Homework 2

Will Orser Nolan Meyer Rohit Shah

Due Thursday, September 30 at 11:59pm CST on Moodle

Deliverables: Please use this template to knit an HTML document. Convert this HTML document to a PDF by opening the HTML document in your web browser. *Print* the document (Ctrl/Cmd-P) and change the destination to "Save as PDF". Submit this one PDF to Moodle.

Alternatively, you may knit your Rmd directly to PDF if you have LaTeX installed.

Project Work

Instructions

Goal: Begin an analysis of your dataset to answer your **regression** research question.

Collaboration: Form a team (2-3 members) for the project and this part can be done as a team. Only one team member should submit a Project Work section. Make sure you include the full names of all of the members in your write up.

Data cleaning: If your dataset requires any cleaning (e.g., merging datasets, creation of new variables), first consult the R Resources page to see if your questions are answered there. If not, post on the #rcode-questions channel in our Slack workspace to ask for help. *Please ask for help early and regularly* to avoid stressful workloads.

Required Analyses

1. **Initial investigation: ignoring nonlinearity (for now)**
    a. Use ordinary least squares (OLS) by using the `lm` engine and LASSO (`glmnet` engine) to build a series of initial regression models for your quantitative outcome as a function of the predictors of interest. (As part of data cleaning, exclude any variables that you don't want to consider as predictors.)
        - You'll need two model specifications, `lm_spec` and `lm_lasso_spec` (you'll need to tune this one).
    b. For each set of variables, you'll need a `recipe` with the `formula`, `data`, and pre-processing steps
        - You may want to have steps in your recipe that remove variables with near zero variance (`step_nzv()`), remove variables that are highly correlated with other variables (`step_corr()`), normalize all quantitative predictors (`step_normalize(all_numeric_predictors())`) and add indicator variables for any categorical variables (`step_dummy(all_nominal_predictors())`).
        - These models should not include any transformations to deal with nonlinearity. You'll explore this in the next investigation.
    c. Estimate the test performance of the models using CV. Report and interpret (with units) the CV metric estimates along with a measure of uncertainty in the estimate (`std_error` is readily available when you used `collect_metrics(summarize=TRUE)`).
        - Compare estimated test performance across the models. Which models(s) might you prefer?
    d. Use residual plots to evaluate whether some quantitative predictors might be better modeled with nonlinear relationships.
    e. Which variables do you think are the most important predictors of your quantitative outcome? Justify your answer. Do the methods you've applied reach consensus on which variables are most important? What insights are expected? Surprising?
        - Note that if some (but not all) of the indicator terms for a categorical predictor are selected in the final models, the whole predictor should be treated as selected.

Your Work a & b.

```r
# library statements
# read in data
library(dplyr)
library(readr)
library(broom)
library(ggplot2)
library(tidymodels)
library(tidyverse)
tidymodels_prefer() # Resolves conflicts, prefers tidymodel functions

set.seed(123)

basic_stats <- read_csv("Basic_Stats.csv")
qb_stats <- read_csv("Career_Stats_Passing.csv")
punt_return <- read_csv("Career_Stats_Punt_Return.csv")
punting <- read_csv("Career_Stats_Punting.csv")
receiving <- read_csv("Career_Stats_Receiving.csv")
kickers <- read_csv("Game_Logs_Kickers.csv")
oline_logs <- read_csv("Game_Logs_Offensive_Line.csv")
punters_logs <- read_csv("Game_Logs_Punters.csv")
qb_logs <- read_csv("Game_Logs_Quarterback.csv")
rb_logs <- read_csv("Game_Logs_Runningback.csv")
wrandte_logs <- read_csv("Game_Logs_Wide_Receiver_and_Tight_End.csv")
```

```r
# data cleaning
qb_stats <- qb_stats %>%
    select(-Position)

qb_stats <- qb_stats %>%
    rename("Rating" = "Passer Rating")

basic_stats <- basic_stats %>%
    rename("Team" = "Current Team")

basic_stats <- basic_stats %>%
    rename("hsloc" = "High School Location")

basic_stats <- basic_stats %>%
    rename("Weight" = "Weight (lbs)")

basic_stats <- basic_stats %>%
    rename("Height" = "Height (inches)")

basic_stats <- basic_stats %>%
    filter(Position == "QB")

qb_comb <- merge(qb_stats, basic_stats, by="Name")


qb_comb$`Passes Attempted`[which(qb_comb$`Passes Attempted` == '--')] = NA
qb_comb <- qb_comb %>%
  mutate(`Passes Attempted` = as.numeric(`Passes Attempted`))
```

```r
# creation of cv folds
qbcomb_cv <- vfold_cv(qb_comb, v = 10)
```

```r
# model spec

#OLS spec
lm_spec <-
    linear_reg() %>%
    set_engine(engine = 'lm') %>%
    set_mode('regression')

#LASSO spec
lm_lasso_spec <-
  linear_reg() %>%
  set_args(mixture = 1, penalty = 0) %>% ## mixture = 1 indicates Lasso, we'll choose penalty later
  set_engine(engine = 'glmnet') %>%
  set_mode('regression')
```

```r
# recipes & workflows

#OLS recipe
lm_rec <- recipe(Rating ~ Height + Age + Weight + Experience + College + hsloc + `Passes Attempted` + `
    update_role(`Passes Attempted`, new_role = 'Info') %>%
    update_role(`Games Played`, new_role = 'Info') %>%
    step_filter(`Passes Attempted` >= 50, `Games Played` > 0) %>% #removes potential outliers with few
    step_lincomb(all_numeric_predictors()) %>% # removes predictors that are linear combos of others
    step_corr(all_predictors()) %>% #removes highly correlate variables
    step_novel(all_nominal_predictors()) %>% # important if you have rare categorical variables
    step_nzv(all_numeric_predictors()) %>% # removes predictors with near zero variability
    step_dummy(all_nominal_predictors()) # creates indicator variables for categorical variables

#OLS workflow
lm_wf <- workflow() %>%
  add_recipe(lm_rec) %>%
  add_model(lm_spec)



#LASSO recipe
lasso_rec <- recipe(Rating ~ Height + Age + Weight + Experience + College + hsloc + `Passes Attempted` +
    update_role(`Passes Attempted` , new_role = 'Info') %>%
    update_role(`Games Played` , new_role = 'Info') %>%
    step_filter(`Passes Attempted` >= 50, `Games Played` > 0) %>% #removes potential outliers with few
    step_nzv(all_predictors()) %>% # removes variables with the same value
    step_novel(all_nominal_predictors()) %>% # important if you have rare categorical variables
    step_normalize(all_numeric_predictors()) %>%  # important standardization step for LASSO
    step_dummy(all_nominal_predictors())  # creates indicator variables for categorical variables

#lasso_rec %>% prep(qb_comb) %>% juice()

#LASSO workflow
lasso_wf <- workflow() %>%
  add_recipe(lasso_rec) %>%
  add_model(lm_lasso_spec)
```

```r
# fit & tune models
lm_lasso_spec_tune <-
  linear_reg() %>%
  set_args(mixture = 1, penalty = tune()) %>% ## tune() indicates that we will try a variety of values
  set_engine(engine = 'glmnet') %>%
  set_mode('regression')

lasso_wf <- workflow() %>%
  add_recipe(lasso_rec) %>%
  add_model(lm_lasso_spec_tune)

penalty_grid <- grid_regular(
  penalty(range = c(-5, 3)), #log10 transformed 10^-5 to 10^3
  levels = 50)

tune_res <- tune_grid( # new function for tuning hyperparameters
  lasso_wf, # workflow
  resamples = qbcomb_cv, # folds
  metrics = metric_set(rmse),
  grid = penalty_grid # penalty grid
)

autoplot(tune_res)

collect_metrics(tune_res) %>%
  filter(.metric == 'rmse') %>%
  select(penalty, rmse = mean)


best_penalty <- select_best(tune_res, metric = 'rmse') # choose best penalty value

final_wf <- finalize_workflow(lasso_wf, best_penalty) # incorporates penalty value to workflow

final_fit <- fit(final_wf, data = qb_comb)

tidy(final_fit)
```

c.

```r
#  calculate/collect CV metrics
mod1_cv <- fit_resamples(final_fit,
  resamples = qbcomb_cv,
  metrics = metric_set(rmse, rsq, mae)
)

mod1_cv %>% collect_metrics()
```

d.

```r
# visual residuals
mod1_output <- qb_comb %>%
  bind_cols(predict(final_fit, new_data = qb_comb)) %>%
    mutate(resid = Rating - .pred)

#Height
```

```
ggplot(mod1_output, aes(x = Height, y = resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, color = "red") +
    theme_classic()

#Age
ggplot(mod1_output, aes(x = Age, y = resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, color = "red") +
    theme_classic()

#Weight
ggplot(mod1_output, aes(x = Weight, y = resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, color = "red") +
    theme_classic()

#Experience
ggplot(mod1_output, aes(x = Experience, y = resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, color = "red") +
    theme_classic()

#College
ggplot(mod1_output, aes(x = College, y = resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, color = "red") +
    theme_classic()

#High School Location
ggplot(mod1_output, aes(x = hsloc, y = resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, color = "red") +
    theme_classic()
```

e.

2. **Summarize investigations**
   - Decide on an overall best model based on your investigations so far. To do this, make clear your analysis goals. Predictive accuracy? Interpretability? A combination of both?

The overall best model based on our investigations so far is our model titled “final fit.” This model has age, weight, height, college, high school location, and experience. The penalty for this model is 1 x 10ˆ-5. Some of the key predictors we started with became 0 for this model, more specifically age and weight. Certain levels of the categorical variables got close to 0, namely in experience, high school location, and college.

We want a combination of predictive accuracy and interpretability. Our main goal is predictive accuracy, as we want to find what predictors best predict quarterback performance/rating. However, we still want the the coefficients to be meaningful and interpretable, so we decided on a LASSO algorithm.

3. **Societal impact**
    - Are there any harms that may come from your analyses and/or how the data were collected?
    - What cautions do you want to keep in mind when communicating your work?

We do not believe that there are any real harms that may come from our analyses or the manner in which the data for our project was collected. Our goal for this project is to determine which variables are predictive of an NFL quarterback's passer rating (other than the statistics directly used to calculate passer rating: passing attempts, completions, yards, touchdowns, and interceptions). Our analysis has applications for predicting the end of season passer ratings of current NFL quarterbacks. So, in theory, our analysis could cause harm if it were used by an NFL coaching staff to select which QBs to roster and which to cut (the harm being done to the QBs cut on the basis of our analysis). However, the probability of our analysis ever being used by an NFL coaching staff and the probability of our analysis being the only means by which NFL coaches determine which QBs to roster are both extremely low, mitigating the potential for our analysis to cause harm. The data we are using for our analysis was scraped using Python code from the official NFL website (www.nfl.com), a widely-accessible site, in 2017. When communicating our work, we want to keep in mind the limitations of our analysis. In any sport, and especially in football, there is much more that goes into predicting the performance of a player than quantifiable variables like height, weight, age, or experience. Insofar as this is true, we want our audience to understand that our analysis is only a starting point for predicting the performance of NFL QBs and not an ironclad rule for who will be good and who will not. We would also like to keep in mind the limitations of our data when communicating our work. While our dataset is broad and rich, it does not include all the variables that could potentially predict the passer ratings of NFL QBs. Examples of a few such variables that are not included in our data are: handedness, a measure of the strength of a QB's offensive weapons in a given season, and a measure of the average strength of the defenses faced by a QB in a given season. In summary, we want to ensure our audience understands the limitations of our analysis for predicting NFL QB passer rating as well as the limitations of our data for identifying relevant predictors of passer rating.

Portfolio Work

Length requirements: Detailed for each section below.

Organization: To help the instructor and preceptors grade, please organize your document with clear section headers and start new pages for each method. Thank you!

Deliverables: Continue writing your responses in the same Google Doc that you set up for Homework 1. Include that URL for the Google Doc in your submission.

Note: Some prompts below may seem very open-ended. This is intentional. Crafting good responses requires looking back through our material to organize the concepts in a coherent, thematic way, which is extremely useful for your learning.

Revisions:

- Make any revisions desired to previous concepts. **Important note:** When making revisions, please change from "editing" to "suggesting" so that we can easily see what you've added to the document since we gave feedback (we will "accept" the changes when we give feedback). If you don't do this, we won't know to reread that section and give new feedback.
- General guideance for past homeworks will be available on Moodle (under the Solutions section). Look at these to guide your revisions. You can always ask for guidance in office hours as well.

New concepts to address:

- **Subset selection:**
    - Algorithmic understanding: Look at Conceptual exercise 1, parts (a) and (b) in ISLR Section 6.8. **What are the aspects of the subset selection algorithm(s) that are essential to answering these questions, and why?** (Note: you'll have to try to answer the ISLR questions to respond to this prompt, but the focus of your writing should be on the question in bold here.)

  - Bias-variance tradeoff: What “tuning parameters” control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: What computational time considerations are relevant for this method (how long the algorithms take to run)?
  - Interpretation of output: What parts of the algorithm output have useful interpretations, and what are those interpretations? **Focus on output that allows us to measure variable importance. How do the algorithms/output allow us to learn about variable importance?**
- **LASSO:**
  - Algorithmic understanding: Come up with your own analogy for explaining how the penalized least squares criterion works.
  - Bias-variance tradeoff: What tuning parameters control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: What computational time considerations are relevant for this method (how long the algorithms take to run)?
  - Interpretation of output: What parts of the algorithm output have useful interpretations, and what are those interpretations? **Focus on output that allows us to measure variable importance. How do the algorithms/output allow us to learn about variable importance?**
- **KNN:**
  - Algorithmic understanding: Draw and annotate pictures that show how the KNN (K = 2) regression algorithm would work for a test case in a 2 quantitative predictor setting. Also explain how the curse of dimensionality affects KNN performance. (5 sentences max.)
  - Bias-variance tradeoff: What tuning parameters control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: The KNN algorithm is often called a “lazy” learner. Discuss how this relates to the model training process and the computations that must be performed when predicting on a new test case. (3 sentences max.)
  - Interpretation of output: The “lazy” learner feature of KNN in relation to model training affects the interpretability of output. How? (3 sentences max.)

Reflection

Ethics: Read the article Automated background checks are deciding who’s fit for a home. Write a short (roughly 250 words), thoughtful response about the ideas that the article brings forth. What themes recur from last week’s article (on an old Amazon recruiting tool) or movie (Coded Bias)? What aspects are more particular to the context of equity in housing access?

Reflection: Write a short, thoughtful reflection about how things went this week. Feel free to use whichever prompts below resonate most with you, but don't feel limited to these prompts.

- How are class-related things going? Is there anything that you need from the instructor? What new strategies for watching videos, reading, reviewing, gaining insights from class work have you tried or would like to try?
- How is group work going? Did you try out any new collaboration strategies with your new group? How did they go?
- How is your work/life balance going? Did you try out any new activities or strategies for staying well? How did they go?

Self-Assessment: Before turning in this assignment on Moodle, go to the individual rubric shared with you and complete the self-assessment for the general skills (top section). After "HW2:", assess yourself on each of the general skills. Do feel like you've grown in a particular area since HW1?

Assessing yourself is hard. We must practice this skill. These "grades" you give yourself are intended to have you stop and think about your learning as you grow and develop the general skills and deepen your understanding of the course topics. These grades do not map directly to a final grade.