Your Informative Title Here

Case Study 2

Radiah Khan Catherine Weeks Mercer Mercer
Nafisa Mohammad

2001-11-14

Introduction

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)

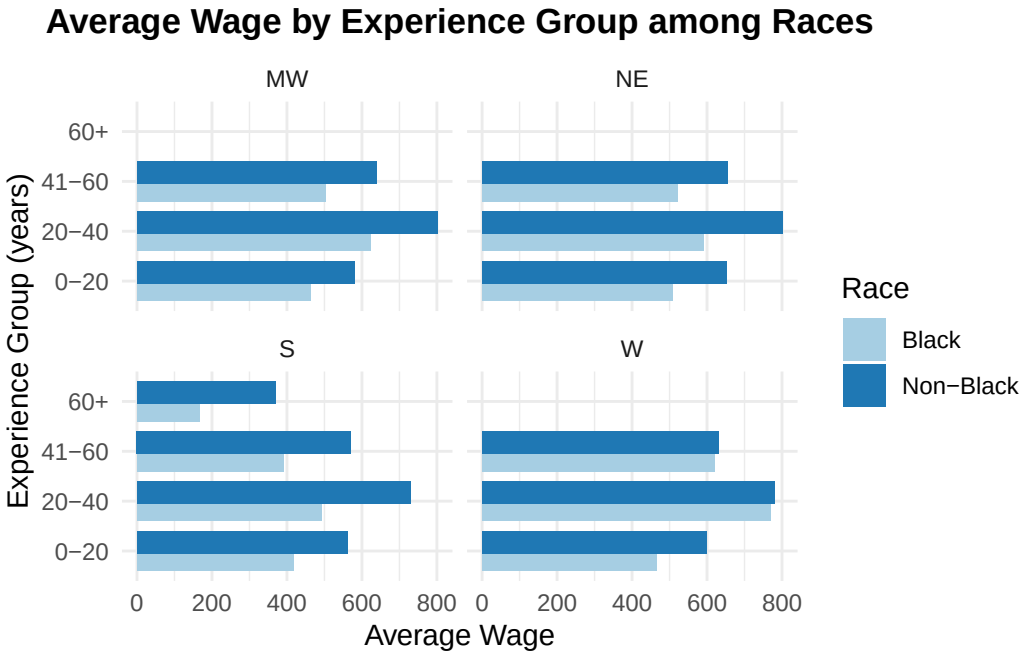
Data Wrangling

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, a new categorical variable, `exp_group` was created by grouping `Experience` into four intervals: 0-20 years, 21-40 years, 41-60 years, and 60+ years of experience.

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

The visualization below shows how average wages increase with experience across all regions and that Black respondents consistently earn lower wages than non-Black respondents.



Preparing for Regression Analysis

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.

```
#factoring the categorical variables in the wage data set
wages$Black <- factor(wages$Black)
wages$Region <- factor(wages$Region)

#setting the new reference level
wages$Black <- relevel(wages$Black, ref = "No")
```

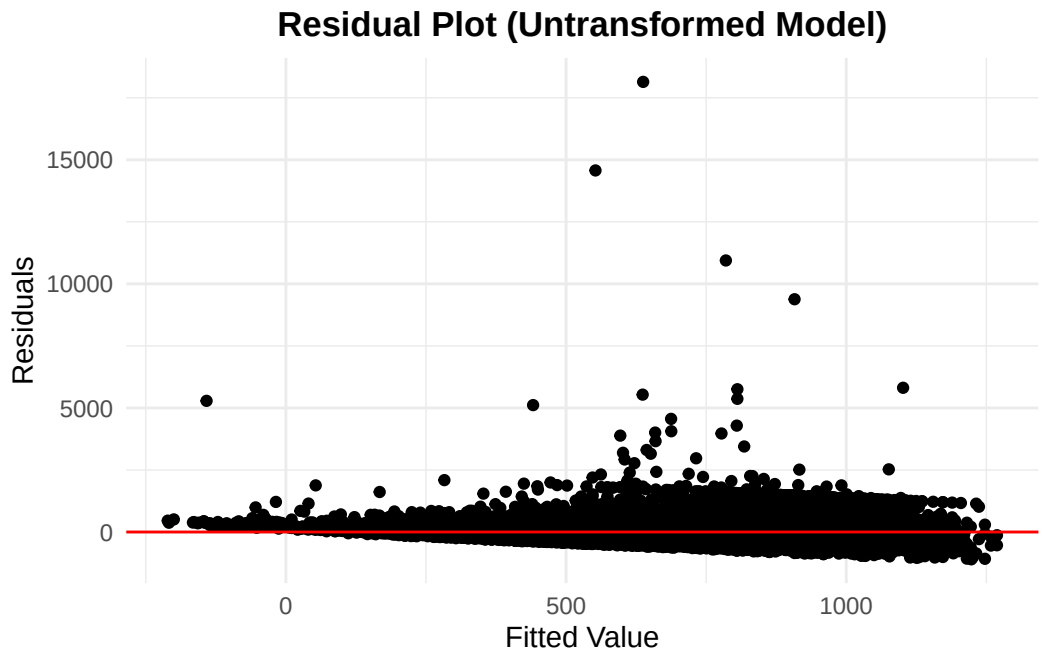
We fits a multiple regression model predicting Wage using Education, Experience, Black, Region, and the Black-by-Region interaction. Then prints the summary of the model results.

term	estimate	std.error	statistic
(Intercept)	-370.52	14.44	-25.65
Education	62.28	0.90	69.56
Experience	10.72	0.21	51.44
BlackYes	-104.62	22.74	-4.60
RegionNE	35.94	7.40	4.85
RegionS	-20.51	7.07	-2.90

term	estimate	std.error	statistic
RegionW	20.38	7.52	2.71
BlackYes:RegionNE	-10.07	32.03	-0.31
BlackYes:RegionS	-22.46	25.99	-0.86
BlackYes:RegionW	0.15	38.13	0.00

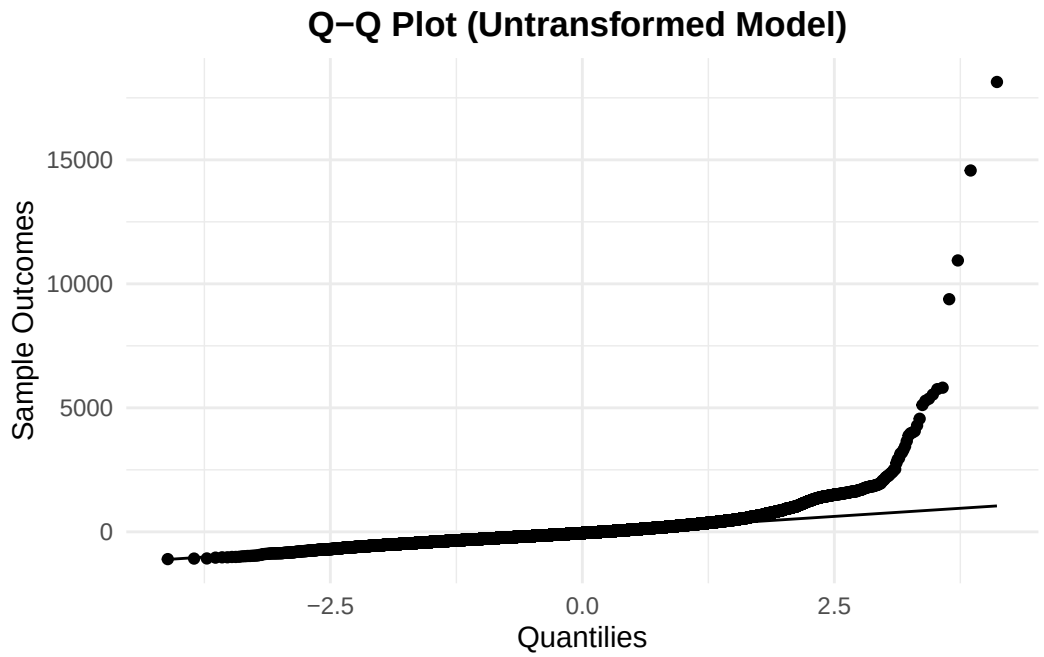
The following two graphs are checking key regression assumptions to see if our model violates the usual conditions for valid inference. The residuals vs. fitted plot checks linearity and equal variance: we want the residuals scattered randomly around zero with no clear curve and roughly the same vertical 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"))
```



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 the

```r
# 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 = "Quantilies", y = "Sample Outcomes") +
  theme_minimal()+
    theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



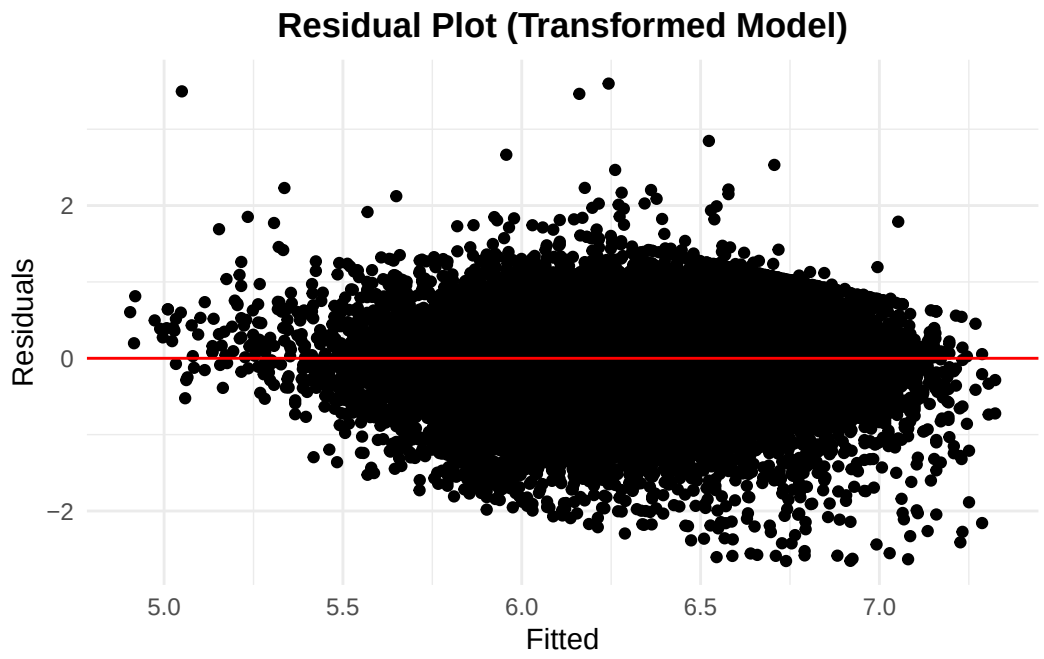

```
Call:
lm(formula = log(Wage) ~ Education + Experience + Black * Region,
    data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6555 -0.3022  0.0444  0.3511  3.5975

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.6601287  0.0194813 239.211  < 2e-16 ***
Education          0.0989226  0.0012075  81.926  < 2e-16 ***
Experience         0.0182709  0.0002811  64.992  < 2e-16 ***
BlackYes          -0.1928006  0.0306687  -6.287 3.30e-10 ***
RegionNE           0.0606917  0.0099847   6.078 1.23e-09 ***
RegionS           -0.0547966  0.0095287  -5.751 8.99e-09 ***
RegionW            0.0033789  0.0101459   0.333    0.739
BlackYes:RegionNE -0.0035277  0.0431955  -0.082    0.935
```

```
BlackYes:RegionS  -0.0417635  0.0350495  -1.192    0.233
BlackYes:RegionW   0.0382073  0.0514264   0.743    0.458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5357 on 25621 degrees of freedom
Multiple R-squared:  0.2711,    Adjusted R-squared:  0.2709
F-statistic:  1059 on 9 and 25621 DF,  p-value: < 2.2e-16
```

```r
#Prettifying it
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

```r
# 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 = "Quantilies", 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 refernce 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 proofing our research question.

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

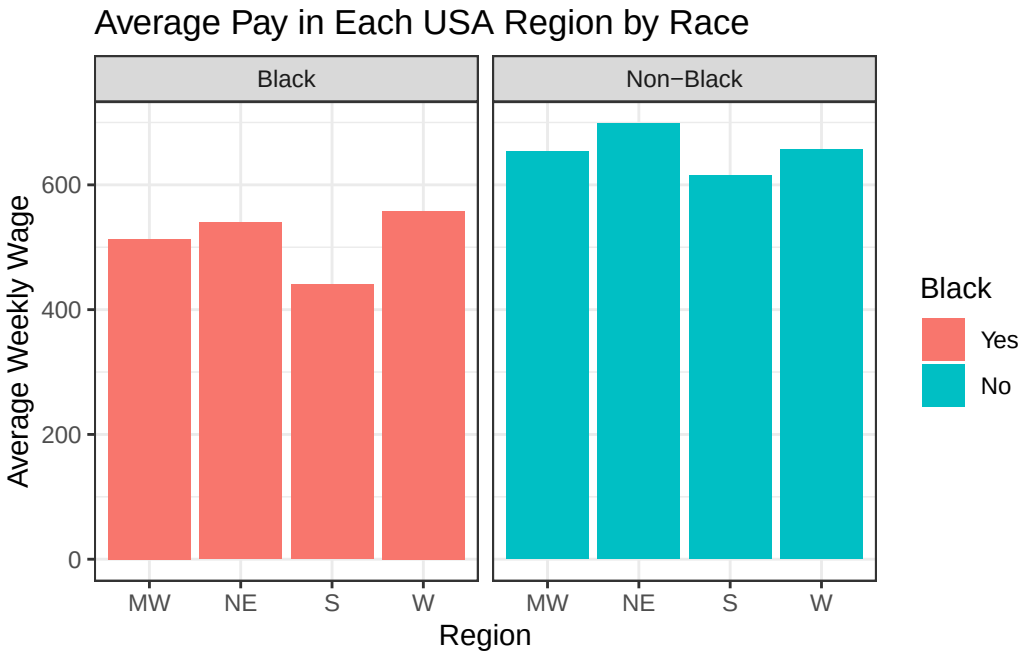
```
f_results <- anova(wages_reduced_model,mlm_model)
f_results
```

```
Analysis of Variance Table

Model 1: log(Wage) ~ Education + Experience + Black + Region
Model 2: log(Wage) ~ Education + Experience + Black * Region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  25624 7353.4
2  25621 7352.1  3    1.2473 1.4489 0.2264
```

```
anova_df <- as.data.frame(f_results)
flextable(anova_df)|>
  set_caption("F-test Result Summary")|>
  autofit()|>
  theme_apa()
```

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

```
#creating new labels for the facet wrapped variables
graph.labels <- c("Yes"="Black","No" = "Non-Black")

#plotting a graph exploring Avg pay by region faceted by race
wages_region_avg_wage |> ggplot(aes(x=avg_wage_region, y = Region, fill = Black)) +
  geom_col()+
  facet_grid(~Black,labeller = labeller(Black = graph.labels) )+
  coord_flip() +labs(title = "Average Pay in Each USA Region by Race", x= "Average Weekly Wage
  theme_bw()
```



#Region does not affect wages in races –> then the question is not if race affects wage, it is if region affects the wage gap

```
library(knitr)
#Saving summary table of wages data
wages_summary <-summary(wages)
wages_df <- as.data.frame(wages_summary)


#printing the summary table created above with captions.
#flextable(wages_df)|>
  #set_caption("Wage Summary")|>
```

```
  #autofit()|>
  #theme_apa()
```

Graph for Average Education by Region and Race, USA

```
#get averages of education per region
wages_edu_group_region <- wages |>
  group_by(Region, Black) |>
  summarise(avg_education = mean(Education)) |>
  ungroup()

graph.labels <- c("Yes"="Black","No" = "Non-Black")

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"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.line.y = element_blank(),
    axis.line.x = element_blank(),
  ) + coord_flip()
```