Dongmin Wu

918110396

STA108 Linear Regression

Instructor: Lingfei Cui

Introduction:

The goal of this report is to compare the relationship between **Income:** *Total Personal Income (dollars)* and **Degree:** *Percent of Adult population (persons 25 years old or older) with bachelor's degree* in different **Regions** (North East, North Central, South, West) by using dataset **'CDI.csv'** that includes variables 'income', 'degree', 'region'. We will fit linear regression models for each region and trying to find the best model that explains the impact of education level on income by comparing the models.

Summary:

Overall Summary:

-Income: The income ranges from 8,899 to 37,541 dollars. The mean income is 18,561 dollars. The median income is 17,759 dollars.

-Degree: The degree percentage ranges from 8.1% to 52.3%, with a mean of 21.08% and a median of 19.70%.

-Region: The data is categorized by four regions - North East (NE), North Central (NC), South (S), and West (W).

Summary by Regions:

-West (W): Income ranges from 11,379 to 37,541 dollars, with a mean of 18,323 dollars. Degree percentage ranges from 8.2% to 44%, with a mean of 22%.

-North Central (NC): Income ranges from 12,597 to 27,378 dollars, with a mean of 18,301 dollars. Degree percentage ranges from 8.5% to 41.9%, with a mean of 19.8%.

-South (S): Income ranges from 8,899 to 31,699 dollars, with a mean of 17,487 dollars. Degree percentage ranges from 9.1% to 52.3%, with a mean of 21%.

-North East (NE): Income ranges from 12,704 to 33,330 dollars, with a mean of 20,599 dollars. Degree percentage ranges from 8.1% to 38.3%, with a mean of 21.8%.

Detailed table summary including standard deviation and median will be included in table section (1.1)

Income vs. Degree Percentage scatter plot will be included (1.2)

Data Preparation:

We identified and removed outliers for model fitting. Outliers were determined by using IQR method. Total of 30 outliers include observations with extreme values in either income or degree percentage were identified and removed from the dataset.

After removing the outliers, new dataset was saved as "cleaneddata" which consists of 410 observations. Income now ranges from 10,190 to 26,248 dollars with mean of 17,943 dollars and a median of 17,488 dollars. Degree now ranges from 8.1% to 39.1% with a mean of 20.12% and a median of 19.3%.

A scatter plot without outliers was created to illustrate the relationship between percentage of individuals with a bachelor's degree and personal income. *(1.3)*

Outliers that were removed will be demonstrated (1.4)

Model fitting:

We will fit simple linear regression model for each region (NE, NC, S, W). We use the percentage of individuals with a bachelor's degree as the explanatory variable, and we use total personal income as the response variable. We will plot the estimated regression lines with scattered data points on

separated graphs *(1.5)* to determine the best predictor or region. And we will do model diagnostics and interpret the best model.

The models were compared using variance of the errors σ^2 and coefficient of determination R^2

NE: $\sigma^2 = 4988728$ $R^2 = 0.5280093$

NC: $\sigma^2 = 3388625$ $R^2 = 0.2907494$

S: $\sigma^2 = 5400081$ $R^2 = 0.2911495$

W: $\sigma^2 = 5764397$ $R^2 = 0.4692425$

Which the highest R^2 value indicates the best fit region, which is NE

Model Diagnostics:

We used several diagnostic tests on the Northeast region to evaluate the assumptions of the normal regression model.

1. Q-Q Plot of Residuals: In our plot, most points lied closed to the line, means that the residuals are approximately normally distributed. However, there are some minor departures at the tails.
2. Residuals vs. Fitted Values: In our plot, the residuals seem to be randomly scattered around the line. Which means variance of the residuals is roughly constant. However, a few possible outliers. *Graphs (1.6)*

Interpretation:

The linear regression model for NE region which is the best predictor can be represented by equation:

Income = 9223.8156 + 522.1588 × Degree Percentage

β0=9223.8156 represents the estimated average income when degree percentage is zero.

β1=522.1588 is the slope which represents that for each 1% increase in the degree percentage, the average income is expected to increase by approximately 522.1588 dollars.

Confidence Interval: **β0**: (7534.1318, 10913.4995), **β1**: (448.5001, 595.8176) means that we are 95% confident that true β0 and β1 lies in the given interval.

$R^2 = 0.5280093$ for NE region means approximately 52.8% of the variability in income can be explained by the degree percentage.

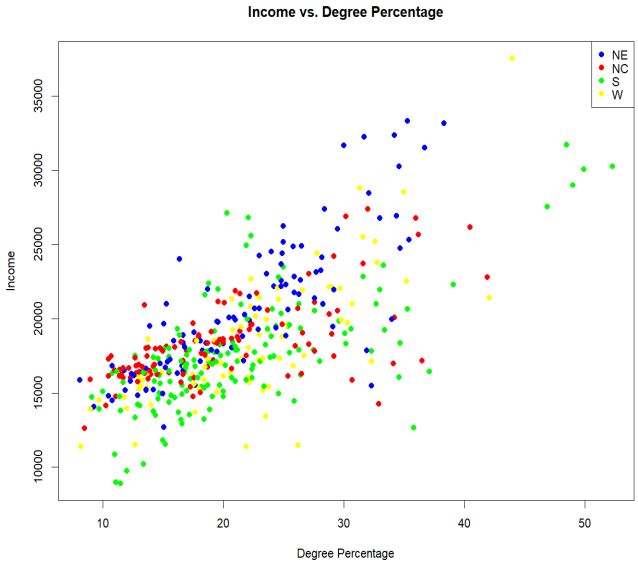
Conclusion:

We found out that the degree percentage is a significant predictor of income in different regions. In our dataset, Northeast region has the strongest relationship. However, one limitation is we assumed it's a linear relationship between degree percentage and income, which won't capture any nonlinear relation that might exist.

Tables/Graphs:

region	mean_income	median_income	sd_income	max_income	min_income	mean_degree	median_degree	sd_degree	max_degree	min_degree
<chr>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 NC	18301.	17817	2745.	27378	12597	19.8	18.2	7.46	41.9	8.5
2 NE	20599.	19785	4635.	33330	12704	21.8	21.6	7.22	38.3	8.1
3 S	17487.	17110	3844.	31699	8899	21.0	19.2	8.20	52.3	9.1
4 W	18323.	17268	4276.	37541	11379	22.0	22.1	7.25	44	8.2

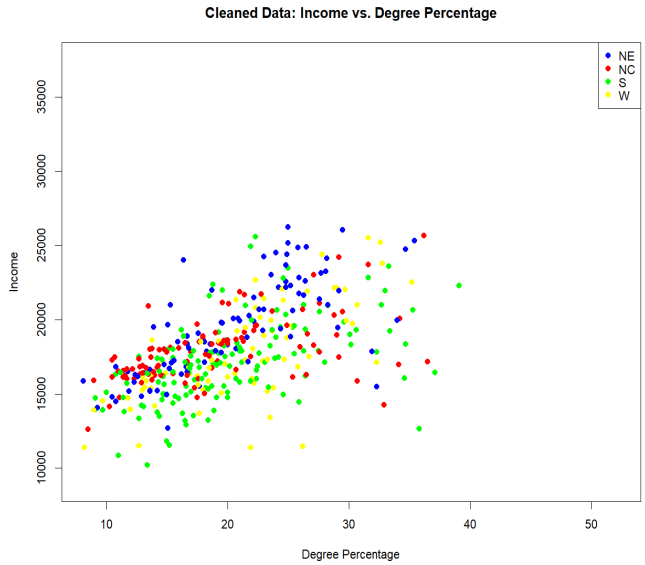
(1.2) Summary by Region



(1.2) Income vs. Degree Percentage scatter plot

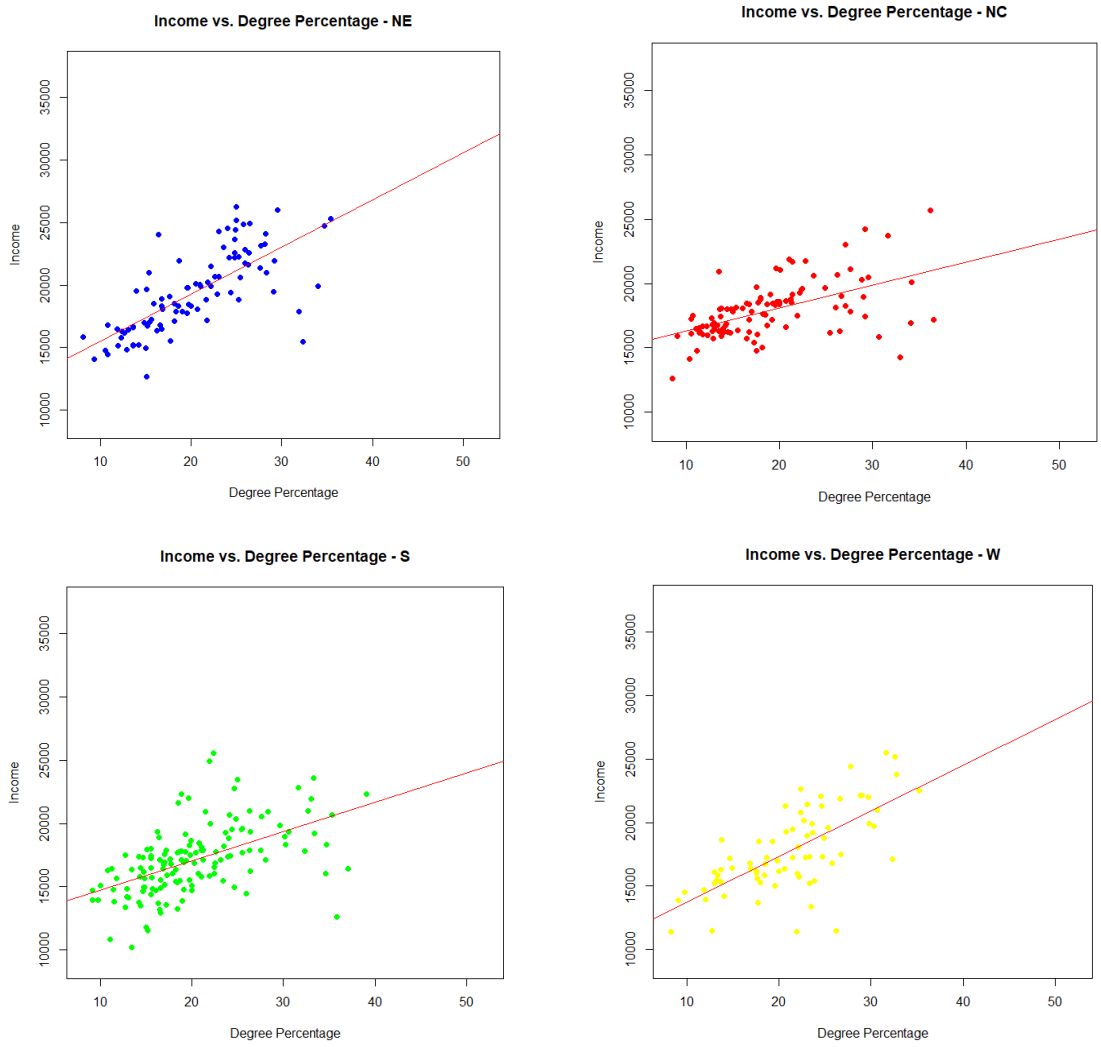
	income	degree	region
19	31679	30.0	NE
25	26884	30.2	NC
32	33330	35.3	NE
34	26798	22.1	S
39	32342	34.2	NE
41	32230	31.7	NE
42	28999	49.0	S
46	26772	36.0	NC
48	30081	49.9	S
53	28532	35.0	W
58	28462	32.1	NE
69	28819	31.3	W
72	26909	34.4	NE
81	27391	28.4	NE
87	27378	32.0	NC
117	31520	36.7	NE
128	8899	11.5	S
141	26156	40.5	NC
168	22782	41.9	NC
180	26757	33.0	NE
188	9728	12.0	S
201	33180	38.3	NE
206	37541	44.0	W
214	21421	42.1	W
248	27546	46.9	S
272	30242	52.3	S
337	8973	11.1	S
396	31699	48.5	S
410	30255	34.6	NE
437	27125	20.3	S

(1.4) Outliers that were removed

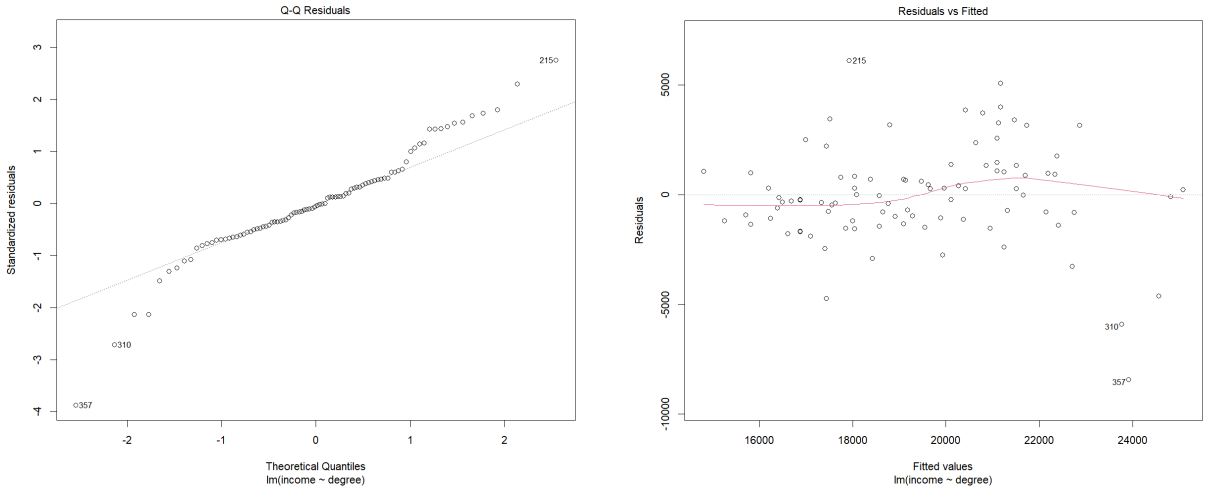


(1.3) Income vs. Degree Percentage scatter plot without outliers

Tables/Graphs:



(1.5) Estimated regression lines with scattered data points on North East, North Central, South, West



(1.6) Model diagnostics' graphs

R appendix:

```
library(ggplot2)
library(dplyr)
data <- read.csv("E:/Study/STA108/Project/CDI.csv")
print("Data Structure:")
str(data)
print("First Few Rows of Data:")
head(data)
print("Summary of Data:")
summary(data)
#scatter plot
colors <- c("NE" = "blue", "NC" = "red", "S" = "green", "W" = "yellow")
plot(data$degree, data$income, main="Income vs. Degree Percentage", xlab="Degree Percentage", ylab="Income",
col=colors[data$region], pch=19)
legend("topright", legend=names(colors), col=colors, pch=19)
#summary by region
data %>%
  group_by(region) %>%
  summarise(
    mean_income = mean(income),
    median_income = median(income),
    sd_income = sd(income),
    max_income = max(income),
    min_income = min(income),
    mean_degree = mean(degree),
    median_degree = median(degree),
    sd_degree = sd(degree),
    max_degree = max(degree),
    min_degree = min(degree)
  )
# IQR
incomeIQR <- IQR(data$income)
degreeIQR <- IQR(data$degree)
#upper lower bound
incomeUpper <- quantile(data$income, 0.75) + 1.5 * incomeIQR
incomeLower <- quantile(data$income, 0.25) - 1.5 * incomeIQR
degreeUpper <- quantile(data$degree, 0.75) + 1.5 * degreeIQR
degreeLower <- quantile(data$degree, 0.25) - 1.5 * degreeIQR
# remove outliers and print them
cleaneddata <- subset(data, income > incomeLower & income < incomeUpper & degree > degreeLower & degree < degreeUpper)
outliers <- subset(data, income <= incomeLower | income >= incomeUpper | degree <= degreeLower | degree >= degreeUpper)
print("Removed Outliers:")
print(outliers)
summary(cleaneddata)
```

```
x_range <- range(data$degree)
y_range <- range(data$income)
#plot without outliers
plot(cleaneddata$degree, cleaneddata$income, main="Cleaned Data: Income vs. Degree Percentage", xlab="Degree Percentage",
ylab="Income", col=colors[cleaneddata$region], pch=19, xlim=x_range, ylim=y_range)
legend("topright", legend=names(colors), col=colors, pch=19)
################################################################################
#Model fitting
# get data for each region
dataNE <- subset(cleaneddata, region == "NE")
dataNC <- subset(cleaneddata, region == "NC")
dataS <- subset(cleaneddata, region == "S")
dataW <- subset(cleaneddata, region == "W")
# function
plotfunc <- function(data, region, color) {
  model <- lm(income ~ degree, data = data)
  summary_model <- summary(model)
  # draw data and fit line
  plot(data$degree, data$income, main=paste("Income vs. Degree Percentage -", region),
       xlab="Degree Percentage", ylab="Income", col=color, pch=19, xlim=x_range, ylim=y_range)
  abline(model, col="red")
  # sigma2
  sigma2 <- sum(residuals(model)^2) / model$df.residual
  # ret
  return(list(model = model, sigma2 = sigma2, R2 = summary_model$r.squared))
}
# draw for each region
resNE <- plotfunc(dataNE, "NE", colors["NE"])
resNC <- plotfunc(dataNC, "NC", colors["NC"])
resS <- plotfunc(dataS, "S", colors["S"])
resW <- plotfunc(dataW, "W", colors["W"])
# print sigma2 and R2
print(paste("NE Region - Sigma2:", resNE$sigma2, "R2:", resNE$R2))
print(paste("NC Region - Sigma2:", resNC$sigma2, "R2:", resNC$R2))
print(paste("S Region - Sigma2:", resS$sigma2, "R2:", resS$R2))
print(paste("W Region - Sigma2:", resW$sigma2, "R2:", resW$R2))
# best predictor
R2_values <- c(NE = resNE$R2, NC = resNC$R2, S = resS$R2, W = resW$R2)
bestRegion <- names(which.max(R2_values))
print(paste("The best predictor/region is:", bestRegion))
bestModel <- switch(bestRegion,
                    "NE" = resNE$model,
                    "NC" = resNC$model,
                    "S" = resS$model,
```

```
                                "W" = resW$model)
plot(bestModel)
# LR model
modelNE <- lm(income ~ degree, data = subset(data, region == "NE"))
summary(modelNE)
# 95% CI
coefNE <- coef(modelNE)
confintNE <- confint(modelNE, level = 0.95)
print(coefNE)
print(confintNE)
```