Exam 1

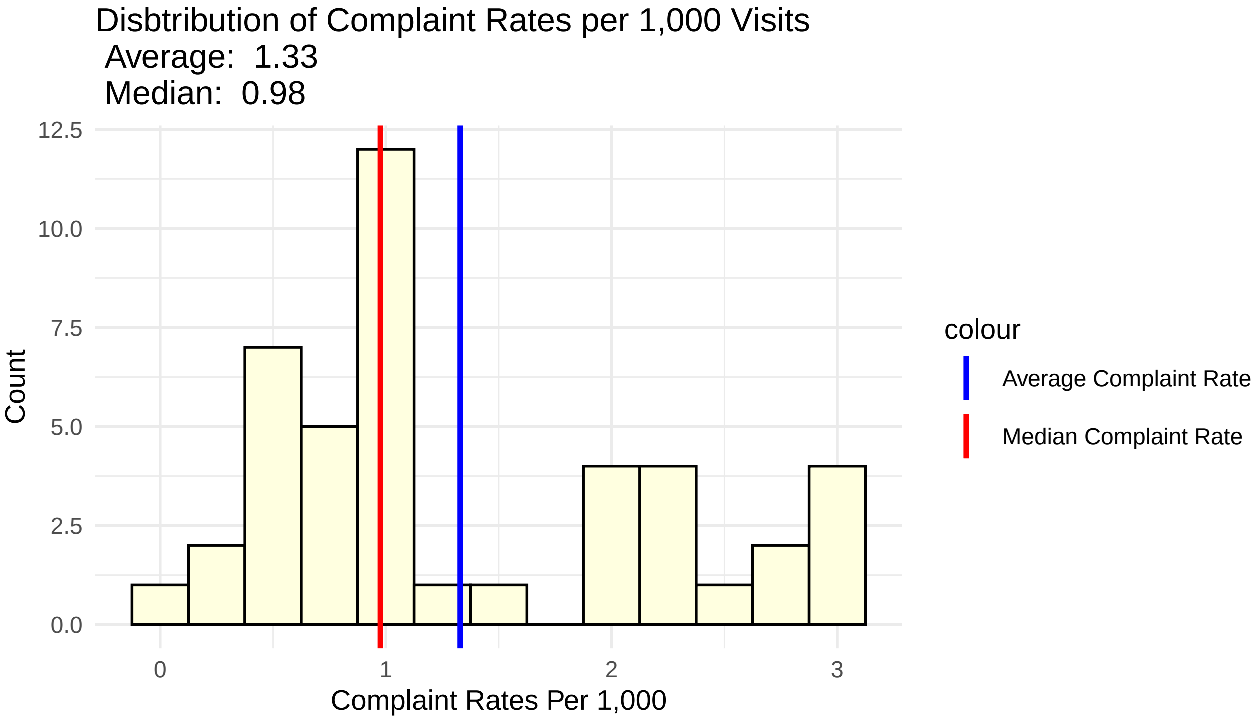
Denis Ostroushko

2022-10-28

Problem 1

1- A

Before fitting the model I like to explore the distribution shpare of the response variable and collect some fundamental summary statistics. Knowing the shape and the spread of the response variable will help us manage the expectation regarding model fit and variance of residuals.



The distribution of complaints per 1,000 visits somewhat balanced without extreme outliers. The mean is pretty close to the median, suggesting again that more extreme values on the upper end of complaints per 1,000 do not knew the mean very much.

The two tables below describe the distribution of numeric variables in the data set, as well as correlation between the three of them.

We can see that the scales of predictors and complaints per 1,000 vary greatly, so we should expect that the coefficients are going to be very small, probably in the 0.001 to 0.0001 range.

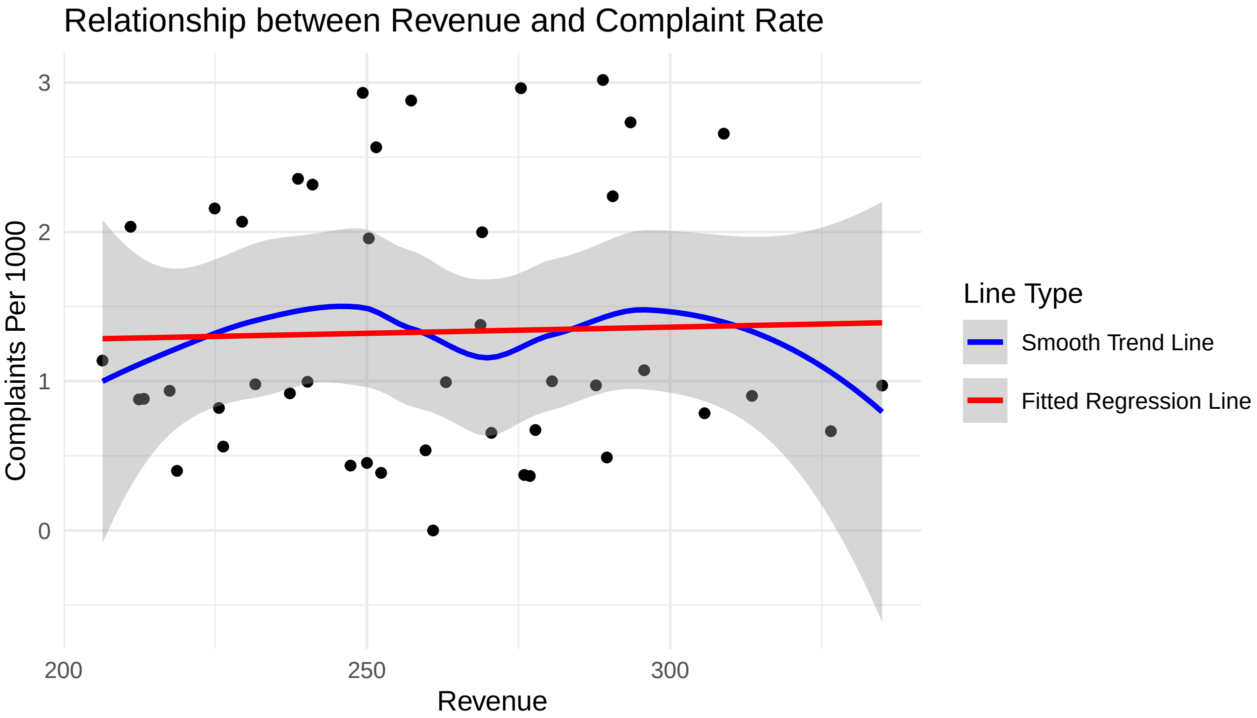
Table 2: Correaltion of Numeric Covariates

	Complaint Rate per 1,000	Revenue	Hours Worked
Complaint Rate per 1,000	1.0000000	0.0305876	0.2788799
Revenue	0.0305876	1.0000000	-0.0405506
Hours Worked	0.2788799	-0.0405506	1.0000000

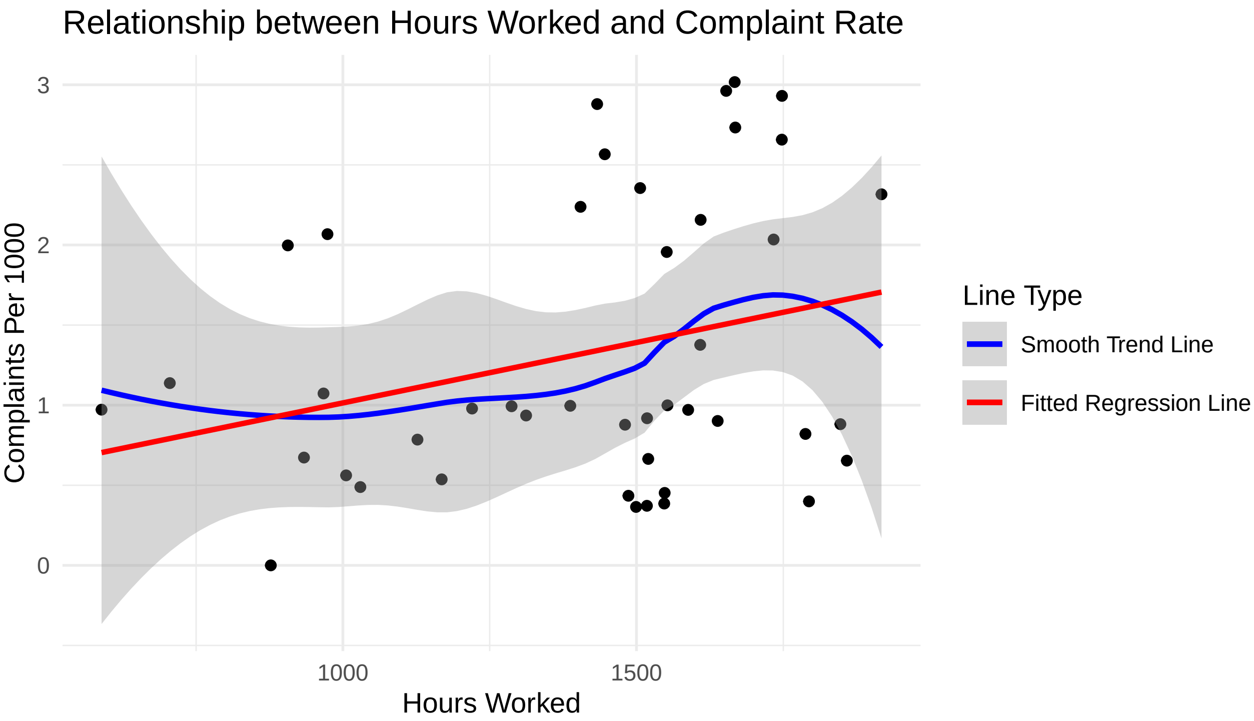
Table 1: Summary of Numeric Variables

Variables	Min	Max	Mean	S.D
complaint_rate_1000	0.00	3.02	1.33	0.88
revenue	206.42	334.94	260.14	32.64
hours	589.00	1917.25	1417.40	326.98

We continue to perform explanatory data analysis in this section by looking at the scatter plots of predictors versus complaint rates. It does not appear that revenue is related to complaint rate at all.



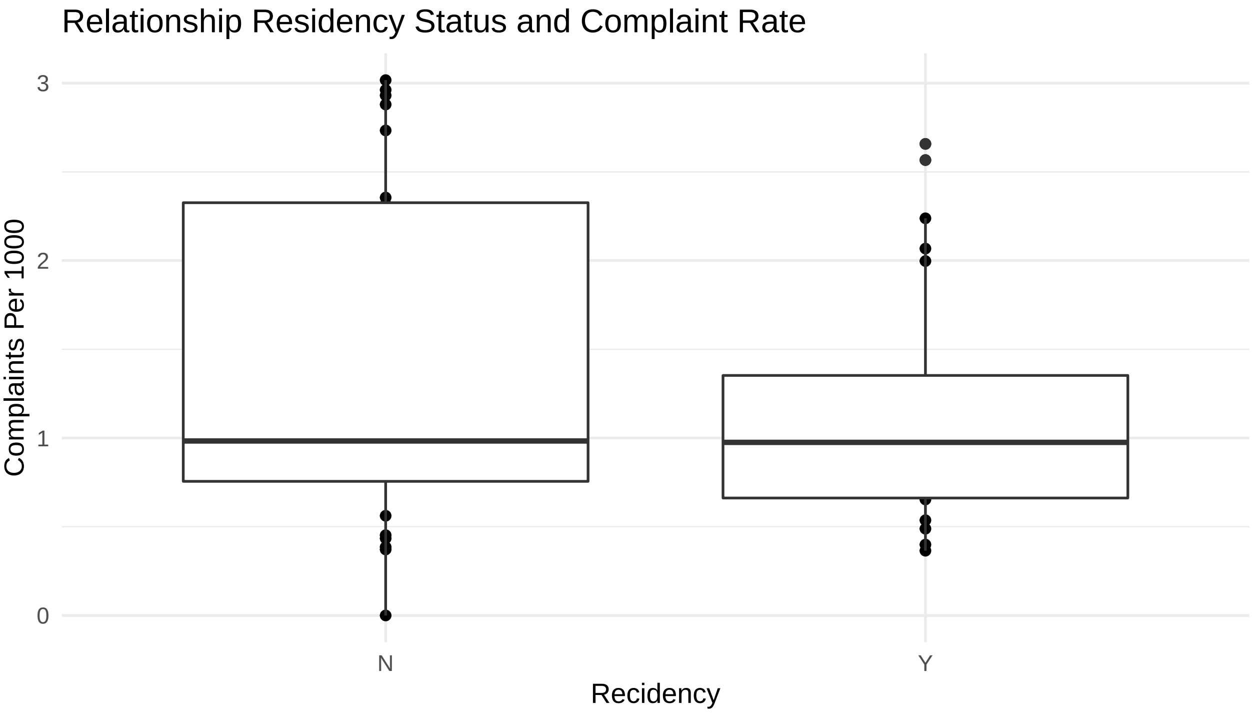
We can see that the number of hours worked is somewhat linearly related to the complaint rates, suggesting the practitioners who work more hours tend to accumulate higher complaint rate, however, the the variance of values is very large around the suggested regression line, so we might not be able to detect a statistically significant relationship when fitting the model.



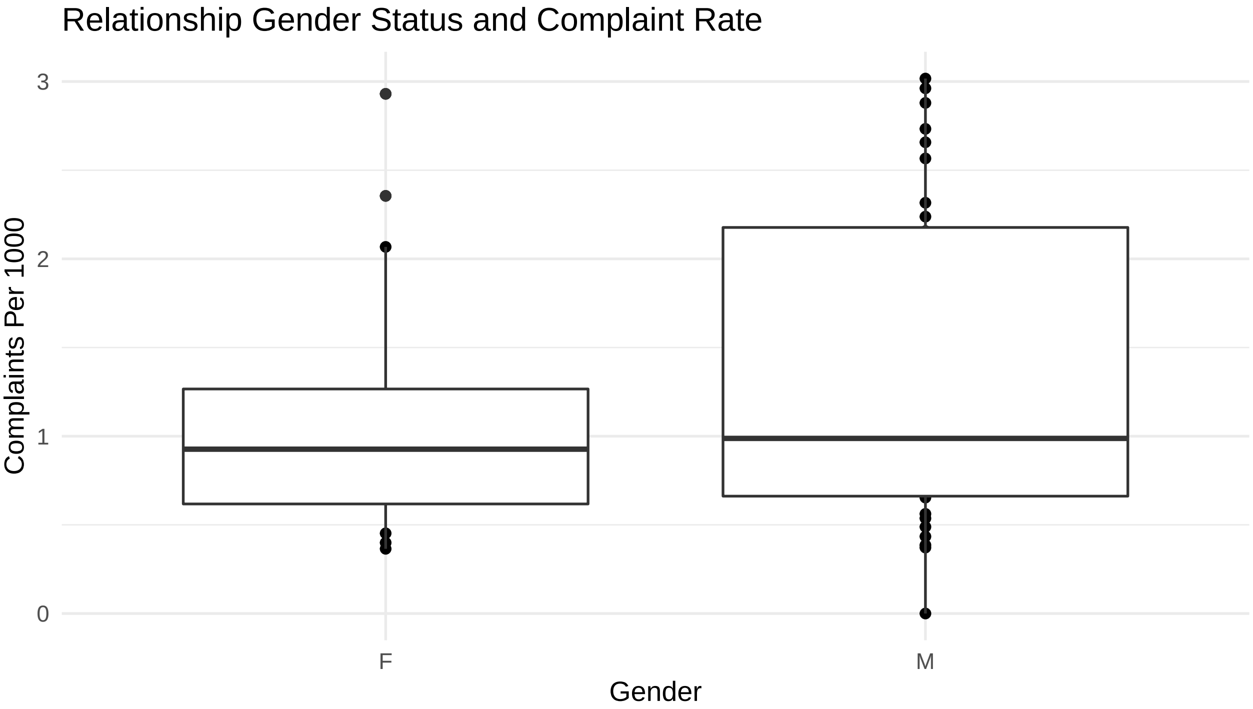
Overall, both plots suggest that linear fit is appropriate for both of these variables. Smooth Loess function does not show any consistent curvature in the data, but rahter randomly fluctuates around the fitted regression line.

We have categorical predictors also:

Residency has two levels: Y, N with 54.55%, 45.45% class presence respectively. It does not appear that the median and mean values are different across the two residency levels.



Gender has two levels: F, M with 27.27%, 72.73% class presence respectively. It does not appear that the median and mean values are different across the two gender levels.



Now we are ready to fit and examine the Normal Error Regression Model. When fitting any kind of a model, we need to be careful with the assumptions we take on. **Model Assumptions** are listed below

1. Residuals, Error Terms, are normally distributed with mean $\mu = 0$ and constant variance σ^2.
2. Since fitted values depend on model parameters $\hat{\beta}_i$ and errors e_i , we assume each outcome Y_i comes from a normal distribution with mean $\mu = E[Y_i]$ and variance σ^2.
3. We assume that variance of residuals is constant.
4. Errors are independent and each unit of interest, a data point, is also independent of other observations in the sample.
5. The model is linear because $\hat{Y}_i$ can be expressed as a linear combination of weights, coefficients, $\hat{\beta}_i$ and constant observed data points X_i.
6. Predictors are not correlated or weakly correlated.

After listing model assumptions, we can state the mode:

$$E[Complaint\ \ Rate] = \hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2 + \hat{\beta}_3 * X_3 + \hat{\beta}_4 * X_4 =$$

$$E[Complaint\ \ Rate] = \hat{\beta}_0 + \hat{\beta}_1 * Revenue + \hat{\beta}_2 * Hours\ \ Worked + \hat{\beta}_3 * Gender + \hat{\beta}_4 * Residency$$

Overall ANOVA

Before investigating individual coefficients and t-test for predictors, we want to look at the overall ANOVA table, and overall F-test. We want to see if the set of all predictors is helpful at explaining the variance of complaint rates per 1,000, and therefore we will know if some of all coefficients are statistically different from 0.

ANOVA table for the F-test is given below:

Source	SSR	DF	MS	F Statistic	P(F* > F)
Regression	3.254294	4	0.8135735	1.04	0.3969
Error	30.386120	39	0.7791313	NA	NA
Total	33.640414	43	NA	NA	NA

- Null Hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1}$
- Alternative Hypothesis: H_a : Not all coefficients β_i are zero
- $F-$statistic: 1.04
- Cutoff F^*-statistic: 2.6123
- So, $F < F^*$, therefore we do not have enough evidence to reject the null hypothesis to conclude that some or all coefficients β_i are consistently different from zero.
- Moreover, $P(F^* > F) = 0.3969$
- Conclusion: There is not enough statistical evidence that every predictor has a coefficient different from 0. Therefore, we can't reject the null hypothesis. When we look at the individual t tests for coefficients, we might see some suggestive relationships, supported by the somewhat big values of the t-statistic and small p-values, but none of them should be statistically significant.

Table below shows **Regression Coefficients** and model summary. Like we expected, these coefficients are small becuase the scale and range of predictors and response vairable are not the same.

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	-0.064405	1.250366	-0.051509	0.959183
revenue	0.001351	0.004610	0.293122	0.770983
hours	0.000676	0.000461	1.467079	0.150373
genderM	0.197338	0.314907	0.626654	0.534537
residencyY	-0.132728	0.329286	-0.403077	0.689093

- R square and 0.0967
- Adjusted R Square 0.0041
- Coefficients explanation:
  - `revenue` is revenue in dollars per hour is a continuous predictor. When revenue increases by 1 dollar per hour, we expect the number of complaints to increase by 0.001351, after adjusting for other predictors.
  - `hours` is the number of hours worded, and is a continuous predictor. With each additional hour of work, we expect the number of complaints to increase by 0.001351, after adjusting for other predictors.
  - `genderM` is a coefficient for the group of men practitioners, when compared with women practitioners, which is a reference level here. Physicians who are men on average are expected to have 0.197338 more complaints per 1,000 when compared with the women physicians after adjusting for other predictors.
  - `residencyY` is a coefficient for the group of practitioners who have a residency fellowship, compared with practitioners who do not participate in such program, which is a reference level here. Physicians with fellowship are expected to have 0.132728 less complaints per 1,000 when compared with the women physicians after adjusting for other predictors.

- As expected, none of these predictors show any evidence of statistical significance, but the results are not contradictory. Doctors who are trained to work in emergency medicine should be able to do their work better, and therefore should have less complaints. Additional hours may result in extra complaints, if the doctor is overworked and their ability to perform reduces.

1- B

A special interest is to investigate how the extra hours of work impact the average complaint rate for each practitioner, given their characteristics we adjust for.

We begin the inference of this variable with a formal t-test.

- Null Hypothesis: $H_0 : \hat{\beta}_4 = 0$
- Alternative Hypothesis: $H_a : \hat{\beta}_4 \neq 0\$$
- Test statistic $T : 1.467079$
- $P(t^* > t) = 0.150373$
- Conclusion: p-value is above 0.05 so there is not enough statistical evidence to reject the null hypothesis to conclude that the additional hour of work consistently results in the average increase of complaint rates. However, the p-value is not greatly far for the accepted significance level, so, this relationship is suggestive. Perhaps, with mode data, or a better statistical model we will be able to verify that this relationship is in fact consistent. My recommendation to the managers and decision makers would be to pay close attention to this factor, because even thoght the test shows no significance, the relationship is perhaps still real, and can't be detected from this sample.

One additional Hour worked results in 0.000676 additional complaints on average.

However, it makes more sense to, say, look at 20 hours. So, an average increase in complaint rates per 1,000 is 0.000676 * 20 = 0.01352.

It is reasonable to expect that a practitioner who is overworked will have an extra 20 hours of work on top of regular hours in one week, especially in a busy or underfunded facility.

C.I.

Using formula $C.I.\ bounds = Estimate \pm 1.96 * Standard\ Error$

C.I. for the estimate 0.000676 with a 0.000461 standard error is (-0.000256, 0.001609)

Similar to the coefficient, we can perform a linear transformation of the lower and upper bounds, and obtain a confidence interval for the effect of extra 20 hours of work. So, an average increase in complaint rates per 1,000 for the extra 20 hours of work is 0.000676 * 20 = 0.01352, with a confidence interval (-0.00512, 0.03218).

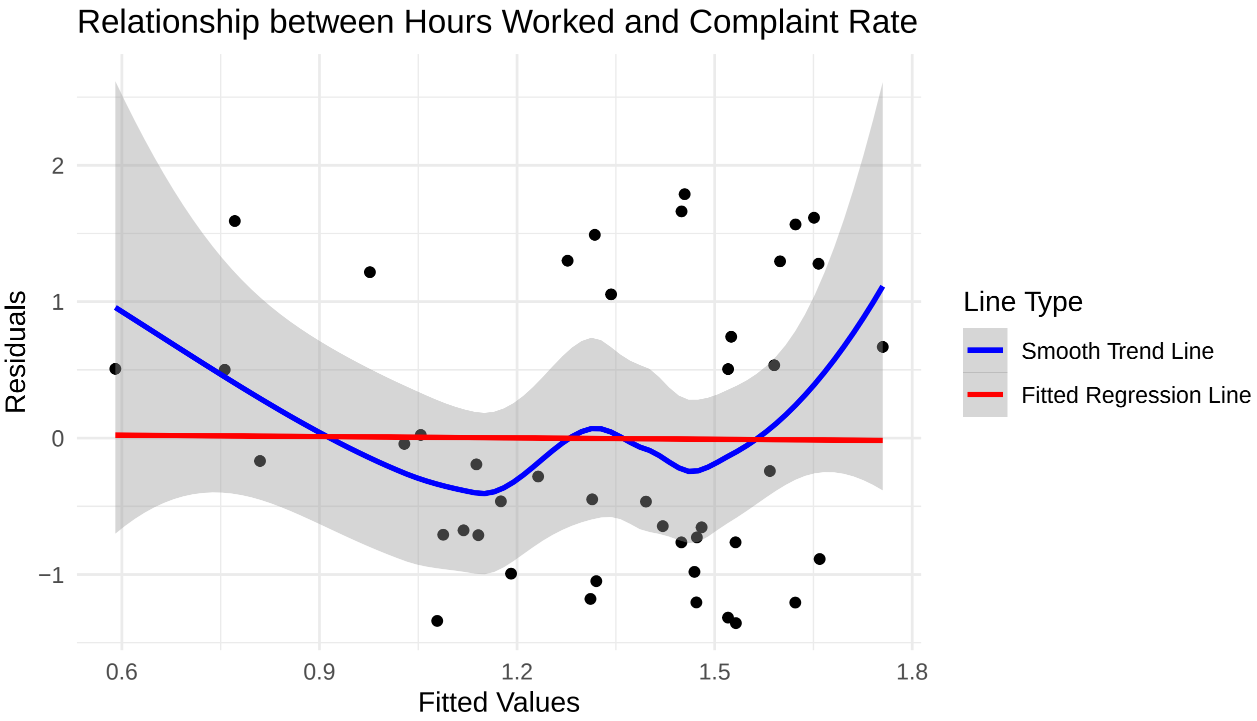
It appears that most of the confidence interval is above 0, in fact, 86.27% of values in the confidence interval are above 0. So, even though this evidence is pretty weak, we would still pay attention to this variables as a source of Y variance explanation.

1- C

The plot below shows the relationship between fitted values and studentized residuals from the regression model we built and evaluated in the previous two sections. There is no linear trend, as evidenced by the flat fitted regression line.

Smooth trend line suggests that either there is some violation of assumptions at the lower and upper ends of the fitted values, or there is simply a small number of values there.

In any case, variance appears to be somewhat constant, we do not see a megaphone or a violin shape. However, residuals above 0 tend to have an upper bound of around 2, whereas residuals below 0 tend to have a lower bound of around 1.5. Overall, this is not a huge cause for concern, but something we should keep an eye.

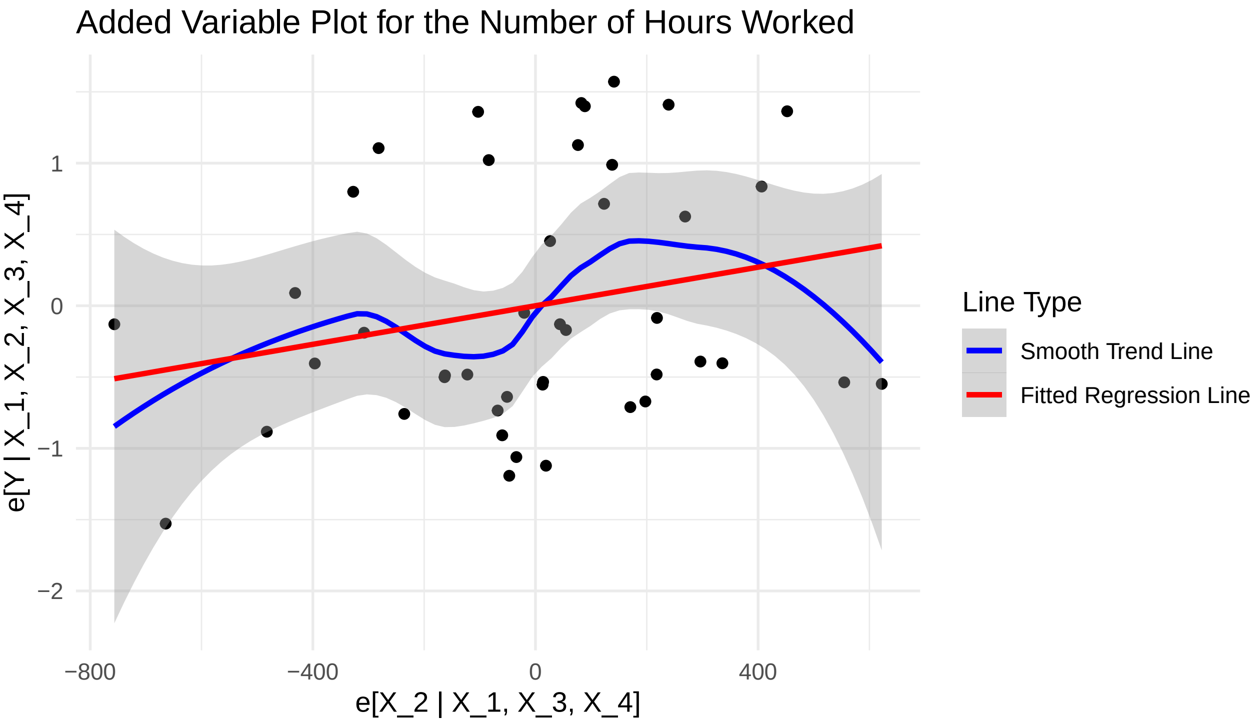


1- D

In order to evaluate the nature of the relationship between complaint rates per 1,000 and the number of hours worked after adjusting for the other 3 predictors we use an added variable plot. In order to do that we will need to obtain two sets of residuals from the two models:

- Model 1: obtains residuals for \$Y = \$ Complaints per 1,000 visits. We denote these residuals as $\epsilon_Y = e(Y|X_1, ..., X_4)$:
  - $Y = \hat{\beta}_0 + \hat{\beta}_1 * Revenue + \hat{\beta}_2 * Hours\ Worked + \hat{\beta}_3 * Gender + \hat{\beta}_4 * Residency + \epsilon_Y$
- Model 2: obtains residuals for \$X_2 = \$ Number of Hours Worked. We denote these residuals as $\epsilon_X = e(X_2|X_1, X_3, X_4)$:
  - $X_2 = \hat{\beta}_0 + \hat{\beta}_1 * Revenue + \hat{\beta}_3 * Gender + \hat{\beta}_4 * Residency + \epsilon_x$

Plot below shows the relationship between the two sets of predictors:



This plots gives us two pieces fo evidence that we sue to describe the marginal relationship of X_2 and Y, after adjusting for three other predictors:

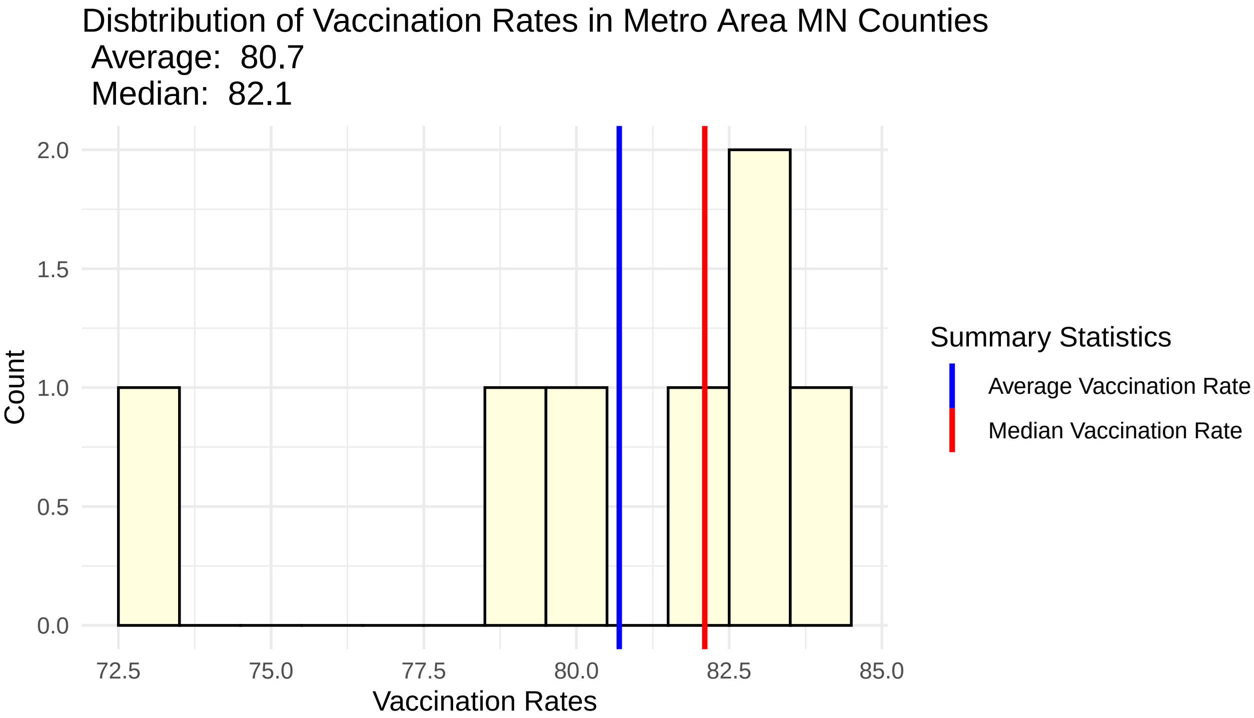
1. The relationship between X_2 and Y, after accounting for other predictors, is linear in its nature. We can see that the smooth trend line fluctuates randomly around the fitted regression line, suggesting that there is no consistent curved, or other non-linear relationship between the number of hours worked and the complaint rate.
2. The fitted regression line that confirm linear relationship has a positive, upward facing, slope, suggesting that the number of hours worked can be used a potentially useful predictor that help increase the percentage of variation of in the complaint rate. This supports our previous conclusion that X_2 may be employed as a useful predictor of Y, but this model does not gives us enough sufficient evidence to make such a claim.

Problem 2

2 - A

In order to pick between the tow-sample T-test and Wilcoxon test we need to understand the shape of the distribution, in particular the spread of values, and the effect that extreme values and outliers can have on the t-test. While t-test is robust and produces that we can rely on, it is known that in heavily skewed distributions non-parametric methods that rely on rank of observations will be more effective. On the other hand, if we do not see a heavily skewed distribution, but instead see a distribution that is approximately normal, we want to use a t-test, because for such data Wilcoxon has only 95% of statistical power of the the T-test.

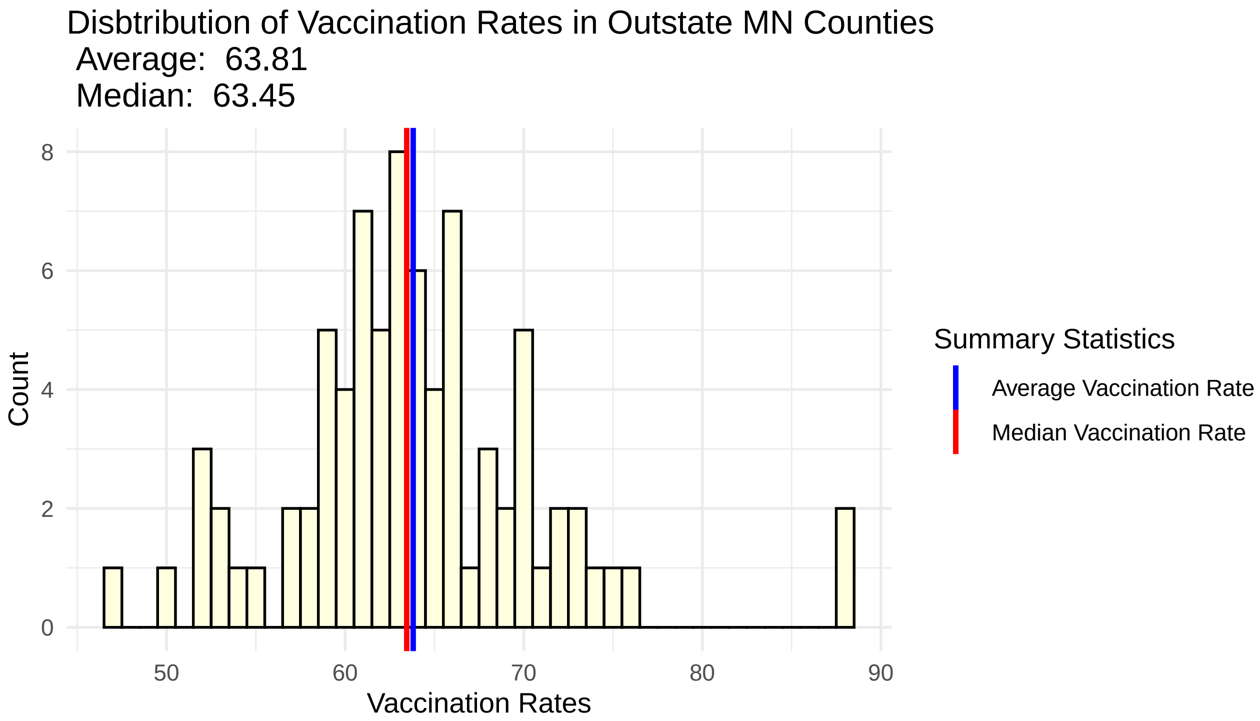
The plot below shows vaccination rates for the metro area counties. It is pretty card to make any conclusions from this plot, and this group of observations, since there are only 7 counties that make up Metro area.



On the other hand, there are 80 Outstate counties. These counties produce a balanced bell-shaped distribution that looks normal. However, there are a few outliers. Outstate counties with unusually high vaccination rates are: Olmsted and Cook with 88.5 and 87.9 vaccination rates, respectively.

Olmsted county includes Rochester, a pretty big city by the outstate standards. Moreover, Mayo clinic is located there, so we can speculate that more people should have more trusting relationship with medicine and public health there.

Cook county is located by the Canadian Border, I am not sure what conclusion we can draw from this fact.



The table below summarized two distributions in terms of most common summary statistics. This table gives me the impression that we will be able to conclude that the two sample means are in fact different because the we have small standard deviations, while the two means are quite different.

Table 3: Vaccination Rates Summary by County Type

Type	N	Mean	Median	S.D.
Outstate	80	63.81	63.45	7.08
Metro	7	80.70	82.10	3.63

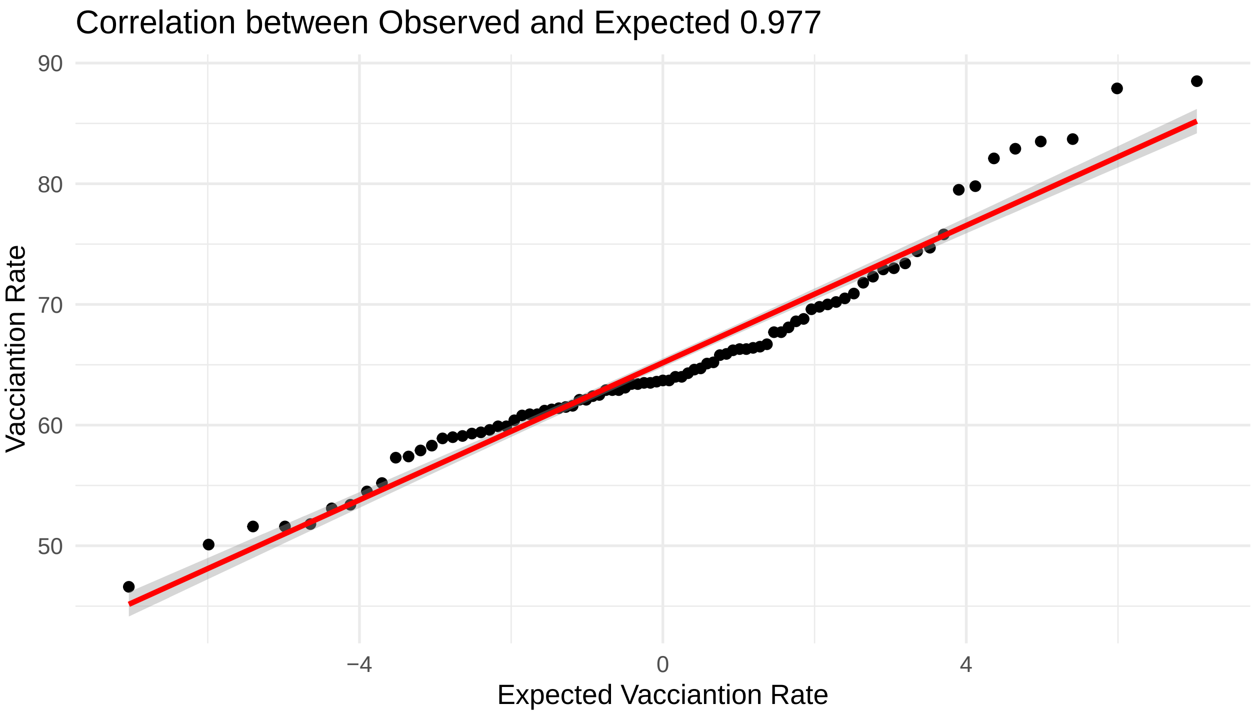
Before conducting the test, we also wish to see if the overall distribution of the two samples combined is normal. Recall, there are only 7 counties in the metro area, so we should combine the two samples for this verification.

We can test the Normality of Vaccination Rates distribution against the expected quantiles of standard normal distribtuion.

We can calculate these expected values using the formula:

$$\sqrt{Variance} \times z(\frac{Value - .375}{N + .25})$$

It appears that the sample of data we have is approximately normally distributed.

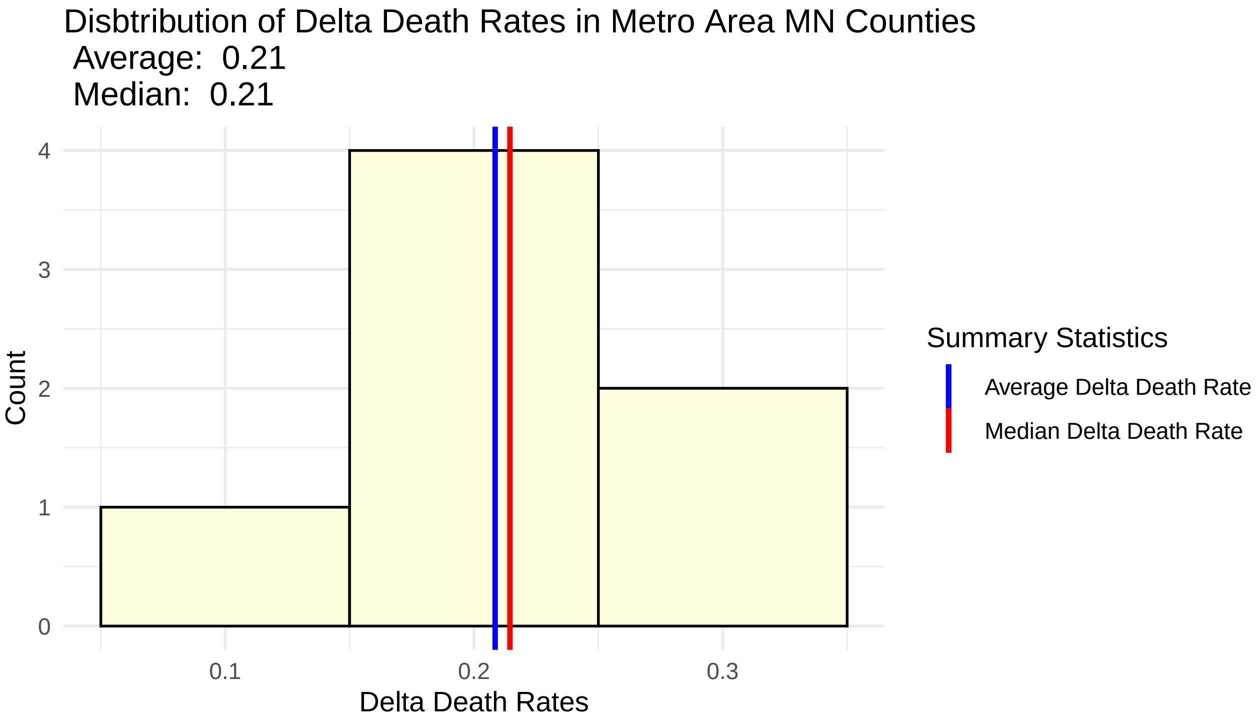


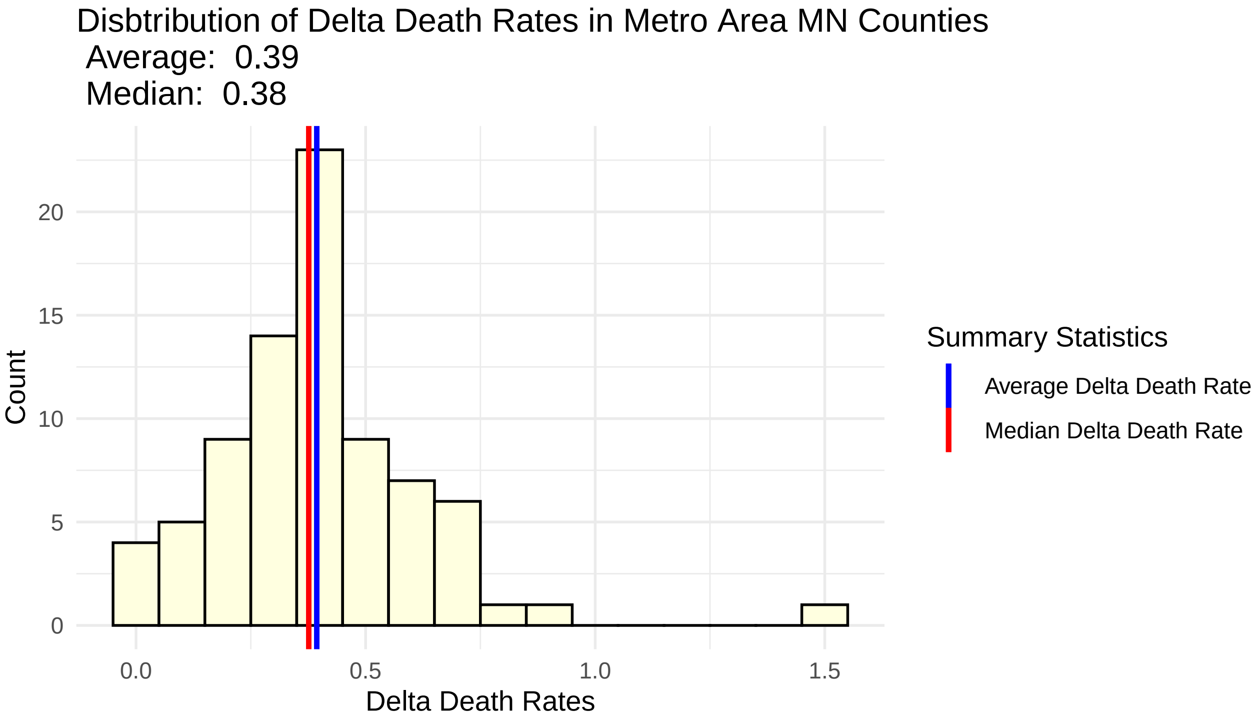
Therefore, we will use T test here.

Test results summary and interpretation are given below:

- Null Hypothesis: $H_0 : \mu_{metro\ area} = \mu_{outstate}$
- Test statistic: $H_a : \mu_{metro\ area} \neq \mu_{outstate}$
- Metro area mean vaccination rate is 80.7, while outstate median vaccination mean is 63.81
- Estimated difference is -16.89, bounded by (-20.3959 , -13.3841)
- Test statistic T: -10.6576166
- $P(T^* > T) = 0.000001$
- Conclusion: P-value is small, so we can reject the null hypothesis and conclude that the average difference in vaccination rates on the county level is statistically significant between metro and rural areas. On average, we can expect metro area counties to have -16.89 vaccines per 1,000 county residents.

2 - B





Outlier: Faribault

Faribault county is kind of an outlier

Death Rates for Outleir counties Olmsted and Cook 0.13, 0

88.5, 87.9

Table 4: Death Rates Summary by County Type

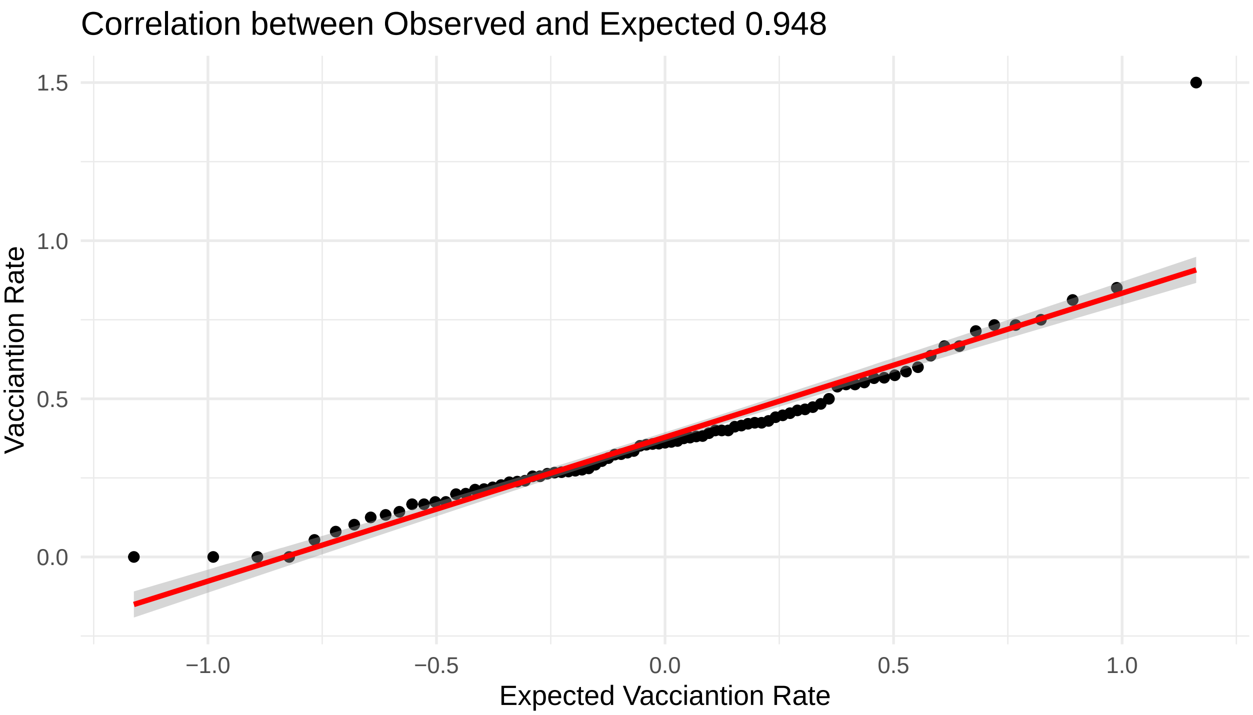
Type	N	Mean	Median	S.D.
Outstate	80	0.3936199	0.3761792	0.23
Metro	7	0.2085040	0.2144522	0.04

- **Normality of Death Rates**

In order to test outliers for normality we plot the residuals against expected values of residuals in a normally distributed random sample.

We can calculate these expected values using the formula:

$$\sqrt{Variance} \times z(\frac{DeathRate - .375}{N + .25})$$



Test results summary and interpretation:

- Null Hypothesis: $H_0 : \mu_{metro\ area} = \mu_{outstate}$
- Test statistic: $H_a : \mu_{metro\ area} \neq \mu_{outstate}$
- Metro area mean vaccination rate is 0.208504, while outstate median vaccination mean is 0.3936199
- Estimated difference is 0.185116, bounded by (0.1252 , 0.2451)
- Test statistic T: 6.1920794
- $P(T^* > T) = 0$
- Conclusion:

2 - C

Model Specificantion

$$E[DeathRate] = \hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2 =$$

$$E[DeathRate] = \hat{\beta}_0 + \hat{\beta}_1 * Vaccination\ Rate + \hat{\beta}_2 * Metro\ Area\ County\ Indicator$$

Overall ANOVA test

Source	SSR	DF	MS	F Statistic	P(F* > F)
Regression	0.5740292	2	0.2870146	6.37	0.0027
Error	3.7854492	84	0.0450649	NA	NA
Total	4.3594784	86	NA	NA	NA

- Null Hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1}$
- Alternative Hypothesis: H_a : Not all coefficients β_i are zero
- $F-$statistic: 6.37
- Cutoff F^*-statistic: 3.1052
- So, $F < F^*$, therefore we do not have enough evidence to reject the null hypothesis to conclude that some or all coefficients β_i are consistently different from zero.
- Moreover, $P(F^* > F) = 0.0027$
- Conclusion:

Model Estimates

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	0.990663	0.214503	4.618412	0.000014
v_rate	-0.009357	0.003341	-2.800576	0.006329
region	-0.027083	0.100922	-0.268358	0.789081

- R square and 0.1317
- Adjusted R Square 0.111
- Null Hypothesis: $H_0 : \hat{\beta}_2 = 0$
- Alternative Hypothesis: $H_a : \hat{\beta}_2 \neq 0$$
- Test statistic T : -0.268358
- $P(t^* > t) = 0.789081$
- Conclusion

Interpretation of coefficient

Metro Area expected to have 0.0271 deaths per 1,000

C.I.

Using formula $C.I.\ bounds = Estimate \pm 1.96 * Standard\ Error$

C.I. for the estimate -0.027083 with a 0.100922 standard error is (-0.227779, 0.173612)

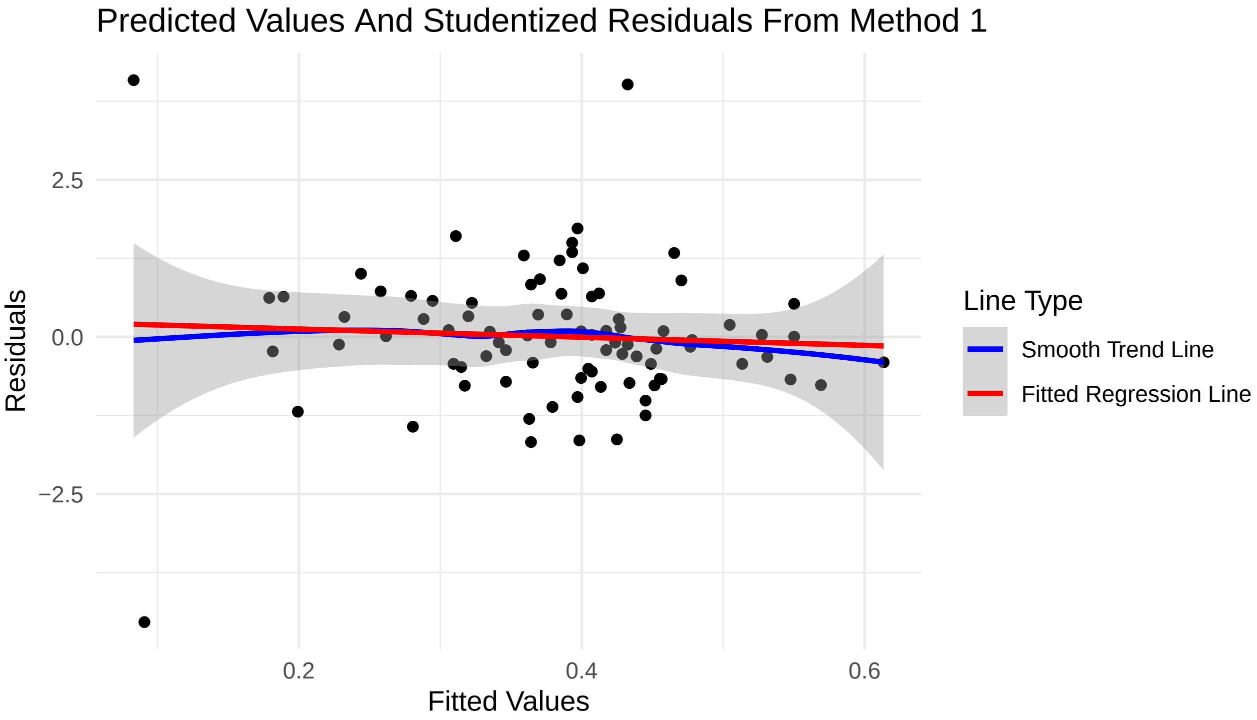
- Conclusion on the effects of two predictors

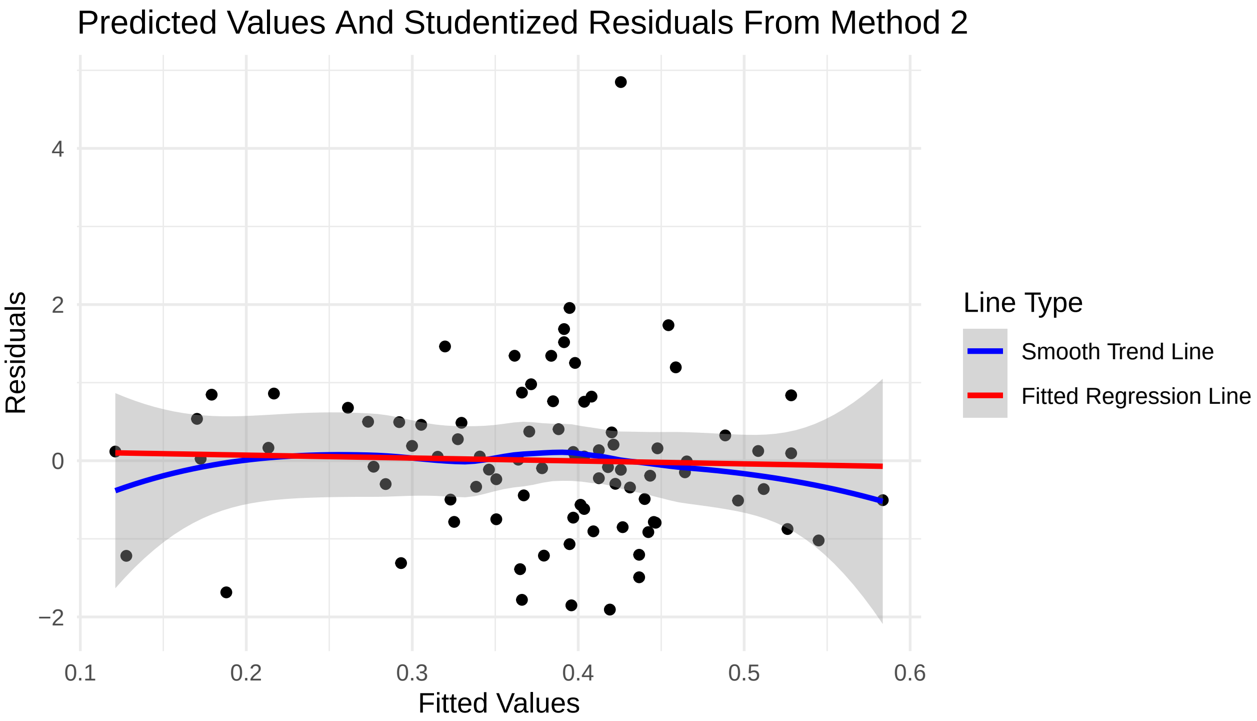
2 - D

We definitely have non-constant variance

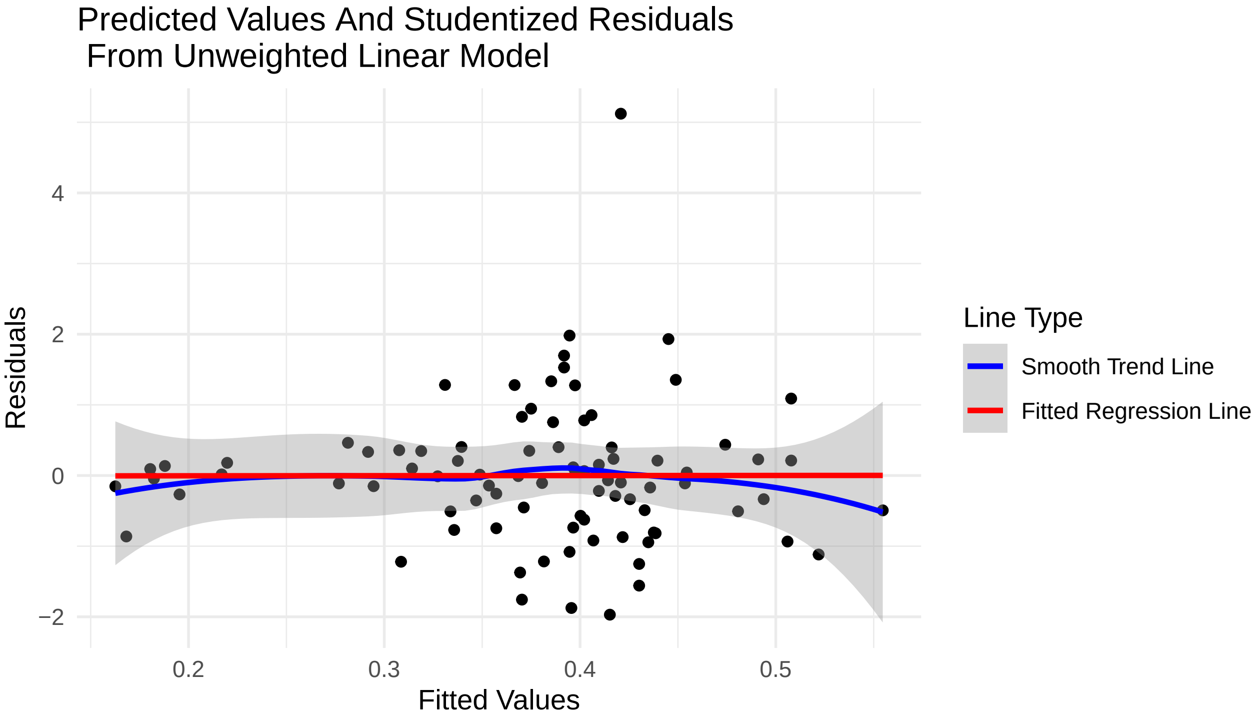
Procedure: Reg 14 Slide 62

Book pages 421-431





Predicted Values And Studentized Residuals From Method 2



Predicted Values And Studentized Residuals
From Unweighted Linear Model