ST 538 Project 1

Group 3

2024-04-19

Question of Importance

We are interested in studying the relationship between a teachers salary and several possible factors, which are the most important and how do they affect the salary of a teacher.

Our Question: What factors are most important in determining a teachers salary, and can we use these to predict what a teacher should be making? Below you will find all the factors that we are considering in our analysis.

Exploratory Analysis

Variable Name	Description
SOCP	Occupation code
WAGP	Wages or salary past 12 months
SEX	Sex of person(m/f)
AGEP	Age of person
SCHL	Educational attainment
SCIENGP	Field of degree(na-less than bachelor, 1 science/engineering, 0 not)
LANX	language spoken at home(1-speaks other, 2-only english)
RACWHT	white(0-no, 1-yes)
WKHP	Usual hours worked per week past 12 months

Occupation code for teachers are 251000, 252010, 252020, 252030, 252050, we are not including "other teachers and instructors" as it is not specific enough to determine if they are teaching at a school or not, this might include things like substitute teachers, which will inherently have a lower salary as they may not be working full time.

We filter out teachers with reported wages less than $27456 as this is the minimum wage working full time in non-urban counties. We are also filtering out teachers making more than 150,000K as this is close to the maximum wage + add ons (coaching, etc) for a teacher in the Lake Oswego area. We finally filter down to only teachers that are public (COW == 3) which is employees of the local government, this will help to get rid of any private school teachers that have different salary structures than public teachers.

Transformations: SOCP: create column with readable teacher level names
SEX: 0 for female, 1 for male
DEGREE: create column with readable degree names
SCIENGP: 1 - stem, 2 - not_stem, 3 - no_bachelor
LANX: 1 - speaks other, 0 - only english

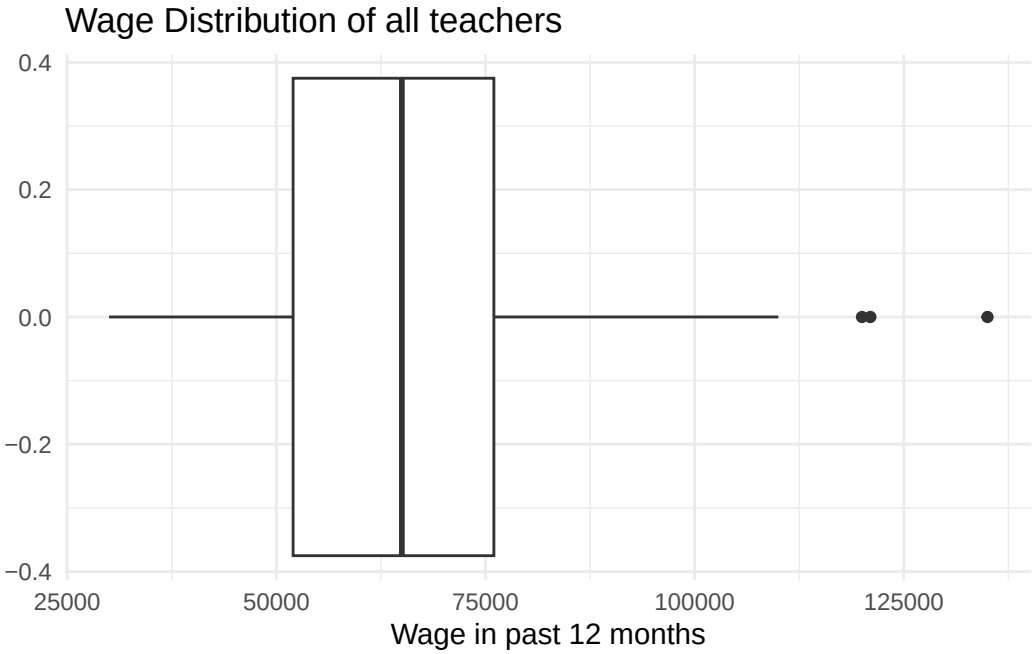
Get an idea for the distribution of the WAGP target variable. From the plots and summary statistics, we can see that we still have a large standard deviation in the wages of teachers across the board, the meadian and mean appear to

be a good measure of center, and we still have a few outliers that we may need to address in the future. Particulary teachers that are making over 100,000K.

Visualize The Data

Grade 1-12 Teacher Wage Summary Statistics

min	median	mean	stdev	q25	q75	q99	max
30000	65000	64046.18	16309.37	52000	76000	97190	135000



Model Selection and Checking

We started by building a linear model with all of the explanatory variables, the model is:

$$WAGP = \beta_o + \beta_1 AGEP + \beta_2 SEX + \beta_3 DEGREE + \beta_4 TEACH_T YPE + \beta_5 LANX + \beta_6 RACWHT + \beta_7 WKHP$$

We then checked the assumptions of the linear model, that the residuals were normally distributed, and random. We noticed that there was a funnel trend with the residuals which suggests heteroscadasticity. After this we checked the studentized residuals for outliers, removing any data points that fell above 2, or below -2. We then re-ran the model with the new data set, and checked the residuals, this residual plot looked much better as the points were randomly distributed around 0 and showed much lest trending. However they will still vastly over and underestimating the salary of teachers (residuals between 10,000 and 20,000).

Because we weren't able to obtain a good model with a general linear model we attempted to use regsubsets to find a model of less variables that accounted for the variation in our response variable, however there wasn't a good model selection from this technique, as all the residuals showed heterscadicity.

Our final approach was to use a ridge regression model in hopes that adding the penalization term to the explantory variables would improve the performance of our model. However this attempt also showed a large amount of heteroscadisticity in the residuals.

Results

From our exploratory analysis and model selection process we can see a couple things, first the assumptions of fitting a linear model to the data are certainly violated as we were not able to produce a model that had normally distributed residuals. This is likely due to the fact that the data is not linear, and there are likely other factors that are not accounted for in the model that are affecting the salary of teachers. Because of the nonlinear relationship between the explanatory variables and the response variable we were not able to produce a well fit linear model that explained the variation in the data. The next steps for this exploration would be to try a different model that can account for the non-linear relationship between the explanatory and the response variables.

Appendix

```
r <- getOption("repos")
r["CRAN"] <- "https://cloud.r-project.org/"
options(repos=r)

if(!require(readxl)) {
    install.packages("readxl")
}
if(!require(reader)) {
    install.packages("reader")
}
if(!require(car)) {
    install.packages("car")
}
if(!require(gt)) {
  install.packages("gt")
}
if(!require(numform)) {
  install.packages("numform")
}
library(car)
library(ggplot2)
library(glmnet)
library(gt)
library(leaps)
library(MASS)
library(numform)
library(readxl)
library(reader)
library(tidyverse)


temp <- tempfile()
download.file("https://www2.census.gov/programs-surveys/acs/data/pums/2022/5-Year/csv_por.zip", temp
data <- read_csv(unz(temp, "psam_p41.csv"))
unlink(temp)

proj1 <- data |>
  subset(SOCP %in% c("251000", "252010", "252020", "252030", "252050")) |>
  filter(WAGP > 27456 & WAGP < 150000) |>
  filter(SCHL > 20 & SCH < 24) |>
  filter(COW ==3) |>
  filter(WKHP > 35) |>
  mutate(TEACH_TYPE = case_when(
    SOCP == "251000" ~ "postsecondary",
    SOCP == "252010" ~ "pre_elementary",
    SOCP == "252020" ~ "elementary/middle",
    SOCP == "252030" ~ "secondary",
    SOCP == "252050" ~ "special_ed"
  )) |>
  filter(TEACH_TYPE %in% c("elementary/middle", "secondary")) |>
```

```
  mutate(SEX = ifelse(SEX == 2, 0, SEX)) |>
  mutate(DEGREE = case_when(
    SCHL < 20 ~ "no degree",
    SCHL == 20 ~ "associates",
    SCHL == 21 ~ "bachelors",
    SCHL == 22 ~ "masters",
    SCHL == 23 ~ "professional",
    SCHL == 24 ~ "doctorate"
  )) |>
  mutate(SCIENGP = case_when(
    is.na(SCIENGP) ~ "no_bachelor",
    SCIENGP == 1 ~ "stem",
    SCIENGP == 2 ~ "not_stem"
  )) |>
  mutate(LANX = ifelse(LANX == 1, 1, 0)) |>
  mutate_at(c(
    "TEACH_TYPE",
    "SEX",
    "DEGREE",
    "SCIENGP",
    "LANX",
    "RACAIAN",
    "RACBLK",
    "RACWHT"
    ),
    as.factor
  ) |>
  select(c(
    "TEACH_TYPE",
    "SEX",
    "DEGREE",
    # "SCIENGP", There seems to be some colinearity between SCIENGP and DEGREE
    "WAGP",
    "AGEP",
    "LANX",
    "RACWHT",
    "WKHP"
    )
  )

summary_df <- proj1 |>
  summarise(
    min=min(WAGP),
    median=median(WAGP),
    mean=mean(WAGP),
    stdev=sd(WAGP),
    q25=quantile(WAGP, 0.25),
    q75=quantile(WAGP, 0.75),
    q99=quantile(WAGP, 0.99),
    max=max(WAGP)
  )
summary_df |>
  gt() |>
```

```
  tab_header(title="Grade 1-12 Teacher Wage Summary Statistics")

proj1 |>
  ggplot(aes(x=WAGP)) +
    geom_histogram(bins=50, fill="dodgerblue") +
    geom_vline(xintercept = mean(proj1$WAGP)) +
    geom_vline(xintercept = median(proj1$WAGP)) +
    labs(
      title="Wage distribution of all teachers",
      x="Wages in past 12 months",
      y="Count"
    )
```

```
proj1 |>
  ggplot(aes(x=WAGP)) +
    geom_boxplot() +
    labs(
      title="Wage Distribution of all teachers",
      x="Wage in past 12 months"
    ) +
    theme_minimal()
```

```
proj1 |>
  ggplot(aes(x=WAGP, y=TEACH_TYPE)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    scale_y_discrete(
      limits=c(
        "secondary",
        "elementary/middle"
      )
    ) +
    scale_x_continuous(labels=ff_denom(mix.denom=TRUE, prefix="$", pad.char="")) +
    labs(
      title="Wage distribution among different type of teachers",
      y="Type of Teacher",
      x="Wage in past 12 months"
    ) +
    theme_minimal()
```

```
proj1 |>
  ggplot(aes(x=WAGP, y=SEX)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    labs(
      title="Wage distribution among Sex of teachers",
      y="Sex",
      x="Wage in past 12 months"
    ) +
    scale_y_discrete(labels=(c("female", "male"))) +
    scale_x_continuous(labels=ff_denom(mix.denom=TRUE, prefix="$", pad.char="")) +
    theme_minimal()
```

```r
proj1 |>
  ggplot(aes(x=WAGP, y=DEGREE)) +
    geom_violin() +
    geom_boxplot(width=0.1) +
    labs(
      title="Wage distribution among degrees teachers",
      y="Degree",
      x="Wage in past 12 months"
    ) +
    scale_y_discrete(
      limits=c(
        "doctorate",
        "professional",
        "masters",
        "bachelors",
        "associates"
        )
      ) +
    scale_x_continuous(labels=ff_denom(mix.denom=TRUE, prefix="$", pad.char="")) +
    theme_minimal()
```

```r
proj1$WKHP <- as.integer(proj1$WKHP)
lin_mod <- lm(WAGP ~ AGEP + SEX + DEGREE + TEACH_TYPE + LANX + RACWHT + WKHP, data = proj1)
summary(lin_mod)

qplot(fitted(lin_mod),resid(lin_mod), data=proj1)
```

```r
qqnorm(resid(lin_mod))
qqline(resid(lin_mod))
```

```r
plot(density(resid(lin_mod)))
```

```r
proj1.test <- proj1 |>
                mutate(studres = studres(lin_mod)) |>
                filter(studres < 2 & studres > -2)

lin_mod.test <- lm(WAGP ~ AGEP + SEX + DEGREE + TEACH_TYPE + LANX + RACWHT + WKHP, data = proj1.test
summary(lin_mod.test)
qplot(fitted(lin_mod.test),resid(lin_mod.test), data=proj1.test)
```

```r
proj1.test <- cbind(proj1.test, fitted = lin_mod.test$fitted.values)

testdat <- fortify(lin_mod.test,proj1.test)

qplot( rownames(testdat) , .hat, data=testdat)
```

```r
qplot( rownames(testdat), .stdresid, data=testdat)
```

```r
qplot( rownames(testdat), .cooksd, data=testdat)
```

```r
regfit_full <- regsubsets(WAGP ~ ., data=proj1, nvmax=10)
reg_summary <- summary(regfit_full)
reg_summary

which.min(reg_summary$rss)
which.max(reg_summary$adjr2)
which.min(reg_summary$cp)
which.min(reg_summary$bic)

# plot the four tests for goodness of fit
par(mfrow=c(2, 2))
plot(reg_summary$rss, xlab="Number of Variables", ylab="RSS", type="l")
points(
    which.min(reg_summary$rss),
    reg_summary$rss[which.min(reg_summary$rss)],
    col="red",
    cex=2,
    pch=20
)
plot(reg_summary$adjr2, xlab="Number of Variables", ylab="Adjusted RSq", type="l")
points(
    which.max(reg_summary$adjr2),
    reg_summary$adjr2[which.max(reg_summary$adjr2)],
    col="red",
    cex=2,
    pch=20
)
plot(reg_summary$cp, xlab="Number of Variables", ylab="Cp", type="l")
points(
    which.min(reg_summary$cp),
    reg_summary$cp[which.min(reg_summary$cp)],
    col="red",
    cex=2,
    pch=20
)
plot(reg_summary$bic, xlab="Number of Variables", ylab="BIC", type="l")
points(
    which.min(reg_summary$bic),
    reg_summary$bic[which.min(reg_summary$bic)],
    col="red",
    cex=2,
    pch=20
)
```

```r
predict.regsubsets <- function(object, newdata, id, ...) {
  form <- as.formula(object$call[[2]])
  mat <- model.matrix(form, newdata)
  coefi <- coef(object, id=id)
  xvars <- names(coefi)
  mat[, xvars] %*% coefi
}
```

```
best_model_index <- which.min(reg_summary$bic)
best_model_coef <- coef(regfit_full, best_model_index)
pred <- predict(regfit_full, proj1, best_model_index)
residuals <- proj1$WAGP - pred
plot_data <- cbind(proj1$WAGP, residuals)

plot_data |>
  as_tibble() |>
  ggplot(aes(x=V1, y=V2)) +
    geom_point()
```

```
best_model_index <- which.max(reg_summary$adjr2)
best_model_coef <- coef(regfit_full, best_model_index)
pred <- predict(regfit_full, proj1, best_model_index)
residuals <- proj1$WAGP - pred
plot_data <- cbind(proj1$WAGP, residuals)

plot_data |>
  as_tibble() |>
  ggplot(aes(x=V1, y=V2)) +
    geom_point()
```

```
set.seed(289)
train <- sample(c(TRUE, FALSE), nrow(proj1), rep=TRUE)
test <- (!train)
regfit_best <- regsubsets(WAGP ~ ., data=proj1[train, ], nvmax=9)
test_mat <- model.matrix(WAGP ~ ., data=proj1[test, ])
val_errors <- rep(NA, 10)

for (i in 1:9) {
    coefi <- coef(regfit_best, id=i)
    pred <- test_mat[, names(coefi)] %*% coefi
    val_errors[i] <- mean((proj1$WAGP[test] - pred)^2)
}
val_errors
model_num <- which.min(val_errors)
coef(regfit_best, model_num)
model_num

X <- model.matrix(WAGP ~ ., data=proj1)[, -1]
y <- proj1$WAGP
grid <- 10^seq(10, -2, length=100)
ridge_model <- glmnet(X[train, ], y[train], alpha=0, lambda=grid)
coef(ridge_model)

cv_out <- cv.glmnet(X[train, ], y[train], alpha=0)
best_lambda <- cv_out$lambda.min
pred <- predict(ridge_model, s=best_lambda, newx=X[test, ])
residuals <- y[test] - pred
plot_data <- cbind(y[test], residuals)
plot_data |>
  as_tibble() |>
```

```
  ggplot(aes(x=V1, y=s1)) +
    geom_point()
```

```
forward <- regsubsets(WAGP ~ ., data=proj1,
                      nvmax = 10,
                      method = 'forward')
for_summary <- summary(forward)
for_summary
which.max(for_summary$adjr2)
which.min(for_summary$cp)

backward <- regsubsets(WAGP ~ ., data=proj1,
                       nvmax = 10,
                       method='backward')
back_summary <- summary(backward)
back_summary
which.max(back_summary$adjr2)
which.min(back_summary$cp)
```