Money Talks:A Statistical Analysis of Weekly Earnings Variables

Tianyi
Wang

31 Nov 2024

1 Introduction

In today's dynamic labor market, understanding the determinants of weekly earnings remains a fundamental challenge for economists, policymakers, and organizational leaders. This report presents a comprehensive statistical analysis examining how geographical location, age, educational attainment, and racial demographics influence income patterns. The relationship between earnings and these socioeconomic factors is multifaceted and continues to evolve with changing economic conditions. Our analysis employs robust statistical methods to examine these relationships using linear models, aiming to provide insights that can inform policy initiatives to reduce income gaps and promote economic opportunity among different demographic groups.

2 Data

We have 4903 different observations, where the amount of the wage earned by a specific person during a week is defined as the response variable. The 5 available explanatory variables are: the **Age** of the person, the only integer variable ranging from 16 to 85; the **Region** of this person, a categorical variable with levels {West, Midwest, South, Northeast}; the **Location** of this person, a categorical variable with 2 levels {Metro, NotMetro}; the **Education Level(EducCateg)** of this person, a categorical variable with 16 levels and its corresponding EducCode from low to high {LessThanFirstGrade (Code: 1), FirstSecondThirdOrFourthGrade (Code: 2), FifthorSixthGrade (Code: 3), SeventhOrEighthGrade (Code: 4), NinthGrade (Code: 5), TenthGrade (Code: 6), EleventhGrade (Code: 7) , TwelthButNoDiploma (Code: 8), HighSchoolDiploma (Code: 9), SomeCollegeButNoDegree (Code: 10), AssocDegOccupVocat (Code: 11), AssocDegAcadem (Code: 12), BachelorsDegree (Code: 13), MastersDegree (Code: 14) ProfSchoolDegree (Code: 15), DoctorateDegree (Code: 16)} ; and the **Race** of this person, a categorical variable with 2 levels {Black, White}.

The plots shown in Figure 1 illustrate that the distribution of wages and the relationship between it and age. Firstly, the density curve suggests that the distribution is positively skewed, with the majority of individuals earning wages clustered at a lower level, while gradual tapering of the curve toward higher earnings indicates few people earning at the upper end of the scale. Secondly, the scatter plot reveals a general trend where wages increase with age up to a certain point, likely reflecting the accumulation of experience and career advancement. Meanwhile,the points are widely dispersed, highlighting significant variability in earnings across all age groups.

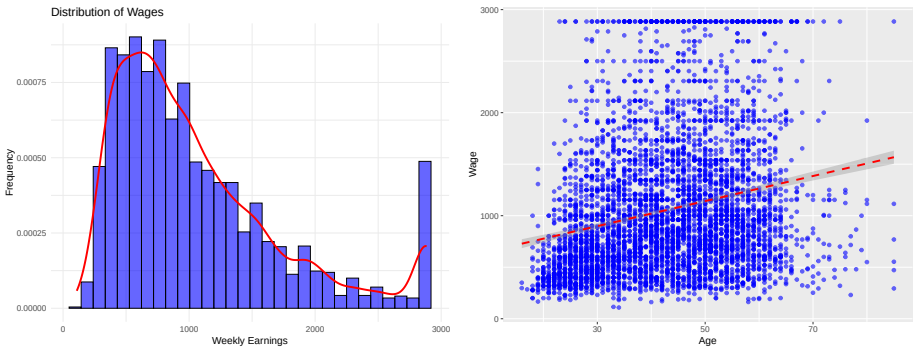


Figure 1: (LEFT) Distribution of Wage, (RIGHT) Relationship Between Age and Wage

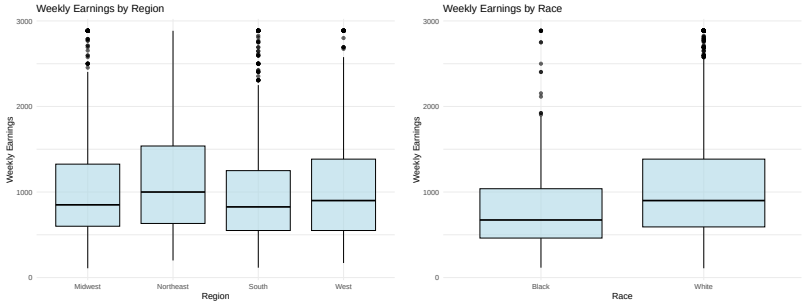


Figure 2: Box-plots of the number of wages within the levels of two factors

Box-plots shown in Figure 2 illustrate that people who are white tend to make more money than people who are black, and also that highest weekly earnings are made in Northeast region. The mean number of weekly earnings made by people from RaceBlack, White are 822 and 1078 respectively. The mean number of weekly earnings made by people from RegionMidwest, Northeast, South, West are 1023, 1176, 1001 and 1063, respectively. Box-plot shown in Figure 3 illustrates that a clear trend that people with higher education levels tend to earn more on average. For example, those with advanced degrees such as Doctorate or Master's degrees exhibit significantly higher median wage while individuals with education levels below a high school diploma generally earn the least. The interaction plot shown in Figure 3 suggests that the effect of the two factors, region and location, on weekly earnings is not likely additive, as the lines for "Metro" and "NotMetro" are not relatively parallel.

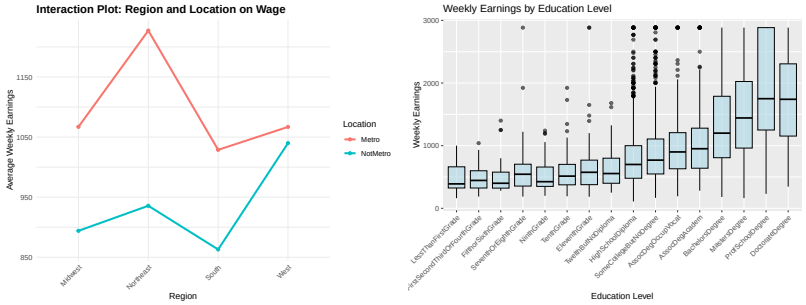


Figure 3: Box-plot of the number of wages regarding to education level and Interaction Plot :Region and Location on Wage

3 Modelling

We start by fitting a normal linear model which includes the main effects. We assume that the errors are independent and normally distributed with constant variance. We can report the results of a fitted normal linear model:

`Model 1 : Wage \sim Region + Location + Age + Race + EducCateg`

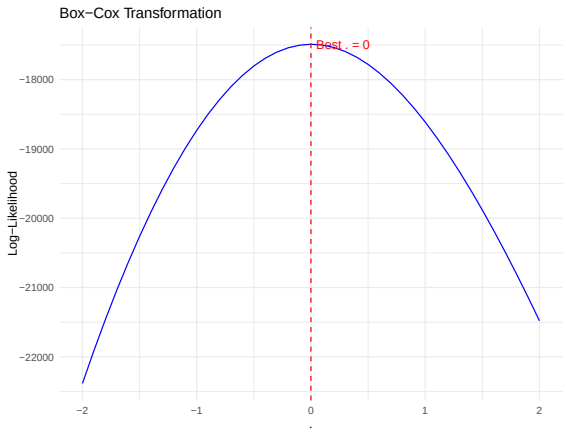


Figure 4: Box-Cox Transformation

To identify the most appropriate transformation in this case we can use the Box-Cox family of transformations. The profile log-likelihood plot for parameter λ shown in Figure 4 suggests setting $\lambda = 0$ which corresponds to the log-transformation. Hence, we can renew our model:

`Model 2 : log(Wage) ~ Region + Location + Race + Age + EducCateg`

The full model is fitted to the log-transformed response and the diagnostic graphs are acceptable; in Figure 5(Right) the points are randomly scattered in a relatively even band across the x-axis and in Figure 5(Left) the points tend to follow the straight line through the origin. There are very few studentized residuals outside the (-3, 3) range.

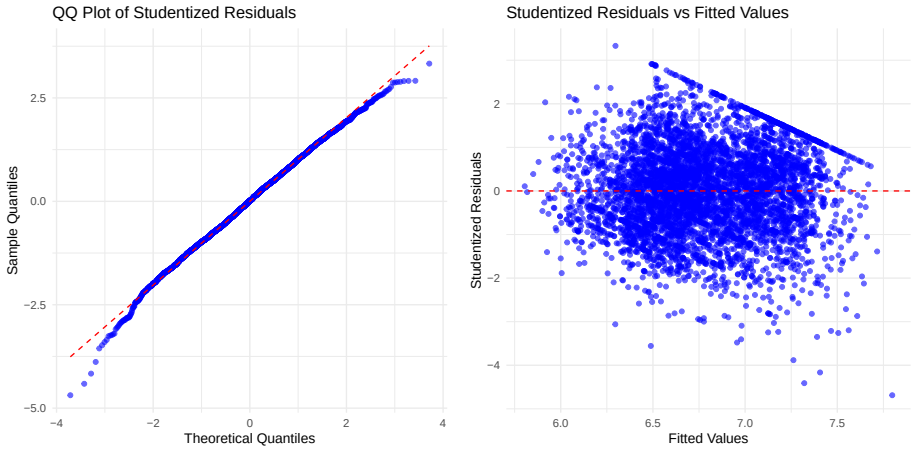


Figure 5: (LEFT) QQ Plot of Studentized Residuals, (RIGHT) Studentized Residuals vs Fitted Values

Proceeding with model selection, we use the F-test to test H_0: the interaction between **Region** and **Location** is not significant, versus the alternative that it is significant. The F-statistic of this test is equal to 3.67, which is in the 5% critical region of an F-distribution with 3 and 4863 degrees of freedom (p-value = 0.01175). Therefore, we reject the null hypothesis and conclude that the interaction between **Region** and **Location** is significant. This result indicates that the interaction term should be retained in the model to improve the model's fitness.

Similarly, we use the F-test to test H_0: the interaction between **Age** and **EducCateg** is not significant, versus the alternative that it is significant. The F-statistic of this test is equal to 2.86, which is in the 5% critical region of an F distribution with 15 and 4863 degrees of freedom (p-value = 0.00017). Therefore, we reject the null hypothesis and conclude that the interaction between **Age** and **EducCateg** is significant. This result indicates that the interaction term should also be retained in the model to improve the model's fitness.

The model is then fitted to the log-transformed response and the F-tests for dropping any of the main effects from the model are significant, as the ANOVA table below demonstrates.

```
Analysis of Variance Table
Response: log(Wage)
                  Df Sum Sq Mean Sq F value  Pr(>F)
Region             3     15     5.1   20.07 6.4e-13 ***
Location           1     13    12.7   49.81 1.9e-12 ***
Race               1     25    24.6   96.37 < 2e-16 ***
Age                1    119   118.7  464.28 < 2e-16 ***
EducCateg         15    390    26.0  101.59 < 2e-16 ***
Region:Location    3      3     0.9    3.67 0.01175 *
Age:EducCateg     15     11     0.7    2.86 0.00017 ***
Residuals       4863   1243     0.3
```

Therefore, the model we chose is:

```
Model 3 : log(Wage) ~ Region * Location + Race + Age * EducCateg
```

The diagnostic plots for the chosen model (not shown) are similar to those obtained for the corresponding model without interaction terms included (Figure 5) and are therefore acceptable. Moreover, there are very few studentized residuals outside the (-3, 3) range.

4 EducCateg v.s EducCode

In this section, we compare the two normal linear models, each of which includes the main effects and specified interaction terms. These models assume that the errors are independent and normally distributed with constant variance. Both models include an interaction between Region and Location, as well as the main effect of Race. However, there is a difference in the variable representing the education level:

```
Model 3.1 : log(Wage) ~ Region * Location + Race + Age * EducCateg
```

```
Model 3.2 : log(Wage) ~ Region * Location + Race + Age * EducCode
```

We use AIC to compare two models in Table 1. Model 3.1 has 41 degrees of freedom, indicating a significant complexity compared to Model 3.2, which has only 13 degrees of freedom. In terms of AIC, Model 3.1 achieves a slightly lower value (7268.04) than Model

3.2 (7293.77), suggesting a better balance between model fit and complexity. However, the slight difference in AIC indicates that the improvement in fit might not justify the substantially greater complexity of Model 3.1, as it could risk overfitting. As a result, we prefer Model 3.2.

	df	AIC
wage.lmfit3.1	41.00	7268.04
wage.lmfit3.2	13.00	7293.77

Table 1: AIC and df Comparison of 2 Models

5 Interpretation and Conclusions

A summary of our selected model is presented below:

```
Call:
lm(formula = log(Wage) ~ Region * Location + Race + Age * EducCode,
    data = wage)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1697 -0.3435  0.0062  0.3485  1.9034

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       4.660578   0.113798   40.95  < 2e-16 ***
RegionNortheast                   0.100614   0.024184    4.16  3.2e-05 ***
RegionSouth                      -0.006119   0.021912   -0.28  0.78005
RegionWest                        0.020391   0.023667    0.86  0.38896
LocationNotMetro                 -0.095156   0.033104   -2.87  0.00406 **
RaceWhite                         0.188979   0.026896    7.03  2.4e-12 ***
Age                               0.018391   0.002412    7.62  2.9e-14 ***
EducCode                          0.139592   0.010177   13.72  < 2e-16 ***
RegionNortheast:LocationNotMetro -0.091865   0.053332   -1.72  0.08504 .
RegionSouth:LocationNotMetro      0.030457   0.047693    0.64  0.52311
RegionWest:LocationNotMetro       0.100587   0.054686    1.84  0.06592 .
Age:EducCode                     -0.000742   0.000222   -3.34  0.00084 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.508 on 4891 degrees of freedom
Multiple R-squared:  0.305, Adjusted R-squared:  0.303
F-statistic:  195 on 11 and 4891 DF,  p-value: <2e-16
```

The final model evaluates the log-transformed weekly earnings (log(Wage)) as a function of region, location, race, age, and education level, along with their interactions.

Race : With the coefficient of **RaceWhite** and exp(0.188979) = 1.208 (95% CI: (1.146, 1.27)), which indicates that the expected weekly earnings of White workers are approximately 20.8% higher than those of Black workers.

Region : With the coefficient of Northeast and exp(0.100614) = 1.106 (95% CI: (1.055,

1.16)), which indicates that the expected weekly earnings of workers in the Northeast are approximately 10.6% higher than those in the Midwest (baseline), accounting for other factors.

Location : With the coefficient of LocationNotMetro and $\exp(-0.095156) \approx 0.909$ (95% CI: (0.852, 0.97)), which indicates that the expected wages in NotMetro areas are about 9.1% less than in Metro,accounting for other factors.

Interactions of Region and Location : With the interaction term RegionNortheast:LocationNotMetro has a coefficient of -0.091865 and $\exp(-0.091865) \approx 0.912$ (95% CI (0.822, 1.01)), suggesting that the expected wage of workers in NotMetro areas of the Northeast is slightly lower compared to the baseline. Other region-location interaction terms are not statistically significant, indicating no meaningful differences between these subgroups.

Age : The coefficient for Age is 0.018391, meaning that for every additional year of age, weekly earnings increase by a factor of $\exp(0.018391) \approx 1.019$, or 1.9%.

Education : The coefficient for EducCode is 0.139592 and $\exp(0.139592) \approx 1.150$, indicating a 15.0% increase in weekly earnings for each additional unit of education.

Interaction of Age and Education : The coefficient of the interaction term Age:EducCode is -0.000742, and $\exp(-0.000742) \approx 0.999$. This indicates that the wage premium associated with education is negligible as age increases, with a 95% CI (0.999, 1.00).

```
                     (Intercept)                      RegionNortheast
                         105.697                               1.106
                     RegionSouth                          RegionWest
                           0.994                               1.021
                LocationNotMetro                           RaceWhite
                           0.909                               1.208
                             Age                            EducCode
                           1.019                               1.150
RegionNortheast:LocationNotMetro     RegionSouth:LocationNotMetro
                           0.912                               1.031
     RegionWest:LocationNotMetro                    Age:EducCode
                           1.106                               0.999
```

	2.5 %	97.5 %
(Intercept)	84.562	132.11
RegionNortheast	1.055	1.16
RegionSouth	0.952	1.04
RegionWest	0.974	1.07
LocationNotMetro	0.852	0.97
RaceWhite	1.146	1.27
Age	1.014	1.02
EducCode	1.127	1.17
RegionNortheast:LocationNotMetro	0.822	1.01
RegionSouth:LocationNotMetro	0.939	1.13
RegionWest:LocationNotMetro	0.993	1.23
Age:EducCode	0.999	1.00

R code is attached below

R code

1073354

2024-11-30

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
##data
# Load necessary libraries
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/96/fcz5qq0j1s90r885t2jfg8zh0000gn/T//RtmpgWCtFP/downloaded_packages
```

```
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/96/fcz5qq0j1s90r885t2jfg8zh0000gn/T//RtmpgWCtFP/downloaded_packages
```

```
install.packages("stargazer")
```

```
##
## The downloaded binary packages are in
##  /var/folders/96/fcz5qq0j1s90r885t2jfg8zh0000gn/T//RtmpgWCtFP/downloaded_packages
```

```
install.packages("xtable")
```

```
##
## The downloaded binary packages are in
##  /var/folders/96/fcz5qq0j1s90r885t2jfg8zh0000gn/T//RtmpgWCtFP/downloaded_packages
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics
Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ggplot2)
library(dplyr)
```

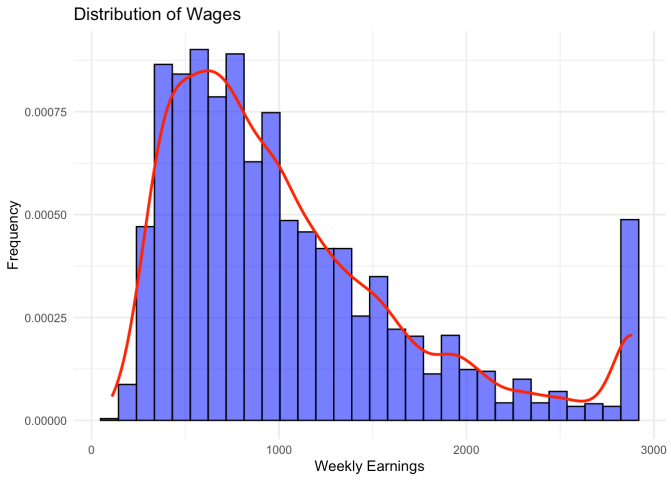


```
##
## Attaching package: 'dplyr'
```

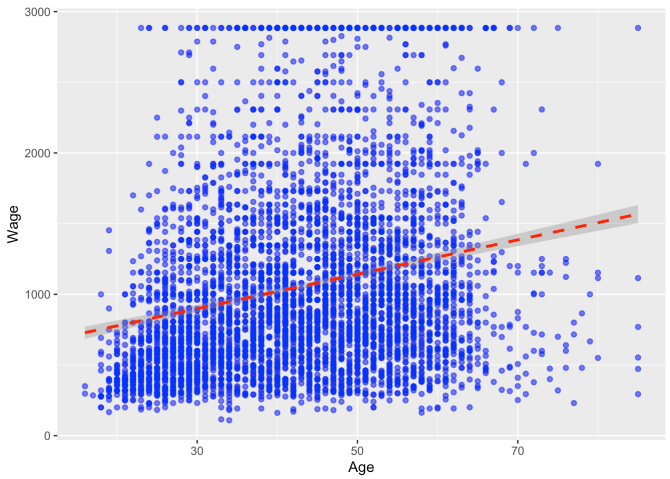
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(xtable)

# Load the dataset (adjust the file path as needed)
wage <- read.csv("https://www.stats.ox.ac.uk/~laws/SB1/data/wages.csv", stringsAsFact
ors = TRUE)
attach(wage)

# exploratory plots
# 1. Distribution of Wage
ggplot(wage, aes(x = Wage)) +
   geom_histogram(aes(y = after_stat(density)),bins = 30, fill = "blue", color = "bla
ck", alpha = 0.6) +
   geom_density(color = "red", linewidth = 1) +
   labs(title = "Distribution of Wages", x = "Weekly Earnings", y = "Frequency") +
   theme_minimal()
```



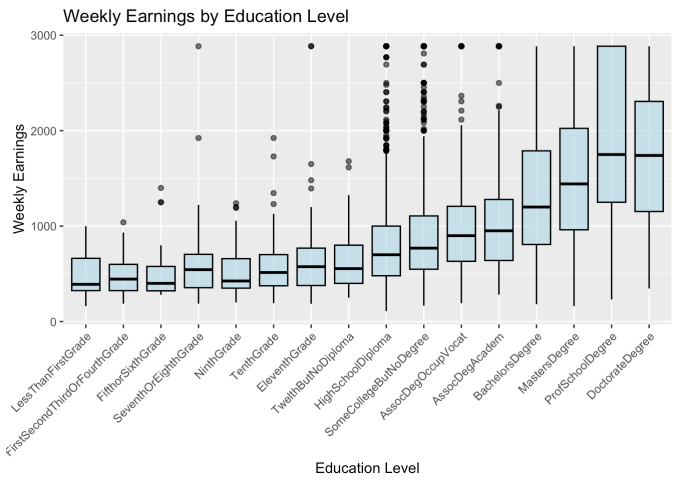

```
# 2. Relationship Between Age and Wage
ggplot(wage, aes(x = Age, y = Wage)) +
   geom_point(color = "blue", alpha = 0.6) +
   geom_smooth(method = "lm",formula = y ~ x, color = "red", linetype = 2)
```



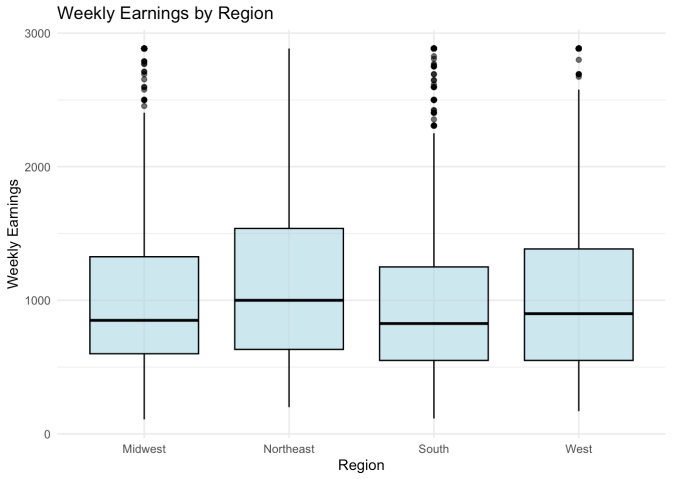

```
   labs(title = "Relationship Between Age and Wage", x = "Age", y = "Weekly Earning
s") +
   theme_minimal()
```

```
## NULL
```

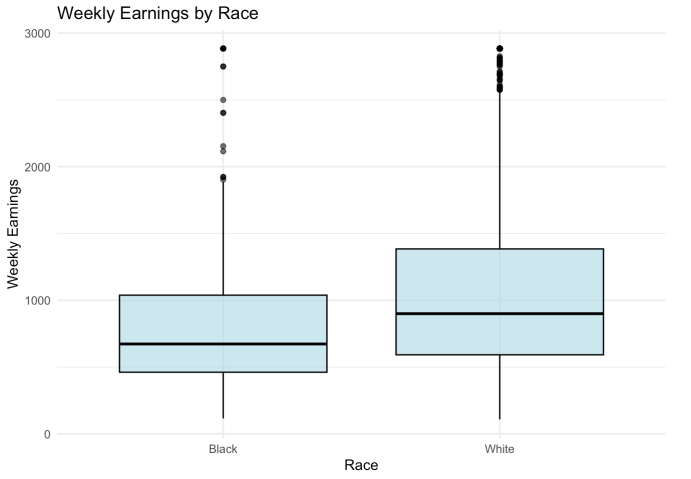
```
# 3. Weekly Earnings by Education Level
wage$EducCateg <- factor(wage$EducCateg, levels = c(
"LessThanFirstGrade", "FirstSecondThirdOrFourthGrade", "FifthorSixthGrade",
"SeventhOrEighthGrade", "NinthGrade", "TenthGrade", "EleventhGrade",
"TwelfthButNoDiploma", "HighSchoolDiploma", "SomeCollegeButNoDegree",
"AssocDegOccupVocat", "AssocDegAcadem", "BachelorsDegree", "MastersDegree",
"ProfSchoolDegree", "DoctorateDegree"
))
ggplot(wage, aes(x = EducCateg, y = Wage)) +
geom_boxplot(fill = "lightblue", color = "black", alpha = 0.6) +
labs(title = "Weekly Earnings by Education Level", x = "Education Level", y = "Weekly
Earnings") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
# 4. Weekly Earnings by Region
ggplot(wage, aes(x = Region, y = Wage)) +
geom_boxplot(fill = "lightblue", color = "black", alpha = 0.6) +
labs(title = "Weekly Earnings by Region", x = "Region", y = "Weekly Earnings") +
theme_minimal()
```




```
# 5. Weekly Earnings by Race
ggplot(wage, aes(x = Race, y = Wage)) +
geom_boxplot(fill = "lightblue", color = "black", alpha = 0.6) +
labs(title = "Weekly Earnings by Race", x = "Race", y = "Weekly Earnings") +
theme_minimal()
```


```
## numerical summaries
tapply(Wage, Race, mean)
```

```
##     Black     White
##  822.2913 1077.5200
```

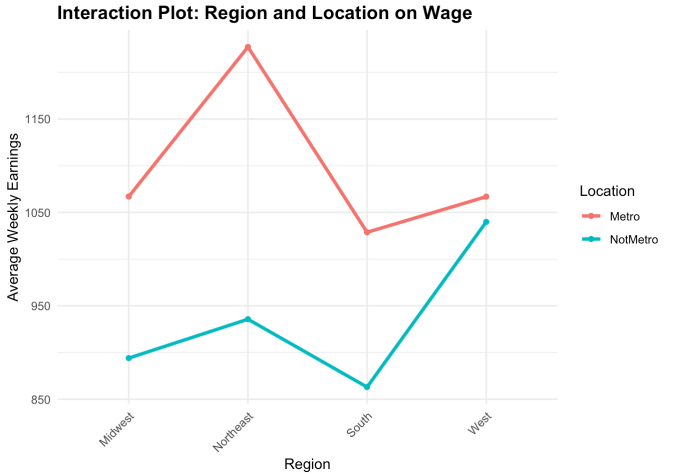
```
tapply(Wage, Region, mean)
```

```
##   Midwest Northeast     South      West
##  1023.303  1175.813  1000.814  1062.855
```

```
ggplot(wage, aes(x = Region, y = Wage, color = Location, group = Location)) +
  stat_summary(fun = mean, geom = "line", linewidth = 1.2) +
  stat_summary(fun = mean, geom = "point", linewidth = 3) +
  labs(
    title = "Interaction Plot: Region and Location on Wage",
    x = "Region",
    y = "Average Weekly Earnings",
    color = "Location"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(size = 14, face = "bold")
  )
```



```
## Warning in stat_summary(fun = mean, geom = "point", linewidth = 3): Ignoring
## unknown parameters: `linewidth`
```




```
#initial model
wage.lmfit1 <- lm(Wage ~ Region + Location + Age + Race + EducCateg)
```

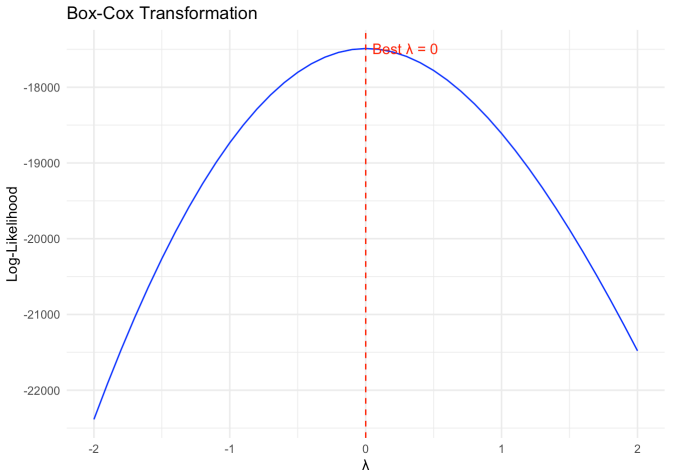
```
# Load necessary libraries - Perform Box-Cox transformation
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

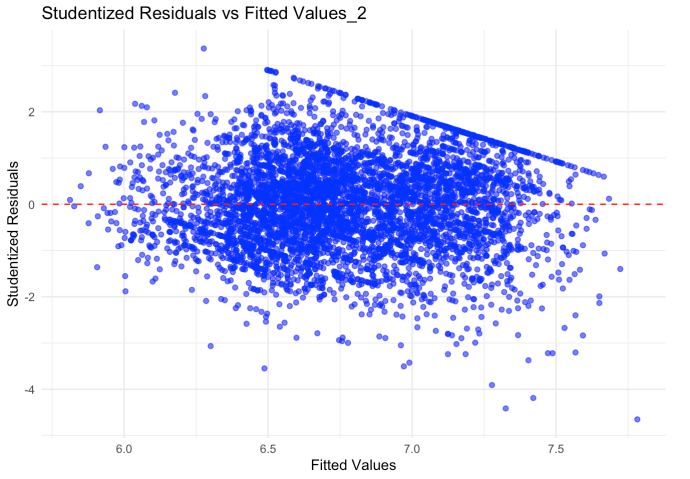
```
## The following object is masked from 'package:dplyr':
##
##     select
```



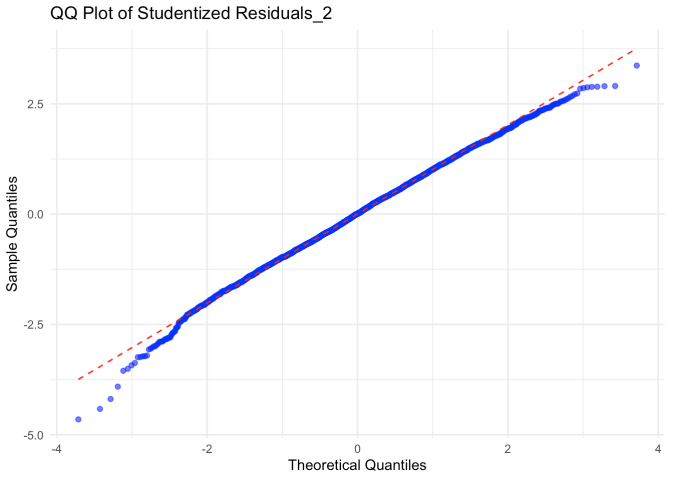
```
library(ggplot2)

# Fit the initial linear model
wage.lmfit1 <- lm(Wage ~ Region*Location + Age + Race + EducCateg, data = wage)

# Perform Box-Cox transformation
boxcox_results <- boxcox(wage.lmfit1, lambda = seq(-2, 2, by = 0.1), plotit = FALSE)

# Extract lambda and log-likelihood values
lambda_values <- boxcox_results$x
log_likelihoods <- boxcox_results$y

# Create a data frame for ggplot
boxcox_df <- data.frame(Lambda = lambda_values, LogLikelihood = log_likelihoods)

# Find the maximum likelihood point (best lambda)
best_lambda <- lambda_values[which.max(log_likelihoods)]

# Plot using ggplot
ggplot(boxcox_df, aes(x = Lambda, y = LogLikelihood)) +
  geom_line(color = "blue") +
  geom_vline(xintercept = best_lambda, linetype = "dashed", color = "red") +
  labs(
    title = "Box-Cox Transformation",
    x = "λ",
    y = "Log-Likelihood"
  ) +
  theme_minimal() +
  annotate(
    "text", x = best_lambda, y = max(log_likelihoods),
    label = paste("Best λ =", round(best_lambda, 2)),
    hjust = -0.1, color = "red"
  )
```




```
# models and plots with log-transformed response
wage.lmfit2 <- lm(log(Wage) ~ Region + Location + Age + Race + EducCateg, data = wag
e)
# Plot Studentized Residuals vs. Fitted Values & QQ Plot of Studentized Residuals
fitted_values <- fitted(wage.lmfit2)
studentized_residuals <- rstudent(wage.lmfit2)
residuals_df <- data.frame(Fitted = fitted_values, Residuals = studentized_residuals)

ggplot(residuals_df, aes(x = Fitted, y = Residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Studentized Residuals vs Fitted Values_2",
    x = "Fitted Values",
    y = "Studentized Residuals"
  ) +
  theme_minimal()
```


```
# save pdf
ggsave("Studentized Residuals vs Fitted Values_2.pdf", height = 5, width = 5)


ggplot(data = data.frame(Residuals = studentized_residuals), aes(sample = Residuals))
+
  stat_qq(color = "blue", alpha = 0.6) +
  stat_qq_line(linetype = "dashed", color = "red") +
  labs(
    title = "QQ Plot of Studentized Residuals_2",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_minimal()
```

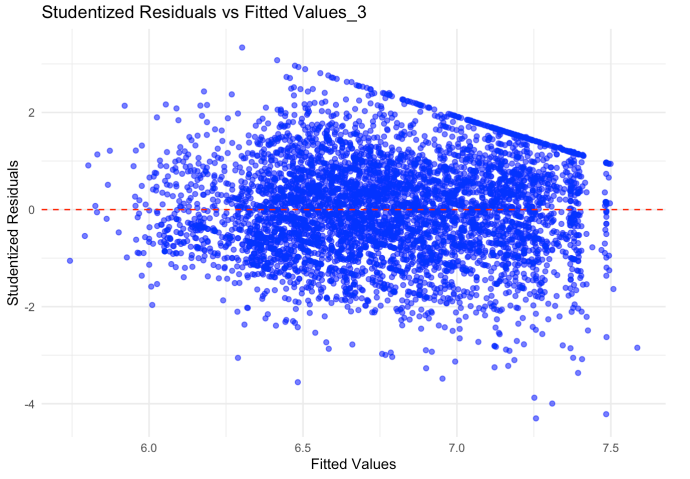

```
# Save the QQ plot
ggsave("QQ Plot of Studentized Residuals_2.pdf", height = 5, width = 5)
```

```
# model with variable interactions
wage.lmfit3 <- lm(log(Wage) ~ Region * Location + Race + Age * EducCateg, data = wag
e)
anova(wage.lmfit3)
```

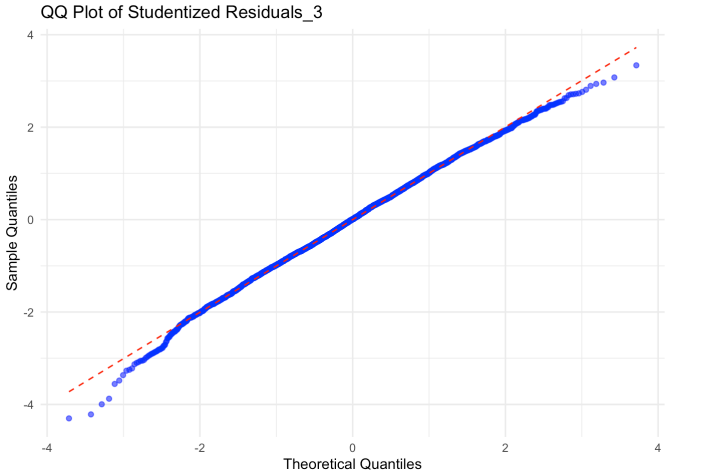
```
## Analysis of Variance Table
##
## Response: log(Wage)
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## Region             3   15.39   5.130  20.0674 6.363e-13 ***
## Location           1   12.73  12.733  49.8094 1.933e-12 ***
## Race               1   24.63  24.634  96.3695 < 2.2e-16 ***
## Age                1  118.68 118.682 464.2816 < 2.2e-16 ***
## EducCateg         15  389.54  25.969 101.5915 < 2.2e-16 ***
## Region:Location    3    2.81   0.938   3.6690 0.0117489 *
## Age:EducCateg     15   10.98   0.732   2.8632 0.0001681 ***
## Residuals       4863 1243.11   0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
# models and plots with log-transformed response
wage.lmfit3 <- lm(log(Wage) ~ Region * Location + Race + Age * EducCateg, data = wag
e)
# Plot Studentized Residuals vs. Fitted Values & QQ Plot of Studentized Residuals
fitted_values <- fitted(wage.lmfit3)
studentized_residuals <- rstudent(wage.lmfit3)
residuals_df <- data.frame(Fitted = fitted_values, Residuals = studentized_residuals)

ggplot(residuals_df, aes(x = Fitted, y = Residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Studentized Residuals vs Fitted Values_3",
    x = "Fitted Values",
    y = "Studentized Residuals"
  ) +
  theme_minimal()
```



```
# save pdf
ggsave("Studentized Residuals vs Fitted Values_3.pdf", height = 5, width = 5)


ggplot(data = data.frame(Residuals = studentized_residuals), aes(sample = Residuals))
+
  stat_qq(color = "blue", alpha = 0.6) +
  stat_qq_line(linetype = "dashed", color = "red") +
  labs(
    title = "QQ Plot of Studentized Residuals_3",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_minimal()
```




```
# Save the QQ plot
ggsave("QQ Plot of Studentized Residuals_3.pdf", height = 5, width = 5)
```

```
## EducCateg v.s EducCode
wage.lmfit3.1 <- lm(log(Wage) ~ Region * Location + Race + Age * EducCateg, data = wa
ge)
wage.lmfit3.2 <- lm(log(Wage) ~ Region * Location + Race + Age * EducCode, data = wag
e)

# Compare models directly
aic_comparison <- AIC(wage.lmfit3.1, wage.lmfit3.2)
print(aic_comparison)
```



```
##              df      AIC
## wage.lmfit3.1 41 7268.043
## wage.lmfit3.2 13 7293.768
```

```
aic_table <- xtable(aic_comparison, caption = "AIC Comparison of Models")
print(aic_table, type = "latex", include.rownames = TRUE, sanitize.text.function = id
entity)
```

```
## % latex table generated in R 4.4.1 by xtable 1.8-4 package
## % Sun Dec  1 00:10:16 2024
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & df & AIC \\
##   \hline
## wage.lmfit3.1 & 41.00 & 7268.04 \\
##   wage.lmfit3.2 & 13.00 & 7293.77 \\
##    \hline
## \end{tabular}
## \caption{AIC Comparison of Models}
## \end{table}
```

```
## final
wage.lmfit3.2 <- lm(log(Wage) ~ Region * Location + Race + Age * EducCode, data = wag
e)

summary(wage.lmfit3.2)
```



```
##
## Call:
## lm(formula = log(Wage) ~ Region * Location + Race + Age * EducCode,
##     data = wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16973 -0.34346  0.00618  0.34850  1.90338
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.6605777  0.1137978  40.955  < 2e-16 ***
## RegionNortheast                   0.1006143  0.0241842   4.160 3.23e-05 ***
## RegionSouth                      -0.0061195  0.0219125  -0.279 0.780051
## RegionWest                        0.0203905  0.0236666   0.862 0.388965
## LocationNotMetro                 -0.0951559  0.0331042  -2.874 0.004065 **
## RaceWhite                         0.1889789  0.0268963   7.026 2.41e-12 ***
## Age                               0.0183906  0.0024124   7.623 2.95e-14 ***
## EducCode                          0.1395917  0.0101766  13.717  < 2e-16 ***
## RegionNortheast:LocationNotMetro -0.0918646  0.0533324  -1.722 0.085044 .
## RegionSouth:LocationNotMetro      0.0304573  0.0476930   0.639 0.523106
## RegionWest:LocationNotMetro       0.1005874  0.0546855   1.839 0.065920 .
## Age:EducCode                     -0.0007419  0.0002221  -3.341 0.000841 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5084 on 4891 degrees of freedom
## Multiple R-squared:  0.3047, Adjusted R-squared:  0.3031
## F-statistic: 194.8 on 11 and 4891 DF,  p-value: < 2.2e-16
```

```
coef(wage.lmfit3.2)
```

```
##                       (Intercept)                   RegionNortheast
##                     4.6605777142                     0.1006143046
##                       RegionSouth                        RegionWest
##                     -0.0061194600                     0.0203905050
##                   LocationNotMetro                        RaceWhite
##                     -0.0951559143                     0.1889788924
##                               Age                           EducCode
##                      0.0183906092                     0.1395916773
## RegionNortheast:LocationNotMetro     RegionSouth:LocationNotMetro
##                     -0.0918645577                     0.0304572804
##       RegionWest:LocationNotMetro                     Age:EducCode
##                      0.1005873638                    -0.0007419095
```

```
confint(wage.lmfit3.2)
```



```
##                                       2.5 %       97.5 %
## (Intercept)                      4.437482820  4.883672608
## RegionNortheast                  0.053202316  0.148026293
## RegionSouth                     -0.049077723  0.036838804
## RegionWest                      -0.026006652  0.066787662
## LocationNotMetro                -0.160055044 -0.030256785
## RaceWhite                        0.136250082  0.241707703
## Age                              0.013661296  0.023119923
## EducCode                         0.119641030  0.159542324
## RegionNortheast:LocationNotMetro -0.196419973  0.012690858
## RegionSouth:LocationNotMetro     -0.063042468  0.123957029
## RegionWest:LocationNotMetro      -0.006620827  0.207795555
## Age:EducCode                     -0.001177242 -0.000306577
```

```
exp(coef(wage.lmfit3.2))
```

```
##                       (Intercept)                   RegionNortheast
##                     105.6971272                       1.1058500
##                       RegionSouth                        RegionWest
##                       0.9938992                       1.0205998
##                   LocationNotMetro                        RaceWhite
##                       0.9092312                       1.2080155
##                               Age                           EducCode
##                       1.0185608                       1.1498042
## RegionNortheast:LocationNotMetro     RegionSouth:LocationNotMetro
##                       0.9122287                       1.0309258
##       RegionWest:LocationNotMetro                     Age:EducCode
##                       1.1058202                       0.9992584
```

```
exp(confint(wage.lmfit3.2))
```

```
##                                       2.5 %      97.5 %
## (Intercept)                      84.5618162 132.1149806
## RegionNortheast                   1.0546430   1.1595434
## RegionSouth                       0.9521071   1.0375258
## RegionWest                        0.9743286   1.0690684
## LocationNotMetro                  0.8520969   0.9701964
## RaceWhite                         1.1459684   1.2734219
## Age                               1.0137550   1.0233893
## EducCode                          1.1270922   1.1729739
## RegionNortheast:LocationNotMetro  0.8216671   1.0127717
## RegionSouth:LocationNotMetro      0.9389036   1.1319672
## RegionWest:LocationNotMetro       0.9934010   1.2309615
## Age:EducCode                      0.9988235   0.9996935
```