# Homework 7

Brady Lamson

2022-04-03

## Appendix E

### Problem 5

```
helprtc <-
    mosaicData::HELPrct %>%
    mutate(
        homeless01 =
            case_when(
                homeless == "housed" ~ 0,
                homeless == "homeless" ~ 1
            )
    ) %>%
    select(where(is.numeric))
```

Before we start I want to `skimr::skim()` the data to get a general overview of what I'm working with.

```
helprtc %>%
    skimr::skim()
```

Table 1: Data summary

| | |
|---|---|
| Name | Piped data |
| Number of rows | 453 |
| Number of columns | 22 |
| Column type frequency: | |
| numeric | 22 |
| Group variables | None |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| age | 0 | 1.00 | 35.65 | 7.71 | 19.00 | 30.00 | 35.00 | 40.00 | 60.00 | |
| anysubstatus | 207 | 0.54 | 0.77 | 0.42 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | |
| cesd | 0 | 1.00 | 32.85 | 12.51 | 1.00 | 25.00 | 34.00 | 41.00 | 60.00 | |

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| d1 | 0 | 1.00 | 3.06 | 6.19 | 0.00 | 1.00 | 2.00 | 3.00 | 100.00 | |
| daysanysub | 209 | 0.54 | 75.31 | 79.24 | 0.00 | 5.00 | 33.00 | 164.25 | 268.00 | |
| dayslink | 22 | 0.95 | 255.61 | 151.02 | 2.00 | 74.00 | 361.00 | 365.00 | 456.00 | |
| drugrisk | 1 | 1.00 | 1.89 | 4.34 | 0.00 | 0.00 | 0.00 | 1.00 | 21.00 | |
| e2b | 239 | 0.47 | 2.50 | 2.52 | 1.00 | 1.00 | 2.00 | 3.00 | 21.00 | |
| female | 0 | 1.00 | 0.24 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | |
| i1 | 0 | 1.00 | 17.91 | 20.02 | 0.00 | 3.00 | 13.00 | 26.00 | 142.00 | |
| i2 | 0 | 1.00 | 24.55 | 28.02 | 0.00 | 4.00 | 18.00 | 33.00 | 184.00 | |
| id | 0 | 1.00 | 233.40 | 134.75 | 1.00 | 119.00 | 233.00 | 348.00 | 470.00 | |
| indtot | 0 | 1.00 | 35.73 | 7.15 | 4.00 | 32.00 | 38.00 | 41.00 | 45.00 | |
| linkstatus | 22 | 0.95 | 0.38 | 0.49 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | |
| mcs | 0 | 1.00 | 31.68 | 12.84 | 6.76 | 21.68 | 28.60 | 40.94 | 62.18 | |
| pcs | 0 | 1.00 | 48.05 | 10.78 | 14.07 | 40.38 | 48.88 | 56.95 | 74.81 | |
| pss_fr | 0 | 1.00 | 6.71 | 4.00 | 0.00 | 3.00 | 7.00 | 10.00 | 14.00 | |
| sexrisk | 0 | 1.00 | 4.64 | 2.80 | 0.00 | 3.00 | 4.00 | 6.00 | 14.00 | |
| avg_drinks | 0 | 1.00 | 17.91 | 20.02 | 0.00 | 3.00 | 13.00 | 26.00 | 142.00 | |
| max_drinks | 0 | 1.00 | 24.55 | 28.02 | 0.00 | 4.00 | 18.00 | 33.00 | 184.00 | |
| hospitalizations | 0 | 1.00 | 3.06 | 6.19 | 0.00 | 1.00 | 2.00 | 3.00 | 100.00 | |
| homeless01 | 0 | 1.00 | 0.46 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | |

There are a few big takeaways from this skimming.

- First is a large number of NAs. I'm going to outright remove the variables with 200+ missing values as that's nearly half the data set. For the other NAs I'll do median imputation, that is replacing the NA with the median value of that variable.
- Second are many variables that do not appear to follow a normal distribution. Sadly the `skim()` functions console histograms don't show up on pdf, but trust me here! A few of these numeric variables seem log-linear though, I can log transform those to help the model work better.
- Finally is the scale of our numeric variables is all over the place. I'll need to normalize all of these variables if I want my model to be able to glean any important information from the data. There is also have an **id** column tucked away in there, I'll need to handle that.
- There's another problem, Looking at some variables there are a few that are identical. I'm not sure what the best way to handle this is outside of manually selecting out the ones I catch. Below are all the variables I caught.

```
helprtc %>%
    select(avg_drinks, i1, max_drinks, i2, hospitalizations, d1) %>%
    head()
```

```
##   avg_drinks i1 max_drinks i2 hospitalizations d1
## 1         13 13         26 26                3  3
## 2         56 56         62 62               22 22
## 3          0  0          0  0                0  0
## 4          5  5          5  5                2  2
## 5         10 10         13 13               12 12
## 6          4  4          4  4                1  1
```

Here we setup a recipe, this a convenient way to tackle a lot of the data pre-processing I want to do. This makes life a whole lot easier when working with a data set that's a little moody.

```r
homeless_recipe <-
    recipe(homeless01 ~., data = helprtc) %>%
    # Make homeless01 a factor ----
    step_mutate(homeless01 = homeless01 %>% as.factor()) %>%

    # Remove duplicate variables and variables w/ over 200+ NAs ----
    step_select(-c(i1, i2, d1, anysubstatus, daysanysub, e2b)) %>%

    # Set id column to be an id, not a predictor ----
    update_role(id, new_role = "id") %>%

    # Do median imputation for variables with missing values ----
    step_impute_median(dayslink, drugrisk, linkstatus) %>%

    # Normalize numeric predictors so they're on the same scale ----
    step_normalize(all_numeric_predictors()) %>%

    # Log transform variables that appear log-normal to help it approach a normal dist ----
    step_log(avg_drinks, max_drinks, indtot, age, signed = TRUE)

homeless_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##         id          1
##    outcome          1
##  predictor         20
##
## Operations:
##
## Variable mutation for homeless01 %>% as.factor()
## Variables selected -c(i1, i2, d1, anysubstatus, daysanysub, e2b)
## Median imputation for dayslink, drugrisk, linkstatus
## Centering and scaling for all_numeric_predictors()
## Signed log transformation on avg_drinks, max_drinks, indtot, age
```

Next we'll create our model and pass both it and our recipe into a workflow!

```r
homeless_model <-
    logistic_reg(mode = "classification") %>%
    set_engine("glm")

homeless_workflow <-
    workflow() %>%
    add_model(homeless_model) %>%
    add_recipe(homeless_recipe)

homeless_workflow
```

```
## == Workflow ====================================================================
```

```
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 5 Recipe Steps
##
## * step_mutate()
## * step_select()
## * step_impute_median()
## * step_normalize()
## * step_log()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Computational engine: glm
```

Workflows are fantastic, they help organize the modeling process and encourage good methodology. Essentially you can bind modeling and pre-processing objects together!

```
fit(homeless_workflow, helprtc) %>%
    broom::tidy() %>%
    arrange(p.value)
```

```
## # A tibble: 15 x 5
##    term            estimate std.error statistic p.value
##    <chr>              <dbl>     <dbl>     <dbl>   <dbl>
##  1 pss_fr           -0.292      0.103    -2.83  0.00471
##  2 (Intercept)      -0.236      0.107    -2.22  0.0267
##  3 avg_drinks        1.18       0.604     1.96  0.0503
##  4 female           -0.211      0.110    -1.93  0.0537
##  5 sexrisk           0.173      0.103     1.69  0.0912
##  6 age               0.504      0.399     1.26  0.207
##  7 indtot            0.584      0.476     1.23  0.220
##  8 linkstatus        0.329      0.305     1.08  0.282
##  9 pcs              -0.0988     0.111    -0.892 0.372
## 10 dayslink          0.244      0.305     0.801 0.423
## 11 drugrisk          0.0634     0.104     0.609 0.542
## 12 cesd              0.0582     0.145     0.400 0.689
## 13 max_drinks        0.164      0.541     0.304 0.761
## 14 hospitalizations  0.0167     0.108     0.154 0.878
## 15 mcs               0.0181     0.142     0.127 0.899
```

At the risk of interpreting these results incorrectly, it appears that, according to our p-values, that we only really have 3 predictors that we have significant evidence for. It is important to note, that with our `homeless0` column, 0 is for housed individuals and 1 is for unhoused individuals. Our first predictor, `pss_fr` is a quantifier for an individuals perceived social support by friends with higher scores indicating more support. The negative coefficient indicates that higher support from friends is a predictor of not being homeless. We see this negative coefficient with `female` as well, which indicates that being female may make someone less likely to be homeless. `avg_drinks` is the last of the predictors with a p.value less than or close to 0.05, and it shows a pretty strong relationship between a larger number of drinks consumed per day and homelessness.