Project 1

Cassandra Morgan

2023-09-24

For my dataset, I chose the prevailing wage rates from 2018 from Illinois. The prevailing wage rate is the "average wage paid to similarly employed workers in a specific occupation in the area of intended employment". Wages are important to everyone, and figuring out the prevailing wage rate can affect choices you make of where to live and what jobs to choose. The dataset itself was obtained from the Illinois government through https://data.illinois.gov/. The question I seek to answer is: what effects do training and county location have on wage rates. The variables in the data set are: effective date, county, trade title, region, type, class, base wage, foreman wage, Overtime M-F, Overtime Sa, Overtime Sunday, Overtime holiday, Health/welfare benefit, pension, vacation, and training.

Link: https://data.illinois.gov/dataset/idol-2018-prevailing-wage-rates/resource/0c95f063-aed9-4db7-adc3-c224acee8fc2?view_id=0be4d3c7-3e39-4c06-854d-bc5ac8669bc1

I used the tidyverse, ggplot, and dplyr libraries to perform the data analysis.

```
wagedata <- read.csv("prevailingwagerates.csv")




basewage_and_county <- select(wagedata, County, Base.Wage)


basewage_grouped_county <- group_by(basewage_and_county, County)

mean_basewage_by_county <- summarize(basewage_grouped_county,  mean1 = mean(Base.Wage))

arrange(mean_basewage_by_county, mean1)
```

```
## # A tibble: 102 x 2
##    County      mean1
##    <chr>       <dbl>
##  1 Rock Island  31.6
##  2 Mercer       31.6
##  3 Crawford     32.0
##  4 Wabash       32.0
##  5 Clark        32.4
##  6 Edwards      32.6
##  7 Massac       32.6
##  8 Hamilton     32.7
##  9 Edgar        32.8
## 10 Franklin     32.9
## # i 92 more rows
```

To create the table above, I selected down to base wage and county. Then, I grouped by county. Next, I use summarize to find the mean base wage by county. Finally, I arranged by wage. From the table: the county with the lowest mean base wage was Rock Island at $31.56. The county with the highest mean base wage was Will at $45.70. The difference in mean base wage rate between the highest earning county and the lowest was $14.14. This difference shows that county of employment can possibly have a large effect on how much you are paid.

```
median(mean_basewage_by_county$mean1)
```

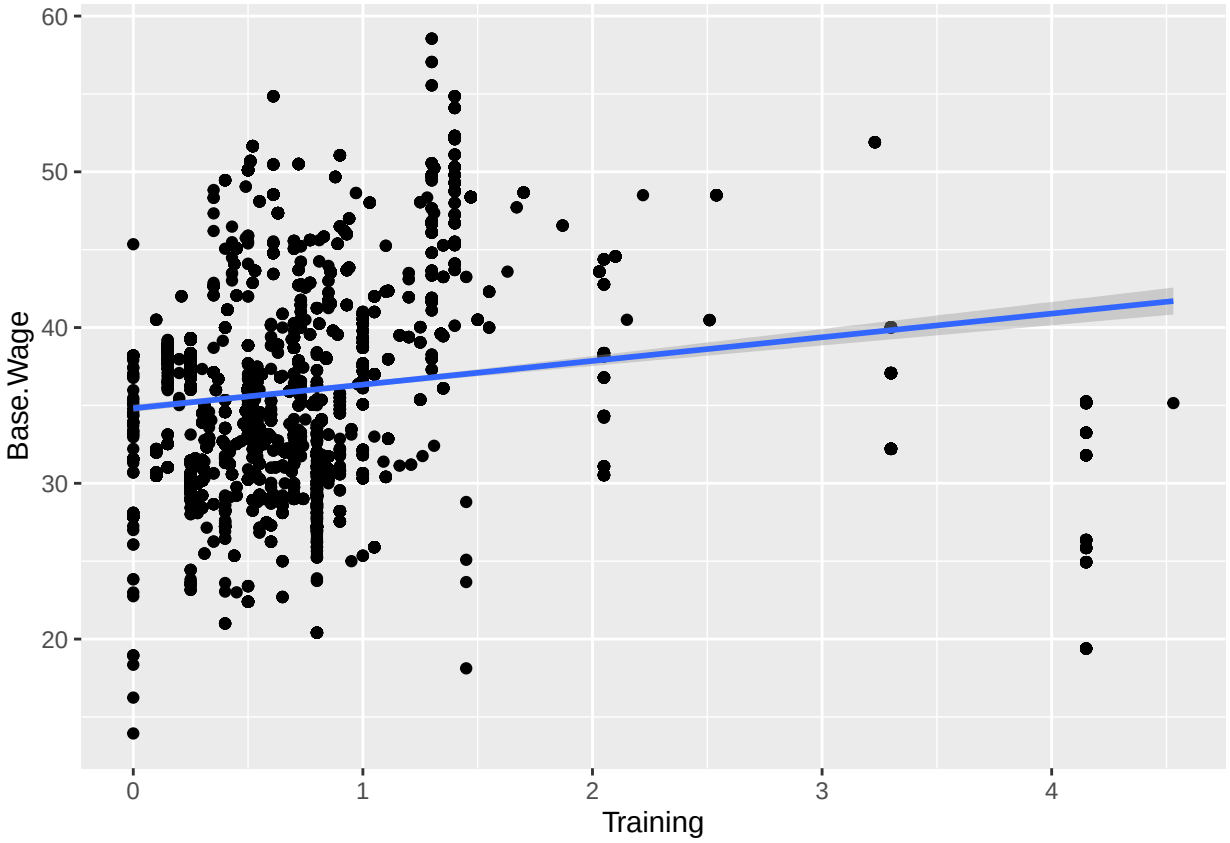
```
## [1] 34.65538
```

The median base wage of the counties was $34.66. Taken together with the range, this implies that there are rich counties bringing the average up by a considerable margin.

```
basewage_and_training <- select(wagedata, Training, Base.Wage)

ggplot(basewage_and_training, aes(x=Training, y=Base.Wage)) + geom_point() +
geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



To determine the effect of training on wages, I used a scatterplot. First I selected for training and base wage. Then, I used ggplot to plot base wage vs training. The line on the plot represents the correlation between training and base wage. The shaded region is the confidence interval. From the plot, it appears that there is a positive correlation between training and base wage. However, this does not prove that training causes higher base wage. To analyze further, I will look at a summary of the data.

```r
model1 <- lm(basewage_and_training$Base.Wage ~ basewage_and_training$Training)

summary(model1)
```

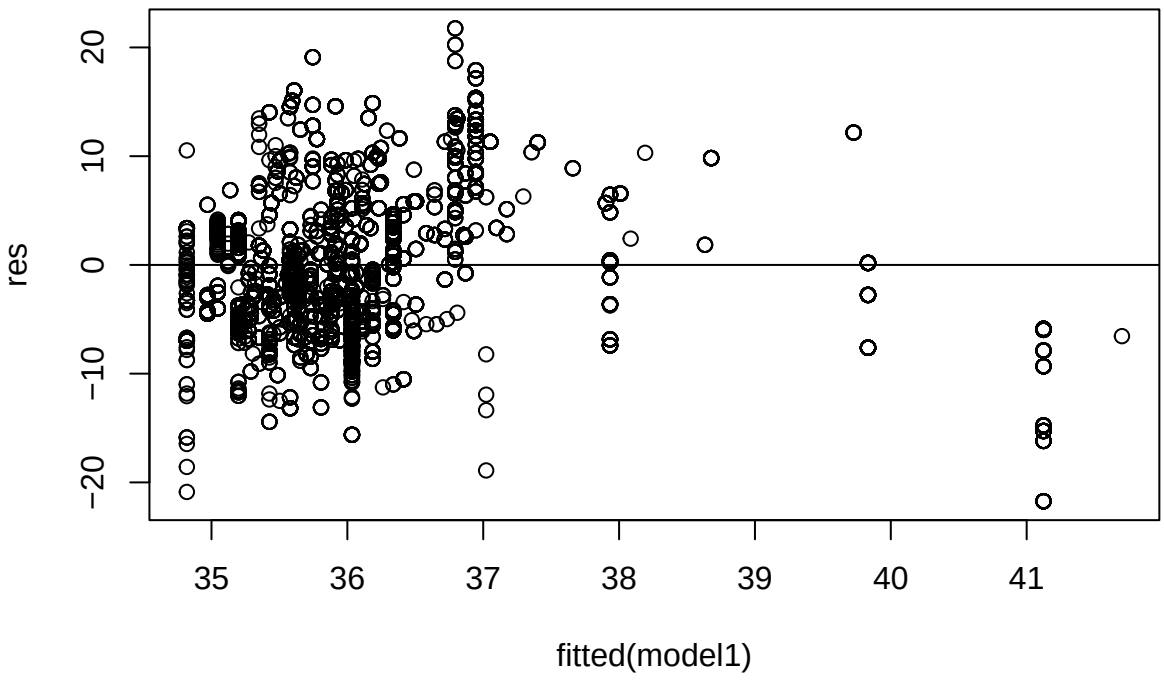
```
##
## Call:
## lm(formula = basewage_and_training$Base.Wage ~ basewage_and_training$Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.734  -4.731  -0.913   3.452  21.756
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     34.8186     0.1220  285.49   <2e-16 ***
## basewage_and_training$Training   1.5194     0.1163   13.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.511 on 6665 degrees of freedom
## Multiple R-squared:  0.02499,    Adjusted R-squared:  0.02484
## F-statistic: 170.8 on 1 and 6665 DF,  p-value: < 2.2e-16
```

Using the same linear regression model as the plot above, I used the summary function to learn information about the model itself. Linear regression models always have a predictor variable and a response variable. The predictor variable here is training (in years) and the response variable is base wage (in dollars). Residuals are the difference between observed response values and the predictions of the model. To analyze the residuals, we look at symmetry of the 5 residual summary points around the mean. This will help us determine how well the model fits the data. To help visualize this, I will use a plot of the residuals vs the fitted model:

```r
res <- resid(model1)

plot(fitted(model1), res)

abline(0,0)
```



Ideally, the points in the residual vs fitted plot would be randomly distributed around the residual = 0 line. From the residuals vs fitted plot I generated, it appears that the residuals are randomly oriented above and below the residual = 0 line. This implies that this model might be a good fit.

Next, I will talk about the coefficients, from the summary above, it appears that the intercept is 34.8186 and the slope is 1.5194. This evidence implies there might be a correlation between years of training and base wage. The slope here is positive which would imply that more training might correlate positively with higher base wage.

Ideally, the standard errors would be low, and the standard error for both the intercept and the slope in this case were low. This is a good sign for our model.

The t-values (which are 285.49 for the intercept and 13.07 for the slope), are relatively far from 0. This increases the odds that we can reject the null hypothesis.

Smaller p-values are better, and the p-values of the model are very close to 0. This further evidence that indicates that we can reject the null hypothesis.

The residual standard error is 6.511 which represents a percentage error of 18.7%. This means that we expect to be about 18.7% off with our predictions.

In total, due to the standard errors, t-values, p-values, and residual standard error, the summary implies that the model is a good fit. Taken together with the coefficients, this implies that the there is a positive correlation between training years and base wage level.

Previously, we have examined mean base wage by county and base wage by training. Finally we will look at training by county. This will allow us to see if different counties have higher or lower mean training.

```r
training_and_county <- select(wagedata, County, Training)


training_grouped_county <- group_by(training_and_county, County)

mean_training_by_county <- summarize(training_grouped_county,  mean2 = mean(Training))

arrange(mean_training_by_county, mean2)
```

```
## # A tibble: 102 x 2
##    County     mean2
##    <chr>      <dbl>
##  1 Crawford   0.502
##  2 Clark      0.505
##  3 Douglas    0.513
##  4 Vermilion  0.513
##  5 Edgar      0.522
##  6 Moultrie   0.529
##  7 Jasper     0.559
##  8 Edwards    0.577
##  9 Champaign  0.581
## 10 Cumberland 0.581
## # i 92 more rows
```

I selected for county and training this time. Then I grouped by county. Then, I summarized to find the mean training years by county. Finally, I arranged by this new mean. From the table, the top base wage county was Will, which shows up at number 10 on mean years of training by county. Rock island, which was the lowest base wage county, is the 31st from the bottom in terms of mean years of training. This relationship shows that there is likely a correlation between years of training and where one lives.

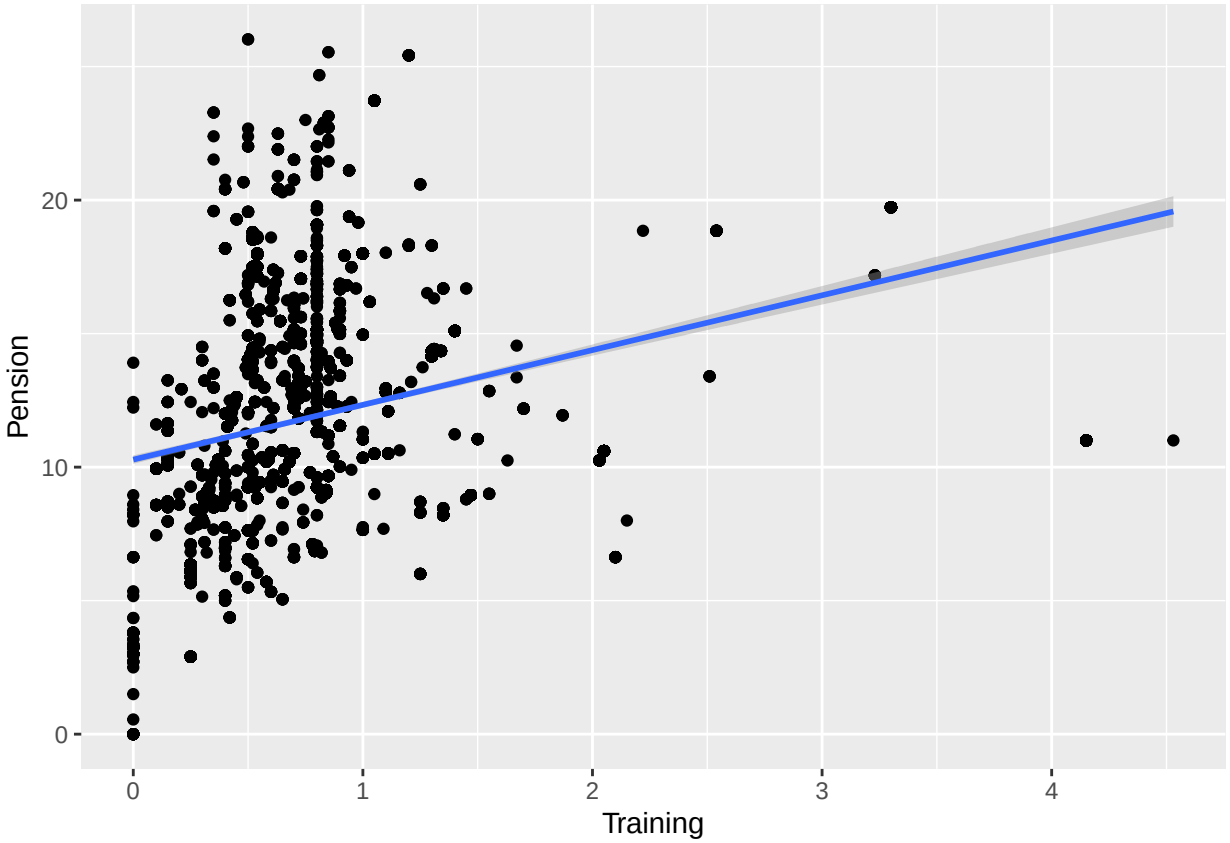
Taken together With the analysis from earlier, it appears that there is likely a correlation between where one lives, how many years of training one has, and what their base wage is.

My next question is whether or not people with higher levels of training will have greater pensions.

```r
pension_and_training <- select(wagedata, Training, Pension)

ggplot(pension_and_training, aes(x=Training, y=Pension)) + geom_point() +
geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



To determine the effect of training on pension, I used a scatterplot. First I selected for training and pension. Then, I used ggplot to plot pension vs training. The line on the plot represents the correlation between training and pension. The shaded region is the confidence interval. From the plot, it appears that there is a positive correlation between training and pension. However, this does not prove that training causes higher pensions. To analyze further, I will look at a summary of the data.

```
model2 <- lm(pension_and_training$Pension ~ pension_and_training$Training)

summary(model2)
```

```
##
## Call:
## lm(formula = pension_and_training$Pension ~ pension_and_training$Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.281  -3.694  -0.602   2.886  14.713
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   10.28101    0.08040  127.88   <2e-16 ***
## pension_and_training$Training  2.05124    0.07664   26.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.292 on 6665 degrees of freedom
```

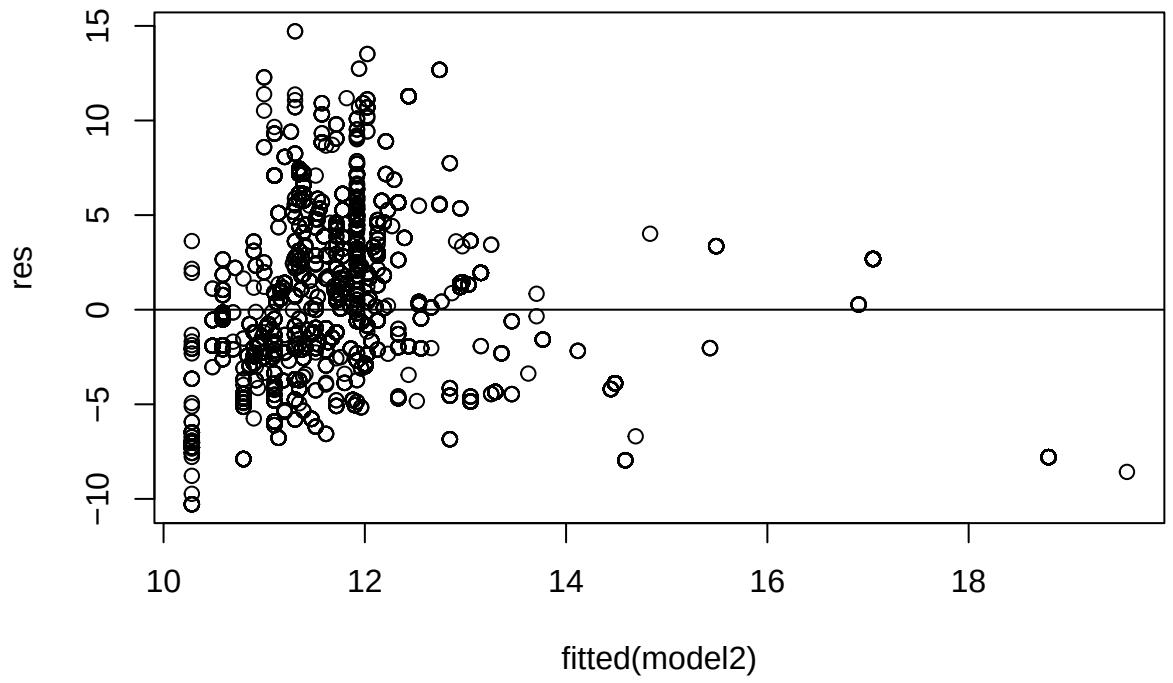
```
## Multiple R-squared:  0.09706,    Adjusted R-squared:  0.09692
## F-statistic: 716.4 on 1 and 6665 DF,  p-value: < 2.2e-16
```

Using a similar linear regression model as the one earlier, I used the summary function to learn information about the model itself. The predictor variable here is training and the response variable is pension. Residuals are the difference between observed response values and the predictions of the model. To analyze the residuals, we look at symmetry of the 5 residual summary points around the mean. I will now use a plot of the residuals vs the fitted model:

```
res <- resid(model2)

plot(fitted(model2), res)

abline(0,0)
```



From the residuals vs fitted plot I generated, it appears that the residuals are mostly above and below the residual = 0 line. This implies that this model might be not a good fit.

Next, I will talk about the coefficients, from the summary above, it appears that the intercept is 10.28101 and the slope is 2.05124. This evidence implies there might be a correlation between years of training and pensions level if the model is a good fit. The slope here is positive which would imply that more training might correlate positively with higher pension.

Ideally, the standard errors would be low, and the standard error for both the intercept and the slope in this case were low. This is a good sign for our model.

The t-values (which are 127.88 for the intercept and 26.77 for the slope), are relatively far from 0. This increases the odds that we can reject the null hypothesis.

Smaller p-values are better, and the p-values of the model are very close to 0. This further evidence that indicates that we can reject the null hypothesis.

The residual standard error is 4.292 which represents a percentage error of 3.36%. This means that we expect to be about 3.36% off with our predictions.

Finally, the R-squared value is 0.09706 which is low. This also implies that the model might not be a good fit.

In total, residuals and R-squared, the summary implies that the model might not be a good fit for the data.

As the last part, we will be looking to see if county of employment effects pension.

```
pension_and_county <- select(wagedata, County, Pension)


pension_grouped_county <- group_by(pension_and_county, County)

mean_pension_by_county <- summarize(pension_grouped_county,  mean3 = mean(Pension))

arrange(mean_pension_by_county, mean3)
```

```
## # A tibble: 102 x 2
##    County    mean3
##    <chr>     <dbl>
##  1 Wabash     8.98
##  2 Edwards    8.98
##  3 Massac     9.46
##  4 Richland   9.54
##  5 White      9.57
##  6 Lawrence   9.64
##  7 Alexander  9.74
##  8 Union      9.76
##  9 Pulaski    9.76
## 10 Johnson    9.76
## # i 92 more rows
```

I selected for county and pension this time. Then I grouped by county. Then, I summarized to find the mean pension by county. Finally, I arranged by this new mean. From the table, the top pension county was DuPage. The lowest pension county was Wabash This relationship shows that there is likely a correlation between pension and the county where one lives.

Taken altogether: there appears to be a correlation between base wage and county as well as base wage and training as well as training and county and finally county and pension.