Case Study #2: Disparities in wage between Black and white workers across regions of America

Anisha Jain

2024-11-06

Introduction

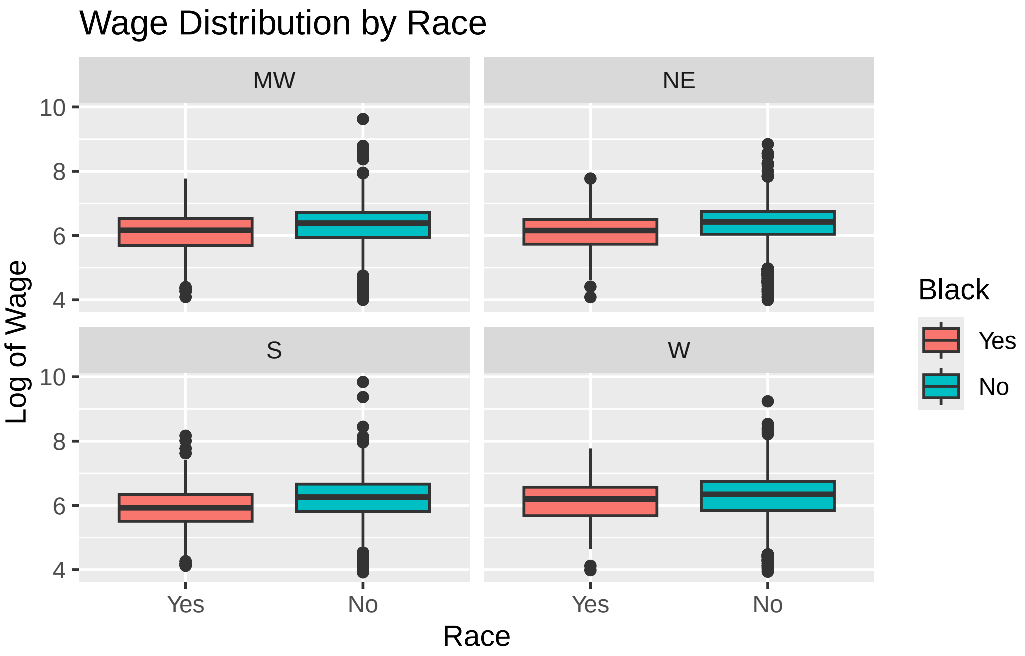
We are investigating the disparities in pay based on race across different regions of America. The U.S. Census bureau believes that the difference in pay may change based on region. We want to understand if across different regions, the pay gap changes for Black and white workers given the same education and experience. Next we want an estimation for the pay gaps for different regions. Finally, we want to evaluate if, given the same region, experience, and education, Black males are paid less than non-Black males. In our analysis, we consider the wage in dollars received by an individual the response variable and the race (Black, non-Black) and location based on region (mid-west, northeast, south, west) as the explanatory variables. We consider levels of education and work experience to be confounding variables.

Exploratory Data Analysis

This data set comes from a data problem in *The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*, written by Ramey, F.L. and Schafer, D.W. in 2002. There are 25,631 observations in the data.

Race	n	Median_Wage	Mean_Wage	SD
Black	1988	412.29	478.7273	307.8136
Non-Black	23643	569.80	653.7366	451.2722

From the table we see that both the median and mean of wages are higher for non-Black individuals. We can further assess the wages across regions in the form of box plots.



In the above figure we are plotting the wages against race for each of the four regions in America. Across the different regions, the differences in ranges and medians for non-Black and Black individual changes across each region. Yet, the median wage for white workers is always higher than the median wage for black workers. With that said, it's unclear how stark the difference is across regions. Note that we have taken the natural log of wages to reduce visual clutter.

Methods

We used a linear regression model to predict the wage (`W`) from race (`NB` for a response of non-Black) and region (`NE`, `S`, `MW`). Note that we did not include an interaction term between the race and the region as we do not expect the wage differential by race to change significantly across regions. We included terms for our confounding variables, level of education (`Ed`) and years of work experience (`Ex`). Through normal quantile quantile and fitted vs. residuals plots, we found that a log transformation on the wage variable supported the necessary linear regression model conditions. Our model for the population mean is

$$E[\log(W)|NB, NE, S, W, Ed, Ex] = \beta_0 + \beta_1(Ed) + \beta_2(Ex) + \beta_3(NB) + \beta_4(NE) + \beta_5(S) + \beta_6(W)$$

Results

The estimated model parameters, their standard errors, and 95% confidence intervals can be found below of the linear regression model for parallel slopes.

Table 2: Coefficients and Standard Errors

	Estimate	Std_Error	Lower	Upper
(Intercept)	4.4473460	0.0222449	4.4037448	4.4909472
BlackNo	0.2131223	0.0127313	0.1881683	0.2380763
RegionNE	0.0605616	0.0097147	0.0415203	0.0796029
RegionS	-0.0589539	0.0091429	-0.0768746	-0.0410333
RegionW	0.0042020	0.0099373	-0.0152756	0.0236796
Education	0.0989679	0.0012070	0.0966021	0.1013337
Experience	0.0182769	0.0002811	0.0177259	0.0188279

The `BlackNo` coefficient, corresponding to β_3 tells us that, while holding region, experience and education levels constant, the median wage of a white male is $e^{0.1928006} = 1.23$ times the median wage of a black male.

Transformations

When modelling with no transformations, we found that our fitted vs. residuals plot had a funnel shape, and our normal quantile-quantile plot had an extreme tail on the right side. This violated the variance and normality conditions. After administering the log transformation on the wage variable, we found that variance and normality conditions were improved. The plots describing these conditions can be found in the R appendix.

Hypothesis tests

First we wanted to understand whether the difference in pay for Black and white workers differed across regions. We constructed a linear regression model with an interaction term with the race and region variables. The population model for the model with interaction terms is:

$$E[\log(W)|NB, NE, S, W, Ed, Ex] = \beta_0 + \beta_1(Ed) + \beta_2(Ex) + \beta_3(NB) + \beta_4(NE) + \beta_5(S) + \beta_6(W) +$$

$$\beta_7(NE * NB) + \beta_8(S * NB) + \beta_9(W * NB).$$

We conducted a F-test to understand if the interaction term between race and region had any contribution to the regression model. In symbols, $H_0 : \beta_7 = 0, \beta_8 = 0, \beta_9 = 0$, and $H_A : \beta_7 \neq 0, \beta_8 \neq 0, \text{or } \beta_9 \neq 0$.

In the F-test we found a F-value of 1.4489 and a p-value 0.2264. With an $\alpha = .05$ we fail to reject the null hypothesis. This tells us that the interaction term between race and region does not contribute meaningfully to the model, and that the pay gap likely does not vary significantly based on region.

Next, based on the interaction term model, we will find an estimate for the pay gap in each region as well a 95% confidence interval for these estimates.

Region	Gap_Estimate	Lower	Upper
Northeast	1.216926	1.157232	1.276621
South	1.264358	1.231014	1.297701
West	1.167183	1.086284	1.248082
Midwest	1.212641	1.130266	1.295016

Due to the results of our F-test, we know that a parallel slopes model is a more appropriate model for the population mean. We can conduct a final t-test to determine whether the pay gap between Black and white workers is statistically significant, given the region, education, and experience level. In symbols, $H_0 : \beta_3 = 0$, and $H_A : \beta_3 \neq 0$. We found a t value of 16.740 and a p value of 2e-16. With $\alpha = .05$ we reject our null hypothesis. We can conclude that there exists a significant pay gap associated with race, holding region, education, and experience levels constant.

Conclusion

This report investigated the disparities in wage for Black and white males across different regions in America. We found that there is no significant difference in the pay gaps across regions. We also found estimates for the size of the pay gaps in the four regions. In the end we concluded that, holding region, education, and experience constant, there is a statistically significant pay gap between Black and white workers.

One limitation in the data analysis is that we did not include information about the proximity to a city as a variable of interest or confounder in our analysis, though it may have an effect in understanding the relationship between race and pay gaps. Another limitation is the conditions for linear regression. The normal quantile quantile plot still has some tails, but because there are so many observations its possible that we can relax these conditions.

R Appendix

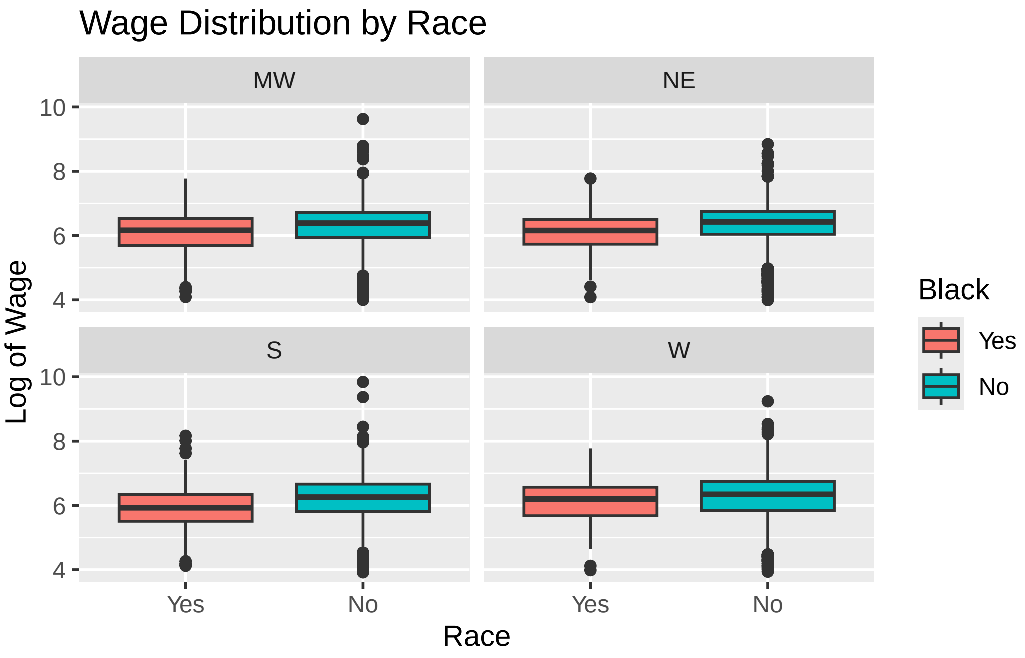
```
# Loading necessary packages
library(tidyverse)
library(Sleuth2)      # the package containing the data for the case study
library(kableExtra)   # for creating nicely formatted tables in Quarto
library(performance)
library(see)
library(broom)

# Loading the Statistical Sleuth data library
library(Sleuth2)
# Reading in and saving the data
wages <- Sleuth2::ex1029
wages <- wages |> mutate(LogWage = log(Wage))
```

```r
# creating a table for the descriptive statistics for wages based on race
wages_black <- wages |> filter(Black == "Yes") |>
  summarize(Race = "Black",
            n = n(),
            Median = median(Wage),
            Mean = mean(Wage),
            SD = sd(Wage))
wages_nonblack <- wages |> filter(Black == "No") |>
  summarize(Race = "Non-Black",
            n = n(),
            Median = median(Wage),
            Mean = mean(Wage),
            SD = sd(Wage))


# joining the tables of descriptive statistics
wages_table <- wages_black %>%
  full_join(wages_nonblack, by = c("Race", "n", "Median", "Mean", "SD"))
```

```r
# printing wages table
kable(wages_table)
```

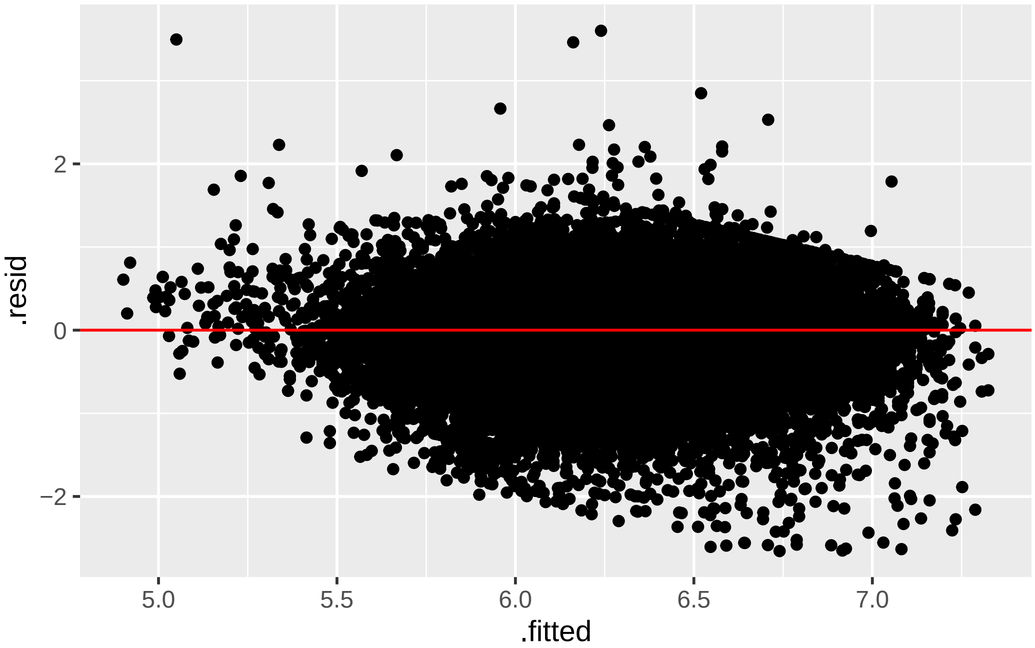
Race	n	Median	Mean	SD
Black	1988	412.29	478.7273	307.8136
Non-Black	23643	569.80	653.7366	451.2722

```r
# box plot for each region for the wage differential
ggplot(wages, aes(x = Black, y = log(Wage), fill = Black)) +
  geom_boxplot() +
  labs(
    title = "Wage Distribution by Race",
    x = "Race",
    y = "Log of Wage"
  )+
  facet_wrap(~Region)
```

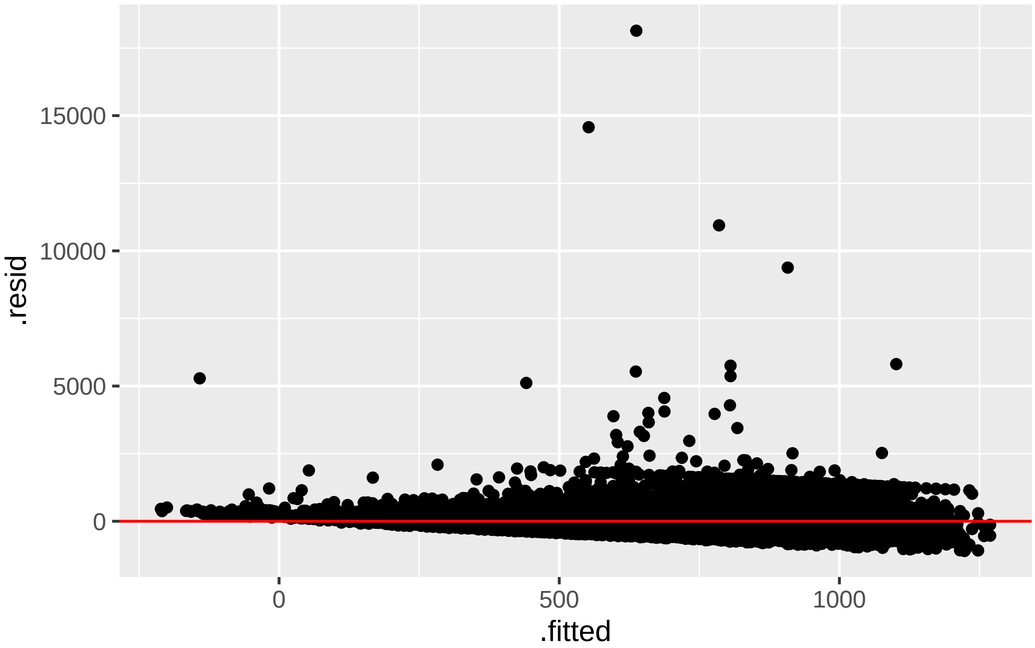


```
# constructing the regression model with log(Wage) and interaction term
wage_lm <- lm(log(Wage) ~ Black * Region + Education + Experience, data = wages)

# creating the wage model for parallel slopes
wage_parallel_lm <- lm(log(Wage) ~ Black + Region + Education + Experience, data
 ↪   = wages)

# constructing the regression model with no transformations
wage_nolog_lm <- lm((Wage) ~ Black * Region + Education + Experience, data =
 ↪   wages)
```

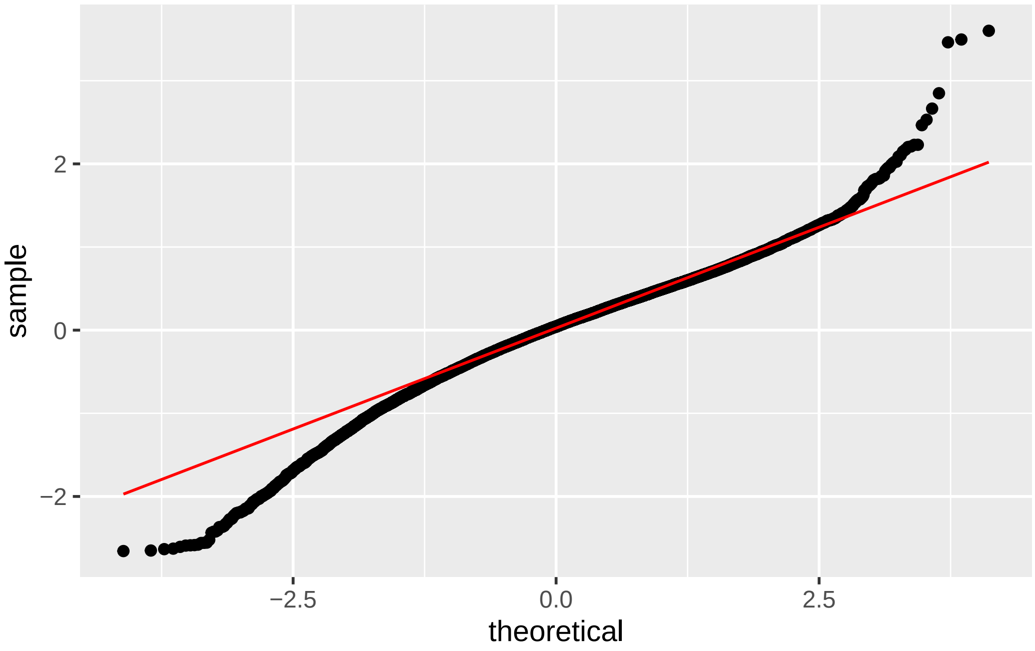
```
# comparison of fitted vs. residuals
wage_parallel_lm |> ggplot(aes(x = .fitted, y = .resid)) + geom_point() +
 ↪   geom_hline(yintercept = 0, col = "red")
```



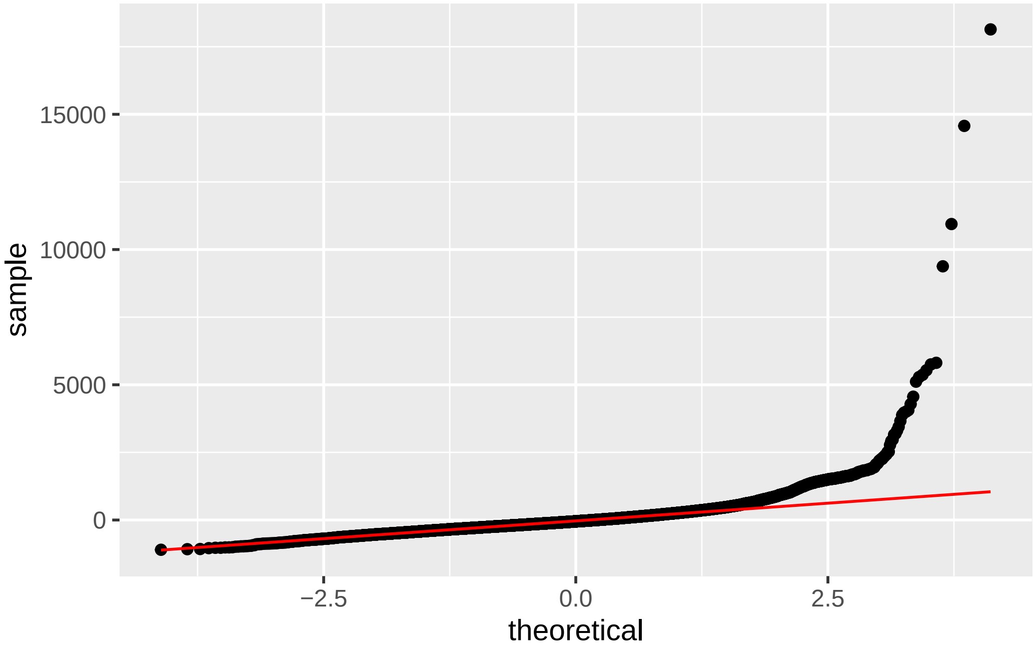
```
wage_nolog_lm |> ggplot(aes(x = .fitted, y = .resid)) + geom_point() +
↪  geom_hline(yintercept = 0, col = "red")
```



```
# comparison of normal quantile quantile plot

wage_parallel_lm |> ggplot(aes(sample = .resid)) + geom_qq() + geom_qq_line(col =
↪  "red") + xlab("theoretical") + ylab("sample")
```



```r
wage_nolog_lm |> ggplot(aes(sample = .resid)) + geom_qq() + geom_qq_line(col =
↪  "red") + xlab("theoretical") + ylab("sample")
```



```r
# Representing the regression table as a dataframe (i.e., tidying the summary()
↪  output)

t = qt(1- .05/2, df=25631 - 7)
model_summary <- summary(wage_parallel_lm)
coefficients_table <- data.frame(
  Estimate = coef(model_summary)[, "Estimate"],
  Std_Error = coef(model_summary)[, "Std. Error"],
  Lower = coef(model_summary)[, "Estimate"] - t * coef(model_summary)[, "Std.
   ↪  Error"],
  Upper = coef(model_summary)[, "Estimate"] + t * coef(model_summary)[, "Std.
   ↪  Error"]
)
```

```
#printing out coefficients table
kable(coefficients_table, caption = "Coefficients and Standard Errors")
```

Table 5: Coefficients and Standard Errors

	Estimate	Std_Error	Lower	Upper
(Intercept)	4.4473460	0.0222449	4.4037448	4.4909472
BlackNo	0.2131223	0.0127313	0.1881683	0.2380763
RegionNE	0.0605616	0.0097147	0.0415203	0.0796029
RegionS	-0.0589539	0.0091429	-0.0768746	-0.0410333
RegionW	0.0042020	0.0099373	-0.0152756	0.0236796
Education	0.0989679	0.0012070	0.0966021	0.1013337
Experience	0.0182769	0.0002811	0.0177259	0.0188279

```
#anova test for including interaction terms
anova(wage_lm)
```

```
Analysis of Variance Table

Response: log(Wage)
                Df Sum Sq Mean Sq   F value Pr(>F)
Black            1  170.7  170.73  594.9801 <2e-16 ***
Region           3   91.8   30.60  106.6532 <2e-16 ***
Education        1 1257.7 1257.72 4382.9684 <2e-16 ***
Experience       1 1213.2 1213.23 4227.9287 <2e-16 ***
Black:Region     3    1.2    0.42    1.4489 0.2264
Residuals    25621 7352.1    0.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 #this code block finds the estimates and confidence intervals for the pay gap

 #calculating standard deviation
 wages_sd <- function(c1, c2) {
   g_sd <- sqrt(vcov(wage_lm)[c1, c1] + vcov(wage_lm)[c2, c2] +
 ↪  2*vcov(wage_lm)[c1, c2])
```

```
  return (g_sd)
}

#get sd
sd_ne <- wages_sd(2, 8)
sd_w <- wages_sd(2, 10)
sd_s <- wages_sd(2, 9)
sd_mw <- 0.0420270 #from summary of wage_lm

#pt estimates
y_ne <- exp(coef(wage_lm)[2] + coef(wage_lm)[8])
y_w <- exp(coef(wage_lm)[2] + coef(wage_lm)[10])
y_s <- exp(coef(wage_lm)[2] + coef(wage_lm)[9])
y_mw <- exp(coef(wage_lm)[2])

#t value for interaction term model
t_star = qt(1 - .05/2, 25631 - 10)

#creating table
wages_NE <- wages |> summarize(Region = "Northeast",
            Gap_Estimate = y_ne,
            Lower = y_ne - t * sd_ne,
            Upper = y_ne + t * sd_ne)
wages_W <- wages |> summarize(Region = "West",
            Gap_Estimate = y_w,
            Lower = y_w - t * sd_w,
            Upper = y_w + t * sd_w)
wages_S <- wages |> summarize(Region = "South",
            Gap_Estimate = y_s,
            Lower = y_s - t * sd_s,
            Upper = y_s + t * sd_s)
wages_MW <- wages |> summarize(Region = "Midwest",
            Gap_Estimate = y_mw,
            Lower = y_mw - t * sd_mw,
            Upper = y_mw + t * sd_mw)


 wages_region_table <- wages_NE |>
  full_join(wages_S, by = c("Region","Gap_Estimate", "Lower", "Upper")) |>
  full_join(wages_W, by = c("Region","Gap_Estimate",  "Lower", "Upper")) |>
  full_join(wages_MW, by = c("Region","Gap_Estimate",  "Lower", "Upper"))
```

```
kable(wages_region_table)
```

Region	Gap_Estimate	Lower	Upper
Northeast	1.216926	1.157232	1.276621
South	1.264358	1.231014	1.297701
West	1.167183	1.086284	1.248082
Midwest	1.212641	1.130266	1.295016