MLDS 401/IEMS404: Homework 7
Due: November 30, 15:00
Professor Malthouse

1. Use the estimates from the toxicity problem. Generate an ROC curve and find the area under the curve. You have summarized data and I would like for you to generate the ROC curve "by hand." Hint: there are $g = 6$ values of $x = 1, \dots, 6$. Let $\hat{p}_x$ be the predicted probability for x using the logistic regression model. *Answer: I have computed the table exactly below. Note that there are 669 positives and* $1500-669 = 831$ *negatives, which provide the denominators for the TPR and FPR columns.*

Cut value	# yes (Cum)	TPR	FPR	Area
$0 \le c < 0.123$	0 (0)	$1 - \frac{0}{669} = 1$	$1 - \frac{0}{831} = 1$	
$0.123 \le c < 0.215$	28 (28)	$1 - \frac{28}{669} = 0.958$	$1 - \frac{222}{831} = 0.733$	0.2616
$0.215 \le c < 0.349$	53 (81)	$1 - \frac{81}{669} = 0.879$	$1 - \frac{419}{831} = 0.496$	0.2178
$0.349 \le c < 0.512$	93 (174)	$1 - \frac{174}{669} = 0.740$	$1 - \frac{576}{831} = 0.307$	0.1529
$0.512 \le c < 0.673$	126 (300)	$1 - \frac{300}{669} = 0.552$	$1 - \frac{700}{831} = 0.158$	0.0964
$0.673 \le c < 0.801$	172 (472)	$1 - \frac{472}{669} = 0.295$	$1 - \frac{778}{831} = 0.064$	0.0397
$0.801 \le c \le 1$	197 (197)	$1 - \frac{669}{669} = 0$	$1 - \frac{831}{831} = 0$	0.0094
Total	669			0.77768

Answer: See table above for .77768. For example, $.2616 = \frac{1}{2}(1 + .9581)(1 - .7329)$. *Here is my R code.*

```
# this is used to set things up for the plot.roc function
toxlong = data.frame(
  x = c(rep(1,250), rep(2,250), rep(3,250), rep(4,250),
    rep(5,250), rep(6,250)),
  y = c(
    rep(1, 28), rep(0, 250-28), rep(1, 53), rep(0, 250-53),
    rep(1, 93), rep(0, 250-93), rep(1, 126), rep(0, 250-126),
    rep(1, 172), rep(0, 250-172), rep(1, 197), rep(0, 250-197)
  )
)
fit2 = glm(y~x, binomial, toxlong) # estimates and SE match prob 1
summary(fit2)
library(pROC)
plot.roc(toxlong$y, fit2$fitted.values, print.auc=T)

myroc = data.frame(
  tpr = c(1, 1-28/669, 1-89/669, 1-174/669, 1-300/669, 1-472/669, 0),
  fpr = c(1, 1-222/831, 1-419/831, 1-576/831, 1-700/831, 1-778/831, 0)
)
plot(myroc$fpr, myroc$tpr, type="l")
```

2. Suppose we have a sample of size n where observation i consists of dependent variable Y_i, a multinomial RV taking values $\{1, \ldots, K\}$, and $(p+1)$-vector of predictor variables $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\mathsf{T}$. Let $\boldsymbol{\alpha}_k$ be a $(p+1)$-vector of regression coefficients. Let $\pi_{ik} = \mathsf{P}(Y_i = k)$ for $k = 1, \ldots, K$ and

$$\log \pi_{ik} = \boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i - \log Z, \qquad (k = 1, \ldots, K)$$

where log is the natural log function and the term $\log Z$ ensures that the probabilities sum to one, i.e., $\sum_{k=1}^K \pi_{ik} = 1$.

(a) Show that $Z = \sum_{k=1}^K \exp(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i)$. *Answer: Exponentiating both sides of the* $\log \pi_{ik}$ *expression we get* $\pi_{ik} = \exp(\alpha_k^\mathsf{T} \mathbf{x}_i)/Z$.

$$1 = \sum_{k=1}^K \pi_{ik} = \sum_{k=1}^K \frac{\exp(\alpha_k^\mathsf{T} \mathbf{x}_i)}{Z} = \frac{1}{Z} \sum_{k=1}^K \exp(\alpha_k^\mathsf{T} \mathbf{x}_i).$$

The result follows by multiplying both sizes by Z.

(b) Show that $\pi_{ik} = \exp(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i)/Z$. This is called the softmax function. *Answer: From the previous part,*

$$\pi_{ik} = \frac{\exp(\alpha_k^\mathsf{T} \mathbf{x}_i)}{Z}$$

Substitute the value Z *from the previous part to get the result.*

(c) The usual formulation of the multinomial logit from class picks a base category (WLOG class 1) and assumes:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \boldsymbol{\beta}_k^\mathsf{T} \mathbf{x}_i, \qquad (k = 2, \ldots, K)$$

How is $\boldsymbol{\beta}_k$ related to $\boldsymbol{\alpha}_k$? You will see that multinomial and softmax are just reparameterizations of each other. *Answer: Exponentiate both sides and substitute the result from part b to see that* $\beta_k = \alpha_k - \alpha_1$:

$$\exp(\boldsymbol{\beta}_k^\mathsf{T} \mathbf{x}_i) = \frac{\pi_{ik}}{\pi_{i1}} = \frac{\exp(\alpha_k^\mathsf{T} \mathbf{x}_i)}{Z} \cdot \frac{Z}{\exp(\alpha_1^\mathsf{T} \mathbf{x}_i)} = \frac{\exp(\alpha_k^\mathsf{T} \mathbf{x}_i)}{\exp(\alpha_1^\mathsf{T} \mathbf{x}_i)} = \exp\left[(\alpha_k - \alpha_1)^\mathsf{T} \mathbf{x}_i\right]$$

3. This problem studies news deserts. You have data for (nearly) every county in the US:

- `numPub23`: number of newspapers published for the county in 2023. This count is the **dependent variable**.
- `numPub18`: number of newspapers published for the county in 2018. With a five-year period this is a lagged version of the dependent variable.
- `age`: average age in county in 2021

- `SES21`: socioeconomic status (average of income and education)
- `Lpopdens2021`: population density of the county in 2021
- `Lblack2021`: percent of county that is black in 2021
- `Lhisp2021`: percent of county that is Hispanic in 2021

The goal is to build a predictive model forecasting which counties are likely to be news deserts in five years. We will consider two models

- **demographic** use `age, SES21, Lpopdens2021, Lblack2021 and Lhisp2021` as predictors
- **AR1+** Use the `log(numPub18+1)` and the demographics as predictors.

```
> dat = read.csv("NewsDesert.csv") %>%
+   mutate(atrisk=as.integer(Cpub2023>=1),
+     pub3.2023 = cut(Cpub2023, c(-.5, .5, 1.5, 999), c("0", "1", "2+")))
> fit1 = glm(atrisk ~ age + SES21 + Lpopdens2021 + Lblack2021 + Lhisp2021,
    family=binomial, data=dat)
> summary(fit1)
              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 ***
Lblack2021    0.260243   0.041879   6.214 5.16e-10 ***
Lhisp2021    -0.120036   0.046345  -2.590   0.0096 **

    Null deviance: 4265.8  on 3139  degrees of freedom
Residual deviance: 3818.3  on 3134  degrees of freedom

> fit2 = glm(atrisk ~ log(Cpub2018+1) + age + SES21 + Lpopdens2021 + Lblack2021
    + Lhisp2021, family=binomial, data=dat)
> summary(fit2)
                    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
Lblack2021          0.020765   0.075617   0.275   0.7836
Lhisp2021           0.148021   0.089649   1.651   0.0987 .
---
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

> plot.roc(dat$atrisk, fit1$fitted.values, print.auc = T, print.auc.y=.7,
    legacy.axes=T, xlab = "False Positive Rate",
    ylab = "True Positive Rate (Recall)")
> plot.roc(dat$atrisk, fit2$fitted.values, print.auc=T, print.auc.y=.9,
    print.auc.x=0.8, col=2, add=T)
> legend("bottomright", c("Demographic", "AR1+"), col=1:2, lty=1)
```

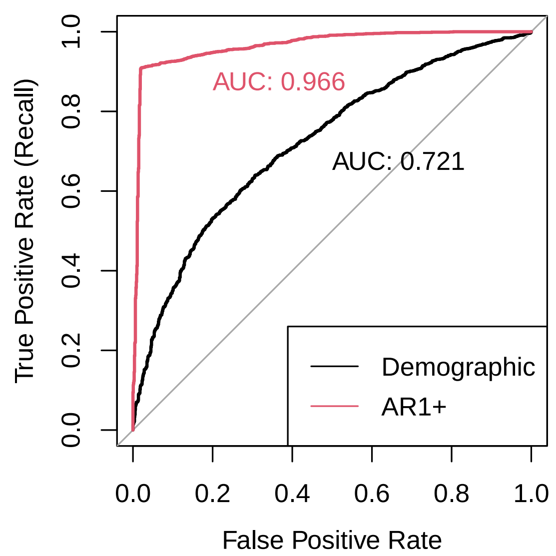
(a) Create a new variable `atrisk` that equals 1 if the county is at risk (`numPub23` ≤ 1) and 0 otherwise.

(b) Use a logistic regression to predict `atrisk` from the demographics only. Which variables increase the probability of being at risk? Which decrease the probability? *Answer: In descending order of the absolute magnitude of the z statistics, population density has a negative association with being at risk ($z = -11.5$), higher SES is associated with lower levels of at risk ($z = -8.4$), black is associated with being at risk (6.2), higher levels of Hispanic are associated with lower levels of at risk (-2.6), and age has a positive association with being at risk (2.0).*

(c) Use a logistic regression to predict `atrisk` from the AR1+ variables. Interpret the model. How do you explain the difference in significant variables? *Answer: The only variable that is significant is `log(Cpub2018+1)`. We have a pipe situation.*

(d) Create an ROC curve showing the predicted values from the two models on the same plot. Find AUC for each of the two models.

Answer: see plot



4. This problem also uses the news desert data.

```
> table(dat$pub3.2023) # part a

   0    1   2+
 203 1628 1309

> rbind(
+   coef(mult1),
+   twoplusover1 = apply(coef(mult1), 2, diff) )
     (Intercept)           age     SES21 Lpopdens2021  Lblack2021   Lhisp2021
1      0.5566863  0.0183136912 0.0998117    0.4935043 -0.09776332 -0.10510627
2+     0.1046673  0.0001901162 0.4838967    0.8487873 -0.34891350  0.03309927
2+/1  -0.4520190 -0.0181235750 0.3840850    0.3552830 -0.25115018  0.13820555

> mult2 <- multinom(pub3.2023~ log(Cpub2018+1) + age + SES21 + Lpopdens2021
    + Lblack2021 + Lhisp2021, data = dat, maxit = 1000)
> rbind(
+   coef(mult2),
+   twoplusover1 = apply(coef(mult2), 2, diff) )
     (Intercept) log(Cpub2018 + 1)         age     SES21 Lpopdens2021
1      -6.291909          10.77356  0.03482058 0.1363311    0.2676928
2+    -16.013669          21.29050  0.02034290 0.2748213    0.2572602
2+/1   -9.721761          10.51694 -0.01447768 0.1384902   -0.0104326
        Lblack2021    Lhisp2021
1       0.04279924 -0.009514818
2+      0.02170045 -0.157327352
```

```
2+/1   -0.02109880 -0.147812533
```

(a) Create a variable pub3.2023 that takes three values: 0 newspapers, 1 newspaper, or 2+ newspapers. Submit a frequency distribution (table).

(b) Use a multinomial regression to predict pub3.2023 from the demographics. Find the missing logit. Interpret all three logits (0 vs. 1, 1 vs. 2+ and 0 vs. 2+). *Answer: See output above. The base category is 0. We don't have z statistics. Population density, age and SES all have positive associations, and black and Hispanic has a negative associations, with the logit of 1 newspaper verusus 0. Population density, age, SES, and Hispanic have positive associations, and black has a negative association, with the logit of 2+ versus 0 newspapers. Population density, SES and Hispanic have positive associations, and age and black have negative associations, with the logit of 2+ versus 1 newspaper.*

(c) Use a multinomial regression to predict pub3.2023 from the AR1+ variables. Find the missing logit and interpret the model. *Answer: the coefficients for lagged newspaper counts are very large.*

(d) For your two models, find accuracy; per-class precision, recall and F_1; and macro precision, recall and F_1. What do you conclude about which classes can be easily distinguished versus those that are more difficult to predict? *Answer: With model 1 (demographics) no cases are predicted to have 0 NPs and consequently precision is undefined, since we cannot divide by a column total of 0. With precision undefined F_1 is also undefined. Per-class recall suggest that it is easier to identify the 1's (recall=0.77) than the 2+'s (recall=0.49). For AR1+, all of the counties classified as $Y = 0$ actually have no newspapers giving perfect per-class precision, although $48 + 4$ actual 0's are misclassified. Class 0 is smaller and the threshhold to be predicted a 0 is high, giving the high precision. Predictions of 1 have precision 0.95 and predictions of 2+ have precision 0.88. As for recall, the 0's are the most difficult to identify (recall=0.74), 1's have recall=0.90 and 2+'s have recall=0.98—it's easist to classify places with a lot of NPs and most difficult to classify the actual 0s, answering the question about thisch classes are most difficult to predict.*

```
> # Evaluate mult1
> predicted = factor(apply(mult1$fitted.values, 1, which.max), 1:3, c("P0", "P1", "P2
> (cm=table(actual=dat$pub3.2023, predicted)) # confusion matrix
      predicted
actual   P0   P1  P2+
    0     0  175   28
    1     0 1261  367
    2+    0  666  643
> (rowsums = apply(cm, 1, sum)) # number of instances per class
   0    1   2+
```

```
 203 1628 1309
> (colsums = apply(cm, 2, sum)) # number of predictions per class
  P0   P1  P2+
   0 2102 1038
> (recall = diag(cm) / rowsums) # per-class recall
        0         1        2+
0.0000000 0.7745700 0.4912147
> mean(recall) # macro Recall
[1] 0.4219282

> # Evaluate mult2
> predicted = factor(apply(mult2$fitted.values, 1, which.max), 1:3, c("P0", "P1", "P2
> (cm=table(actual=dat$pub3.2023, predicted)) # confusion matrix
      predicted
actual   P0   P1  P2+
    0   151   48    4
    1     0 1459  169
    2+    0   25 1284
> (rowsums = apply(cm, 1, sum)) # number of instances per class
   0    1   2+
 203 1628 1309
> (colsums = apply(cm, 2, sum)) # number of predictions per class
  P0   P1  P2+
 151 1532 1457
> (precision = diag(cm) / colsums) # per-class precision
       P0        P1       P2+
1.0000000 0.9523499 0.8812629
> (recall = diag(cm) / rowsums) # per-class recall
        0         1        2+
0.7438424 0.8961916 0.9809015
> (f1 = 2 * precision * recall / (precision + recall) ) # per-class f
       P0        P1       P2+
0.8531073 0.9234177 0.9284165
> mean(precision) # macro Precision
[1] 0.9445376
> mean(recall) # macro Recall
[1] 0.8736452
> mean(f1) # macro F1
[1] 0.9016472
```

5. Return to problem 4 from homework 5 using data from the German book company.

   (a) You estimated a model in part d using the logs of tof, r, f and $m+1$, and in part e you applied it to the test set. Compute a gains table using the test-set data.

(b) How much money do you expect to make per customer if you used this model to select 40% of the names to be contacted? *Answer: 6.04*

```
> fit = lm(logtarg ~ log(tof) + log(r) + log(ford) + log(m+1), all[train, ])
> yhat = predict(fit, all[!train,])
> gains(yhat, as.integer(all$target[!train]>0), all$target[!train]) # from class notes
# A tibble: 5 × 13
  qtile     n Nrespond    amt RespRate AvgAmt  CumN CumResp CumAmt CumRespRate CumAvgAmt liftResp liftAmt
  <fct> <int>    <int>  <dbl>    <dbl>  <dbl> <int>   <int>  <dbl>       <dbl>     <dbl>    <dbl>   <dbl>
1 Q1     2246      397 19461.   0.177   8.66   2246     397 19461.      0.177      8.66     2.47    2.66
2 Q2     2246      188  7667.   0.0837  3.41   4492     585 27128.      0.130      6.04     1.82    1.86
3 Q3     2246      105  4779.   0.0467  2.13   6738     690 31907.      0.102      4.74     1.43    1.45
4 Q4     2246       63  2466.   0.0280  1.10   8984     753 34373.      0.0838     3.83     1.17    1.18
5 Q5     2246       52  2179.   0.0232  0.970 11230     805 36552.      0.0717     3.25     1       1
```

(c) What fraction of customers will respond if you use this model to select 40% of the names? *Answer: 0.13*

(d) The next two parts estimate a two-step model using the training data only. This part estimates the **response model**. Create a variable `buy` that equals 1 if the customer bought (i.e., `target`> 0). Estimate a logistic regression predicting `buy` from any variables you wish. This estimates **conversion probabilities**, $\hat{\pi}_i$. What variables are predictive in this model? *Answer: b*

```
> all$buy = as.integer(all$target>0)
> fit2 = glm(buy ~ log(tof) + log(r) + log(ford), binomial, all[train, ])
> summary(fit2)

Call:
glm(formula = buy ~ log(tof) + log(r) + log(ford), family = binomial,
    data = all[train, ])

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.60239    0.11119 -32.399  < 2e-16 ***
log(tof)    -0.41898    0.07973  -5.255 1.48e-07 ***
log(r)      -0.25022    0.04540  -5.511 3.56e-08 ***
log(ford)    0.81564    0.08329   9.793  < 2e-16 ***
```

(e) Now estimate a **conditional demand model** using the training data only. To do so, regress `logtarg` on some predictor variables using only buyers in the training set. This estimates the log spending amount of buyers, $\hat{y}_i$. What variables are predictive? *Answer: b*

```
> fit3 = lm(logtarg ~ log(ford) +log(m+1), all, subset=train&buy)
> summary(fit3)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.10822    0.21180   5.233 2.68e-07 ***
log(ford)   -0.65907    0.06910  -9.537  < 2e-16 ***
log(m + 1)   0.67446    0.05907  11.417  < 2e-16 ***
```

(f) Apply the response and conditional demand models to the test set and multiply $\hat{\pi}_1 e^{\hat{y}_i}$ and use this score to create a gains table. Which model is better @40%? The one from homework 5 or the twostep? *Answer: The amount goes up to 6.39 (from 6.04), but the response rate drops slightly to 0.127 (from 0.130).*

```
> yhat = predict(fit2, all[!train,], type="resp") * exp(predict(fit3, all[!train,]))
> gains(yhat, as.integer(all$target[!train]>0), all$target[!train])
# A tibble: 5 × 13
  qtile     n Nrespond    amt RespRate AvgAmt  CumN CumResp CumAmt CumRespRate CumAvgAmt liftResp liftAmt
  <fct> <int>    <int>  <dbl>    <dbl>  <dbl> <int>   <int>  <dbl>       <dbl>     <dbl>    <dbl>   <dbl>
1 Q1     2246      368 20195.   0.164   8.99   2246     368 20195.      0.164      8.99     2.29    2.76
2 Q2     2246      203  8523.   0.0904  3.79   4492     571 28717.      0.127      6.39     1.77    1.96
3 Q3     2246      115  3331.   0.0512  1.48   6738     686 32048.      0.102      4.76     1.42    1.46
4 Q4     2246       70  2815.   0.0312  1.25   8984     756 34863.      0.0841     3.88     1.17    1.19
5 Q5     2246       49  1689.   0.0218  0.752 11230     805 36552.      0.0717     3.25     1       1
```