Assignment 5

```
## Warning: package ’tidyverse’ was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package ’ggplot2’ was built under R version 4.0.5
```

```
## Warning: package ’tibble’ was built under R version 4.0.5
```

```
## Warning: package ’tidyr’ was built under R version 4.0.5
```

```
## Warning: package ’readr’ was built under R version 4.0.5
```

```
## Warning: package ’purrr’ was built under R version 4.0.5
```

```
## Warning: package ’dplyr’ was built under R version 4.0.5
```

```
## Warning: package ’stringr’ was built under R version 4.0.5
```

```
## Warning: package ’forcats’ was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Warning: package ’tidymodels’ was built under R version 4.0.5
```

```
## Registered S3 method overwritten by ’tune’:
##   method                   from
##   required_pkgs.model_spec parsnip
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.1.3 --
```

```
## v broom        0.7.9      v rsample      0.1.0
## v dials        0.0.10     v tune         0.1.6
## v infer        1.0.0      v workflows    0.2.3
## v modeldata    0.1.1      v workflowsets 0.1.0
## v parsnip      0.1.7      v yardstick    0.0.8
## v recipes      0.1.17
```

```
## Warning: package ’broom’ was built under R version 4.0.5
```

```
## Warning: package ’dials’ was built under R version 4.0.5
## Warning: package ’scales’ was built under R version 4.0.5

## Warning: package ’infer’ was built under R version 4.0.5

## Warning: package ’modeldata’ was built under R version 4.0.5

## Warning: package ’parsnip’ was built under R version 4.0.5

## Warning: package ’rsample’ was built under R version 4.0.5

## Warning: package ’tune’ was built under R version 4.0.5

## Warning: package ’workflows’ was built under R version 4.0.5

## Warning: package ’workflowsets’ was built under R version 4.0.5

## Warning: package ’yardstick’ was built under R version 4.0.5

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.

## Warning: package ’plotly’ was built under R version 4.0.5

##
## Attaching package: ’plotly’

## The following object is masked from ’package:ggplot2’:
##
##     last_plot

## The following object is masked from ’package:stats’:
##
##     filter

## The following object is masked from ’package:graphics’:
##
##     layout
```

```r
ad<-readRDS("area_data.Rds")
```

1. Run a model on a training subset of the data which predicts the percent of the population in the labor force (dependent variable) as a function of the percent of the population that is insured (independent variable).

Split the data between training and testing

```r
split_data<-ad%>%initial_split(prop=.5)

ad_train<-training(split_data)

ad_test<-testing(split_data)
```

Specify the model

```r
lm_fit <-
  linear_reg() %>%
  set_engine("lm")%>%
  set_mode("regression")
```

Specify the formula

```r
lf_formula<-as.formula("perc_in_labor_force~perc_insured")
```

Specify the recipe

```r
lf_rec<-recipe(lf_formula,ad)
```

Add the model and the recipe to the workflow

```r
lf_wf<-workflow()%>%
  add_model(lm_fit)%>%
  add_recipe(lf_rec)
```

##Fit the model

```r
lf_results<-fit(lf_wf,ad_train)
```

See the results

```r
lf_results%>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   63.1       7.33       8.61  1.17e-16
## 2 perc_insured  -0.0363    0.0766    -0.474 6.36e- 1
```

2. Summarize the coefficient for percent of the population that is insured in a sentence or two.

The coefficient for percent insured is .0293, but it is not statistically significant. There's no observable relationship between the percent of the population insured and the percent of the population in the labor force.

3. Calculate the model fit by calculating the rmse in the *testing* data.

```
ad_test<-
  predict(lf_results,ad_test)%>%
  rename(pred1=.pred)%>%
  bind_cols(ad_test)
```

```
rmse_1<-ad_test%>%yardstick::rmse(truth=perc_in_labor_force,estimate=pred1)
rmse_1
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        6.17
```

4. Add variables for the census division and the percent of the population with commutes above 30 minutes. Comment on the estimates for both of these variables. (*N.B: the census division is a categorical variable*).

```
lf_formula<-as.formula("perc_in_labor_force~perc_insured+
                       division+
                       perc_commute_30p")
```

Specify the recipe

```
lf_rec<-recipe(lf_formula,ad)%>%
  step_dummy(division)
```

How do we know which is the 'reference' category

```
ad%>%group_by(division)%>%count()
```

```
## # A tibble: 9 x 2
## # Groups:   division [9]
##   division               n
##   <fct>              <int>
## 1 East North Central   159
## 2 West North Central   120
## 3 Mid-Atlantic          66
## 4 New England           26
## 5 East South Central    95
## 6 South Atlantic       153
## 7 West South Central   131
## 8 Mountain              94
## 9 Pacific               82
```

```
lf_wf<-lf_wf%>%
  update_recipe(lf_rec)
```

```
lf_results<-fit(lf_wf,ad_train)
```

```
lf_results%>%tidy()
```

```
## # A tibble: 11 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                    106.       7.62      13.9   9.66e-37
##  2 perc_insured                    -0.410    0.0774    -5.30  1.84e- 7
##  3 perc_commute_30p                -0.151    0.0349    -4.34  1.78e- 5
##  4 division_West.North.Central     3.43     1.07       3.22  1.39e- 3
##  5 division_Mid.Atlantic           -0.131    1.22      -0.107 9.15e- 1
##  6 division_New.England            4.70     2.11       2.23  2.61e- 2
##  7 division_East.South.Central    -4.98     1.08      -4.60  5.61e- 6
##  8 division_South.Atlantic        -5.91     1.08      -5.46  7.83e- 8
##  9 division_West.South.Central    -5.11     1.10      -4.63  4.74e- 6
## 10 division_Mountain              -1.17     1.05      -1.12  2.64e- 1
## 11 division_Pacific               -0.827    1.21      -0.683 4.95e- 1
```

```
ad_test<-
  predict(lf_results,ad_test)%>%
  rename(pred2=.pred)%>%
  bind_cols(ad_test)
```

5. Calculate the model fit by examining the rmse in the testing data. Comment in a sentence on what the rmse means and how it compares to the rmese in step 3.

```
rmse_2<-ad_test%>%yardstick::rmse(truth=perc_in_labor_force,estimate=pred2)
rmse_2
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        5.33
```

The rmse decreased from 6.17 to 5.33,indicating that the second model fits the testing data better.

6. Create another model by adding at least two other variables. Answer the following questions in a sentence for each:

```
lf_formula<-as.formula("perc_in_labor_force~perc_insured+
                       division+
                       perc_commute_30p+
                       college_educ+
                       perc_moved_in")
```

Specify the recipe

```
lf_rec<-recipe(lf_formula,ad)%>%
  step_dummy(division)
```

```
lf_wf<-lf_wf%>%
  update_recipe(lf_rec)
```

```
lf_results<-fit(lf_wf,ad_train)
```

```
lf_results%>%tidy()
```

```
## # A tibble: 13 x 5
##    term                        estimate std.error statistic  p.value
##    <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                  92.7       6.56      14.1   9.39e-38
##  2 perc_insured                 -0.383     0.0661    -5.80  1.23e- 8
##  3 perc_commute_30p             -0.0824    0.0307    -2.68  7.53e- 3
##  4 college_educ                  0.374     0.0300    12.4   8.96e-31
##  5 perc_moved_in                -0.242     0.200     -1.21  2.26e- 1
##  6 division_West.North.Central   3.09      0.913      3.38  7.83e- 4
##  7 division_Mid.Atlantic        -1.24      1.04      -1.19  2.33e- 1
##  8 division_New.England          0.348     1.82       0.191 8.49e- 1
##  9 division_East.South.Central  -4.32      0.942     -4.59  5.87e- 6
## 10 division_South.Atlantic      -5.83      0.949     -6.14  1.78e- 9
## 11 division_West.South.Central  -3.68      0.957     -3.84  1.40e- 4
## 12 division_Mountain            -2.58      0.975     -2.64  8.52e- 3
## 13 division_Pacific             -2.06      1.04      -1.98  4.83e- 2
```

```
ad_test<-
  predict(lf_results,ad_test)%>%
  rename(pred3=.pred)%>%
  bind_cols(ad_test)
```

```
rmse_3<-ad_test%>%
  rmse(truth=perc_in_labor_force,estimate=pred3)
rmse_3
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        4.80
```

- What is the rmse from your new model? How does it compare to the rmse from the previous model?

The rmse from the third model is 4.8. This indicates that or predictions are off, on average, by about 5 percentage points. This is an improvement from our first model.

- Which predictors appear to be related to the outcome? How do you know?

The percent of the population that has a college education is positively related to the percent of the population in the labor force. For each additional percent of the population with a bachelor's degree, the percent of the population in the workforce is predicted to increase by .3 percentage points. This result is statistically significant. There are significant differences in the percent of the population in the labor force across census divisons.