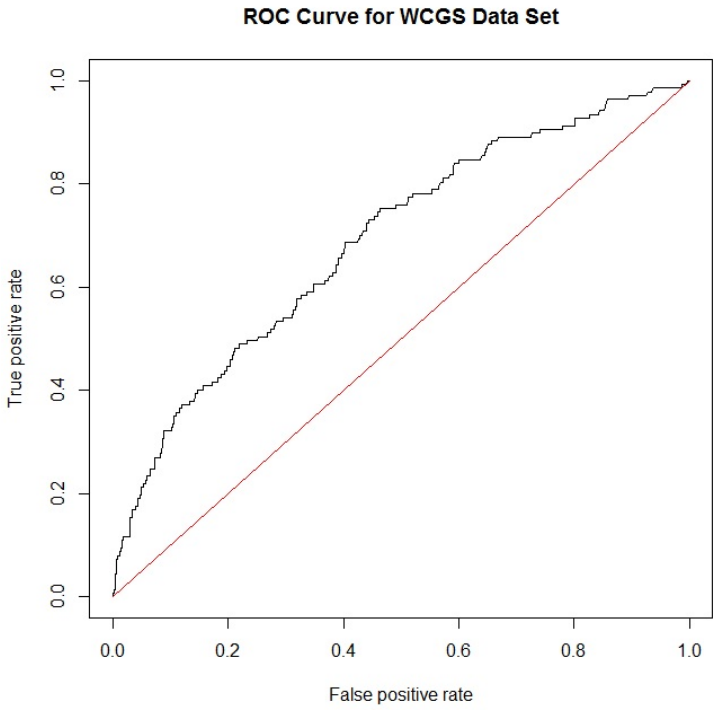
Stat 6021: Guided Question Set 10 Solutions

1. (a) From the previous guided question set, we went with the model with age, sbp, and ncigs as the predictors, dropping dbp from the model, as it was the only insignificant predictor in the model.

   (b) The ROC curve is shown below. Since the curve is in the top-left side of the diagonal line, our model does better than random guessing.



   (c) From R, the AUC is 0.6899. An AUC of 0.5 indicates a model that is as effective as random guessing. An AUC of 1 indicates a model that is 100% effective.

   Note about set.seed(): There has been a change in how random numbers are being generated. The previous method of generating random numbers was not was not as uniform as thought. So a change was made in R version 3.6.0.

   - For R versions 3.6.0 and later, you are able to specify if you want to use the "Rejection" sampler (the current one) or the old "Rounding" sampler.

- Type `RNGkind(sample.kind = "Rejection")` or `RNGkind(sample.kind = "Rounding")` for the needed sampler. Type this on a line before using the `set.seed()` function.
- You should use the rejection sampler. The rounding sampler should be used only if you want to reproduce results that used R versions before 3.6.0.
- Based on the documentation, the rejection sampler should be the default for versions 3.6.0 and later. However, about one-third of my students in my in-residence classes who use version 3.6.0 or later report having the rounding sampler as the default, but they were all able to use the RNGkind() function to switch to the rejection sampler.
- If you use R versions before 3.6.0, I advise you to update. These versions do not have the capability to choose the sampler. Type `version` in your console to see which version of R you are using.
- All my examples are worked using version 3.6.0 with the rejection sampler.
- If you are using the old rounding sampler, the AUC will be 0.6764.

(d) The confusion matrix with 0.5 as the cutoff:

```
> table(test$chd69, preds>0.5)
    FALSE
  0  1440
  1   137
```

We note that

- Among males who do not have heart disease, 0 out of 1440 are incorrectly classified as having heart disease. We have a false positive rate of 0.
- Among males who do have heart disease, 0 out of 119 are correctly classified as having heart disease. We have a true positive rate of 0.
- Overall error rate of $\frac{137}{1440} = 0.095$.
- What this table informs us is that everyone was classified as not having heart disease. This is equivalent to a classifier that decides to classify everyone as not having heart disease regardless of the person's values on the predictors.

The confusion matrix with 0.1 as the cutoff:

```
> table(test$chd69, preds>0.1)
    FALSE TRUE
  0  1160  280
  1    78   59
```

We note that

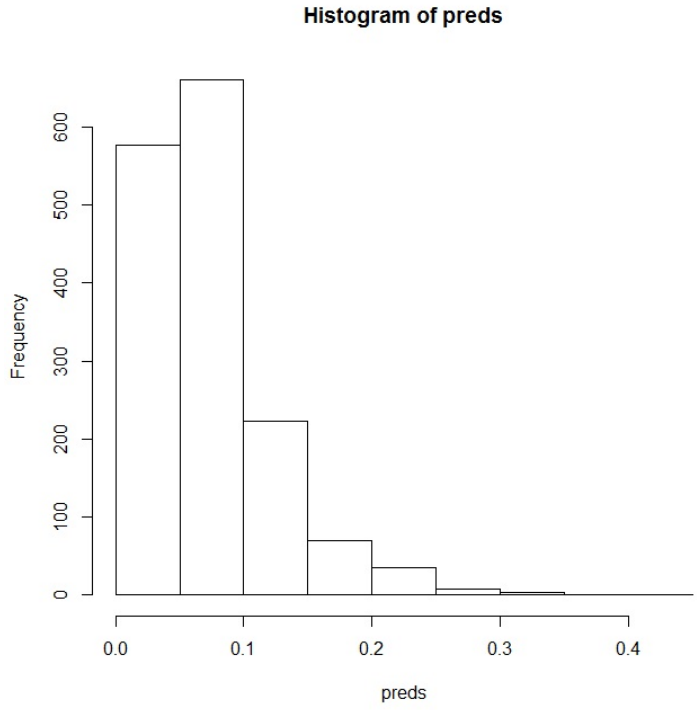
- Among males who do not have heart disease, 280 out of 1440 (0.194) are incorrectly classified as having heart disease. We have a small false positive rate, but the false positive rate has increased.

- Among males who do have heart disease, 59 out of 137 (0.431) are correctly classified as having heart disease. The true positive rate has improved (but still isn't great).
- Overall error rate of $\frac{280+78}{1440} = 0.249$. Notice how the overall error rate has increased.

The table below gives the counts for the test data. Notice the small proportion of males who have heart disease.

```
> table(test$chd69)
   0    1
1440  137
```

The histogram below gives the distribution of the predicted probabilities of having heart disease for the test data. Notice how very few observations have predicted probabilities that are greater than 0.5, or even greater than 0.2.



A few comments:

- The predicted probability of having heart disease is small among middle-aged men.
- Using a cutoff of 0.5 to classify an observation as having heart disease may be unrealistic in this context, since we are modeling a rare event.
- Lowering the cutoff may be more reasonable, especially if the consequence of having a large false negative rate is more serious (incorrectly classifying a male with heart disease as not having heart disease).
- Lowering the cutoff will result in more subjects being classified as having heart disease, thus increasing the true positive rates and false positive rates.

- Using a cutoff of 0.5 will minimize the overall error rate (number of false positives and false negatives over sample size). Is our consideration overall error rate, or reducing the false positive rate or reducing the false negative rate? Consultation with a subject matter expert will be needed.

2. (a) See R script

   (b)

```
Call:
multinom(formula = crime_level ~ dis + ptratio)

Coefficients:
       (Intercept)      dis    ptratio
low       8.639437 2.253808 -0.7487965
medium   10.969841 1.501224 -0.7551356

Std. Errors:
       (Intercept)       dis   ptratio
low       2.409789 0.2479895 0.1202263
medium    2.350523 0.2409497 0.1161566
```

(c)

```
> z
       (Intercept)      dis   ptratio
low       3.585143 9.088322 -6.228224
medium    4.666980 6.230445 -6.501011

> p
        (Intercept)          dis      ptratio
low    3.368931e-04 0.000000e+00 4.717531e-10
medium 3.056592e-06 4.651117e-10 7.978196e-11
```

(d) All the estimated coefficients are statistically significant. The baseline class is high-crime.

- The coefficients for distance are positive, and the coefficients for ptratio are negative.
- The estimated relative risk of being in a low-crime area versus being in a high-crime area multiplied by a factor of $\exp(2.25) = 9.5$ as the weighted distance of the tract from the employment centers increases by one-unit, while keeping student-teacher ratio constant.
- The estimated relative risk of being in a low-crime area versus being in a high-crime area multiplied by a factor of $\exp(1.50) = 4.5$ as the weighted distance of the tract from the employment centers increases by one-unit, while keeping student-teacher ratio constant.
- The estimated relative risk of being in a low-crime area versus being in a high-crime area multiplied by a factor of $\exp(-0.75) = 0.47$ as the student-teacher ratio increases by one-unit, while keeping the weighted distance of the tract from the employment centers constant.
- The estimated relative risk of being in a low-crime area versus being in a high-crime area multiplied by a factor of $\exp(-0.76) = 0.47$ as the student-teacher ratio increases by one-unit, while keeping the weighted distance of the tract from the employment centers constant.