A Regression Analysis of Suicide Rates

Group 5:

Melissa Russell, Jonathan Voth, & Nick Wawee

Watch on YouTube

Introduction

Problem → Thousands of people die each year from suicide.

Goal → Aid suicide preventive measures by identifying the factors most significant in predicting suicides.

Predictions →

1. Males will have higher suicide rates than women.
2. Elderly adults will not have higher suicide rates than younger adults.

Introduction

Our Dataset: "Suicide Rates Overview 1985 to 2016"

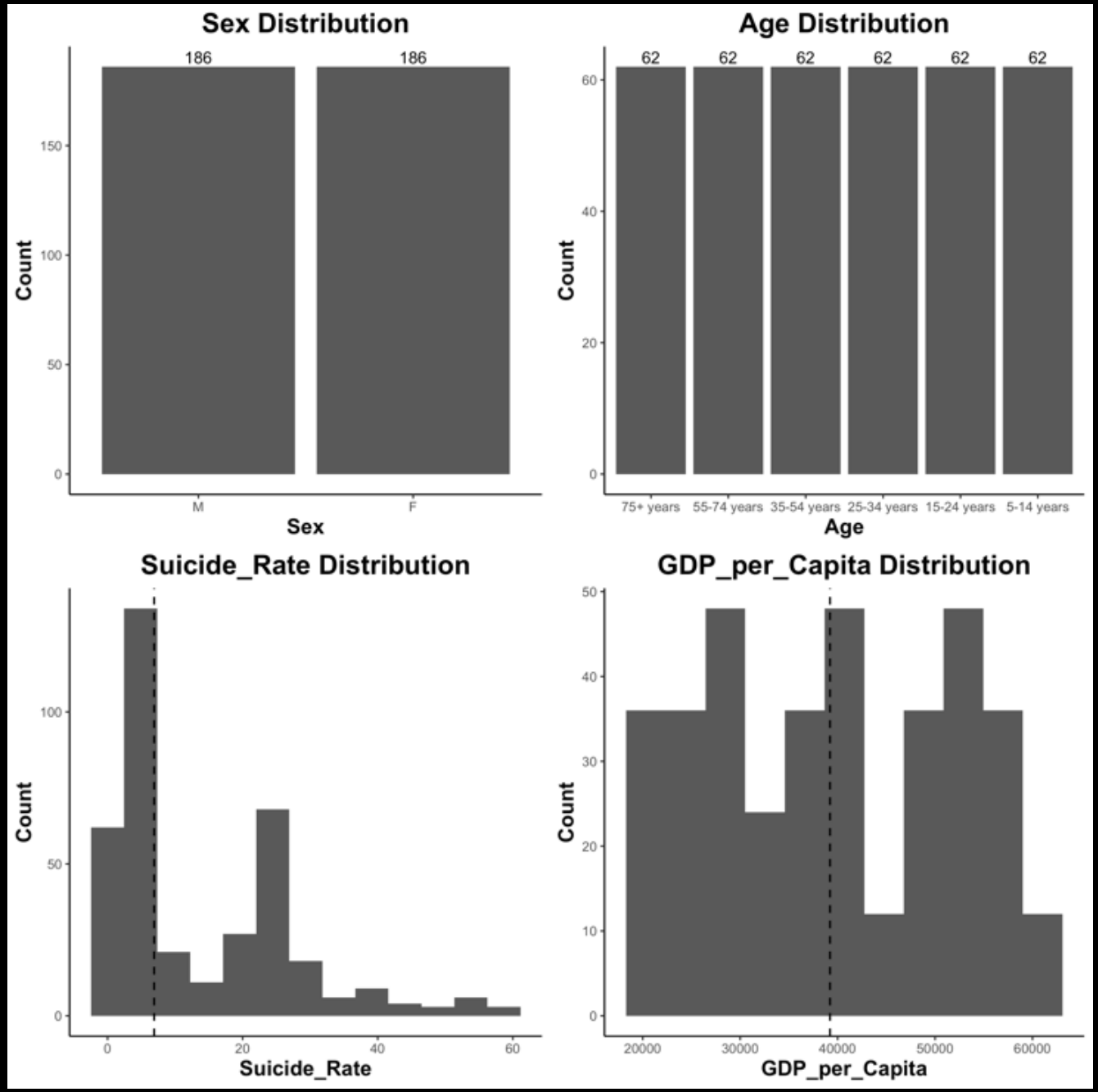
- Found on Kaggle.com
- Reports the suicide rate, sex, year, age group, population, GDP, and generation for 101 countries across the world
- Original dataset contains 27,821 rows
  - We narrow this down by only analyzing rates in the United States (372 rows)

Variables Included in our Analysis:

Variable	Response/Regressor	Type
Suicide Rate	Response	Numeric
Year	Regressor	Integer
Sex	Regressor	Factor (2 levels)
Age	Regressor	Factor (6 levels)
GDP	Regressor	Integer
GDP per Capita	Regressor	Integer

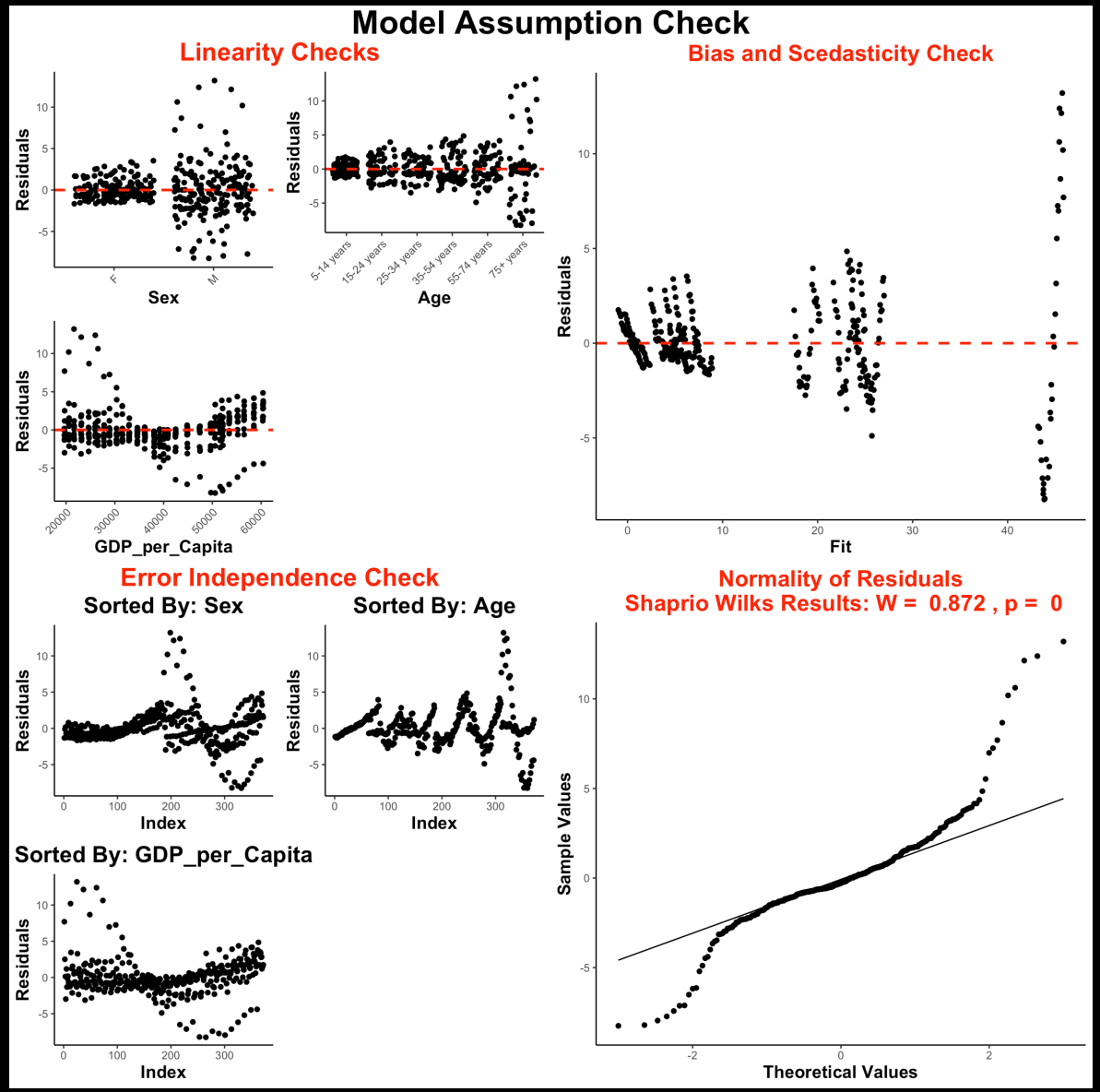
Model Fitting

- GDP was removed due to collinearity
- Interaction term was included between age and sex
- **Distributions:**
  - Uniform distribution of categorical regressors
  - Response variable shows a right skew (Rate = 25-30)
  - GDP per capita shows a left skew



Checking Model Assumptions

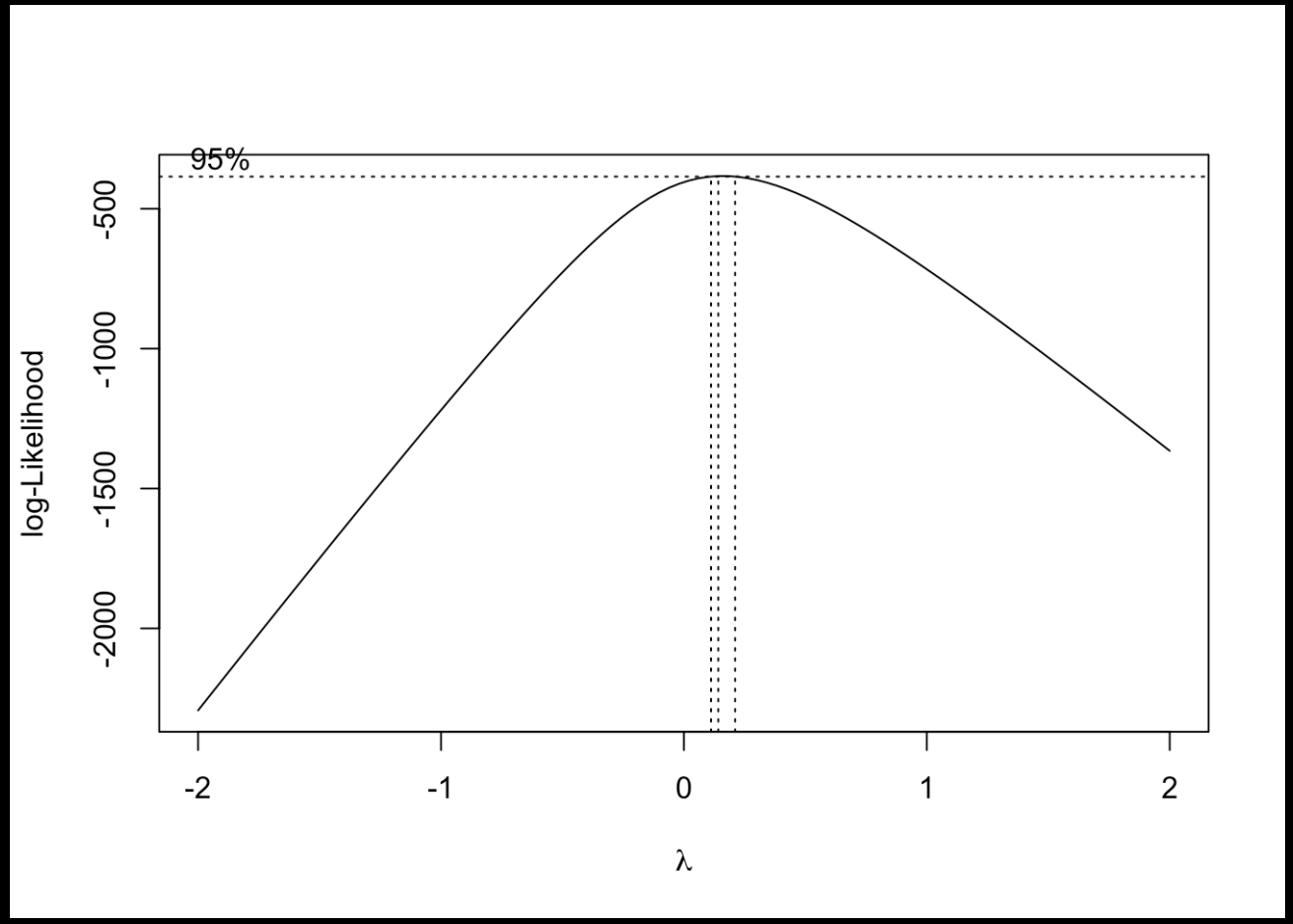
- Linearity assumption does not hold across different regressors– uneven distributions around zero
- Model appears to generally have a constant mean, but is heteroscedastic
- Errors appear to be in distinct groups above and below zero
- Points on the tails of the distribution that largely deviate from qqline, reject Shapiro Wilks[1] test



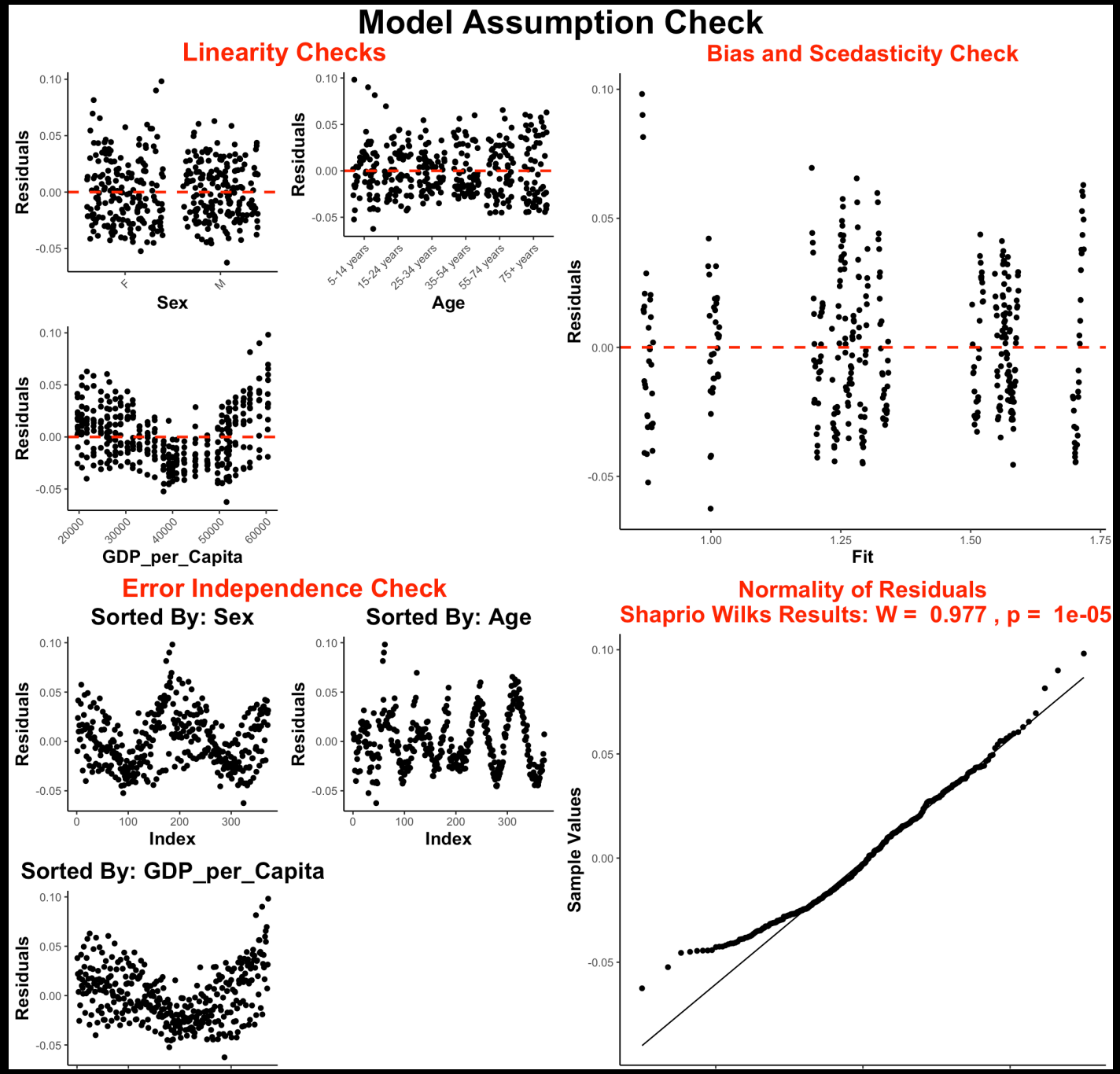
[1]Royston, Patrick (1982). An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 31, 115–124. doi: 10.2307/2347973.

Power Transformation

- Suicide rates were transformed by finding the optimal power using the BoxCox method[1]
- The response variable is iteratively transformed with different powers and the log-likelihood function is evaluated
- The power ($\hat{\lambda}$) at which the function is maximized is optimal
- $\hat{\lambda} = 0.1414$ for suicide rates

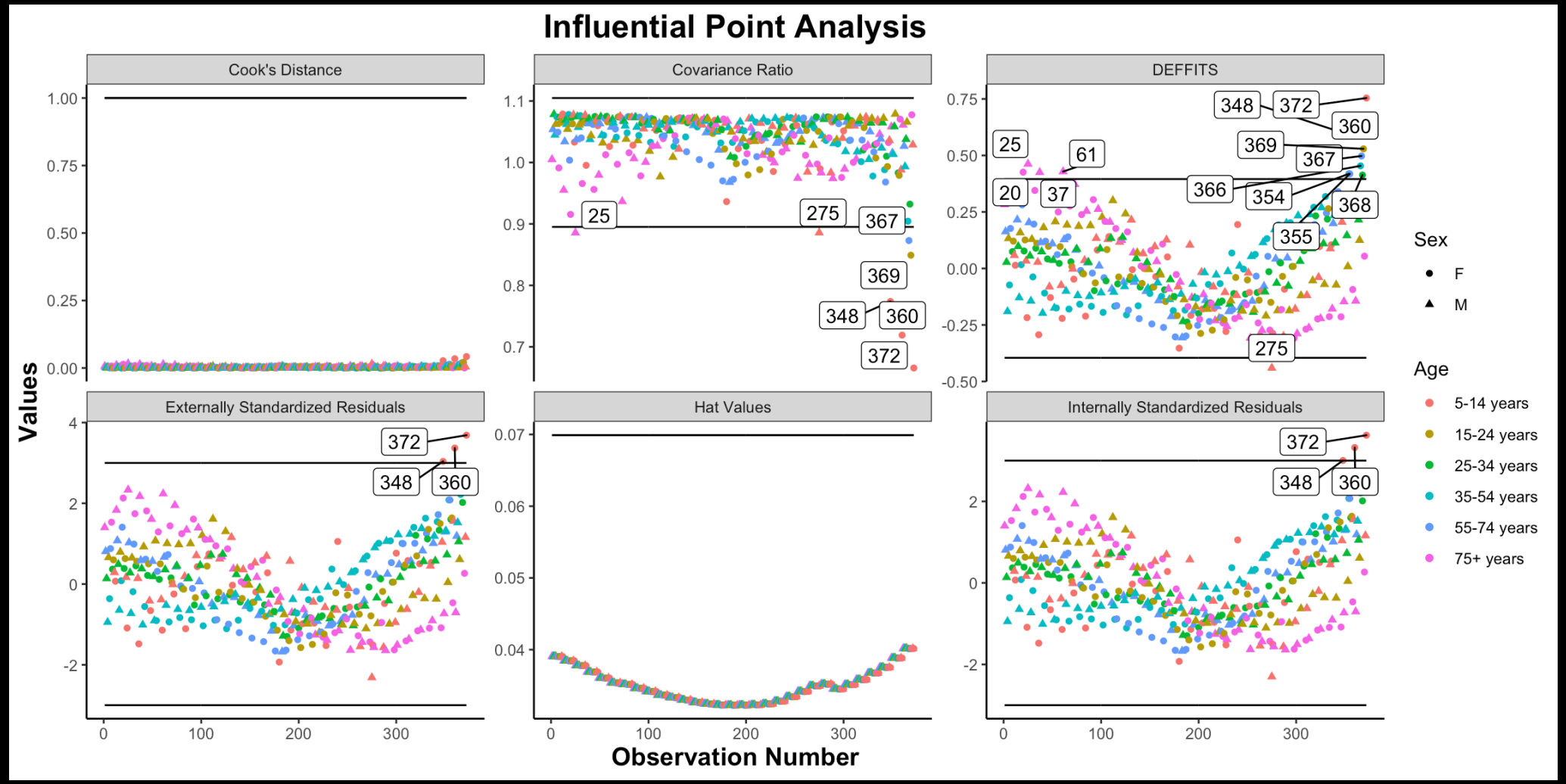


[1]Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). Journal of the Royal Statistical Society B, 26, 211–252.



Transformation Validation

- Linearity assumption violation alleviated, less grouping above and below zero
- Model is now homoscedastic ☺
- Errors appear to be less correlated than before, but are still present
- Points deviate less from the qqline, residuals appear to be more normal
  - Shapiro Wilks test still rejects the hypothesis residuals are normal



- Observation numbers 348, 360, and 372 are both considered as outliers and influential
  - All appear to be in the 5-14-year age range and female
- Males 75+ are considered influential (Observations 25, 37, and 61)

Hypothesis Testing

- Do males have a significantly higher suicide rate than females?
  - H_o: ß1 = 0 vs. H_A: ß1 ≠ 0
  - p-value = < 2e-16; We reject the null hypothesis. There is sufficient evidence to suggest ß1 is not equal to 0, meaning the effect of males is significant.
- Does age significantly affect suicide rates?
  - H_o: ß2 = ß3 = ß4 = ß5 = ß6 = 0 vs. H_A: At least one ≠ 0
  - p-value = < 2.2e-16; We reject the null hypothesis. There is sufficient evidence to suggest at least one of ß2, ß3, ß4, ß5, ß6 is not equal to 0, meaning the effect of age is significant.
- Is there evidence of interaction between sex and age?
  - H_o: ß8 = ß9 = ß10 = ß11 = ß12 = 0 vs. H_A: At least one ≠ 0
  - p-value = < 2.2e-16; We reject the null hypothesis. There is sufficient evidence to suggest at least one of ß8, ß9, ß10, ß11, ß12 is not equal to 0, suggesting evidence of interaction.

Hypothesis Testing (cont.)

- $\beta 2$, $\beta 3$, $\beta 4$, $\beta 5$, and $\beta 6$ represent the effect of each sex and age combination for females.
  - H_o: $\beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6$ vs. H_A: $\beta 2 \neq \beta 3 \neq \beta 4 \neq \beta 5 \neq \beta 6$
  - Each test resulted in a p-value < 0.05. We have enough evidence to reject the null hypothesis and conclude that suicide rates amongst the different age groups for females is significantly different.

- The difference in the male suicide rates amongst the different age groups is represented by $\beta 2+\beta 8$, $\beta 3+\beta 9$, $\beta 4+\beta 10$, $\beta 5+\beta 11$, and $\beta 6+\beta 12$.
  - H_o: $\beta 2+\beta 8 = \beta 3+\beta 9 = \beta 4+\beta 10 = \beta 5+\beta 11 = \beta 6+\beta 12$ vs. H_A: $\beta 2+\beta 8 \neq \beta 3+\beta 9 \neq \beta 4+\beta 10 \neq \beta 5+\beta 11 \neq \beta 6+\beta 12$
  - Only two tests resulted in a p-value > 0.05. The difference between males 25-34 years and males 35-54 years is not significant (p-value = 0.05593). Additionally, the difference between the suicide rates in males 35-54 years and males 55-74 years resulted in a p-value of 0.1359. This difference is not significant. This ultimately means that males in these age groups commit suicide at very similar rates.

Coefficient	Estimate	Test Statistic	P-Value	95% CI
Intercept	0.8987	133.140	< 2e-16	(0.885, 0.912)
sexmale	0.1274	18.134	< 2e-16	(0.114, 0.141)
age15-24 years	0.3267	46.515	< 2e-16	(0.313, 0.341)
age25-34 years	0.3868	55.068	< 2e-16	(0.373, 0.401)
age35-54 years	0.4530	64.494	< 2e-16	(0.439, 0.467)
age55-74 years	0.4133	58.842	< 2e-16	(0.400, 0.427)
age75+ years	0.3656	52.046	< 2e-16	(0.352, 0.379)
GDP_per_capita	-5.207e-7	-4.473	1.04e-5	(-7.5e-7, -2.9e-7)
sexmale: age15-24 years	0.1815	18.271	< 2e-16	(0.162, 2.010)
sexmale: age25-34 years	0.1652	16.633	< 2e-16	(0.146, 0.185)
sexmale: age35-54 years	0.1125	11.324	< 2e-16	(0.093, 0.132)
sexmale: age55-74 years	0.1627	16.377	< 2e-16	(0.143, 0.182)
sexmale: age75+ years	0.3364	33.868	< 2e-16	(0.317, 0.356)

Results

- $R^2 = 0.9872$
- Adjusted $R^2 = 0.9868$

Results (cont.)

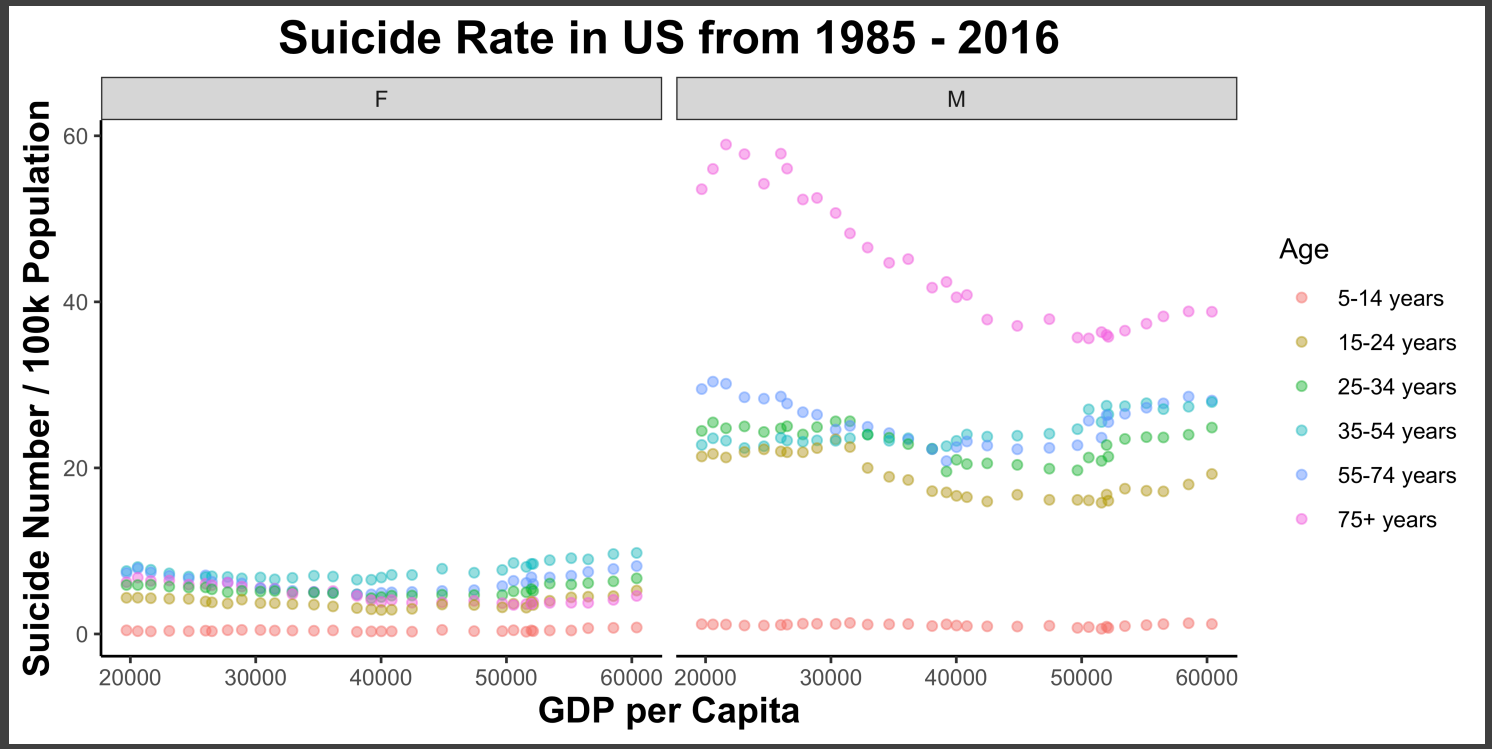
- Response = (Suicide Rate)$^{0.1414}$
- Slope = -5.207e-7(GDP_per_capita)
- Sex and age variables affect the intercept in the model

Trends:

- Male suicide rates increase with age (not all differences are significant)
- Female suicide rates peak in the 35-54 years age group
- Males commit suicide at a higher rate than females

Group	Intercept
Male, 5-14 years	1.0261
Male, 15-24 years	1.5343
Male, 25-34 years	1.5781
Male, 35-54 years	1.5916
Male, 55-74 years	1.6021
Male, 75+ years	1.7218
Female, 5-14 years	0.8987
Female, 15-24 years	1.2254
Female, 25-34 years	1.2855
Female, 35-54 years	1.3517
Female, 55-74 years	1.3120
Female, 75+ years	1.2643

Discussion



Predictions:

1. Males will have higher suicide rates than women.
2. Elderly adults will not have higher suicide rates than younger adults.

Results:

- ☑ Correct; males have higher suicide rates than women.
- ☑ Suicide rates peak for women at age group 35-54 years, but increase with age for men, with men being 75+ years having the highest rates.
- ☒

Conclusion

- Very strong model ($R^2 > 98\%$)
- Significant regressors:
  - Sex
  - Age
  - GDP per capita

- Based on these results, we can better target suicide preventive measures
  - 75+ years men

Thank you!