Categorical Project 2

Maya Elshweikhy

Installing packages

```
library(tinytex)
library(carData)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MASS)
```

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

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(pscl)
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
library(ggplot2)
library(mlogit)
```

```
## Loading required package: dfidx
```

```
##
## Attaching package: 'dfidx'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

Loading the Heating data from the mlogit package

```
data("Heating", package = "mlogit")
str(Heating)
```

```
## 'data.frame':    900 obs. of  16 variables:
##  $ idcase: num  1 2 3 4 5 6 7 8 9 10 ...
##  $ depvar: Factor w/ 5 levels "gc","gr","ec",..: 1 1 1 4 4 1 1 1 1 1 ...
##  $ ic.gc : num  866 728 599 835 756 ...
##  $ ic.gr : num  963 759 783 793 846 ...
##  $ ic.ec : num  860 797 720 761 859 ...
##  $ ic.er : num  996 895 900 831 986 ...
##  $ ic.hp : num  1136 969 1048 1049 883 ...
##  $ oc.gc : num  200 169 166 181 175 ...
##  $ oc.gr : num  152 169 138 147 139 ...
##  $ oc.ec : num  553 520 439 483 404 ...
##  $ oc.er : num  506 486 405 425 390 ...
##  $ oc.hp : num  238 199 171 223 178 ...
##  $ income: num  7 5 4 2 2 6 4 6 5 7 ...
##  $ agehed: num  25 60 65 50 25 65 35 20 60 20 ...
##  $ rooms : num  6 5 2 4 6 7 2 7 6 2 ...
##  $ region: Factor w/ 4 levels "valley","scostl",..: 4 2 4 2 1 2 2 1 2 2
...
```

Question 1

```
# Prepare the data for mlogit
Heating_mlogit <- mlogit.data(Heating, choice = "depvar", shape = "wide",
                              varying = c(3:12))

# Run the multinomial logit model with installation cost and operating cost
model1 <- mlogit(depvar ~ ic + oc, data = Heating_mlogit)

summary(model1)
```

```
##
## Call:
## mlogit(formula = depvar ~ ic + oc, data = Heating_mlogit, method = "nr")
##
## Frequencies of alternatives:choice
##       ec       er       gc       gr       hp
## 0.071111 0.093333 0.636667 0.143333 0.055556
##
## nr method
## 6 iterations, 0h:0m:0s
```

```
## g'(-H)^-1g = 9.58E-06
## successive function values within tolerance limits
##
## Coefficients :
##                    Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept):er  0.19459102  0.20424212  0.9527 0.3407184
## (Intercept):gc  0.05213336  0.46598878  0.1119 0.9109210
## (Intercept):gr -1.35058266  0.50715442 -2.6631 0.0077434 **
## (Intercept):hp -1.65884594  0.44841936 -3.6993 0.0002162 ***
## ic             -0.00153315  0.00062086 -2.4694 0.0135333 *
## oc             -0.00699637  0.00155408 -4.5019 6.734e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1008.2
## McFadden R^2:  0.013691
## Likelihood ratio test : chisq = 27.99 (p.value = 8.3572e-07)
```

Interpretation As shown in the R output, the Frequencies of alternatives outputs the prorportion of each heating system choice in which ec (electric central) represents 7.1%, er (electric room) represents 9.3%, gc (gas central) represents 63.7%, gr (gas room) represents 14.3%, and hp (heat pump)represents 5.6%.The model took 6 iterations to converge with 9.58E-06 convergence criterion.

Coefficients The coefficients for each alternative (relative to the base category, ec by default) are provided along with their standard errors, z-values, and p-values:

Intercepts:

er: 0.195 (not significant, p = 0.341) gc: 0.052 (not significant, p = 0.911) gr: -1.351 (significant, p = 0.008) hp: -1.659 (significant, p < 0.001)

Variables:

ic (installation cost): -0.00153 (significant, p = 0.014) oc (operating cost): -0.00700 (highly significant, p < 0.001)

Model Fit

Log-Likelihood: -1008.2 **McFadden R^2**: 0.013691, indicating that the proportion of the variance explained by the model is relatively low. **Likelihood Ratio Test:** Chi-squared statistic = 27.99 p-value = 8.3572e-07, indicating that the model is statistically significant.

```r
# Fit the model with only the operating cost
model2 <- mlogit(depvar ~ oc, data = Heating_mlogit)
summary(model2)
```

```
##
## Call:
## mlogit(formula = depvar ~ oc, data = Heating_mlogit, method = "nr")
##
```

```
## Frequencies of alternatives:choice
##       ec       er       gc       gr       hp
## 0.071111 0.093333 0.636667 0.143333 0.055556
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 1.05E-05
## successive function values within tolerance limits
##
## Coefficients :
##                   Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):er -0.0485091  0.1787929 -0.2713  0.786149
## (Intercept):gc  0.0647232  0.4642576  0.1394  0.889124
## (Intercept):gr -1.5585841  0.4988178 -3.1246  0.001781 **
## (Intercept):hp -2.0427010  0.4202429 -4.8608 1.169e-06 ***
## oc             -0.0071930  0.0015502 -4.6400 3.484e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1011.3
## McFadden R^2:  0.010696
## Likelihood ratio test : chisq = 21.868 (p.value = 2.9212e-06)

# Comparing models using AIC
AIC(model1)

## [1] 2028.457

AIC(model2)

## [1] 2032.58

# Perform likelihood ratio test between two models
lrtest(model1, model2)

## Likelihood ratio test
##
## Model 1: depvar ~ ic + oc
## Model 2: depvar ~ oc
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -1008.2
## 2   5 -1011.3 -1 6.1223    0.01335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

Log-Likelihood Model 1: -1008.2 Model 2: -1011.3 Model 1 has a higher log-likelihood so it has a better fit.

McFadden's R-squared Model 1: 0.013691 Model 2: 0.010696 Model 1 accounts for a slightly higher proportion of the variance in the outcome variable compared to Model 2.

Likelihood Ratio Test To formally compare the models, we use the likelihood ratio test results:

Chi-squared statistic: 6.1223 Degrees of freedom: 1 (difference in the number of parameters between the two models) p-value: 0.01335 We reject the null hypothesis since the p-value is less than alpha (0.05) in which Model 2 is sufficient. we cann conclude that Model 1 provides a significantly better fit.

Considering the log-likelihood, McFadden's R-squared, and the likelihood ratio test, Model 1, integrating both installation cost (ic) and operating cost (oc), exhibits a superior fit to the dataset compared to Model 2,which includes only the operating cost (oc). The additionn of oc enhances the model.

```
# Fit the null model
null_model <- mlogit(depvar ~ 1, data = Heating_mlogit)

# Perform likelihood ratio test between two models
lrtest(model1, null_model)
```

```
## Likelihood ratio test
##
## Model 1: depvar ~ ic + oc
## Model 2: depvar ~ 1
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   6 -1008.2
## 2   4 -1022.2 -2 27.99  8.357e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

Degrees of Freedom (Df):

Model 1: 6 Model 2: 4 Difference in degrees of freedom: 6 - 4 = 2 Log-Likelihood (LogLik):

Model 1: -1008.2 Model 2: -1022.2 Chi-Squared Statistic (Chisq) = 2 × (LogLik of Model 1 - LogLik of Model 2) = 2 × (-1008.2 - (-1022.2)) = 27.99

P-Value (Pr(>Chisq)):

The p-value for the likelihood ratio test is 8.357 × 10^-7, which is very small.

Significance: Since the p-value (8.357 × 10^-7) is very small , so we reject the null hypothesis that the model with only the intercepts (Model 2) is sufficient. Therefore, the chosen model (Model 1) with both ic and oc is statistically significant and provides a significantly better fit to the data.

```
# checking coefficient significance for model 1
summary(model1)
```

```
##
## Call:
## mlogit(formula = depvar ~ ic + oc, data = Heating_mlogit, method = "nr")
##
## Frequencies of alternatives:choice
##       ec       er       gc       gr       hp
## 0.071111 0.093333 0.636667 0.143333 0.055556
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 9.58E-06
## successive function values within tolerance limits
##
## Coefficients :
##                    Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept):er  0.19459102  0.20424212  0.9527 0.3407184
## (Intercept):gc  0.05213336  0.46598878  0.1119 0.9109210
## (Intercept):gr -1.35058266  0.50715442 -2.6631 0.0077434 **
## (Intercept):hp -1.65884594  0.44841936 -3.6993 0.0002162 ***
## ic             -0.00153315  0.00062086 -2.4694 0.0135333 *
## oc             -0.00699637  0.00155408 -4.5019 6.734e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1008.2
## McFadden R^2:  0.013691
## Likelihood ratio test : chisq = 27.99 (p.value = 8.3572e-07)
```

Interpretation (Intercept):er

Estimate: 0.19459102 Std. Error: 0.20424212 z-value: 0.9527 Pr(>|z|): 0.3407184
Interpretation: The intercept for electric room (er) heating systems shows no statistical significance (p-value > 0.05).

(Intercept):gc

Estimate: 0.05213336 Std. Error: 0.46598878 z-value: 0.1119 Pr(>|z|): 0.9109210
Interpretation: The intercept for gas central (gc) heating systems also lacks statistical significance (p-value > 0.05).

(Intercept):gr

Estimate: -1.35058266 Std. Error: 0.50715442 z-value: -2.6631 Pr(>|z|): 0.0077434 **
Interpretation: The intercept for gas room (gr) heating systems is statistically significant (p-value < 0.01).

(Intercept):hp

Estimate: -1.65884594 Std. Error: 0.44841936 z-value: -3.6993 Pr(>|z|): 0.0002162 ***
Interpretation: The intercept for heat pump (hp) heating systems is highly significant (p-value < 0.001).

ic (Installation Cost)

Estimate: -0.00153315 Std. Error: 0.00062086 z-value: -2.4694 Pr(>|z|): 0.0135333 *
Interpretation: The installation cost is statistically significant (p-value < 0.05). The negative coefficient suggests that higher installation costs decrease the likelihood of choosing a particular heating system.

oc (Operating Cost)

Estimate: -0.00699637 Std. Error: 0.00155408 z-value: -4.5019 Pr(>|z|): 6.734e-06 ***
Interpretation: The operating cost is highly significant (p-value < 0.001). The negative coefficient suggests that higher operating costs decrease the likelihood of choosing a particular heating system.

Conclusion:

For Statistical Significance, both installation cost (ic) and operating cost (oc) are significant predictors of the choice of heating system.

Intercepts:the intercepts for gas room (gr) and heat pump (hp) heating systems are the only ones statistically significant, indicating that baseline levels are different from reference category

Question 2

```
# Load the data
setwd("~/Desktop/Spring 2024/Categorical Data/Project 2")
data <- read.csv("insurance.csv")
str(data)
```

```
## 'data.frame':    665249 obs. of  25 variables:
##  $ customer_ID      : int  10000000 10000000 10000000 10000000 10000000
10000000 10000000 10000000 10000000 10000005 ...
##  $ shopping_pt      : int  1 2 3 4 5 6 7 8 9 1 ...
##  $ record_type      : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ day              : int  0 0 0 0 0 0 0 0 0 3 ...
##  $ time             : chr  "08:35" "08:38" "08:38" "08:39" ...
##  $ state            : chr  "IN" "IN" "IN" "IN" ...
##  $ location         : int  10001 10001 10001 10001 10001 10001 10001 10001
10001 10006 ...
##  $ group_size       : int  2 2 2 2 2 2 2 2 2 1 ...
##  $ homeowner        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ car_age          : int  2 2 2 2 2 2 2 2 2 10 ...
##  $ car_value        : chr  "g" "g" "g" "g" ...
##  $ risk_factor      : int  3 3 3 3 3 3 3 3 3 4 ...
##  $ age_oldest       : int  46 46 46 46 46 46 46 46 46 28 ...
##  $ age_youngest     : int  42 42 42 42 42 42 42 42 42 28 ...
##  $ married_couple   : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ C_previous       : int  1 1 1 1 1 1 1 1 1 3 ...
##  $ duration_previous: int  2 2 2 2 2 2 2 2 2 13 ...
##  $ A                : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ B                : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ C                : int  2 2 2 2 2 2 2 2 2 3 ...
##  $ D                : int  2 2 2 2 2 2 2 2 2 3 ...
##  $ E                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ F                : int  2 2 2 2 2 2 2 2 2 0 ...
##  $ G                : int  2 1 1 1 1 1 1 1 1 2 ...
##  $ cost             : int  633 630 630 630 630 638 638 638 634 755 ...
```

```
data$risk_factor <- factor(data$risk_factor, ordered = TRUE)
model3 <- data[, c("risk_factor", "car_age", "car_value")]

# Remove missing values
model3 <- na.omit(model3)

# Fit the ordinal logistic regression model
model <- polr(risk_factor ~ car_age + car_value, data = model3, method =
"logistic")
# Example of model summary output
summary(model)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = risk_factor ~ car_age + car_value, data = model3,
##     method = "logistic")
##
## Coefficients:
##               Value Std. Error t value
## car_age     0.02637  0.0005263  50.108
## car_valuea -1.26734  0.1055340 -12.009
## car_valueb -0.78612  0.1008112  -7.798
## car_valuec -0.65874  0.0787322  -8.367
## car_valued -0.58091  0.0774143  -7.504
## car_valuee -0.61261  0.0773005  -7.925
## car_valuef -0.61565  0.0773562  -7.959
## car_valueg -0.71009  0.0775082  -9.161
## car_valueh -0.83103  0.0782151 -10.625
## car_valuei -0.85913  0.0839426 -10.235
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -1.6256   0.0775   -20.9717
## 2|3  -0.5831   0.0775    -7.5263
## 3|4   0.6178   0.0775     7.9737
##
## Residual Deviance: 1171644.66
## AIC: 1171670.66
```

Interpretation

Coefficients:

car_age: The coefficient for car_age is 0.02637 with a standard error of 0.0005263. This suggests that for each one-unit increase in car_age, the log odds of being in a higher risk factor category increase by 0.02637 units.

car_value (Categories a to i): These coefficients reflect the impact of each level of car_value relative to a baseline level (let's denote it as car_value level z).This implies that compared to car_value level z, being in car_valuea reduces the log odds of being in a higher risk factor category by 1.26734 units.

Intercepts: 1|2, 2|3, 3|4: These intercepts denote the thresholds between risk factor categories. the intercept for 1|2 is -1.6256. This indicates that the log odds of being in risk factor category 1 versus category 2 are -1.6256 when the predictors which are car_age and car_value are zero.

Residual Deviance and AIC:

Residual Deviance: It measures how effectively the model fits the data. Lower values signify better fit.

AIC (Akaike Information Criterion): It measures the model's goodness of fit while considering the number of parameters. Smaller values suggest better fit.

The model effectively captures the relationship between the predictors and the ordinal response variable . The coefficients give insights into how alterations in the predictors influence the log odds of being in higher risk factor categories.