Problem Set 3

Applied Stats II

Due: March 26, 2023

Question 1

1. **Construct and interpret an unordered multinomial logit with `GDPWdiff` as the output and "no change" as the reference category, including the estimated cutoff points and coefficients.**

```
# Data wrangling

dat <- mutate(dat, GDPWdiff = ifelse(GDPWdiff > 0, "positive", ifelse(
  GDPWdiff < 0, "negative", "no change")))

dat$GDPWdiff <- factor(dat$GDPWdiff,
levels = c("no change", "positive", "negative"),
ordered = FALSE)

# Construct the multinomial logit model
unord_model <- multinom(GDPWdiff ~ OIL + REG, dat)

# Display the model summary
summary(unord_model)

# Get the odds
exp(coef(unord_model))
```

Results:

```
> summary(unord_model)
Call:
multinom(formula = GDPWdiff ~ OIL + REG, data = dat)

Coefficients:
(Intercept)       OIL       REG
```

```
positive    4.533759 4.576321 1.769007
negative    3.805370 4.783968 1.379282

Std. Errors:
(Intercept)      OIL       REG
positive   0.2692006 6.885097 0.7670366
negative   0.2706832 6.885366 0.7686958

Residual Deviance: 4678.77
AIC: 4690.77


> exp(coef(unord_model))
(Intercept)       OIL      REG
positive    93.10789  97.15632 5.865024
negative    44.94186 119.57794 3.972047
```

Interpretation:
The odds of having a positive change in "GDPWdiff" are 93.11 times higher for each one-unit increase in the intercept (holding all other variables constant). The odds of having a positive change in "GDPWdiff" are 97.16 times higher for each one-unit increase in the "OIL" variable (holding all other variables constant). The odds of having a positive change in "GDPWdiff" are 5.87 times higher for each one-unit increase in the "REG" variable (holding all other variables constant).
The odds of having a negative change in "GDPWdiff" are 44.94 times higher for each one-unit increase in the intercept (holding all other variables constant). The odds of having a negative change in "GDPWdiff" are 119.58 times higher for each one-unit increase in the "OIL" variable (holding all other variables constant). The odds of having a negative change in "GDPWdiff" are 3.97 times higher for each one-unit increase in the "REG" variable (holding all other variables constant).

```r

# Now get z and p values

z <- summary(unord_model)$coefficients/summary(unord_model)$standard.
    errors
(p <- (1 - pnorm(abs(z), 0, 1)) * 2)


```

Results:

```
          (Intercept)       OIL        REG
positive           0 0.5062612 0.02109459
negative           0 0.4871792 0.07276308
```

Interpretation:
The results suggest that the predictor variable OIL does not have a statistically significant effect on both the positive and negative categories. The p-values for OIL are 0.5062612 and 0.4871792 for the positive and negative categories, respectively. Since both p-values are greater than the common threshold of 0.05, we cannot reject the null hypothesis that the coefficient for OIL is zero in either category.

The predictor variable REG, on the other hand, appears to have a statistically significant effect only on the positive category. The p-value for REG is 0.02109459 for the positive category, but it is 0.07276308 for the negative category. Since the p-value for the positive category is less than 0.05, we can reject the null hypothesis that the coefficient for REG is zero in the positive category, and conclude that REG has a statistically significant effect on the odds of being in the positive category compared to the reference category. However, we cannot reject the null hypothesis that the coefficient for REG is zero in the negative category.

This suggests that this is not a very helpful model.

```r

# To find cutoff points, generate predicted probabilities
pred_probs <- predict(unord_model, newdata = dat, type = "probs")

head(pred_probs)

# Estimate the cutoff points for each category using the quantile
    function
cutoff_NoChange <- quantile(pred_probs[,1], probs = 0.5)
cutoff_Positive <- quantile(pred_probs[,2], probs = 0.5)
cutoff_Negative <- quantile(pred_probs[,3], probs = 0.5)

# Create table of estimated cutoff points

```

```r
cutoff_table <- data.frame(
Category = c("no change", "positive", "negative"),
Cutoff = c(cutoff_NoChange, cutoff_Positive, cutoff_Negative)
)

print(cutoff_table)
```

Results:

```
   Category       Cutoff
1 no change  0.007191671
2  positive  0.669601291
3  negative  0.323207038
```

The cutoff for "no change" is the smallest of the three, at 0.007. This means that if the predicted probability for an observation being in the "no change" category is less than 0.007, the observation will be assigned to one of the other two categories.

The cutoff for "positive" is the largest of the three, at 0.67. This means that if the predicted probability for an observation being in the "positive" category is greater than 0.67, the observation will be assigned to the "positive" category. Conversely, if the predicted probability for an observation being in the "positive" category is less than or equal to 0.67, the observation will be assigned to one of the other two categories.

The cutoff for "negative" is in the middle, at 0.32. This means that if the predicted probability for an observation being in the "negative" category is greater than 0.32 but less than or equal to 0.67, the observation will be assigned to the "negative" category. If the predicted probability for an observation being in the "negative" category is less than or equal to 0.32, the observation will be assigned to the "no change" category.

2. **Construct and interpret an ordered multinomial logit with `GDPWdiff` as the outcome variable, including the estimated cutoff points and coefficients.**

```r

# Relevel
# set a reference level for the outcome
dat$GDPWdiff_ord <- factor(dat$GDPWdiff,
levels = c("negative", "no change", "positive"),
ordered = FALSE)

ord_model <- polr(GDPWdiff_ord ~ OIL + REG, data = dat, Hess = TRUE)
summary(ord_model)

 # Calculate a p value
ctable <- coef(summary(ord_model))
```

```r
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))

# Calculate confidence intervals
(ci <- confint(ord_model))

# convert to odds ratio
exp(cbind(OR = coef(ord_model), ci))


```

Results:

```
Call:
polr(formula = GDPWdiff_ord ~ OIL + REG, data = dat, Hess = TRUE)

Coefficients:
Value Std. Error t value
OIL  -0.1987    0.11572  -1.717
REG1  0.3985    0.07518   5.300

Intercepts:
Value    Std. Error t value
negative|no change  -0.7312   0.0476   -15.3597
no change|positive  -0.7105   0.0475   -14.9554

Residual Deviance: 4687.689
AIC: 4695.689
```

```r

# Calculate a p value
ctable <- coef(summary(ord_model))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))

# Calculate confidence intervals
(ci <- confint(ord_model))

# convert to odds ratio
exp(cbind(OR = coef(ord_model), ci))

```

Results:

```
                      Value Std. Error    t value      p value
OIL               -0.1987177 0.11571713  -1.717271 8.592967e-02
```

```
REG                  0.3984834 0.07518479   5.300054 1.157687e-07
negative|no change -0.7311784 0.04760375 -15.359680 3.050770e-53
no change|positive -0.7104851 0.04750680 -14.955440 1.435290e-50
> # Calculate confidence intervals
> (ci <- confint(ord_model))


2.5 %     97.5 %
OIL -0.4237548 0.03019571
REG  0.2516548 0.54643410
```

```r

# To estimate cutoff points:
# First, generate predicted probabilities
pred_probs2 <- predict(ord_model, newdata = dat, type = "probs")

head(pred_probs2)

# Estimate the cutoff points for each category using the quantile
    function
cutoff_Negative2 <- quantile(pred_probs2[,1], probs = 0.5)
cutoff_NoChange2 <- quantile(pred_probs2[,2], probs = 0.5)
cutoff_Positive2 <- quantile(pred_probs2[,3], probs = 0.5)


# Create table of estimated cutoff points

cutoff_table2 <- data.frame(
Category = c("no change", "positive", "negative"),
Cutoff = c( cutoff_Negative2, cutoff_NoChange2, cutoff_Positive2)
)

print(cutoff_table2)

```

Results:

```
   Category      Cutoff
 1 no change 0.324936186
 2  positive 0.004555476
 3  negative 0.670508338
```

Interpretation: For OIL, the odds ratio is 0.8197813, with a 95% confidence interval ranging from 0.6545844 to 1.030656. This suggests that for a one-unit increase in OIL, the odds of the dependent variable being in a higher category are reduced by 18% (i.e., the odds of the dependent variable being in a lower category are increased by 23.4%)

compared to when OIL is held constant. However, the confidence interval contains 1, suggesting that this effect may not be statistically significant.

For REG, the odds ratio is 1.4895639, with a 95% confidence interval ranging from 1.2861520 to 1.727083. This suggests that for a one-unit increase in REG, the odds of the dependent variable being in a higher category are increased by 49% compared to when REG is held constant. Moreover, the confidence interval does not contain 1, indicating that this effect is statistically significant.

In summary, these results suggest that REG is a statistically significant predictor of the dependent variable, whereas the effect of OIL on the dependent variable may not be statistically significant.

The cutoff for "no change" is the highest of the three, at 0.325. This means that if the predicted probability for an observation being in the "no change" category is greater than 0.325, the observation will be assigned to the "no change" category.

The cutoff for "positive" is the lowest of the three, at 0.005. This means that if the predicted probability for an observation being in the "positive" category is less than or equal to 0.005, the observation will be assigned to the "negative" category.

The cutoff for "negative" is in the middle, at 0.67. This means that if the predicted probability for an observation being in the "negative" category is greater than 0.67, the observation will be assigned to the "negative" category.

Question 2

1. **[(a)] Run a Poisson regression because the outcome is a count variable. Is there evidence that PAN presidential candidates visit swing districts more? Provide a test statistic and p-value.**

```
model <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.
    governor.06,
data = mexico, family = "poisson")

# Summarize model results
summary(model)
```

```
Call:
glm(formula = PAN.visits.06 ~ competitive.district + marginality.06 +
PAN.governor.06, family = "poisson", data = mexico)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.2309  -0.3748  -0.1804  -0.0804  15.2669
```

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.81023    0.22209 -17.156   <2e-16 ***
competitive.district -0.08135    0.17069  -0.477   0.6336
marginality.06       -2.08014    0.11734 -17.728   <2e-16 ***
PAN.governor.06      -0.31158    0.16673  -1.869   0.0617 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1473.87  on 2406  degrees of freedom
Residual deviance:  991.25  on 2403  degrees of freedom
AIC: 1299.2

Number of Fisher Scoring iterations: 7
```

The resulting z-value for the coefficient of competitive.district is -0.477, and the corresponding p-value is 0.6336. This means that the coefficient is not statistically significant at the 5% level. Therefore, there is no evidence of a difference in the expected number of visits between swing and non-swing districts, after controlling for economic and social marginality in the district, and whether the district has a PAN governor or not.

2. **[(b)] Interpret the `marginality.06` and `PAN.governor.06` coefficients.**

   The marginality.06 coefficient is -2.08014. This means that, holding other variables constant, a one-unit increase in the marginality score (which measures poverty) is associated with a decrease in the expected log-count of PAN visits by 2.08014 units. In other words, the more impoverished a district is, the less likely it is that the PAN presidential candidate visited it.
   The PAN.governor.06 coefficient is -0.31158. This means that, holding other variables constant, being in a state with a PAN-affiliated governor is associated with a decrease in the expected log-count of PAN visits by 0.31158 units. However, this coefficient is only marginally significant (p-value = 0.0617), meaning that we cannot say with certainty that the presence of a PAN-affiliated governor had a significant effect on the number of PAN visits.

3. **[(c)] Provide the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive (`competitive.distr`**

had an average poverty level (`marginality.06 = 0`), and a PAN governor (`PAN.governor.06=1`).

```
# Create a data frame with the predictor values for the hypothetical district
hypothetical_district <- data.frame(competitive.district = 1,
marginality.06 = 0,
PAN.governor.06 = 1)

# Use the predict() function to estimate the mean number of visits for the hypothetical district

pred_mex <- cbind(predict(model, hypothetical_district , type ="response", se.fit = TRUE), hypothetical_district)

# Print the estimated mean number of visits
print(pred_mex)

```

Results:

fit	se.fit	residual.scale	competitive.district	marginality.06	PAN.governo
0.0149	0.0032	1	1	0	1

```
# Alternative model using the model equation:
model.equation = -3.81023 - 0.08135*1 - 2.08014*0 - 0.31158*1
exp(model.equation)

```

The resulting mean is 0.01494827, as in the fit column in the table above.