

Does Region Shape the Wage Gap Between Black and Non-Black Male Workers?

Case Study 2

Radiyah Khan

Catherine Weeks

Mercer Mercer

Nafisa Mohammad

2025-11-14

Introduction

The United States Census Bureau, starting in 1940, has conducted a monthly survey of the labor force known as the Current Population Survey (CPS). The CPS is used to provide current estimates of the economic status. In this Case Study, we aim to address the question. Were Black males paid less than non-Black males in the same region, with the same levels of education and experience? The motivation behind this question is to determine how racial inequality in the labor force is reflected in wage gaps across the regions in the US.

Data Description

In this case study, we are working with the ex1029 data set available in the Sleuth2 library. This data set represents a sample of 25,632 male respondents to the March 1988 U.S. Current Population Survey that is administered monthly by the U.S. Census Bureau. The variables include:

- **Wage** Weekly wage in 1992 USD
- **Education** Years of education
- **Experience** Years of work experience
- **Black** Indicator of whether the respondent was black (yes or no)
- **Region** Region of residence (Northeast, Midwest, South or West)

Data Wrangling

Before analysis, we checked how many missing observations were present in this data set. Using the `is.na()` and `colSums()` functions, we confirmed that there were no missing observations, so no rows were removed and therefore the dataset used in our analysis contains 25,632 complete observations.

To further prepare for visualization and modeling, we created new variables.

Experience Group

A categorical variable `exp_group` was created by grouping Experience into four intervals:

- For people with less than 21 years of experience: 0–20
- For people with less than 41 years of experience: 21–40
- For people with less than 61 years of experience: 41–60
- For people with more than 60 years of experience: 60+

```
#Making the experience years into groups
wages_exp_group <- wages |>
  mutate(exp_group = case_when(
    Experience < 21 ~ "0-20",
    Experience < 41 ~ "21-40",
    Experience < 61 ~ "41-60",
    TRUE ~ "60+"
  ))
head(wages_exp_group, 5)
```

	Wage	Education	Experience	Black	SMSA	Region	exp_group
1	354.94	7	45	No	Yes	NE	41-60
2	370.37	9	9	No	Yes	NE	0-20
3	754.94	11	46	No	Yes	NE	41-60
4	593.54	12	36	No	Yes	NE	21-40
5	377.23	16	22	No	Yes	NE	21-40

Average Wage by Experience Group and Region

A numerical variable `avg_wage_exp` was created to group the dataset by Region, Race and Experience Group and calculate the mean weekly wages among these groups.

```
#Grouping the data for experience group and region
wages_exp_group_region <- wages_exp_group |>
  mutate(Race = ifelse(Black == "Yes", "Black", "Non-Black")) |>
  group_by(Region, Race, exp_group) |>
  summarise(avg_wage_exp = mean(Wage, na.rm = TRUE), .groups = "drop")
head(wages_exp_group_region,5)
```

```
# A tibble: 5 x 4
  Region Race     exp_group avg_wage_exp
  <fct>  <chr>    <chr>          <dbl>
1 MW      Black    0-20           462.
2 MW      Black    21-40          624.
3 MW      Black    41-60          503.
4 MW      Non-Black 0-20           581.
5 MW      Non-Black 21-40          801.
```

Average Years of Education by Race and Region

A numerical variable `avg_education` was created to group the Race and Region group and calculate their average years of education.

```
wages_edu_group_region <- wages |>
  group_by(Region, Black) |>
  summarise(avg_education = mean(Education)) |>
  ungroup()
head(wages_edu_group_region,5)
```

```
# A tibble: 5 x 3
  Region Black avg_education
  <fct>  <fct>        <dbl>
1 MW      Yes            12.5
2 MW      No             13.3
3 NE      Yes            12.2
4 NE      No             13.3
5 S       Yes            12.2
```

Exploratory Data Analysis

Before testing the research question, we conduct an exploratory data analysis on the dataset first using our existing variables and new variables.

To visualize regional differences, we computed the average weekly wage by region and race and the average education level by region and race.

Summary of the Dataset

```
wages_summary_table <-summary(wages)
wages_summary_table
```

Wage	Education	Experience	Black	SMSA
Min. : 50.39	Min. : 0.00	Min. :-4.00	Yes: 1988	Yes:19040
1st Qu.: 356.13	1st Qu.:12.00	1st Qu.: 9.00	No :23643	No : 6591
Median : 567.23	Median :12.00	Median :16.00		
Mean : 640.16	Mean :13.08	Mean :18.59		
3rd Qu.: 826.21	3rd Qu.:16.00	3rd Qu.:27.00		
Max. :18777.20	Max. :18.00	Max. :63.00		

Region

MW:6226

NE:5949

S :7991

W :5465

This summary table shows us that the minimum weekly wage recorded is \$50.39 USD, and the mean is \$640.16 USD. The maximum wage is much higher at \$18777.20 indicating a large variation between wages recorded. The South had the most samples collected, followed by the Midwest, the Northeast, then the West regions.

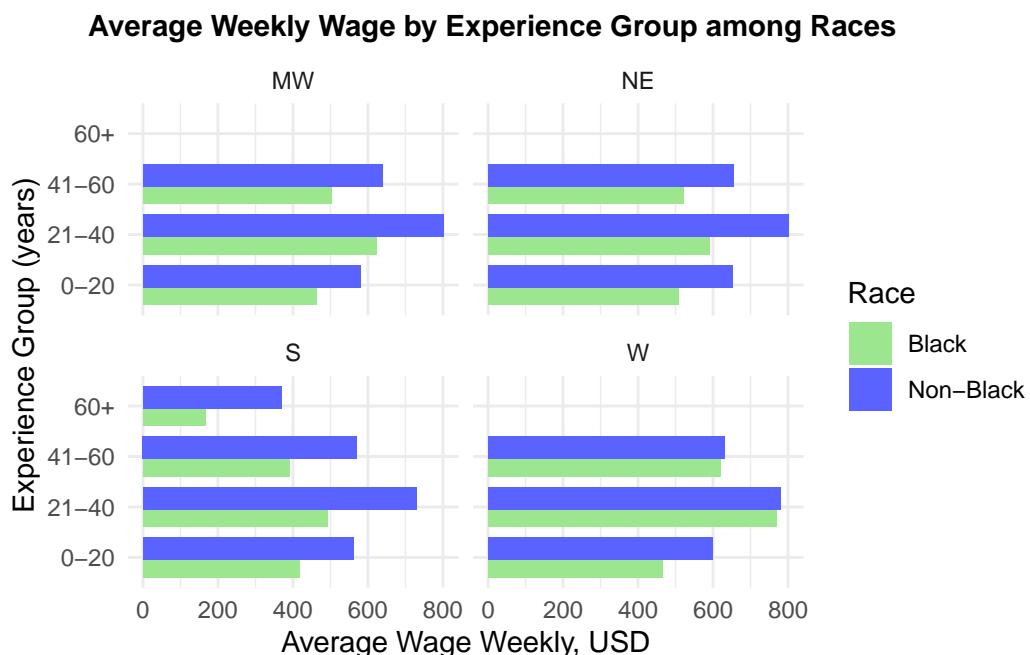
Across all regions, the mean years of education is 13 years, insinuating that all the samples collected had an average education of high school

In the years of experience variable, there are unusual entries of negative years of experience, which might indicate something in the quality of our data set which we are unable to comment on. If we ignore the negative values, our experience ranges from 9 years to 63 years.

Finally, in the Black variable, we can see the amount of people sampled who were black is far less than the non-black population in the sample.

Average Weekly Wage By Experience Group Among Races

```
#Plotting the experience group variable and avg wage by race
ggplot(wages_exp_group_region,
aes(x = exp_group, y = avg_wage_exp, fill = Race)) +
geom_col(position = position_dodge(width = 0.7)) +
facet_wrap(~ Region) +
labs(title = "Average Weekly Wage by Experience Group among Races",
x = "Experience Group (years)",
y = "Average Wage Weekly, USD") +
coord_flip() +
scale_fill_manual(values = c("#9BE68E", "#5A63FF"))+
theme_minimal() +
theme(plot.title = element_text(size = 11, hjust = 0.3, face = "bold"))
```

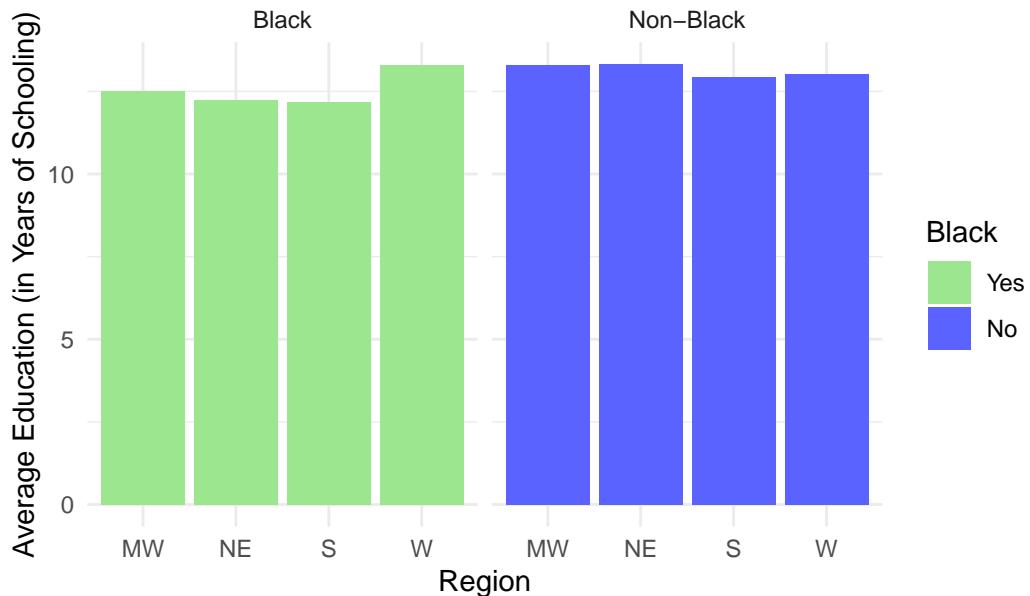


The plot above shows that only the South region has samples of individuals with over 60 years of experience. Within the same experience groups, we can see that Black men had lower weekly wage on average compared to non-Black men which is consistent across all regions.

Average Years of Education by Region and Race

```
#Plot
ggplot(wages_edu_group_region, aes(x = avg_education, y = Region, fill = Black)) +
  geom_col() +
  facet_grid(~Black, labeller = labeller(Black = graph.labels) )+
  labs(title = "Average Education by Region and Race, USA",
       x = "Average Education (in Years of Schooling)",
       y = "Region"
  ) +
  scale_fill_manual(values = c("#9BE68E", "#5A63FF"))+
  theme_minimal() +
  theme(plot.title = element_text(size = 11, hjust = 0.5, face = "bold"))+coord_flip()
```

Average Education by Region and Race, USA



This plot shows that there are only slight differences in average years of education regionally. The one area of the plot that stands out is in the West, Black individuals have a higher ratio of average years of education compared to their non-Black counterparts, which stands out slightly from other regions.

Statistical Methods

For this research question, we are going to use a nested F test model to see if the wage gap is statistically significant among Black and Non black individuals given the same years of experience and education. For this we will set up two multiple linear regression model. The reason we are choosing a multiple linear regression approach is because there are more than one predictors for the wage (Race, Region, Education, Experience). Before conducting the F test we will see if our models adhere to the four inference conditions. If they do not, we will transform the model accordingly using appropriate transformation. Furthermore, we will test the inference conditions again. Finally we will compare two multiple linear regression models: first model is the reduced model that will have no interaction between the race and region and the second model is the full model that has an interaction between Race and Region.

Fitting the Multiple Linear Regression Models

Before modeling, we convert the Black and Region columns into categorical variables. The relevel call sets “No” as the baseline group so model coefficients for Black compare other levels to non-Black respondents.

We fit a multiple regression model predicting Wage using Education, Experience, Black, Region, and the Black-by-Region interaction. Then prints the summary of the model results.

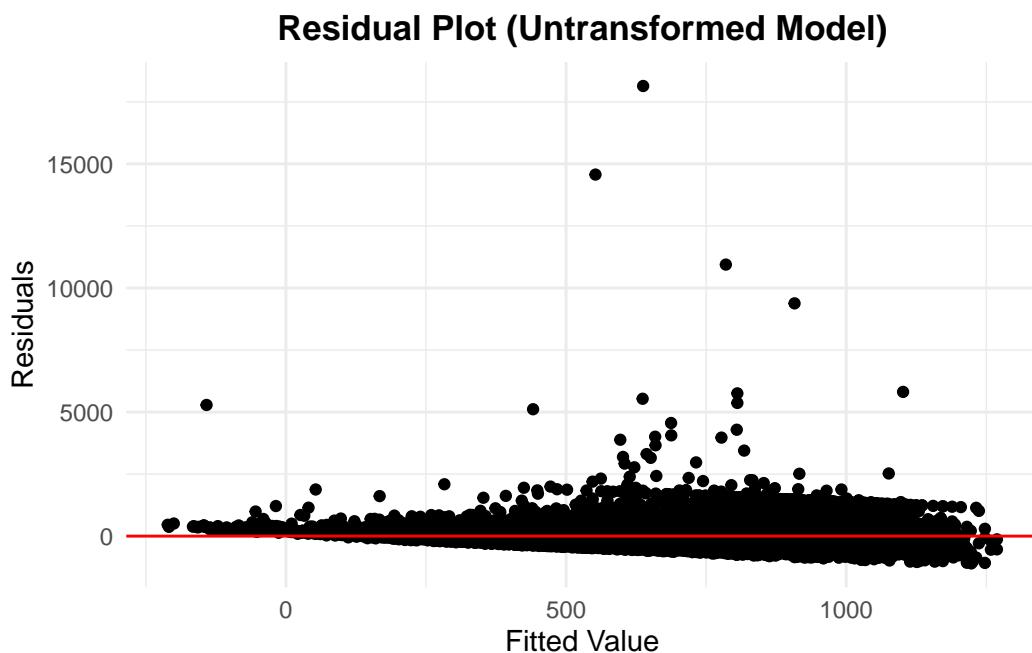
$$\begin{aligned} E[Wage|Education, Experience, Race, Region] = & \beta_0 + \beta_1 Education \\ & + \beta_2 Experience + \beta_3 Race + \\ & \beta_4 NorthEast + \beta_5 South + \beta_6 West + \\ & \beta_7 Race * NorthEast + \beta_8 Race * South + \beta_9 Race * West \end{aligned}$$

Here the variables are Wages representing the weekly wages, Education representing years of education, Experience representing years of experience, Race representing whether a respondent was black our reference level here is non black respondents, and Region being split up into Northeast, Midwest, South, and West with our reference level set as the Midwest.

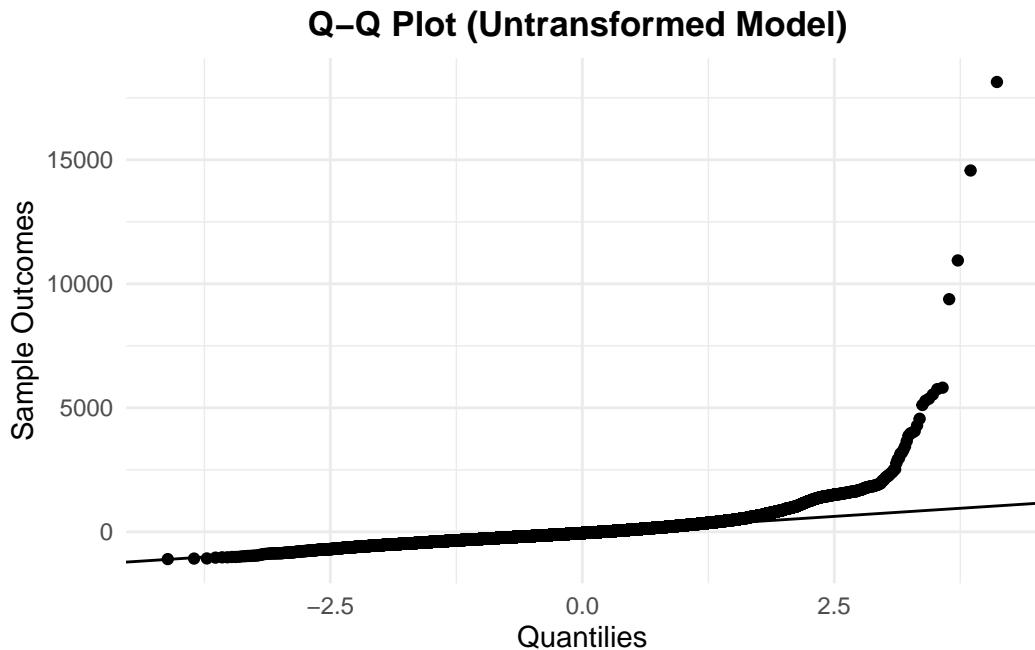
Checking the Conditions for untransformed model

The following two graphs are checking key regression assumptions to see if our model violates the conditions of linearity, normality, and equal variance. The residuals vs. fitted plot checks linearity and equal variance: we want the residuals scattered randomly around zero with no clear curve and roughly the same vertical spread across the range of fitted values (a funnel shape indicates heteroscedasticity).

```
# Checking Linearity and Equal Variance
ggplot(mlm_model_untransformed |> augment(), aes(x = .fitted, y = .resid))+
  geom_point()+
  geom_hline(yintercept = 0, col = "red")+
  labs(title = "Residual Plot (Untransformed Model)" ,
       x = "Fitted Value",
       y = "Residuals")+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



```
# Checking for the normality condition
mlm_model_untransformed |>
  augment() |> ggplot(aes(sample = .resid)) + geom_qq() +
  geom_qq_line() +
  labs(title = "Q-Q Plot (Untransformed Model)", x = "Quantiles", y = "Sample Outcomes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



The Q–Q plot checks whether the residuals are approximately Normal: points should lie close to the reference line; If points pull away from the line, it suggests the residuals aren't following a normal pattern. That matters because our standard errors, confidence intervals, and p-values assume normal residuals. Therefore, if the assumption is false, those numbers can be misleading.

Thus, we will perform log transformation on our response variable in our case will be the Wage.

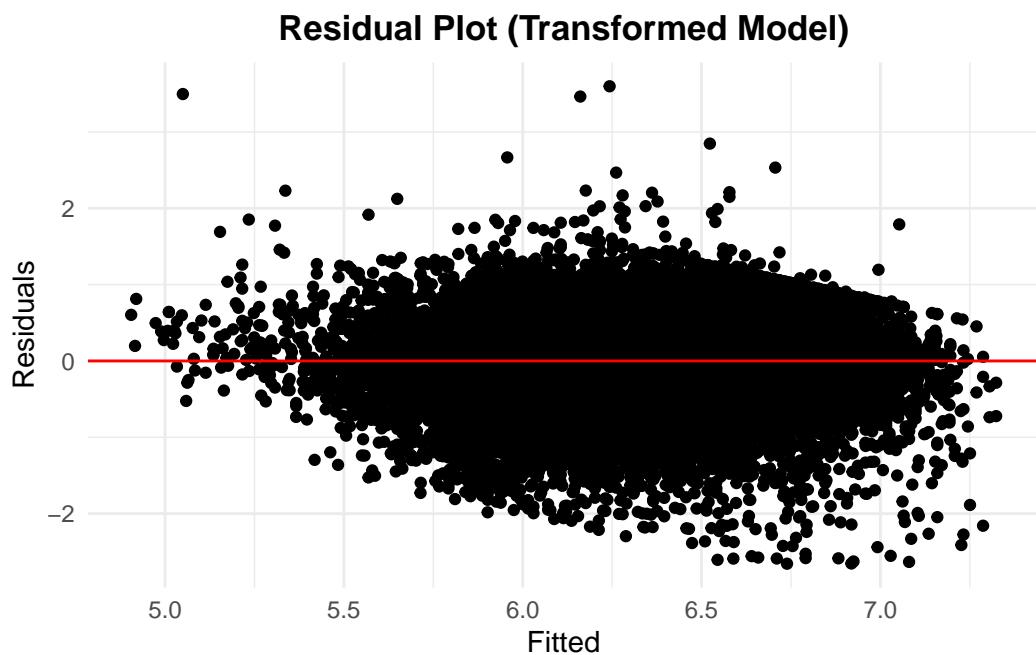
```
#MLM model with log scaled y axis
mlm_model <- lm(log(Wage) ~ Education + Experience + Black * Region, data = wages)
mlm_transformed_df <- tidy(mlm_model, conf.int = TRUE)
flextable(mlm_transformed_df) |>
  set_caption("Regression Summary of the Transformed Model: Predicting Wage") |>
  autofit() |>
  theme_apa()
```

term	estimate	std.error	statistic	p.value
(Intercept)	4.66	0.02	239.21	0.00
Education	0.10	0.00	81.93	0.00
Experience	0.02	0.00	64.99	0.00
BlackYes	-0.19	0.03	-6.29	0.00
RegionNE	0.06	0.01	6.08	0.00
RegionS	-0.05	0.01	-5.75	0.00
RegionW	0.00	0.01	0.33	0.74
BlackYes:RegionNE	-0.00	0.04	-0.08	0.93
BlackYes:RegionS	-0.04	0.04	-1.19	0.23
BlackYes:RegionW	0.04	0.05	0.74	0.46

$$E[\log(wage)|Education, Experience, Race, Region] = \beta_0 + \beta_1 Education + \beta_2 Experience + \beta_3 Race + \beta_4 NorthEast + \beta_5 South + \beta_6 West + \beta_7 Race * NorthEast + \beta_8 Race * South + \beta_9 Race * West$$

Checking the condition on the transformed model

```
# Checking Linearity and Equal Variance
ggplot(mlm_model |> augment(), aes(x = .fitted, y = .resid))+
  geom_point()+
  geom_hline(yintercept = 0, col = "red")+
  labs(title = "Residual Plot (Transformed Model)",
       x = "Fitted",
       y = "Residuals") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



```
# Checking for the normality condition
mlm_model |>
  augment() |> ggplot(aes(sample = .resid)) + geom_qq() +
  geom_qq_line() +
  labs(title = "Q-Q Plot (Transformed model)", x = "Quantiles", y = "Sample Outcomes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Untransformed model: The residuals vs fitted plot had a clear pattern that the residuals followed showing that there was no random scattering for the residuals. This indicates that the residuals got larger as fitted values increased. In addition, a few points stuck way above the rest clearly showing that the residuals were not centered at zero. The Q-Q plot showed a strong right-side hook, meaning the residuals had a heavy right tail. It's important to point out that the residuals that fell above the reference line were big numbers. In summary, the errors were uneven and skewed, so the model was being pulled by a few very large wages.

Transformed model (log Wage): The residuals vs fitted plot now looks evenly distributed with residuals centered on zero with similar spread across fitted values (no obvious shape pattern). The Q-Q plot is also much closer to the straight line, so residuals are more like a Normal distribution. With these improvements our transformed model now meets the usual regression assumptions (linearity, equal variance, normality). With our model now better fitted we can now move on to proving our research question. However, although the transformed model is much better, it is not perfect as it still shows curved tails on extreme sides of the plot.

$$E[\log(wage)|Education, Experience, Race, Region] = \beta_0 + \beta_1 Education + \beta_2 Experience + \beta_3 Race + \beta_4 NorthEast + \beta_5 South + \beta_6 West$$

```
wages_reduced_model <- lm(log(Wage) ~ Education + Experience + Black + Region, data = wages)
wages_reduced_df <- tidy(wages_reduced_model, conf.int = TRUE)
flextable(wages_reduced_df) |>
  set_caption("Regression Summary of the Reduced Model: Predicting Wage") |>
  autofit() |>
  theme_ap()
```

term	estimate	std.error	statistic
(Intercept)	4.66	0.02	239.91
Education	0.10	0.00	81.99
Experience	0.02	0.00	65.02
BlackYes	-0.21	0.01	-16.74
RegionNE	0.06	0.01	6.23
RegionS	-0.06	0.01	-6.45
RegionW	0.00	0.01	0.42

Conducting the hypothesis test

For conducting the hypothesis test we will use the nested F test method.

The full model contains 9 predictor terms and the reduced model contains 6 predictors. So our extra term is $8 - 6 = 3$ terms.

$H_0 : \beta_j = 0$ for all 3 predictor terms being dropped from the full model. $H_A : \beta_j \neq 0$ for at least one of the 3 predictor terms being dropped from the full model.

The significance level() for this test is 0.05.

```
#Conducting the nested F test
f_results <- anova(wages_reduced_model, mlm_model)
anova_df <- as.data.frame(f_results)
flextable(anova_df) |>
  set_caption("F-test Result Summary") |>
  autofit() |>
  theme_ap()
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
25,624.00	7,353.35				
25,621.00	7,352.11	3.00	1.25	1.45	0.23

From the results, the F statistic is 1.145 and the p-value is 0.23. The Black \times Region interaction is not statistically significant because our $p = 0.23 > \alpha = 0.05$. Therefore we fail to reject the null that the interaction coefficients are zero, and there is no evidence that the Black–non-Black wage gap differs across the four regions after adjusting for education and experience.

Analysis

From the F-test results, we can see that the test result was not statistically significant. There is a clear wage gap across the nation that doesn't significantly differ based on region. While Black men in the USA are paid less than Non-Black men, this pay gap is consistent and doesn't expand or shrink regionally. In a lot of the plots, we also saw a difference in which age group participates in the labor force.

Some limitations of this analysis are that this data set only represents one month of weekly wages, and doesn't represent seasonal work. The research question also only looks at the wages of men, and doesn't take women and labor that is traditionally performed by women (childcare, caring for elderly family) into account. Another large limitation is that our data only separates individuals into two groups; Black and non-Black which disregards the potential influence of other races, especially in areas that have high populations of Hispanic individuals, for example. This data also only represents wages that were reported to the government, and excludes any shadow work that may have occurred. Finally, the data set used is very old, and we can't use these findings to draw conclusions about the current state of our economy.