Regression vs Tree based Model

Md Mominul Islam
101009250

About Data

1988

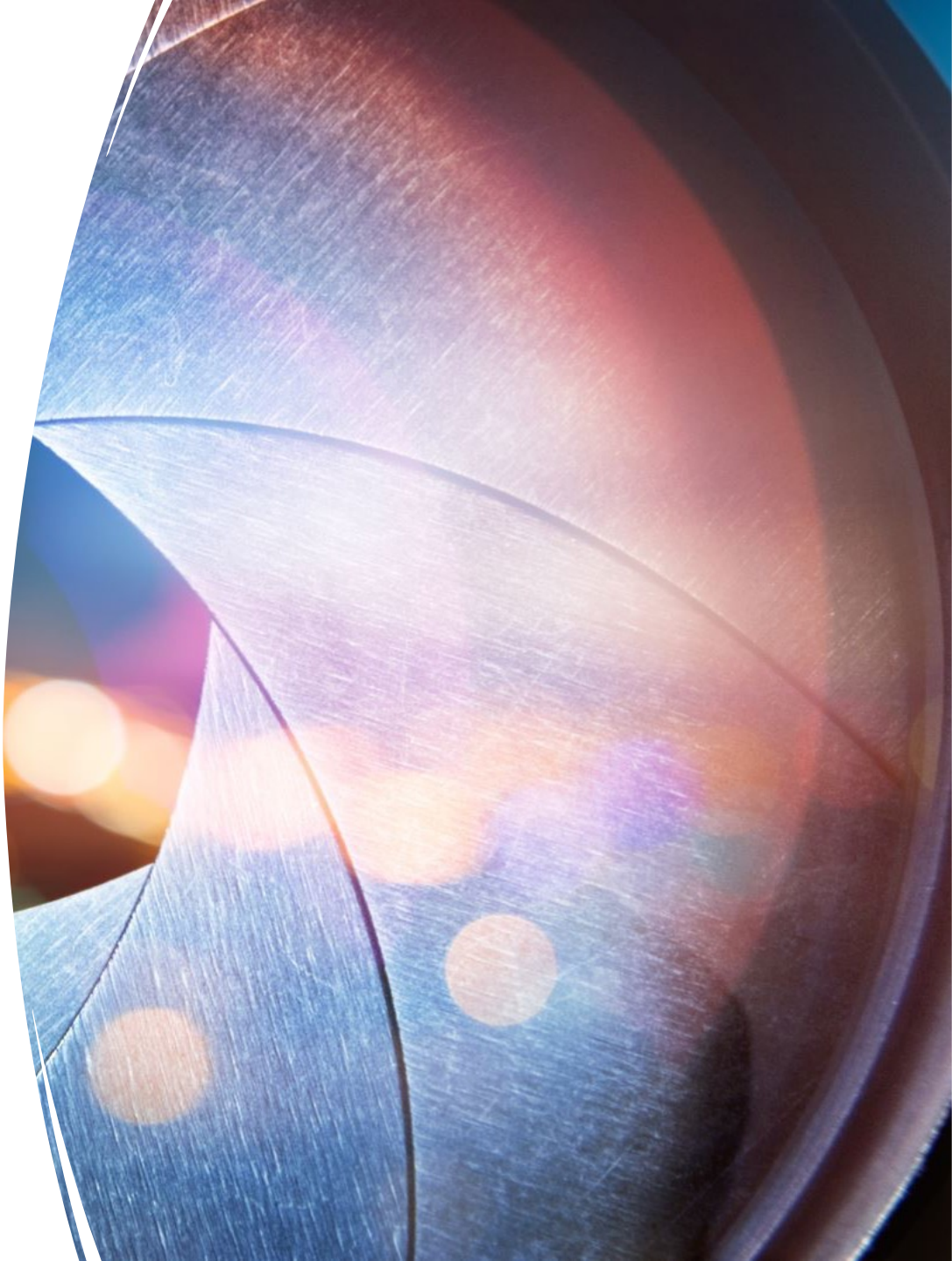
Weekly Wages for US male workers sampled from the Current Population Survey in 1988

Data frame has 2000 rows and 10 columns.

2000

OBJECTIVE

- When the dependent (goal) variable is a continuous number, it is important to understand and compare how error is quantified in OLS and tree regression.



Summary Statistics

Parameters	Wage	Education	Experience
Minimum Value	50.39	0.00	-2.00
First Quartile	308.64	12.00	8.00
Median	522.32	12.00	15.00
Mean	608.12	13.11	18.41
Third Quartile	783.48	16.00	27.00
Maximum Value	7716.05	18.00	59.00

Weekly wages

- Average wage $608.48.
- Minimum value $50.39 and Maximum value $7716.05

Education

0-18 Years

Experience

-2 may be a typo

Skewness

Wage	Education	Experience
3.7872164	-0.6967118	0.6568895

- Wage data is mostly skewed

Target/Dependent Variable

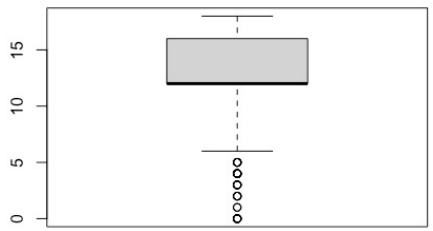
- Wage

Independent (predictor) variables

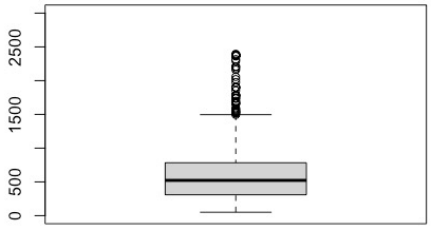
- Education
- Experience

Correlation Matrix

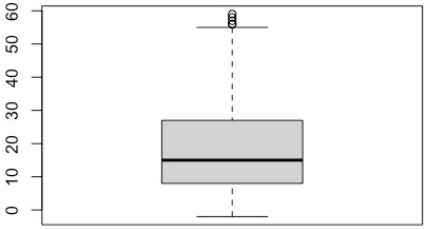
Boxplot Feature to explain skewness and distribution



Boxplot for Education



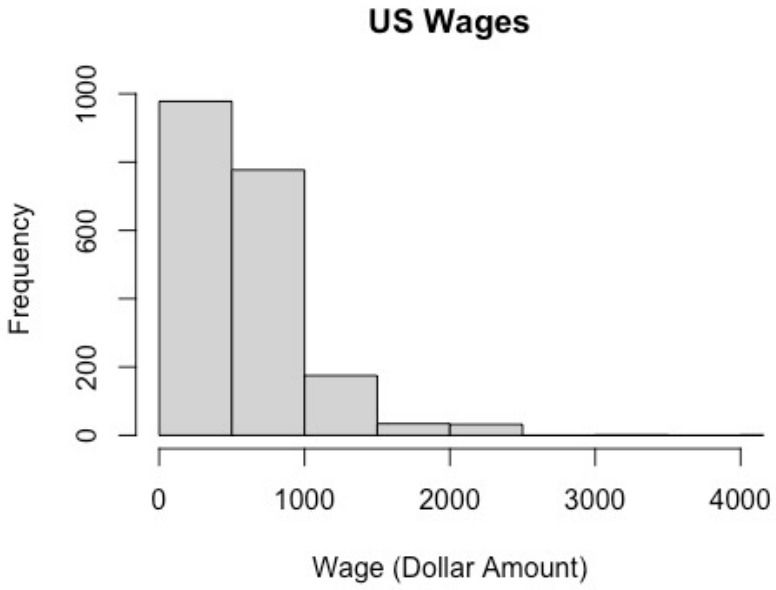
Boxplot for US Wages

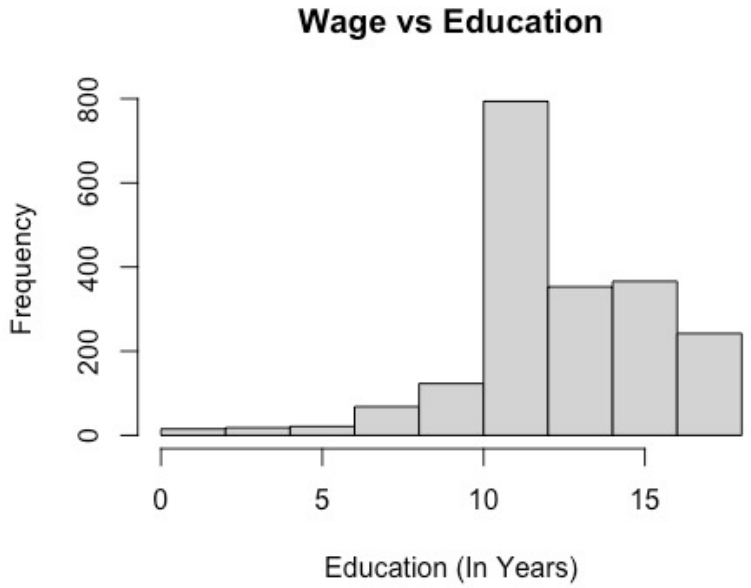


Boxplot for Experience

Histogram

Wage	Education	Experience
3.7872164	-0.6967118	0.6568895





- Wages are right skewed.
- Most people are having income in between $0-1500

- People with high school degree tend to have more wages

Scatterplot

- Hard to tell about trend
- Bucketing would be helpful to discern a pattern

With years of education, wage increases

ORDINARY LEAST SQUARES REGRESSION

Our regression model is defined in the following way:

$$\mathbf{wage} \sim -\mathbf{242.799412} + \mathbf{51.175268}(\mathbf{educ}) + \mathbf{9.774767}(\mathbf{exper}) + \epsilon$$

- With one year increase of education, there will be a raise of $51.17
- With one year increase of experience, there will be a raise of $9.77

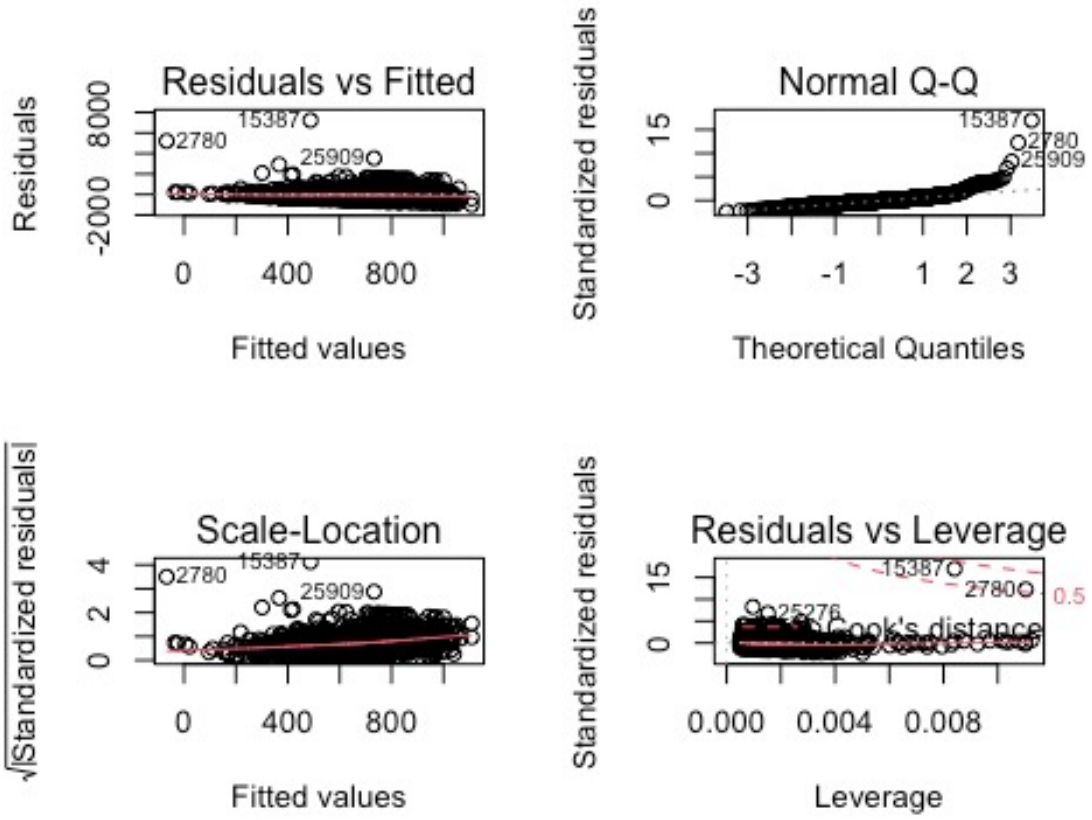
	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-242.7994	50.6816	-4.791	1.78e-06 ***
Education	51.1753	3.3419	15.313	2e-16 ***
Experience	9.7748	0.7506	13.023	2e-16 ***

F-statistics: 156, P-value (F-statistics): $< 2.2^{x}\ 10^{-16,}$

Residual standard error: 427.9

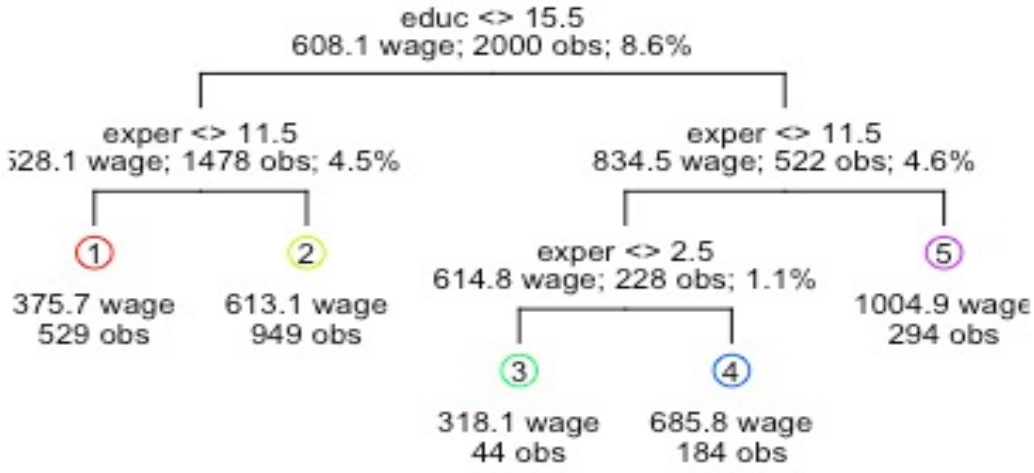
Multiple R-squared: 0.14, Adjusted R-squared: 0.13

REGRESSION MODEL ACCURACY



- Residual vs Fitted: Equally spread residuals around a horizontal line **without distinct patterns**.
- QQ plot: Data is **normally distributed** with some outliers like 15387, 2780 and 25909.

TREE REGRESSION MODEL



Sum of square errors (SSE)

- Tree model, with an SSE of 34298289, was superior at predicting weekly earnings for individuals.
- SSE 365568644 of OLS linear regression was equivalent to this model's performance.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Model	SSE
Regression Model	365568644
Tree Model	342982890