401-hw8-moyi

2024-11-29

Problem 2

Problem 2a

```
# Load the NewsDesert.csv data
news_desert <- read.csv("/Users/homura/Desktop/NewsDesert.csv")

# Corrected categorization using the cut function
news_desert$pub3.2023 <- cut(news_desert$Cpub2023,
                             breaks = c(-1, 0, 1, Inf),
                             labels = c("0 newspapers", "1 newspaper", "2+ newspapers"),
                             right = TRUE)

# Generate a frequency distribution table for the newly created variable
table_pub3_2023 <- table(news_desert$pub3.2023)
table_pub3_2023
```

```
##
##  0 newspapers   1 newspaper 2+ newspapers
##           203          1628          1309
```

Problem 2b

```
# Fit multinomial logistic regression using demographics
library(nnet)
multinom_demo <- multinom(pub3.2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 + Lhisp2021, dat
a = news_desert)
```

```
## # weights:  21 (12 variable)
## initial  value 3449.642586
## iter  10 value 2771.253914
## iter  20 value 2514.660221
## final  value 2514.659879
## converged
```

```
# Assuming multinom_demo is your fitted model
model_summary <- summary(multinom_demo)
print(model_summary)
```

```
## Call:
## multinom(formula = pub3.2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 +
##     Lhisp2021, data = news_desert)
##
## Coefficients:
##               (Intercept)          age     SES21 Lpopdens2021  Lblack2021
## 1 newspaper     0.5566863 0.0183136912 0.0998117    0.4935043 -0.09776332
## 2+ newspapers   0.1046673 0.0001901162 0.4838967    0.8487873 -0.34891350
##                 Lhisp2021
## 1 newspaper   -0.10510627
## 2+ newspapers  0.03309927
##
## Std. Errors:
##               (Intercept)        age      SES21 Lpopdens2021 Lblack2021
## 1 newspaper     0.7283781 0.01439817 0.09106152   0.07496172 0.07208587
## 2+ newspapers   0.7689371 0.01521511 0.09501946   0.07767673 0.07774019
##                Lhisp2021
## 1 newspaper   0.07927516
## 2+ newspapers 0.08440726
##
## Residual Deviance: 5029.32
## AIC: 5053.32
```

```
# Extracting coefficients
coefficients_one_newspaper <- coef(model_summary)["1 newspaper",]
coefficients_two_plus_newspapers <- coef(model_summary)["2+ newspapers",]

# Calculating the missing logit for 1 vs 2
missing_logit <- coefficients_two_plus_newspapers - coefficients_one_newspaper

# Print all
print("0 vs 1")
```

```
## [1] "0 vs 1"
```

```
print(coefficients_one_newspaper)
```

```
##  (Intercept)          age        SES21 Lpopdens2021   Lblack2021    Lhisp2021
##   0.55668631   0.01831369   0.09981170   0.49350431  -0.09776332  -0.10510627
```

```
print("1 vs 2+")
```

```
## [1] "1 vs 2+"
```

```
print(missing_logit)
```

```
##  (Intercept)          age        SES21 Lpopdens2021   Lblack2021    Lhisp2021
##  -0.45201897  -0.01812358   0.38408501   0.35528301  -0.25115018   0.13820555
```

```
print("0 vs 2+")
```

```
## [1] "0 vs 2+"
```

```
print(coefficients_two_plus_newspapers)
```

```
##   (Intercept)           age         SES21  Lpopdens2021    Lblack2021
##  0.1046673437  0.0001901162  0.4838967074  0.8487873203 -0.3489134987
##     Lhisp2021
##  0.0330992750
```

Logit: 0 Newspapers vs. 1 Newspaper
The comparison between "0 newspapers" and "1 newspaper" reveals the influence of predictors on the odds of having at least one newspaper compared to none. **Socioeconomic status (SES21)** has a positive coefficient of 0.0998, indicating that higher SES slightly increases the likelihood of having 1 newspaper compared to none. Similarly, **population density (Lpopdens2021)** has a strong positive effect (0.4935), showing that counties with higher population density are significantly more likely to have 1 newspaper compared to none. **Age** has a positive coefficient of 0.0183, meaning that counties with older average populations are slightly more likely to have 1 newspaper compared to none. In contrast, **percentage of Black residents (Lblack2021)** (-0.0978) and **percentage of Hispanic residents (Lhisp2021)** (-0.1051) negatively impact the odds, suggesting that counties with higher proportions of Black or Hispanic residents are less likely to have 1 newspaper compared to no newspapers.

Logit: 1 Newspaper vs. 2+ Newspapers
When comparing "1 newspaper" to "2+ newspapers," the predictors reveal their impact on moving from a single newspaper to multiple newspapers. **SES21** has a missing logit of 0.3841 ($0.4839 - 0.0998$), indicating that higher socioeconomic status strongly increases the odds of having 2+ newspapers compared to just 1. **Population density (Lpopdens2021)** also positively influences this transition, with a missing logit of 0.3554 ($0.8489 - 0.4935$). **Age** contributes slightly to this transition, with a missing logit of -0.0182 ($0.0001 - 0.0183$), suggesting that younger populations may slightly favor having 2+ newspapers compared to 1. Conversely, **Lblack2021** (-0.2511) shows that higher proportions of Black residents decrease the likelihood of having 2+ newspapers compared to 1, and **Lhisp2021** (0.1382) suggests a slight positive association between higher proportions of Hispanic residents and the odds of having 2+ newspapers compared to 1.

Logit: 0 Newspapers vs. 2+ Newspapers
The comparison between "0 newspapers" and "2+ newspapers" highlights the factors influencing the most substantial difference in newspaper availability. **SES21** has the largest positive impact, with a coefficient of 0.4839, showing that higher SES greatly increases the likelihood of having 2+ newspapers compared to none. **Population density (Lpopdens2021)** (0.8489) is also a strong predictor, indicating that counties with denser populations are far more likely to have 2+ newspapers compared to none. **Age** has a negligible coefficient of 0.0001, suggesting little to no influence of age on this comparison. On the other hand, **Lblack2021** (-0.3489) negatively impacts the odds, showing that counties with higher Black populations are less likely to have 2+ newspapers compared to none. **Lhisp2021** (0.0331) exhibits a slight positive relationship, suggesting a weak association between Hispanic population percentages and having 2+ newspapers compared to none.

Problem 2c

```
# Fit multinomial logistic regression using AR1+ variables
library(nnet)
news_desert$log_numPub18 <- log(news_desert$Cpub2018 + 1)

multinom_ar1 <- multinom(pub3.2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 + Lhisp2021 + log
_numPub18, data = news_desert)
```

```
## # weights:  24 (14 variable)
## initial  value 3449.642586
## iter  10 value 1923.031250
## iter  20 value 888.565585
## iter  30 value 881.476305
## iter  40 value 881.465595
## final  value 881.462890
## converged
```

```
# Assuming multinom_ar1 is your fitted model
model_summary_ar1 <- summary(multinom_ar1)
print(model_summary_ar1)
```

```
## Call:
## multinom(formula = pub3.2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 +
##     Lhisp2021 + log_numPub18, data = news_desert)
##
## Coefficients:
##               (Intercept)        age     SES21 Lpopdens2021 Lblack2021
## 1 newspaper     -6.291909 0.03482058 0.1363311    0.2676928 0.04279924
## 2+ newspapers  -16.013669 0.02034290 0.2748213    0.2572602 0.02170045
##                  Lhisp2021 log_numPub18
## 1 newspaper   -0.009514818     10.77356
## 2+ newspapers -0.157327352     21.29050
##
## Std. Errors:
##               (Intercept)        age     SES21 Lpopdens2021 Lblack2021
## 1 newspaper      1.525641 0.02619604 0.1526711    0.1218633  0.1249710
## 2+ newspapers    1.751381 0.03030737 0.1738604    0.1406284  0.1448944
##               Lhisp2021 log_numPub18
## 1 newspaper   0.1420486    0.9851144
## 2+ newspapers 0.1666341    1.0608268
##
## Residual Deviance: 1762.926
## AIC: 1790.926
```

```
# Extracting coefficients
coefficients_one_newspaper <- coef(model_summary_ar1)["1 newspaper", ]
coefficients_two_plus_newspapers <- coef(model_summary_ar1)["2+ newspapers", ]

# Calculating the missing logit for 1 vs 2+
missing_logit_ar1 <- coefficients_two_plus_newspapers - coefficients_one_newspaper

# Print all
print("0 vs 1")
```

```
## [1] "0 vs 1"
```

```
print(coefficients_one_newspaper)
```

```
##  (Intercept)          age        SES21 Lpopdens2021   Lblack2021    Lhisp2021
## -6.291908593  0.034820583  0.136331098  0.267692848  0.042799243 -0.009514818
## log_numPub18
## 10.773561952
```

```
print("1 vs 2+")
```

```
## [1] "1 vs 2+"
```

```
print(missing_logit_ar1)
```

```
##  (Intercept)          age         SES21 Lpopdens2021   Lblack2021    Lhisp2021
##  -9.72176072  -0.01447768   0.13849021  -0.01043260  -0.02109880  -0.14781253
## log_numPub18
##  10.51694138
```

```
print("0 vs 2+")
```

```
## [1] "0 vs 2+"
```

```
print(coefficients_two_plus_newspapers)
```

```
##  (Intercept)          age         SES21 Lpopdens2021   Lblack2021    Lhisp2021
## -16.01366931   0.02034290   0.27482131   0.25726025   0.02170045  -0.15732735
## log_numPub18
##  21.29050333
```

Logit: 0 Newspapers vs. 1 Newspaper
The coefficients for the comparison between "0 newspapers" and "1 newspaper" highlight the factors influencing the likelihood of having at least one newspaper compared to none. **Log(numPub18)**, with a coefficient of 10.7736, has the largest positive impact, suggesting that counties with a higher number of newspapers in 2018 are far more likely to have at least one newspaper in 2023 compared to none. **Socioeconomic status (SES21)** (0.1363) and **population density (Lpopdens2021)** (0.2677) also have strong positive effects, indicating that counties with higher SES and denser populations are significantly more likely to maintain at least one newspaper. On the other hand, **age** (0.0348) has a smaller but still positive effect, suggesting that counties with older populations are slightly more likely to have one newspaper compared to none. **Percentage of Black residents (Lblack2021)** (0.0428) shows a minor positive association, while **percentage of Hispanic residents (Lhisp2021)** (-0.0095) has a negligible negative effect.

Logit: 1 Newspaper vs. 2+ Newspapers
The missing logit for "1 newspaper" vs. "2+ newspapers" (calculated as 10.5169 for log(numPub18), $0.1389 - 0.1363 = 0.0026$ for SES21, etc.) reveals the factors influencing the likelihood of transitioning from one newspaper to multiple newspapers. **Log(numPub18)** (10.5169) remains the dominant predictor, showing that a strong historical trend of newspaper publication dramatically increases the likelihood of having 2+ newspapers compared to just 1. **SES21** (0.0026) and **Lpopdens2021** (-0.0104) have minimal effects in this comparison, suggesting that socioeconomic status and population density contribute less to differentiating counties with 1 vs. 2+ newspapers. **Age** has a small negative coefficient (-0.0145), indicating that counties with younger populations are slightly more likely to transition from 1 newspaper to 2+. **Percentage of Black residents (Lblack2021)** (-0.0210) and **percentage of Hispanic residents (Lhisp2021)** (-0.1478) show small negative associations, suggesting these demographic factors slightly reduce the likelihood of having multiple newspapers compared to just one.

Logit: 0 Newspapers vs. 2+ Newspapers
The comparison between "0 newspapers" and "2+ newspapers" reveals the most pronounced differences. **Log(numPub18)** (21.2905) has an overwhelmingly large positive coefficient, underscoring the critical role of historical newspaper trends in determining whether a county has multiple newspapers versus none. **SES21** (0.2748) and **Lpopdens2021** (0.2573) also have substantial positive effects, highlighting the importance of socioeconomic conditions and population density in supporting the availability of multiple newspapers. **Age** (0.0203) has a smaller positive effect, showing that older populations are somewhat more likely to have multiple newspapers compared to none. In contrast,

Lblack2021 (-0.1573) and **Lhisp2021** (-0.1572) both have negative coefficients, indicating that higher proportions of Black and Hispanic residents reduce the likelihood of having 2+ newspapers compared to none, though these effects are much smaller in magnitude compared to the influence of past newspaper trends.

Problem 2d

```
# Load required libraries
library(caret) # For confusion matrix and metrics
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
library(nnet)  # For multinomial regression

# Predictions for both models
# For demographic model
pred_demo <- predict(multinom_demo, news_desert)

# For AR1+ model
pred_ar1 <- predict(multinom_ar1, news_desert)

# Create confusion matrices
confusion_demo <- confusionMatrix(factor(pred_demo, levels = levels(news_desert$pub3.2023)),
                                  factor(news_desert$pub3.2023, levels = levels(news_desert$pu
b3.2023)))

confusion_ar1 <- confusionMatrix(factor(pred_ar1, levels = levels(news_desert$pub3.2023)),
                                 factor(news_desert$pub3.2023, levels = levels(news_desert$pub
3.2023)))

# Extract metrics for both models
metrics_demo <- list(
  accuracy = confusion_demo$overall['Accuracy'],
  precision = confusion_demo$byClass[,"Precision"],
  recall = confusion_demo$byClass[,"Recall"],
  F1 = confusion_demo$byClass[,"F1"]
)

metrics_ar1 <- list(
  accuracy = confusion_ar1$overall['Accuracy'],
  precision = confusion_ar1$byClass[,"Precision"],
  recall = confusion_ar1$byClass[,"Recall"],
  F1 = confusion_ar1$byClass[,"F1"]
)

# Macro metrics (average of per-class metrics)
macro_demo <- list(
  macro_precision = mean(metrics_demo$precision, na.rm = TRUE),
  macro_recall = mean(metrics_demo$recall, na.rm = TRUE),
  macro_F1 = mean(metrics_demo$F1, na.rm = TRUE)
)

macro_ar1 <- list(
  macro_precision = mean(metrics_ar1$precision, na.rm = TRUE),
  macro_recall = mean(metrics_ar1$recall, na.rm = TRUE),
  macro_F1 = mean(metrics_ar1$F1, na.rm = TRUE)
)

# Print results for both models
print("Demographic Model Metrics")
```

```
## [1] "Demographic Model Metrics"
```

```r
print(metrics_demo)
```

```
## $accuracy
##  Accuracy
## 0.6063694
##
## $precision
##  Class: 0 newspapers   Class: 1 newspaper Class: 2+ newspapers
##                   NA            0.5999049            0.6194605
##
## $recall
##  Class: 0 newspapers   Class: 1 newspaper Class: 2+ newspapers
##            0.0000000            0.7745700            0.4912147
##
## $F1
##  Class: 0 newspapers   Class: 1 newspaper Class: 2+ newspapers
##                   NA            0.6761394            0.5479335
```

```
print(macro_demo)
```

```
## $macro_precision
## [1] 0.6096827
##
## $macro_recall
## [1] 0.4219282
##
## $macro_F1
## [1] 0.6120365
```

```
print("AR1+ Model Metrics")
```

```
## [1] "AR1+ Model Metrics"
```

```
print(metrics_ar1)
```

```
## $accuracy
##  Accuracy
## 0.9216561
##
## $precision
##  Class: 0 newspapers   Class: 1 newspaper Class: 2+ newspapers
##            1.0000000            0.9523499            0.8812629
##
## $recall
##  Class: 0 newspapers   Class: 1 newspaper Class: 2+ newspapers
##            0.7438424            0.8961916            0.9809015
##
## $F1
##  Class: 0 newspapers   Class: 1 newspaper Class: 2+ newspapers
##            0.8531073            0.9234177            0.9284165
```

```
print(macro_ar1)
```

```
## $macro_precision
## [1] 0.9445376
##
## $macro_recall
## [1] 0.8736452
##
## $macro_F1
## [1] 0.9016472
```

Easier Classes:
The AR1+ model shows that “1 newspaper” and “2+ newspapers” are relatively easier to predict. This is evident from their consistently high precision, recall, and F1-scores in both models, particularly in the AR1+ model. The “1 newspaper” class likely benefits from being the intermediate category, making it easier to identify counties that do not fall into the extremes of “0 newspapers” or “2+ newspapers.” Similarly, the “2+ newspapers” class is strongly linked to predictors like historical newspaper counts and socioeconomic factors, making it more distinguishable.

Harder Classes:
The “0 newspapers” class is the most challenging to predict, especially in the demographic model. The complete absence of predictions for this class in the demographic model highlights the difficulty of identifying counties without any newspapers using only demographic factors. Even in the AR1+ model, while precision is perfect for “0 newspapers,” the recall is lower, suggesting that some true instances of this class are still missed. This difficulty may stem from the smaller representation of this class in the dataset or weaker associations between the predictors and this outcome.

Problem 5

Problem 5a

```
# Fit Poisson regression model with demographic variables
poisson_demo <- glm(Cpub2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 + Lhisp2021,
                    family = poisson(link = "log"),
                    data = news_desert)

# Summary of the model
summary(poisson_demo)
```

```
##
## Call:
## glm(formula = Cpub2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 +
##     Lhisp2021, family = poisson(link = "log"), data = news_desert)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.95154    0.13639  -6.976 3.03e-12 ***
## age           0.01054    0.00273   3.861 0.000113 ***
## SES21         0.14832    0.01462  10.146  < 2e-16 ***
## Lpopdens2021  0.28120    0.01038  27.088  < 2e-16 ***
## Lblack2021   -0.07448    0.01401  -5.317 1.06e-07 ***
## Lhisp2021     0.19242    0.01513  12.719  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5943.5  on 3139  degrees of freedom
## Residual deviance: 3813.3  on 3134  degrees of freedom
## AIC: 11003
##
## Number of Fisher Scoring iterations: 5
```

Key Variables Associated with More News Organizations
The results indicate that several variables are positively associated with an increase in the expected number of newspapers. **Population density (`Lpopdens2021`)** has the largest positive coefficient (0.2812), meaning that counties with higher population density are strongly associated with having more newspapers. Similarly, **socioeconomic status (`SES21`)**, with a coefficient of 0.1483, is another significant predictor, indicating that counties with better socioeconomic conditions are more likely to support news organizations. **Percentage of Hispanic residents (`Lhisp2021`)**, with a coefficient of 0.1924, also positively influences the expected number of newspapers, suggesting that counties with higher Hispanic populations are more likely to sustain local news outlets. Additionally, **age** has a smaller positive effect (0.0105), indicating that counties with slightly older populations tend to have more newspapers, though this effect is not as strong as the others.

Variables Associated with Fewer News Organizations
In contrast, **percentage of Black residents (`Lblack2021`)** is negatively associated with the expected number of newspapers, with a coefficient of -0.0745. This indicates that counties with higher proportions of Black residents tend to have fewer newspapers, after controlling for other demographic factors. The negative association may point to broader structural inequalities affecting access to and sustainability of news organizations in these communities.

Variables with Lesser Importance
All variables in the model are statistically significant ($p < 0.05$), as shown by their very low p-values, meaning they all contribute meaningfully to predicting the number of newspapers. However, the magnitude of the coefficients indicates that population density, socioeconomic status, and percentage of Hispanic residents have the largest impact, while age and percentage of Black residents, though significant, have relatively smaller effects.

Problem 5b

```
# Add log-transformed lagged variable to the model
poisson_ar1 <- glm(Cpub2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 + Lhisp2021 + log(Cpub20
18 + 1),
                   family = poisson(link = "log"),
                   data = news_desert)

# Summary of the model
summary(poisson_ar1)
```

```
##
## Call:
## glm(formula = Cpub2023 ~ age + SES21 + Lpopdens2021 + Lblack2021 +
##     Lhisp2021 + log(Cpub2018 + 1), family = poisson(link = "log"),
##     data = news_desert)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.6723404  0.1420348  -4.734 2.21e-06 ***
## age               -0.0008444  0.0029285  -0.288    0.773
## SES21              0.0372922  0.0157615   2.366    0.018 *
## Lpopdens2021      -0.0001608  0.0123430  -0.013    0.990
## Lblack2021        -0.0048498  0.0147297  -0.329    0.742
## Lhisp2021         -0.0051893  0.0158615  -0.327    0.744
## log(Cpub2018 + 1)  1.1555516  0.0198531  58.205  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5943.47  on 3139  degrees of freedom
## Residual deviance:  650.08  on 3133  degrees of freedom
## AIC: 7841.9
##
## Number of Fisher Scoring iterations: 4
```

The inclusion of `log(Cpub2018 + 1)` significantly improves the model, as evidenced by its highly positive coefficient (1.1556) and an extremely small p-value ($< 2 \times 10^{-16}$). This result highlights that historical newspaper trends are the strongest predictor of current newspaper counts, suggesting that counties with a robust history of newspaper publications are much more likely to sustain news organizations in 2023. While **socioeconomic status (SES21)** retains a positive and statistically significant effect (0.0373, $p = 0.018$), its magnitude is greatly reduced compared to the model in Problem 5a. This indicates that much of the predictive power of SES is explained by historical newspaper trends.

The demographic variables `Lpopdens2021` (population density), `Lblack2021` (percentage of Black residents), and `Lhisp2021` (percentage of Hispanic residents) all lose their statistical significance in this model. Their coefficients are close to zero and non-significant ($p > 0.05$), suggesting that their contributions are now negligible after including the historical trend variable. Similarly, **age** becomes non-significant ($p = 0.773$), with a near-zero coefficient (-0.0008), indicating no meaningful relationship with the number of newspapers in 2023. This shift reflects the overwhelming importance of historical newspaper counts, which overshadow the demographic factors in predicting current newspaper availability.

The inclusion of `log(Cpub2018 + 1)` dramatically reduces the residual deviance (from **3813.3** in Problem 5a to **650.08** in this model) and improves the Akaike Information Criterion (AIC) from **11003** to **7841.9**, indicating a much better fit. This improvement underscores the critical role of historical newspaper trends in predicting current counts, far outweighing the predictive power of demographic variables alone.

Problem 5c

```
# Use the AR1+ Poisson model to estimate the probabilities for Y = 0
# Predicted expected counts (lambda) from the Poisson AR1+ model
poisson_lambda <- predict(poisson_ar1, type = "response")

# Probability of Y = 0 (news desert) using the Poisson distribution
poisson_prob_0 <- dpois(0, lambda = poisson_lambda)

# Predicted probabilities for Y = 0 from the multinomial model
multinom_pred <- predict(multinom_ar1, type = "probs")
multinom_prob_0 <- multinom_pred[, "0 newspapers"]

# Combine probabilities into a data frame
probabilities <- data.frame(
  Poisson_Prob_0 = poisson_prob_0,
  Multinomial_Prob_0 = multinom_prob_0
)

# Create a scatterplot matrix
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(
  probabilities,
  title = "Scatterplot Matrix: Poisson vs Multinomial Probabilities"
)
```