401-hw7

2024-11-18

Problem 3b

```
# Define the FPR (False Positive Rate) and TPR (True Positive Rate) values
fpr <- c(0.000, 0.064, 0.157, 0.307, 0.496, 0.733, 1.000)
tpr <- c(0.000, 0.294, 0.552, 0.740, 0.879, 0.958, 1.000)

# Calculate the differences in FPR (x-axis)
delta_fpr <- diff(fpr)

# Calculate the average TPR values between consecutive points (y-axis)
average_tpr <- (tpr[-length(tpr)] + tpr[-1]) / 2

# Apply the trapezoidal rule: Sum of (delta_fpr * average_tpr)
auc <- sum(delta_fpr * average_tpr)

# Print the AUC value
cat("The Area Under the Curve (AUC) is:", auc, "\n")
```

```
## The Area Under the Curve (AUC) is: 0.77772
```

Problem 4a

```
# Load necessary library
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
```

```
# Assuming your dataset is loaded as `news_data`
news_data <- read.csv("/Users/homura/Desktop/NewsDesert.csv")

# Create a new variable `atrisk` where Cpub2023 <= 1
news_data <- news_data %>%
  mutate(atrisk = ifelse(Cpub2023 <= 1, 1, 0))

# Display the first few rows to confirm the new column
head(news_data)
```

```
##   FIPS State  county Cpub2023 Cpub2018 Lpopdens2021  age Lhisp2021 Lblack2021
## 1 1001    AL Autauga        0        0     3.642994 38.5  1.386294  3.0301337
## 2 1003    AL Baldwin        4        4     3.790036 43.4  1.740466  2.2721259
## 3 1005    AL Barbour        1        1     2.466580 40.2  1.740466  3.8753590
## 4 1007    AL    Bibb        1        1     2.695947 39.7  1.335001  3.1311369
## 5 1009    AL  Blount        1        1     3.581949 41.1  2.351375  0.8329091
## 6 1011    AL Bullock        1        1     2.003454 38.7  2.251292  4.2484952
##        SES21 atrisk
## 1  0.7949343      1
## 2  0.9758666      0
## 3 -1.8064779      1
## 4 -0.9729884      1
## 5 -0.8365552      1
## 6 -2.0483254      1
```

Problem 4b

```
# Logistic regression using demographic variables only
logit_demo <- glm(atrisk ~ age + SES21 + Lpopdens2021 + Lhisp2021 + Lblack2021,
                  data = news_data,
                  family = binomial)

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

```
##
## Call:
## glm(formula = atrisk ~ age + SES21 + Lpopdens2021 + Lhisp2021 +
##     Lblack2021, family = binomial, data = news_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.717358   0.413120   1.736   0.0825 .
## age           0.016118   0.008196   1.967   0.0492 *
## SES21        -0.388993   0.046148  -8.429  < 2e-16 ***
## Lpopdens2021 -0.398728   0.034632 -11.513  < 2e-16 ***
## Lhisp2021    -0.120036   0.046345  -2.590   0.0096 **
## Lblack2021    0.260243   0.041879   6.214 5.16e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4265.8  on 3139  degrees of freedom
## Residual deviance: 3818.3  on 3134  degrees of freedom
## AIC: 3830.3
##
## Number of Fisher Scoring iterations: 4
```

```
# Exponentiated coefficients for interpretation (odds ratios)
exp(coef(logit_demo))
```

```
##  (Intercept)          age        SES21 Lpopdens2021    Lhisp2021   Lblack2021
##    2.0490129    1.0162490    0.6777391    0.6711731    0.8868883    1.2972451
```

Variables that increase the probability of being at risk (odds ratios > 1) are age (OR = 1.016, p < 0.05) and Lblack2021 (OR = 1.297, p < 0.001). This means that for each unit increase in age, the odds of becoming a news desert increase by 1.6%, and for each unit increase in the percentage of Black residents, the odds increase by 29.7%. Variables that decrease the probability of being at risk (odds ratios < 1) are SES21 (OR = 0.678, p < 0.001), Lpopdens2021 (OR = 0.671, p < 0.001), and Lhisp2021 (OR = 0.887, p < 0.01). This indicates that for each unit increase in socioeconomic status, population density, and percentage of Hispanic residents, the odds of becoming a news desert decrease by 32.2%, 32.9%, and 11.3% respectively.

Problem 4c

```
# Logistic regression using AR1+ model
logit_ar1 <- glm(atrisk ~ log(Cpub2018 + 1) + age + SES21 + Lpopdens2021 + Lhisp2021 + Lblack2021,
                 data = news_data,
                 family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# Summary of the model
summary(logit_ar1)
```

```
##
## Call:
## glm(formula = atrisk ~ log(Cpub2018 + 1) + age + SES21 + Lpopdens2021 +
##     Lhisp2021 + Lblack2021, family = binomial, data = news_data)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         9.836441   0.885516  11.108   <2e-16 ***
## log(Cpub2018 + 1) -10.595367   0.389356 -27.213   <2e-16 ***
## age                 0.014042   0.015755   0.891   0.3728
## SES21              -0.139328   0.085911  -1.622   0.1049
## Lpopdens2021        0.007554   0.072437   0.104   0.9169
## Lhisp2021           0.148021   0.089649   1.651   0.0987 .
## Lblack2021          0.020765   0.075617   0.275   0.7836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4265.8  on 3139  degrees of freedom
## Residual deviance: 1282.3  on 3133  degrees of freedom
## AIC: 1296.3
##
## Number of Fisher Scoring iterations: 7
```

```
# Exponentiated coefficients for interpretation (odds ratios)
exp(coef(logit_ar1))
```

```
##       (Intercept) log(Cpub2018 + 1)               age              SES21
##      1.870303e+04      2.503171e-05      1.014141e+00      8.699429e-01
##      Lpopdens2021         Lhisp2021        Lblack2021
##      1.007583e+00      1.159537e+00      1.020982e+00
```

In the AR1+ model, the strongest and most significant predictor is log(Cpub2018 + 1) with a very large negative coefficient (β = -10.59, p < 0.001), indicating that counties with more newspapers in 2018 are much less likely to become news deserts. Most notably, while demographic variables (age, SES21, Lpopdens2021, and Lblack2021) were all significant in the demographics-only model, they become non-significant in the AR1+ model (all p > 0.05). Only Lhisp2021 shows marginal significance (p < 0.1) in the AR1+ model. This dramatic change in significant variables suggests that the current number of newspapers is such a strong predictor that it essentially overshadows the demographic factors - in other words, knowing a county's current newspaper count is far more informative for predicting future news deserts than knowing its demographic characteristics.

Problem 4d

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
# Predict probabilities for both models
prob_demo <- predict(logit_demo, type = "response")
prob_ar1 <- predict(logit_ar1, type = "response")

# Create ROC objects
roc_demo <- roc(news_data$atrisk, prob_demo)
```

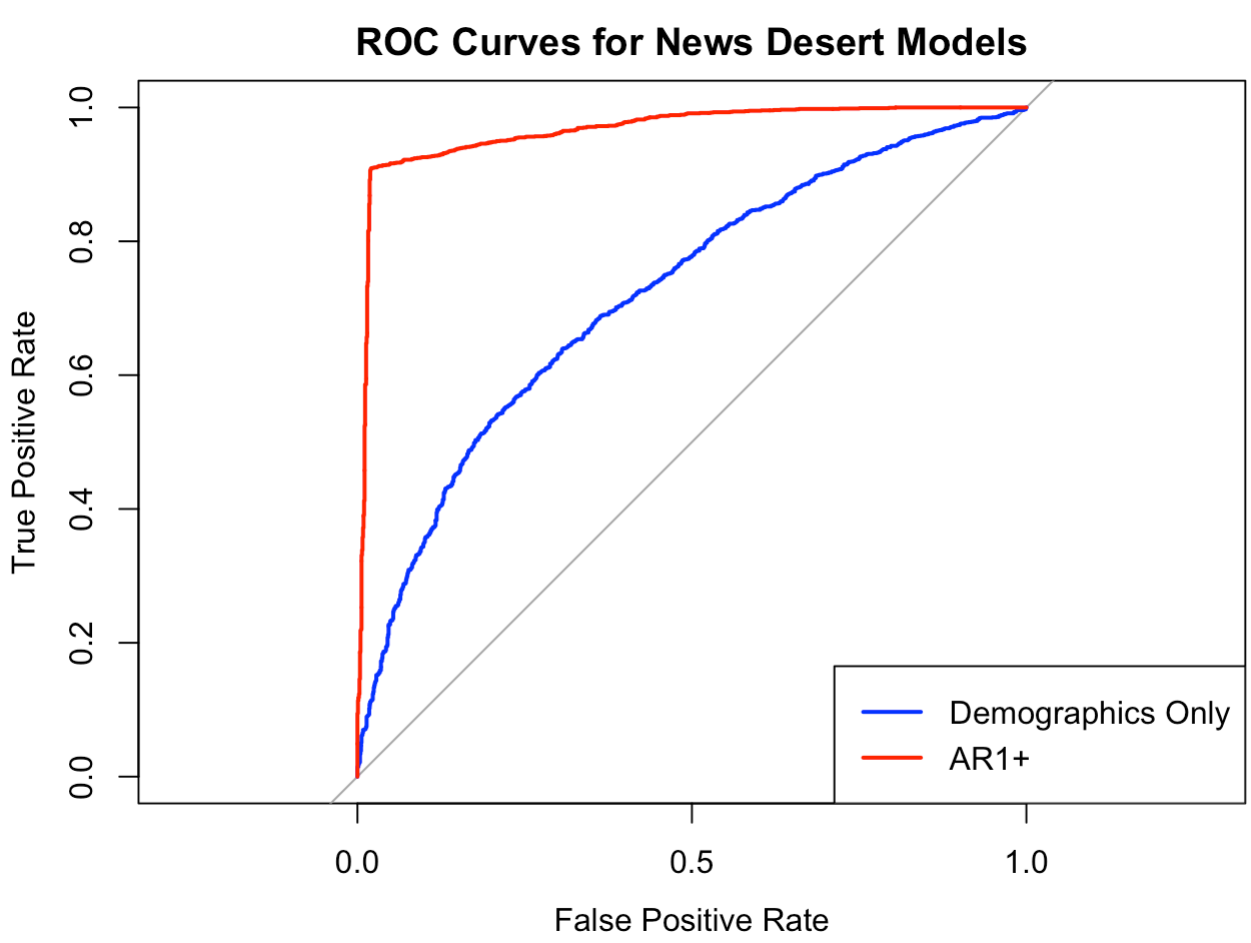
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
roc_ar1 <- roc(news_data$atrisk, prob_ar1)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
# Plot with corrected settings
plot(roc_demo,
     col = "blue",
     main = "ROC Curves for News Desert Models",
     legacy.axes = TRUE,  # This ensures proper FPR axis
     xlab = "False Positive Rate",
     ylab = "True Positive Rate")
lines(roc_ar1, col = "red")
legend("bottomright",
       legend = c("Demographics Only", "AR1+"),
       col = c("blue", "red"),
       lwd = 2)
```




```
# Calculate and display AUC for both models
auc_demo <- auc(roc_demo)
auc_ar1 <- auc(roc_ar1)

cat("AUC for Demographics model:", auc_demo, "\n")
```

```
## AUC for Demographics model: 0.7212801
```

```
cat("AUC for AR1+ model:", auc_ar1, "\n")
```

```
## AUC for AR1+ model: 0.9664045
```