Each part of the problems 5 points

1. Project proposal

2. Agresti 7.1

(a) The prediction equation is
\[
\frac{\hat{\pi}_1}{\hat{\pi}_2} = (\alpha_1 - \alpha_2) + (\beta^{G}_1 - \beta^{G}_2)x + (\beta^{R}_1 - \beta^{R}_2)x_2 \\
= (0.883 + 0.758) + (0.419 - 0.105)x_1 + (0.342 - 0.271)x_2 \\
= 1.641 + 0.314x_1 + 0.071x_2
\]

(b) Controlling for race, the estimated odds for a female are \(e^{0.419} = 1.52\) times the odds for a male.

95% Confidence interval for log odds = 0.419 ± 1.96(0.171) = (0.08384, 0.75416)

95% Confidence Interval for odds = \((e^{0.08384}, e^{0.75416}) = (1.09, 2.13)\) That is, the odds could be a low as 1.09 or as high as 2.13.

(c) For a white female
\[
\frac{\hat{\pi}_1}{\hat{\pi}_3} = \frac{\exp(\alpha_1 + \beta^{G}_1 + \beta^{R}_1)}{1 + \exp(\alpha_1 + \beta^{G}_1 + \beta^{R}_1) + \exp(\alpha_2 + \beta^{G}_2 + \beta^{R}_2)} \\
= \frac{\exp(0.883 + 0.419 + 0.342)}{1 + \exp(0.883 + 0.419 + 0.342) + \exp(-0.785 + 0.105 + 0.271)} \\
= 0.76
\]

(d) For a black male we have
\[
\frac{\hat{\pi}_1}{\hat{\pi}_3} = 0.883 > 0 \implies \frac{\hat{\pi}_1}{\hat{\pi}_3} > 1 \\
\frac{\hat{\pi}_2}{\hat{\pi}_3} = -0.758 < 0 \implies \frac{\hat{\pi}_2}{\hat{\pi}_3} < 1 \\
\implies \hat{\pi}_1 > \hat{\pi}_3 > \hat{\pi}_2
\]

In words the log odds of yes to no are positive, implying that \(\hat{\pi}_1 > \hat{\pi}_3\) / The log odds of undecided to no are negative, implying \(\hat{\pi}_3 > \hat{\pi}_2\).

The log odds for a black women (yes/no) are 0.883 + 0.19 > 0 and for (undecided/no) are -0.758 + 0.105 < 0. The same ordering is implied by the same argument made for black men.

(e) The log odds ratio between yes/no for black females to black males is 0.419 > 0. For undecided/no it is 0.105 > 0. So black females respond yes rather than no more than black males and respond undecided rather than no more than black males. So the respond yes or undecided rather than no more often than do black males. This implies black males respond no rather than yes or undecided more often than do black females which implies \(\hat{\pi}_3\) is larger for black males than for black females. Similar arguments show the same for white females and for white males compared to black males.

Mathematically, we can show this quite easily. Let subscript b denote black, w white, f female and m male. Since \(\beta^{G}_1 > 0\) and \(\beta^{G}_2 > 0\) we have
\[
\frac{\hat{\pi}_{1bf}}{\hat{\pi}_{3bf}} = \frac{\exp(\alpha_1 + \beta^{G}_1)}{\exp(\alpha_1 + \beta^{G}_1 + \beta^{R}_1)} = \frac{\exp(\alpha_1)}{\exp(\alpha_1)} \text{ and } \frac{\hat{\pi}_{1bf}}{\hat{\pi}_{3bf}} = \frac{\exp(\alpha_2 + \beta^{G}_2)}{\exp(\alpha_2)} = \frac{\exp(\alpha_2)}{\exp(\alpha_2)}
\]
\[
\begin{align*}
\frac{\hat{\pi}_{1bf}}{\hat{\pi}_{3bf}} + \frac{\hat{\pi}_{2bf}}{\hat{\pi}_{3bf}} & > \frac{\hat{\pi}_{1bm}}{\hat{\pi}_{3bm}} + \frac{\hat{\pi}_{2bm}}{\hat{\pi}_{3bm}f} \\
\frac{\hat{\pi}_{1bf} + \hat{\pi}_{2bf}}{\hat{\pi}_{3bf}} & > \frac{\hat{\pi}_{1bm} + \hat{\pi}_{2bm}}{\hat{\pi}_{3bm}} \\
1 - \hat{\pi}_{3bf} & > 1 - \hat{\pi}_{3bm} \\
\hat{\pi}_{3bm}(1 - \hat{\pi}_{3bf}) & > \hat{\pi}_{3bf}(1 - \hat{\pi}_{3bm}) \\
\hat{\pi}_{3bm} & > \hat{\pi}_{3bf}
\end{align*}
\]

Similar arguments for white males and females completes the proof.

(f) There are 4 independent multinomials each of length 3. Therefore the saturated model as 4(3−1) = 8 parameters. The model in question has 2 + 2 + 2 = 6 parameters. The difference in the number of parameters between the saturated and model in question is 8−6=2. Deleting the gender effect removes two parameters and so there are 4 parameters in the model without gender. The difference in the residual df between the two models is 6−4 = 2 with a difference in model deviances of 8−0.9 = 7.1. P-value is 0.029. conditional on race, gender is a significant predictor of belief in life after death.

3. Agresti 7.2

(a) 
\[
\ln \frac{\hat{\pi}_R}{\hat{\pi}_D} = \ln \frac{\hat{\pi}_R}{\hat{\pi}_I} - \ln \frac{\hat{\pi}_D}{\hat{\pi}_I} = 1.0 + 0.3x - 3.3 + 0.2x = -2.3 + 0.5x
\]

For every increase of $10,000 in annual income the log odds of preferring Republican as opposed to a Democratic president increase by 0.5 times. The odds increase by exp(0.5)=1.65 times. \(\hat{\pi}_R > \hat{\pi}_D\) when \(x > 2.3/0.5 = 4.6\). This is obtained as follows: If \(\hat{\pi}_R > \hat{\pi}_D\) then

\[
0 < \ln \frac{\hat{\pi}_R}{\hat{\pi}_D} = -2.3 + 0.5x
\]

now solve the inequality in x. So one is more likely to prefer Republican as opposed to Democratic president if annual income is greater than $46,000.

(b) The prediction equation for \(\hat{\pi}_I\) is

\[
\hat{\pi}_I = \frac{1}{1 + \exp(3.3 - 0.2x) + \exp(1.0 + 0.3x)}
\]

(c) As income increases, the probability of preferring a Republican increases, the probability of preferring a Democrat decrease and the probability of preferring an Independent remains relatively constant, with a slight decrease as income increases. If income is less than $46,000 one is more likely to prefer a democrat and if it is greater than $46,000 is more likely to prefer a Republican.
Comparing possible four different models as below, we conclude that there is strong evidence that the additive model with gender and race effects is significantly better than others.

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. df</th>
<th>Resid. Dev</th>
<th>Test</th>
<th>Df</th>
<th>LR stat.</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 gender</td>
<td>20</td>
<td>2162.507</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>2 gender + race</td>
<td>18</td>
<td>2085.782</td>
<td>1 vs 2</td>
<td>2</td>
<td>76.72423</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. df</th>
<th>Resid. Dev</th>
<th>Test</th>
<th>Df</th>
<th>LR stat.</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 race</td>
<td>20</td>
<td>2099.224</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>2 gender + race</td>
<td>18</td>
<td>2085.782</td>
<td>1 vs 2</td>
<td>2</td>
<td>13.44183</td>
<td>0.001205438</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. df</th>
<th>Resid. Dev</th>
<th>Test</th>
<th>Df</th>
<th>LR stat.</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 gender + race</td>
<td>18</td>
<td>2085.782</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>2 gender * race</td>
<td>16</td>
<td>2085.584</td>
<td>1 vs 2</td>
<td>2</td>
<td>0.1982117</td>
<td>0.9056468</td>
</tr>
</tbody>
</table>

Below is the summary of additive mode with gender and race.

```
multinom(formula = identification ~ gender + race, data = X, weights = count)
```

Coefficients:

(Intercept) genderMale racewhite
Independent  1.388246  0.2201885  1.118285
Republican -2.565342  0.5727615  2.278131

Based on the result above, the estimated odds of being republican rather than democrat for whites are $\exp(2.278131) = 9.758$ times the odds of being republican rather than democrat for blacks. Therefore, the estimates odds of being democrat rather than for blacks are 9.758 times than the odds of being democrat rather than for whites.
Also, the estimated odds of being republican rather than democrat for males are $\exp(0.5727615) = 1.773$ times the odds of being republican rather than democrat for females. Therefore, the estimates odds of being democrat rather than for females are 1.773 times than the odds of being democrat rather than for males.

**R code**

```r
X<-data.frame(count=c(132,176,127,42,6,12,172,129,130,56,4,15),
  gender=c(rep("Male",6), rep("Female",6)),
  race=c(rep("white",3), rep("black",3), rep("white",3), rep("black",3)),
  identification=rep(c("Democrat","Republican","Independent")))

library(nnet)
fit1<-multinom(identification~gender*race, weight=count, data=X)
fit2<-multinom(identification~gender+race, weight=count, data=X)
fit3<-multinom(identification~gender, weight=count, data=X)
fit4<-multinom(identification~race, weight=count, data=X)

anova(fit2, fit3)
anova(fit2, fit4)
anova(fit1, fit2)

summary(fit2)
```

5. Agresti 7.4

(a) Comparing non-additive and additive models as below, we conclude that there is no strong evidence that the non-additive model with gender and length effects is significantly better than additive model. Hence we chose the additive model with gender and length.

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. df</th>
<th>Resid. Dev</th>
<th>Test</th>
<th>Df</th>
<th>LR stat.</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Gender + Adult</td>
<td>120</td>
<td>108.9935</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2 Gender * Adult</td>
<td>118</td>
<td>106.5334</td>
<td>1 vs 2</td>
<td>2</td>
<td>2.460165</td>
<td>0.2922685</td>
</tr>
</tbody>
</table>

Below is the summary of additive model with gender and length.

```r
multinom(formula = Choice ~ Gender + Adult, data = Y)
```

Coefficients:

(Intercept) GenderMale AdultSubAdult

I  -1.214950 -1.10874416 2.423343
0  -1.704756 -0.07299331 1.250329

Std. Errors:

(Intercept) GenderMale AdultSubAdult

I  0.6153801 0.6881429 0.7190658
0  0.7315531 0.8148199 0.7608239

Residual Deviance: 108.9935
AIC: 120.9935
In summary,
\[
\ln \hat{\pi}_I = -1.21495 \cdot I\{\text{Gender} = \text{male}\} + 2.423343 \cdot I\{\text{Length} = \text{subadult}\}
\]
\[
\ln \hat{\pi}_O = -1.704756 \cdot I\{\text{Gender} = \text{male}\} + 1.250329 \cdot I\{\text{Length} = \text{subadult}\}
\]

When one is a subadult instead of adult, odds of eating invertebrate instead of fish increases if others remain constant, since \( \beta_{I\text{subadult}} > 0 \). Likewise, when one is a subadult instead of adult, odds of eating other instead of fish increase if others remain constant, since \( \beta_{O\text{subadult}} > 0 \).

Similarly, when one is male instead of female, odds of eating other instead of fish decrease if others remain constant, since \( \beta_{I\text{male}} < 0 \). Likewise, when one is male instead of female, odds of eating other instead of fish decrease if others remain constant, since \( \beta_{O\text{male}} < 0 \).

For adult females, the estimated probabilities of the food-choice categories,
\[
\hat{\pi}_F = \frac{1}{1 + \exp(-1.21495) + \exp(-1.704756)} = 0.67434
\]
\[
\hat{\pi}_I = \frac{\exp(-1.21495)}{1 + \exp(-1.21495) + \exp(-1.704756)} = 0.2006875
\]
\[
\hat{\pi}_O = \frac{\exp(-1.704756)}{1 + \exp(-1.21495) + \exp(-1.704756)} = 0.12297
\]

**R code**

```r
Length<-c(1.30,1.32,1.40, 1.42, 1.42, 1.47, 1.50, 1.52, 1.63, 1.65, 1.65, 1.65, 1.68, 1.70, 1.73, 1.78, 1.80, 1.85, 1.93, 1.98, 2.03, 2.03, 2.31, 2.36, 2.46, 2.59, 2.83, 2.96, 3.01, 3.06, 3.08, 3.16, 3.21, 3.29, 3.41, 3.44, 3.56, 3.58, 3.66, 3.68, 3.71, 3.89, 1.24, 1.30, 1.45, 1.45, 1.55, 1.60, 1.60, 1.65, 1.78, 1.80, 1.88, 2.16, 2.26, 2.31, 2.36, 2.39, 2.41, 2.44, 2.56, 2.67, 2.72, 2.79, 2.84)
Gender<-c(rep("Male",39),rep("Female",24))
Adult<ifelse(Length>1.83, "Adult", "SubAdult")
Y<-data.frame(Length=Length, Adult=Adult, Gender=Gender, Choice=as.factor(Choice))
fit1<-multinom(Choice~Gender*Adult, data=Y)
fit2<-multinom(Choice~Gender+Adult, data=Y)
anova(fit1, fit2)
summary(fit2)
```

(b) Using only observations for which primary food choice was fish or invertebrate, there is no strong evidence that the non-additive model with gender and length effects is significantly better than additive model. Hence we chose the additive model with gender and length.

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. df</th>
<th>Resid. Dev</th>
<th>Test Df</th>
<th>LR stat.</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Gender + Adult</td>
<td>-3</td>
<td>53.95641</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2 Gender * Adult</td>
<td>-4</td>
<td>52.38083</td>
<td>1 vs 2</td>
<td>1</td>
<td>1.575685</td>
</tr>
</tbody>
</table>
Below is the summary of additive model with gender and length.

\[
\text{multinom(formula = Choice ~ Gender + Adult, data = Z)}
\]

Coefficients:

<table>
<thead>
<tr>
<th>Values</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.176860</td>
</tr>
<tr>
<td>GenderMale</td>
<td>-1.290366</td>
</tr>
<tr>
<td>AdultSubAdult</td>
<td>2.544304</td>
</tr>
</tbody>
</table>

Residual Deviance: 53.95641
AIC: 59.95641

In summary,

\[
\ln \frac{\hat{\pi}_I}{\hat{\pi}_F} = -1.17686 - 1.290366 \cdot I\{\text{Gender} = \text{male}\} + 2.544304 \cdot I\{\text{Length} = \text{subadult}\}
\]

To compare parameter estimates and (standard error),

<table>
<thead>
<tr>
<th>Data</th>
<th>Intercept</th>
<th>subAdult</th>
<th>Gender=Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole(F, I, O)</td>
<td>-1.21495(0.615)</td>
<td>-1.10874416 (0.688)</td>
<td>2.423343 (0.719)</td>
</tr>
<tr>
<td>Partial (F,I)</td>
<td>-1.17686 (0.608)</td>
<td>-1.290366(0.744)</td>
<td>2.544304(0.760)</td>
</tr>
</tbody>
</table>

R code

\[
Z<-Y[Y$Choice !="O",]
Z$Choice<-as.character(Z$Choice)
Z$Choice<-as.factor(Z$Choice)
fit1<-multinom(Choice~Gender*Adult, data=Z)
fit2<-multinom(Choice~Gender+Adult, data=Z)
anova(fit1, fit2)
summary(fit2)
\]

(c) Using additive model with continuous measurements of length,

\[
\text{multinom(formula = Choice ~ Gender + Length, data = Y)}
\]

Coefficients:

<table>
<thead>
<tr>
<th>Values</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.719262 -1.2073510 -2.93067974</td>
</tr>
<tr>
<td>GenderMale</td>
<td>0.1586403 -0.08695569</td>
</tr>
</tbody>
</table>

Std. Errors:

<table>
<thead>
<tr>
<th>Values</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.917026 0.724665 0.9548020</td>
</tr>
<tr>
<td>GenderMale</td>
<td>0.1586403 0.08695569</td>
</tr>
</tbody>
</table>

The probability of choosing Invertebrate or Other instead of fish decrease as the length of the body increase, since $\beta^\text{Length}_{IF}$ and $\beta^\text{Length}_{OF} < 0$

To find the estimated length at which the invertebrate and other categories are equally likely, we can try to find that of $\ln \frac{\hat{\pi}_I(\text{Length})}{\hat{\pi}_O(\text{Length})} = 0$ since $\pi_I(\text{Length}) = \pi_O(\text{Length})$
When Gender is male,

\[ \exp(5.719262 - 1.207351 - 2.93067974x) = \exp(-1.102027 + 0.1586403 - 0.08695569x) \]

\[ (5.719262 + 1.102027) + (-1.207351 - 0.1586403) + (-2.93067974 + 0.08695569)x = 0 \]

\[ x = 1.918m \]

When Gender is female,

\[ \exp(5.719262 - 2.93067974x) = \exp(-1.102027 - 0.08695569x) \]

\[ (5.719262 + 1.102027) + (-2.93067974 + 0.08695569)x = 0 \]

\[ x = 2.3987m \]

6. Agresti 7.9

(a) Setting up dummy variable (1,0) for (male, female) and (1,0) for (sequential, alternating), we get
treatment effect = -0.581 (SE=0.212) and gender effect = -0.541 (SE=0.295). The estimated odds
ratios are 0.56 and 0.58. The sequential therapy leads to a better response than the alternating
therapy; The estimates odds of response with sequential therapy below any fixed level are 0.56
times the estimates odds with alternating therapy.

(b) The main effects model fits well \((G^2 = 5.6, df = 7)\), and adding an interaction term does not give
an improved fit (The interaction model has \(G^2 = 4.5, df = 6\)).

subjects, "Would you say that astrology is very scientific, sort of scientific, or not at all scientific?".
The table below cross-classifies their responses with their highest degree. Our goal is to investigate
whether educated people put less credence in astrology.

<table>
<thead>
<tr>
<th>Highest degree</th>
<th>Astrology is scientific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not at all</td>
</tr>
<tr>
<td>0: &lt; High school</td>
<td>98</td>
</tr>
<tr>
<td>1: High school</td>
<td>574</td>
</tr>
<tr>
<td>2: Junior college</td>
<td>122</td>
</tr>
<tr>
<td>3: Bachelor</td>
<td>268</td>
</tr>
<tr>
<td>4: Graduate</td>
<td>148</td>
</tr>
</tbody>
</table>

(a) We view opinion about astrology as the response variable, treat education level as a quantitative
predictor variable with scores \(u_i=0, 1, 2, 3\) and 4, and specify a proportional odds model.

i. State the model and its assumptions, and interpret the parameters.

Answer:

\[ Y_{ij} \overset{iid}{\sim} \text{Multinomial}(\pi_1, \pi_2, \pi_3), \quad \text{where } \logit[P(Y_{ij} \leq j|i)] = \alpha_j + \beta u_i, \quad i = 0, \ldots, 4, \quad j = 1, 2 \]

\(\alpha_j\) is the baseline log-odds of cumulative probability of the levels of astrology for individuals
with ‘< High school’.
\(\beta\) is log-odds ratio of cumulative probabilities that follows a unit change is score \(u_i\), it is
independent of \(j\).
ii. Based on the output below, calculate the estimated cumulative odds ratio comparing the lowest level of education ‘< High school’ with the highest level ‘Graduate’, and interpret your conclusions. (Hint: polr estimates negative coefficients of the parameters of the covariates, i.e. $-\beta u_i$).

Call:
polr(astro ~ edu, data = x, weights = count)

Coefficients:
<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>edu</td>
<td>-0.4614</td>
<td>0.04860722</td>
<td>-9.49191</td>
</tr>
</tbody>
</table>

Intercepts:
<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.0264</td>
<td>0.0853</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2.3137</td>
<td>0.1246</td>
</tr>
</tbody>
</table>

Residual Deviance: 2658.161
AIC: 2664.161

Answer:
\[
\log[OR] = \logit[P(Y_{ij} \leq j \mid \text{Graduate})] - \logit[P(Y_{ij} \leq j \mid < \text{High school})]
= 4 \cdot (-1) \cdot (-0.4614) = 1.8456
\]

On the original scale, $\hat{OR} = e^{1.8456} = 6.3318$. The odds of having a non-scientific view of astrology increases 6.3318 times between individuals with the highest and the lowest education.

iii. Calculate the estimates cell counts for ‘Graduate’ education.

Answer:
\[
\hat{P}(Y \leq 1 \mid \text{Graduate}) = \frac{\hat{p}^{0.0264+4(-1)(-0.4614)}}{1 + \hat{p}^{0.0264+4(-1)(-0.4614)}} = 0.8666
\]
\[
\hat{P}(Y \leq 2 \mid \text{Graduate}) = \frac{\hat{p}^{2.3137+4(-1)(-0.4614)}}{1 + \hat{p}^{2.3137+4(-1)(-0.4614)}} = 0.9845
\]

From these, $\hat{P}(Y = 1) = 0.8666$, $\hat{P}(Y = 2) = 0.9845 - 0.8666 = 0.118$ $\hat{P}(Y = 3) = 1 - 0.9845 = 0.0155$. The corresponding predicted cell counts are 149.05, 20.29 and 2.66.

(b) Now, we fit a second proportional odds model where educational levels are viewed as qualitative predictors.

i. State the model and the assumptions, and interpret the parameters.

Answer:
\[Y_{ij} \overset{iid}{\sim} \text{Multinomial}(\pi_1, \pi_2, \pi_3), \text{ where } \logit[P(Y \leq j \mid i)] = \alpha_j + \tau_i, \ j = 1, 2\]
$\alpha_j$ is the baseline log-odds of cumulative probability of the levels of astrology for individuals with ‘< High school’.
$\tau_i$ is the deviation from the baseline due to the education level $i$. The deviations are not linearly related, and are independent of $j$. 

8
ii. Given the partial output below, conduct a test comparing the models with the continuous and the categorical predictors at the confidence level 95%. Interpret the meaning of the test, and state your conclusions.

**Call:**
\[
polr(formula = astro ~ as.factor(edu), data = x, weights = count)
\]

**Coefficients:**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(edu)1</td>
<td>-0.6497478</td>
<td>0.1511085</td>
<td>-4.299876</td>
</tr>
<tr>
<td>as.factor(edu)2</td>
<td>-1.0657059</td>
<td>0.2164141</td>
<td>-4.924383</td>
</tr>
<tr>
<td>as.factor(edu)3</td>
<td>-1.4746451</td>
<td>0.1917049</td>
<td>-7.692265</td>
</tr>
<tr>
<td>as.factor(edu)4</td>
<td>-1.9439071</td>
<td>0.2581836</td>
<td>-7.529165</td>
</tr>
</tbody>
</table>

**Intercepts:**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>-0.1129 0.1360  -0.8301</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2.1797 0.1603  13.5986</td>
<td></td>
</tr>
</tbody>
</table>

**Residual Deviance:** 2656.409

**AIC:** 2668.409

**Answer:**

We test \( H_0 \): the model in (a), vs \( H_a \): the model in (b). We conduct the likelihood ratio test, which in this case is the same as the deviance goodness of fit test. The test statistic is

\[
2 [\text{logL}(H_a) - \text{logL}(H_0)] = 2 [\text{deviance}(H_0) - \text{deviance}(H_a)]
\]

\[
= 2658.161 - 2656.409 = 1.752 < \chi^2_3(1 - 0.05) = 7.814
\]

We fail to reject \( H_0 \), and conclude that there is no evidence against the model with a linear effect of education level on log odds.

---

8. [Methods qualifying exam, January 2010: use paper and pencil.] A researcher is interested in the relationship between gender and academic performance among high school students. To this end, he uses the following data collected among Minnesota high school graduates of 1938.

<table>
<thead>
<tr>
<th>High School Rank</th>
<th>Lower (L)</th>
<th>Middle (M)</th>
<th>Upper (U)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>1640</td>
<td>3040</td>
<td>3181</td>
<td>7861</td>
</tr>
<tr>
<td>Male</td>
<td>2054</td>
<td>2544</td>
<td>1609</td>
<td>6207</td>
</tr>
</tbody>
</table>

The researcher defines the variables \( X \)

\[
X = \begin{cases} 
0 & \text{Female} \\
1 & \text{Male} 
\end{cases}
\]

\( Y \) = the school rank of a particular student, and \( p_L, p_M \) and \( p_U \) as the probability of graduating with lower, middle and upper school rank respectively. The researcher fits the proportional odds model

\[
\log \frac{P_i}{1 - P_i} = \alpha_i + \beta X, \quad i = 1, 2.
\]

where \( P_1 = p_L \) and \( P_2 = p_L + p_M \). The Maximum Likelihood estimates of the model are: \( \hat{\alpha}_1 = -1.3437 \), \( \hat{\alpha}_2 = 0.3915 \), and \( \hat{\beta} = 0.6479 \).

(a) Provide the rationale for using this model for this dataset, and state the assumptions.
Answer:
The proportional odds model is a popular model for polytomous outcomes because it preserves the meaning of the relationship between the covariate $X$ and the response, regardless of how the outcomes were combined into categories. It assumes that the ratio of odds for graduating with a particular school rank or lower for males versus females is independent of the school rank.

(b) Calculate and interpret the odds ratio of graduating with a middle or a lower school rank for males versus females.

Answer:

$$\frac{odds(Y \leq M | X = 1)}{odds(Y \leq M | X = 0)} = e^\beta = 1.911$$

We conclude that males have a higher odds of graduating with a middle or lower school rank. In other words males have a worse academic performance than females in this dataset.

(c) Calculate the expected counts for female graduates at the Lower, Middle and Upper school level (i.e. for the first row of the table).

Answer:
For females,

$$\hat{p}_L = e^{-1.3437}/(1 + e^{-1.3437}) = .2069$$
$$\hat{p}_L + \hat{p}_M = e^{.3915}/(1 + e^{.3915}) = .5966,$$

so $\hat{p}_M = .5966 - .2069 = .3897$, and $\hat{p}_U = 1 - .5966 = .4034$.

The expected counts are therefore

$$7861 \times (0.2069, 0.3897, 0.4034) = (1626.44, 3063.43, 3171.13).$$

(d) To evaluate the adequacy of the model, the researcher would like to use the $\chi^2$ test.

i. State the null and the alternative hypotheses of the test.

Answer:

$H_0$: the proportional odds model above provides an adequate fit to the data
$H_a$: a different/more complex model is necessary to fit the data

ii. How many terms will be added to calculate the test statistic, and how many degrees of freedom will be used?

Answer:

The $\chi^2$ statistic is a sum over 6 terms, and one would use $1 = 4 - 3$ df for the test.