=poisson, data=vrt)

#summary(ll_glm1)
#summary(ll_glm2)
#summary(ll_glm3)
#summary(ll_glm4)
#summary(ll_glm5)
#summary(ll_glm6)
#summary(ll_glm0)
```

In [23]:

```
#LRT: comparing models

#additive model better than null model
#anova(ll_glm0, ll_glm6, test="LRT")

#joint independence model better than additive model
#anova(ll_glm6, ll_glm4, test="LRT")

#joint independence model equal to conditional independence model
#anova(ll_glm4, ll_glm3, test="LRT")
```

```
#joint independence model equal to homogeneous association model
#anova(ll_glm4, ll_glm2, test="LRT")

#conditional independence model equal to homogeneous association model
#anova(ll_glm3, ll_glm2, test="LRT")

#saturated model better than joint independence model
#saturated model much better than conditional independence model
#saturated model significantly better than homogeneous association model
#anova(ll_glm4, ll_glm1, test="LRT")
#anova(ll_glm3, ll_glm1, test="LRT")
#anova(ll_glm2, ll_glm1, test="LRT")
```

c) GLM to Evaluate Effects

Research Question: Does vaccine type have an effect on immunity/antibody level?

GLM: Ordinal Regression Models

- To answer this research question, it is best to employ an ordinal regression model given the nature of our response variable. Specifically, after determining the optimal cumulative logit models with and without the proportional odds assumption and comparing their residual deviances as well as log-likelihood values (since they have the same number of predictors and are non-nested), we see that the proportional odds assumption may not hold, and so, the more general/flexible cumulative logit model may provide a slightly better fit to the data.

Best Model: Cumulative Logit

Based on the model selection procedure and model comparisons, it is evident that the model which best fits the data is that of the following form:

For immunity/antibody level groups "small", "moderate", and "large", let X^V be the vaccine type random variable of two groups with "placebo" as the reference group (0), and let X^S be the sex random variable of two groups with "female" as the reference group (0):

$$ \text{Model }1: Y = logit(P(\text{small})) = log(\frac{P(\text{small})}{P(\text{moderate or large})}) = X\beta_1 = \beta_{01} + \beta_{11}X^V_v + \beta_{21}X^S_m + \beta_{31}X^V_v X^S_m $$$$ = -0.7340 - 0.9625X^V_v + 0.7575X^S_m - 2.1056X^V_v X^S_m $$$$ \text{Model }2: Y = logit(P(\text{small or moderate})) = log(\frac{P(\text{small or moderate})}{P(\text{large})}) = X\beta_2 = \beta_{02} + \beta_{12}X^V_v + \beta_{22}X^S_m + \beta_{32}X^V_v X^S_m $$$$ = 0.8602 - 0.7191^V_v + 0.7634^S_m - 1.1328X^V_v X^S_m $$
Exponentiated Coefficients:

```
   vaccineV:1          sexM:1      vaccineV:sexM:1
    0.3819444       2.1329365            0.1217759


   vaccineV:2          sexM:2      vaccineV:sexM:2
    0.4871795       2.1456044            0.3221434
```

Given the exponentiated beta coefficients above, it follows that the odds of having a small or moderate immunity/antibody level is approximately two times higher for males than females irrespective of vaccine type, although getting the vaccine does reduce those odds. Moreover, according to the model, getting the vaccine cuts the odds of having a small or moderate immunity/antibody level in half irrespective of sex. But interestingly, adding the factor of being male, makes the same odds one-third times smaller. Although this contradicts our previous intuition that being male increases one's odds of being in the small or moderate immunity groups, it may just be the case that without the introduction of a vaccine, males tend to have (on average) lower antibody levels than females. Yet, evidently, the converse is true if we introduce a vaccine. However, it is not clear whether this observation is purely accidental/circumstantial. What can be said with much more certainty is that vaccine does have a positive effect on immunity level irrespective of its less straightforward interactions with sex. Specifically, aside from the apparent minute difference between small and moderate immunity groups, it is evident that there exists a significant difference in odds between small and large immunity groups.

In [24]:

```
#specifying reference groups and ordering response

#vaccine type
vrt$vaccine <- factor(vrt$vaccine, levels=c("P", "V")) %>% #changing ordered fac
    relevel(vrt$vaccine, ref="P") #reference group = placebo

#sex
vrt$sex <- factor(vrt$sex, levels=c("F", "M")) %>% #changing ordered factor to c
    relevel(vrt$sex, ref="F") #reference group = female

#response: antibody level
vrt$antibody_level_ord <- ordered(vrt$antibody_level, levels=c("S","M","L"))
levels(vrt$antibody_level_ord) #reference group = small (always in the numerator
```

'S' · 'M' · 'L'

In [25]:

```
#PROPORTIONAL ODDS
#model with intercept 1 -> Y=log(P(small)/P(moderate+large))
#model with intercept 2 -> Y=log(P(small+moderate)/P(large))
#different intercepts, same beta coeficcients

#null model
po_glm0 <- vglm(antibody_level_ord ~ 1, family=cumulative(parallel=TRUE), data=v

#model with main effects
po_glm1 <- vglm(antibody_level_ord ~ vaccine + sex, family=cumulative(parallel=T

#model with interraction term -> saturated model
po_glm2 <- vglm(antibody_level_ord ~ vaccine + sex + vaccine*sex, family=cumulat

#po_glm0
#po_glm1
#po_glm2

#Model Selection:
#not many options (3 models to choose from)
#interaction term is statistically significant, we keep it and choose the larges
```

```
#po_glm2 -> smallest residual deviance (~630)

#summary(po_glm2)

#lrtest(po_glm0, po_glm1) #po_glm1 is a better fit than po_glm0
#lrtest(po_glm1, po_glm2) #po_glm2 is a better fit than po_glm1
```

In [26]:

```
#CUMULATIVE LOGIT
#model 1 -> Y=log(P(small)/P(moderate+large))
#model 2 -> Y=log(P(small+moderate)/P(large))
#different intercepts and beta coeficcients

#null model
cl_glm0 <- vglm(antibody_level_ord ~ 1, family=cumulative, data=vrt, weights=fre

#model with main effects
cl_glm1 <- vglm(antibody_level_ord ~ vaccine + sex, family=cumulative, data=vrt,

#model with interraction term -> saturated model
cl_glm2 <- vglm(antibody_level_ord ~ vaccine + sex + vaccine*sex, family=cumulat

#cl_glm0
#cl_glm1
#cl_glm2

#Model Selection:
#not many options (3 models to choose from)
#interaction term is statistically significant, we keep it and choose the larges
#cl_glm2 -> smallest residual deviance (~624)

#summary(cl_glm2)

#lrtest(cl_glm0, cl_glm1) #cl_glm1 is a better fit than cl_glm0
#lrtest(cl_glm1, cl_glm2) #cl_glm2 is a better fit than cl_glm1
```

In [27]:

```
#PROPORTIONAL ODDS or CUMULATIVE LOGIT? (po_glm2 or cl_glm2)

#cl_glm2 has slightly greater log-likelihood than po_glm2
#comparing log-likelihood values instead of LRT:
#both models have the same number of predictors
#they are non-nested
logLik(cl_glm2)
logLik(po_glm2)

#cl_glm2 has a slightly smaller residual deviance than po_glm2
deviance(cl_glm2) #624
deviance(po_glm2) #630

#therefore, the cumulative logit model fits the data slightly better
```

```
-311.793251465997
-315.21877783709
623.586502931993
630.43755567418
```

Question 3:

The dataset “tumor.csv” includes information on tumor responses of patients receiving treatment for small-cell lung cancer by sex. There were two treatment regimes. For the sequential treatment, the same combination of chemotherapeutic agents was administered at each treatment cycle. For the alternating treatment, different combinations were alternated from cycle to cycle (data from Holtbrugger and Schumacher, 1991).

The primary objective is to evaluate the relative effectiveness of the two treatments.

a) Use an appropriate type of GLM to assess the association between treatment, response, and sex.

- i. Perform a model selection procedure to find the model that best fits the data.
- ii. Comment on the results. Is there a strong association between these three variables?

b) Use an appropriate GLM to estimate the relative effectiveness of the two treatments adjusting for sex.

- i. Is the unadjusted effect of treatment the same for any level of response?
- ii. Use an appropriate test to answer this question. [bonus question]
- iii. Does your conclusion from the first part of this problem change after adjusting for sex?
- vi. Based on your conclusions from the previous two parts of this problem, perform a model selection procedure to find the model that best fits the data.
- v. Write the form and interpret the regression coefficients of the “best” model.
- vi. Assess the overall fit of the “best” model using regression diagnostics to identify problems indicating model inadequacy.
- vii. What is your final conclusion about the relative efficacy of the two treatments based on results from the “best” model?

In [28]:

```
#DATA WRANGLING

#importing "vrt" data
tumor <- read.csv("/home/jovyan/AGLM/HW3/tumor.csv")

#renaming values
tumor$treatment[tumor$treatment == "sequential"] <- "S"
tumor$treatment[tumor$treatment == "alternating"] <- "A"

tumor$sex[tumor$sex == "male"] <- "M"
tumor$sex[tumor$sex == "female"] <- "F"

tumor$response[tumor$response == "progressive"] <- "P"
tumor$response[tumor$response == "no change"] <- "NC"
tumor$response[tumor$response == "partial remission"] <- "PR"
tumor$response[tumor$response == "complete remission"] <- "CR"
```

a) GLM to Evaluate Variable Association

Research Question: Is there any association between treatment, response, and sex?

GLM: Log-Linear Model for Contingency Tables

- To answer this research question, it is best to employ a Log-Linear model to the data, given the fact that we are solely interested in testing for association between variables, and not relative effectiveness of treatments.

Best Model: Joint Independence

Based on the model selection procedure and model comparisons (using LR tests and AIC scores), it is evident that the model which best fits the data is that of the following form:

Let $X^A_{:i=1, 2, 3, 4}$ be the response random variable of four groups with $i=1, 2, 3, 4$ representing the "progressive", "no change", "partial remission", and "complete remission" groups, respectively: $X^B_{:j=1, 2}$ be the treatment random variable of two groups with $j=1, 2$ representing the "sequential" and "alternating" groups, respectively; and $X^C_{:k=1, 2}$ be the sex random variable with $k=1, 2$ representing "female" and "male", respectively. Moreover, let $i=4$, $j=2$, and $k=1$ be the reference groups for antibody level, vaccine type, and sex. Then,

$$ log(E[Y_{ij}]) = \beta_0 + \beta^A_i + \beta^B_j + \beta^C_k + \beta^{AB}_{ij} $$
This model not only implies that sex is jointly independent from response and treatment, but it reflects an association between response and treatment. Given the results from the model selection procedure however, it is possible for response and sex to also have some kind of association (implying that treatment and sex could be conditionally independent given response), yet this relationship did not prove to be significant at the 0.05 level (thus, forcing us to choose the joint independence model over the others).

In [29]:

```
#LOG-LINEAR MODEL
#Model Selection: Backward Elimination

#saturated model
ll2_glm1 <- glm(frequency ~ response + treatment + sex + #main effects
                response*treatment + response*sex + treatment*sex + #two-way inte
                response*treatment*sex, #three-way interaction
                family=poisson, data=tumor)

#model with two-way interaction terms <- assumes homogeneous assosiation
ll2_glm2 <- glm(frequency ~ response + treatment + sex + #main effects
                response*treatment + response*sex + treatment*sex, #two-way inter
                family=poisson, data=tumor)

#model without least significant two-way interaction term <- assumes conditional
#treatment and sex are conditionally independent given response
ll2_glm3 <- glm(frequency ~ response + treatment + sex + #main effects
                response*treatment + response*sex, #two-way interactions
                family=poisson, data=tumor)

#model with 1 two-way interaction term <- assumes joint independence (best model
#sex is jointly independent of response and treatment
ll2_glm4 <- glm(frequency ~ response + treatment + sex + #main effects
                response*treatment, #two-way interaction
                family=poisson, data=tumor)

#additive model (main effects) <- assumes mutual independence
ll2_glm5 <- glm(frequency ~ response + treatment + sex, family=poisson, data=tum
```

```r
#null model
ll2_glm0 <- glm(frequency ~ 1, family=poisson, data=tumor)

#summary(ll2_glm1)
#summary(ll2_glm2)
#summary(ll2_glm3)
#summary(ll2_glm4)
#summary(ll2_glm5)
#summary(ll2_glm0)

#ll2_glm3 (95.405) and ll2_glm4 (94.635) have the smallest AIC values (very clos
#based on LRT, ll2_glm4 is better
```

In [30]:

```r
#LRT: comparing models

#additive model better than null model
#anova(ll2_glm0, ll2_glm5, test="LRT")

#joint independence model better than additive model
#anova(ll2_glm5, ll2_glm4, test="LRT")

#joint independence model better than conditional independence model
#both equally good (the former is more parsimoneous)
#anova(ll2_glm4, ll2_glm3, test="LRT")

#joint independence model better than homogeneous association model
#anova(ll2_glm4, ll2_glm2, test="LRT")

#joint independence model better than saturated model
anova(ll2_glm4, ll2_glm1, test="LRT")
```

```
                  A anova: 2 × 5

     Resid. Df     Resid. Dev     Df   Deviance    Pr(>Chi)
         <dbl>          <dbl>  <dbl>      <dbl>       <dbl>
1            7   7.799115e+00     NA         NA          NA
2            0  -8.659742e-15      7   7.799115   0.3506408
```

b) GLM to Evaluate Relative Effectiveness

Research Question: What can be said about the relative effectiveness of the two treatments with and without adjusting for sex?

To assess whether the effect of treatment remains the same for any level of response, it serves us well to determine whether the data assumes proportionality of odds (satisfies the proportional odds assumption). To do this, we compare both the cumulative logit with and without the proportional odds assumption. In doing this, we see that both models fit the data equally well (based on log-likelihood values and residual deviances), indicating that the proportional odds assumption holds, and we may use it to both generate a more parsimonious model and provide us with information about the relationship between response levels for each treatment. Specifically, this assumption implies that the effects of both treatments are proportional and hence, remain (approximately) constant for any level of response. This intuition follows from the fact that the proportional odds approach keeps the slopes (or predictor coefficients) in the linear models for each response level the same (and hence, producing parallel fitted lines). Thus, we conclude that the effect of treatment remains the same for any level of response.

UNADJUSTED Effect of Treatment:

Exponentiated Treatment Coefficient:

```
        treatmentA
           1.78729
```

Given the exponentiated beta coefficients of the unadjusted proportional odds model, it follows that the odds of not achieving complete remission (or displaying a progressive, non-changing, or partial remission response) is roughly 1.8 times higher for those receiving the alternating treatment than it is for those receiving the sequential treatment.

ADJUSTED Effect of Treatment:

Exponentiated Treatment Coefficient:

```
        treatmentA
           1.76809
```

Given the exponentiated beta coefficients of the adjusted proportional odds model, the inferences made about the unadjusted effect of treatment remain the same. That is, the effect of treatment on response remains the same irrespective of one's sex. Thus, it follows that sex neither plays a predictive role on the outcome, nor display any influence on a patient's response treatment. For this reason, our cojecture about the joint independence of sex on treatment and response (part a) is sound. Moreover, in demonstrating that the effect of treatment on response remains the same after adjusting for sex, we prove that sex is not a confounder.

Best Model: Adjusted Proportional Odds

Based on the model selection procedure and model comparisons (using LR tests and AIC scores), it is evident that the model which best fits the data is that of the following form:

For response groups "progressive" (1), "no change" (2), "partial remission" (3), and "complete remission" (4), let X_T be the treatment random variable of two groups with "sequential" as the reference group (0); and let β_{0j} be the varying intercepts for each model $j = {1, 2, 3}$:

$$ Y = logit(P({Y \leq j})) = log(\frac{P(Y \leq j)}{P(Y > j)}) = X\beta_j $$$$ = \beta_{0j} + 0.5699X_T $$

```
 (Intercept)1: −1.2167

 (Intercept)2: 0.3382
```

```
(Intercept)3: 1.3803
```

We may now use these coefficients to obtain the corresponding log-odds for each treatment:

Log-Odds:

Response	Sequential	Alternating
Y <= 1	-1.2167	-0.6468
Y <= 2	0.3382	0.9081
Y <= 3	1.380	1.9499

Odds:

Response	Sequential	Alternating
Y <= 1	0.29621	0.52372
Y <= 2	1.40242	2.47961
Y <= 3	3.97609	7.03009

Given the tables above, we see that the exponentiated intercepts in the model represent the odds of displaying a response that is progressive, up to non-changing, and up to partial remission under the sequential treatment, while the exponentiated sum of intercepts and slope represent the same odds under the alternating treatment. Evidently, the odds are increased for each cumulative response level when the alternating treatment is implemented as opposed to the sequential treatment. This indicates that the latter is more effective, as it reduces the overall odds of not achieving complete remission (or remaining in a category below this one). More specifically, we may interpret the exponentiated slope (treatment coefficient of 0.5699) of approximately 1.8 as the extent to which the sequential treatment reduces these odds for each cumulative response category. That is, as can be observed above, the odds of each cumulative response category under the alternating treatment is approximately 1.8 times higher than those for the sequential treatment. In other words, the odds for each cumulative response category are roughly proportional with respect to treatment.

Final Conclusion:

In the "fitted vs. observed" plot below we see that the chosen model provides an adequate fit for the data. This is further confirmed by individual Wald tests (seen using the summary function), which yield statistically significant p-values for each coefficient in the model. Lastly, the LR test of the null and chosen models, proves that our chosen model fits the data significantly better than the null. Thus, given the now verified implications of this model on the data mentioned previously, our conclusions remain the same about the relative efficacy of the two treatments. That is, the evidence gathered from the model is sound enough to validate the claim that the sequential treatment is preferable to the alternating treatment in reducing the odds of remaining in any of the aforementioned categories (and hence increasing the odds of reaching complete remission). The calculated marginal probabilities below provide further support for this claim.

Probability:

Response	Sequential	Alternating
Y = 1	0.22852	0.34371
Y = 2	0.35524	0.36890
Y = 3	0.21529	0.16286
Y = 4	0.20096	0.12453

In [31]:

```
#specifying reference groups and ordering response

#treatment
tumor$treatment <- factor(tumor$treatment, levels=c("S", "A")) %>% #changing ord
    relevel(tumor$treatment, ref="S") #reference group = sequential

#sex
tumor$sex <- factor(tumor$sex, levels=c("F", "M")) %>% #changing ordered factor
    relevel(tumor$sex, ref="M") #reference group = male

#response
tumor$response <- ordered(tumor$response, levels=c("P","NC","PR","CR"))
levels(tumor$response) #reference group = progressive (always in the numerator)
```

'P' · 'NC' · 'PR' · 'CR'

In [32]:

```
#PROPORTIONAL ODDS
#model with intercept 1 -> Y=log(P(P)/P(NC+PR+CR))
#model with intercept 2 -> Y=log(P(P+NC)/P(PR+CR))
#model with intercept 3 -> Y=log(P(P+NC+PR)/P(CR))
#different intercepts, same beta coeficcients

#-> UNADJUSTED
#null model
po2_glm0 <- vglm(response ~ 1, family=cumulative(parallel=TRUE), data=tumor, wei
#additive model
po2_glm1 <- vglm(response ~ treatment + sex, family=cumulative(parallel=TRUE), d
#model with interraction term -> saturated model
po2_glm2 <- vglm(response ~ treatment + sex + treatment*sex, family=cumulative(p

#po2_glm0
#po2_glm1
#po2_glm2

#Model Selection:
#not many options (3 models to choose from)
#interaction term is NOT statistically significant, we choose the additive model
#po2_glm1 (~789) close in residual deviance to po2_glm2 (~788)

#lrtest(po2_glm0, po2_glm1) #po2_glm1 is a better fit than po2_glm0
#lrtest(po2_glm1, po2_glm2) #po2_glm1 is a better fit than po2_glm2


#-> ADJUSTED FOR SEX
#po2_glmA has residual deviance of ~792
po2_glmA <- vglm(response ~ treatment, family=cumulative(parallel=TRUE), data=tu
lrtest(po2_glmA, po2_glm1) #po2_glmA is a better fit than po2_glm1
```

```
Likelihood ratio test

Model 1: response ~ treatment
Model 2: response ~ treatment + sex
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  44 -396.33
2  43 -394.53 -1 3.5965     0.0579 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In [33]:

```
#CUMULATIVE LOGIT
#model 1 -> Y=log(P(P)/P(NC+PR+CR))
#model 2 -> Y=log(P(P+NC)/P(PR+CR))
#model 3 -> Y=log(P(P+NC+PR)/P(CR))
#different intercepts and beta coeficcients

#-> UNADJUSTED
#null model
cl2_glm0 <- vglm(response ~ 1, family=cumulative, data=tumor, weights=frequency)
#additive model
cl2_glm1 <- vglm(response ~ treatment + sex, family=cumulative, data=tumor, weig
#model with interraction term -> saturated model
cl2_glm2 <- vglm(response ~ treatment + sex + treatment*sex, family=cumulative, 

#cl2_glm0
#cl2_glm1
#cl2_glm2

#Model Selection:
#not many options (3 models to choose from)
#interaction term is NOT statistically significant, we choose the additive model
#cl2_glm1 (~786) somewhat close in residual deviance to cl2_glm2 (~783)

#lrtest(cl2_glm0, cl2_glm1) #cl2_glm1 is a better fit than cl2_glm0
#lrtest(cl2_glm1, cl2_glm2) #cl2_glm1 is a better fit than cl2_glm2

#-> ADJUSTED FOR SEX
#cl2_glmA has residual deviance of ~791
cl2_glmA <- vglm(response ~ treatment, family=cumulative, data=tumor, weights=fr

lrtest(cl2_glmA, cl2_glm1) #cl2_glmA is a better fit than cl2_glm1
```

```
Likelihood ratio test

Model 1: response ~ treatment
Model 2: response ~ treatment + sex
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  42 -395.64
2  39 -392.93 -3 5.4156     0.1438
```

In [34]:

```
#PROPORTIONAL ODDS or CUMULATIVE LOGIT? (po2_glmA or cl2_glmA)
#does the proportional odds assumption hold?

#similar log-likelihood values
#comparing log-likelihood values instead of LRT:
#both models have the same number of predictors
```

```
#they are non-nested
logLik(po2_glmA)
logLik(cl2_glmA)

#similar residual deviances
deviance(po2_glmA)
deviance(cl2_glmA)

#both fit the data equally well, yet po2_glmA is more parsimoneous
#therefore, the proportional odds assumption holds and provides the best model
```

```
-396.326569512947
-395.641299752361
792.653139025894
791.282599504723
```

In [35]:

```
#UNADJUSTED effect of treatment - Proportional Odds
exp(0.5807) #exponentiated treatment coefficient

#ADJUSTED effect of treatment - Proportional Odds
exp(0.5699) #exponentiated treatment coefficient
```

```
1.78728909533313
1.76809023356963
```

In [36]:

```
#BEST MODEL
#additive proportional odds model (adjusted for sex)
summary(po2_glmA)
```

```
Call:
vglm(formula = response ~ treatment, family = cumulative(parallel = TRUE),
    data = tumor, weights = frequency)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -1.2167     0.1717  -7.086 1.38e-12 ***
(Intercept):2   0.3382     0.1564   2.163  0.03054 *
(Intercept):3   1.3803     0.1814   7.611 2.73e-14 ***
treatmentA      0.5699     0.2116   2.694  0.00706 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
logitlink(P[Y<=3])

Residual deviance: 792.6531 on 44 degrees of freedom

Log-likelihood: -396.3266 on 44 degrees of freedom

Number of Fisher scoring iterations: 3

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
treatmentA
  1.768108
```

In [37]:

```
#INTERPRETING COEFFICIENT

#getting marginal probabilities (for alternating treatment)
logit_pYl1 = -1.2167 + 0.5699
logit_pYl2 = 0.3382 + 0.5699
logit_pYl3 = 1.3803 + 0.5699

pYl1 = exp(logit_pYl1)/(1 + exp(logit_pYl1))
pYl2 = exp(logit_pYl2)/(1 + exp(logit_pYl2))
pYl3 = exp(logit_pYl3)/(1 + exp(logit_pYl3))

pY1 = pYl1 - 0
pY2 = pYl2 - pYl1
pY3 = pYl3 - pYl2

#pYl1
#pYl2
#pYl3

#sum(pY1, pY2, pY3)
pY4 = 1 - sum(pY1, pY2, pY3)

#marginal probabilities
pY1 #progressive given alternating treatment
pY2 #no change given alternating treatment
pY3 #partial remission given alternating treatment
pY4 #complete remission given alternating treatment
```

```
0.343711011890575
0.368900194799811
0.162857241448533
0.124531551861081
```

In [38]:

```
#getting marginal probabilities (for sequential treatment)
logit2_pYl1 = -1.2167
logit2_pYl2 = 0.3382
logit2_pYl3 = 1.3803

p2Yl1 = exp(logit2_pYl1)/(1 + exp(logit2_pYl1))
p2Yl2 = exp(logit2_pYl2)/(1 + exp(logit2_pYl2))
p2Yl3 = exp(logit2_pYl3)/(1 + exp(logit2_pYl3))

p2Y1 = p2Yl1 - 0
p2Y2 = p2Yl2 - p2Yl1
p2Y3 = p2Yl3 - p2Yl2

#p2Yl1
#p2Yl2
#p2Yl3

#sum(pY1, pY2, pY3)
p2Y4 = 1 - sum(p2Y1, p2Y2, p2Y3)

#marginal probabilities
p2Y1 #progressive given sequential treatment
p2Y2 #no change given sequential treatment
```

```
p2Y3 #partial remission given sequential treatment
p2Y4 #complete remission given sequential treatment
```

```
0.228517710609371
0.355235504652692
0.215285961980231
0.200960822757706
```

In [39]:

```
#DOUBLE CHECKING PROBABILITIES ARE CORRECT

#cumulative probabilities for each treatment (model 1: y<= progressive)
cp1_tA = pYl1/(1-pYl1)
cp1_tS = p2Yl1/(1-p2Yl1)

#cumulative probabilities for each treatment (model 2: y<= progressive, no chang
cp2_tA = pYl2/(1-pYl2)
cp2_tS = p2Yl2/(1-p2Yl2)

#cumulative probabilities for each treatment (model 3: y<= progressive, no chang
cp3_tA = pYl3/(1-pYl3)
cp3_tS = p2Yl3/(1-p2Yl3)

#log of odds (cumulative probabilities for each treatment)
#should give the same proportional odds beta coefficient (slope) for treatment

#log odds - model 1
log(cp1_tA/cp1_tS) #log((pYl1/(1-pYl1))/(p2Yl1/(1-p2Yl1)))

#log odds - model 2
log(cp2_tA/cp2_tS)

#log odds - model 2
log(cp3_tA/cp3_tS)

#thus, the probabilities above are correct
#assuming proportional odds holds and the model fits the data well

#this means that, under this model, the odds of not achieving complete remission
cp3_tA/cp3_tS #=exp(log(cp3_tA/cp3_tS))
```

```
0.5699
0.5699
0.5699
1.76809023356963
```

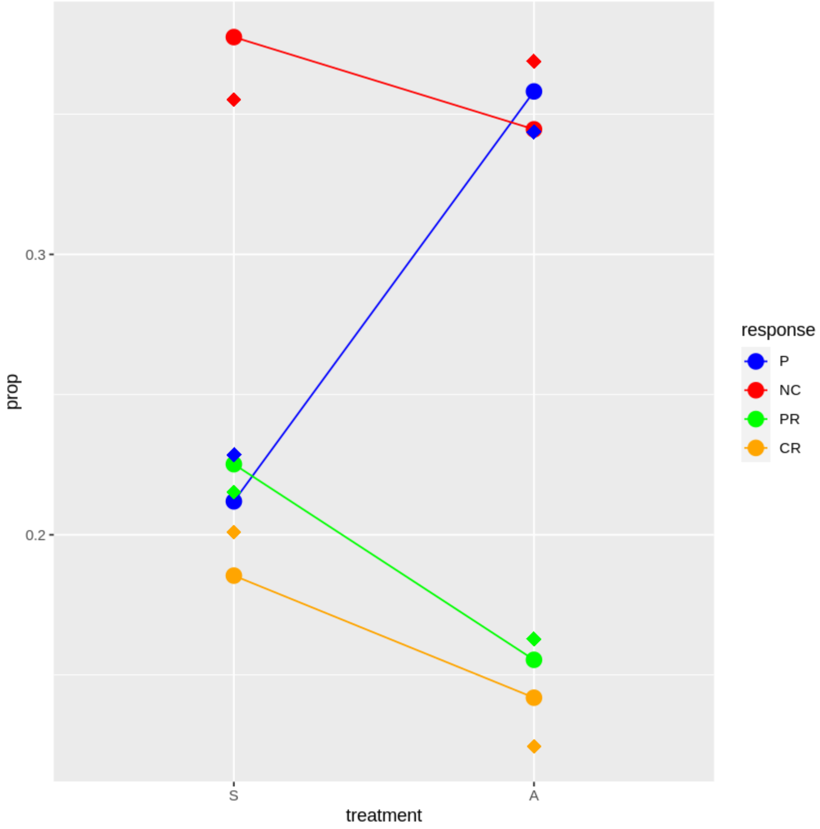
In [40]:

```
#sequential odds
#exp(logit2_pYl1)
#exp(logit2_pYl2)
#exp(logit2_pYl3)

#adjusted odds
#exp(logit_pYl1)
#exp(logit_pYl2)
#exp(logit_pYl3)
```

In [41]:

```
#ASSESSING "BEST" MODEL FIT

#fitted vs. observed
tumor_fit <- as.data.frame(fitted(po2_glmA))

tumor1 <- tumor %>%
            select(-c("sex")) %>%
            group_by(treatment) %>%
            summarize(response = response, frequency = frequency, total = sum(fr
            ungroup() %>%
            group_by(treatment, response) %>%
            mutate(sum_freq = sum(frequency)) %>%
            summarize(response = response, sum_freq = sum_freq, total = total, p
            distinct()

tumor1$fitted <- c(0.2285115, 0.3552443, 0.2152821, 0.2009621, 0.3437053, 0.3689

ggplot(tumor1, aes(x = treatment, y = prop)) +
    geom_line(aes(group = response, color = response)) +
    geom_point(aes(color = response), size = 4) +
    geom_point(aes(y = fitted, color = response), size = 4, shape = 18) +
    scale_color_manual(values=c("blue", "red", "green", "orange"))

#LRT: null vs. chosen model
lrtest(po2_glmA, po2_glm0) #chosen is better
#summary(po2_glmA) #significant coefficients
```

```
`summarise()` has grouped output by 'treatment'. You can override using the `.gr
oups` argument.

`summarise()` has grouped output by 'treatment', 'response'. You can override us
ing the `.groups` argument.

Likelihood ratio test

Model 1: response ~ treatment
Model 2: response ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  44 -396.33
2  45 -399.98  1 7.3148   0.006839 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


0.3
0.2
prop
S
A
treatment
response
P
NC
PR
CR