P8110: Applied Regression II
Homework #6 [20 points]

The "cars.csv" data give the result of a study of motor vehicle safety. In this study, 300 motor vehicle drivers were asked to rate importance of air conditioning and power steering in cars. The columns of variables from left to right are:

sex	=	1 - Women, 2 - Men
age	=	1 - (18-23 years), 2 - (24-40 years), 3 - (> 40 years)
response	=	1 - No or little importance, 2 - Important, 3 - Very important
count	=	Frequency of each response category

1. Fit an ordinal logistic regression model (1) to the data to assess the response to importance of air conditioning and power steering in cars for men and women in the three different age categories.

   - Write down the model. [2 points]

     Model 1:

     $$log\{\frac{Pr(reponse \leq k)}{1 - Pr(reponse \leq k)}\} = \alpha_k + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, k = 1, 2$$

     where

     $$X_1 = \begin{cases} 1, \text{if sex} = 1 \\ 0, \text{if sex} = 2 \end{cases},$$

     $$X_2 = \begin{cases} 1, \text{if age} = 2 \\ 0, \text{otherwise} \end{cases}, \quad X_3 = \begin{cases} 1, \text{if age} = 3 \\ 0, \text{otherwise} \end{cases},$$

     where X_2 and X_3 are the two dummy variables for the age, and the first age group (age in 18-23) is set as the reference group.

   - Test the proportional odds assumption. Show hypotheses, test statistic, degrees of freedom, p-value, and conclusion. [2 point]

     Hypotheses :

     H_0: the slope of the X-variables (β_1, β_2, β_3) are equal across logit equations.

     H_α: different slopes are needed.

     Test statistics and P-values : Here we use the score chi-square test. Based on the SAS output of Model 1, we have

$$\text{Score test:} \quad S = 0.7139, \quad \text{with 3 d.f.}$$

The corresponding p-value is $P(\chi^2_3 \geq 0.7139) = 0.8699$.

Conclusion: According to the result of the Score test, we fail to reject H_0 at a significant level of $\alpha = 0.05$. Therefore, the proportional odds assumption is reasonable.

- Estimate the odds ratio and 95% CI of a lower rating (i.e. rating less important) regarding air conditioning and power steering in cars between men and women. Do women care less or more about the features of air conditioning and power steering in cars? Justify your conclusion. [4 points]

  Odds ratio and 95% CI: Based on the SAS output of Model 1, we have

  $$\text{Odds ratio of women vs men:} \quad 0.562, \quad 95\% \text{ CI:} \quad [0.361, 0.875]$$

  The "Sex women vs men" odds ratio says the women has 0.562 times the odds of giving a lower rating than men. So women care more about the features of air conditioning and power steering in cars.

- Estimate the probability of rating "Very important" regarding the features of air conditioning and power steering in cars for women aged 18-23. [Hint: $\Pr(Y = 3) = 1 - \Pr(Y \leq 2)$] [3 points]

  Probability:

$$\begin{aligned} \hat{Pr}(response \leq 2 | X_1 = 1, X_2 = X_3 = 0) &= \frac{e^{\alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}}{1 + e^{\alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}} \\ &= \frac{e^{2.2308 - 0.5762}}{1 + e^{2.2308 - 0.5762}} \\ &= 0.8395 \end{aligned} \tag{1}$$

  Therefore, $\hat{Pr}(response = 3 | X_1 = 1, X_2 = X_3 = 0) = 1 - 0.8395 = 0.1605$.

SAS output:

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
0.7139	3	0.8699

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	0.6195	0.2156	8.2571	0.0041
Intercept	2	1	2.2308	0.2533	77.5489	<.0001
sex	1	1	-0.5762	0.2261	6.4936	0.0108
age	2	1	-1.1468	0.2773	17.1079	<.0001
age	3	1	-2.2322	0.2904	59.0806	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sex 1 vs 2	0.562	0.361	0.875
age 2 vs 1	0.318	0.184	0.547
age 3 vs 1	0.107	0.061	0.190

2. Fit a multinomial logistic regression model (2) to the data using the "no or little importance" as the reference category.

   - Write down the model. [2 points]

     Model 2:

$$log\{\frac{Pr(response=2)}{Pr(response=1)}\} = \beta_{02} + \beta_{12}X_1 + \beta_{22}X_2 + \beta_{32}X_3$$
$$log\{\frac{Pr(response=3)}{Pr(response=1)}\} = \beta_{03} + \beta_{13}X_1 + \beta_{23}X_2 + \beta_{33}X_3$$

   - Estimate the odds ratio and 95% CI of rating "very important" versus "no or little importance" between men and women. [2 points]

Odds ratio and 95% CI: Based on the SAS output of Model 2, we have that

$$\text{Odds ratio:} \quad 2.254, \qquad 95\%\ \text{CI:} \quad [1.202, 4.230]$$

- Estimate the probability of rating "Very important" regarding the features of air conditioning and power steering in cars for women aged 18-23. [3 points]

Probability:

$$\begin{aligned}
&\hat{Pr}(response = 3 | X_1 = 1, X_2 = X_3 = 0) \\
&= \frac{e^{\beta_{03}+\beta_{13}X_1+\beta_{23}X_2+\beta_{33}X_3}}{1 + e^{\beta_{03}+\beta_{13}X_1+\beta_{23}X_2+\beta_{33}X_3} + e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\beta_{32}X_3}} \\
&= \frac{e^{0.8129-1.8520}}{1 + e^{0.8129-1.8520} + e^{0.3881-0.9789}} \\
&= 0.1854
\end{aligned}$$

SAS output:

Analysis of Maximum Likelihood Estimates

Parameter		response	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		2	1	-0.9789	0.2563	14.5836	0.0001
Intercept		3	1	-1.8520	0.3307	31.3628	<.0001
sex	1	2	1	0.3881	0.3005	1.6677	0.1966
sex	1	3	1	0.8129	0.3210	6.4122	0.0113
age	2	2	1	1.1283	0.3416	10.9059	0.0010
age	2	3	1	1.4780	0.4009	13.5912	0.0002
age	3	2	1	1.5876	0.4029	15.5270	<.0001
age	3	3	1	2.9165	0.4229	47.5594	<.0001

Odds Ratio Estimates

Effect	response	Point Estimate	95% Wald Confidence Limits	
sex 1 vs 2	2	1.474	0.818	2.657
sex 1 vs 2	3	2.254	1.202	4.230
age 2 vs 1	2	3.090	1.582	6.037
age 2 vs 1	3	4.384	1.998	9.620
age 3 vs 1	2	4.892	2.221	10.775
age 3 vs 1	3	18.477	8.066	42.327

– Which model would you choose? Why? [2 points]

I would choose ordinal logistic regression because the response variable has a natural ordering among its three categories. In addition, the proportional odds assumption holds, making ordinal logistic regression the preferred model.

SAS code:

```
data hw6;
infile 'cars.csv' delimiter=',' missover dsd;
input sex age response count;
run;

proc logistic data=hw6;
freq count;
class sex (ref='2') age (ref='1')/param=ref;
model response = sex age;
oddsratio sex;
run;

proc logistic data=hw6;
freq count;
class sex (ref='2') age (ref='1')/param=ref;
model response(ref="1")=sex age / link=glogit;
oddsratio sex;
run;
```