Exam 1

Denis Ostroushko

2022-10-26

Problem 1

1- A

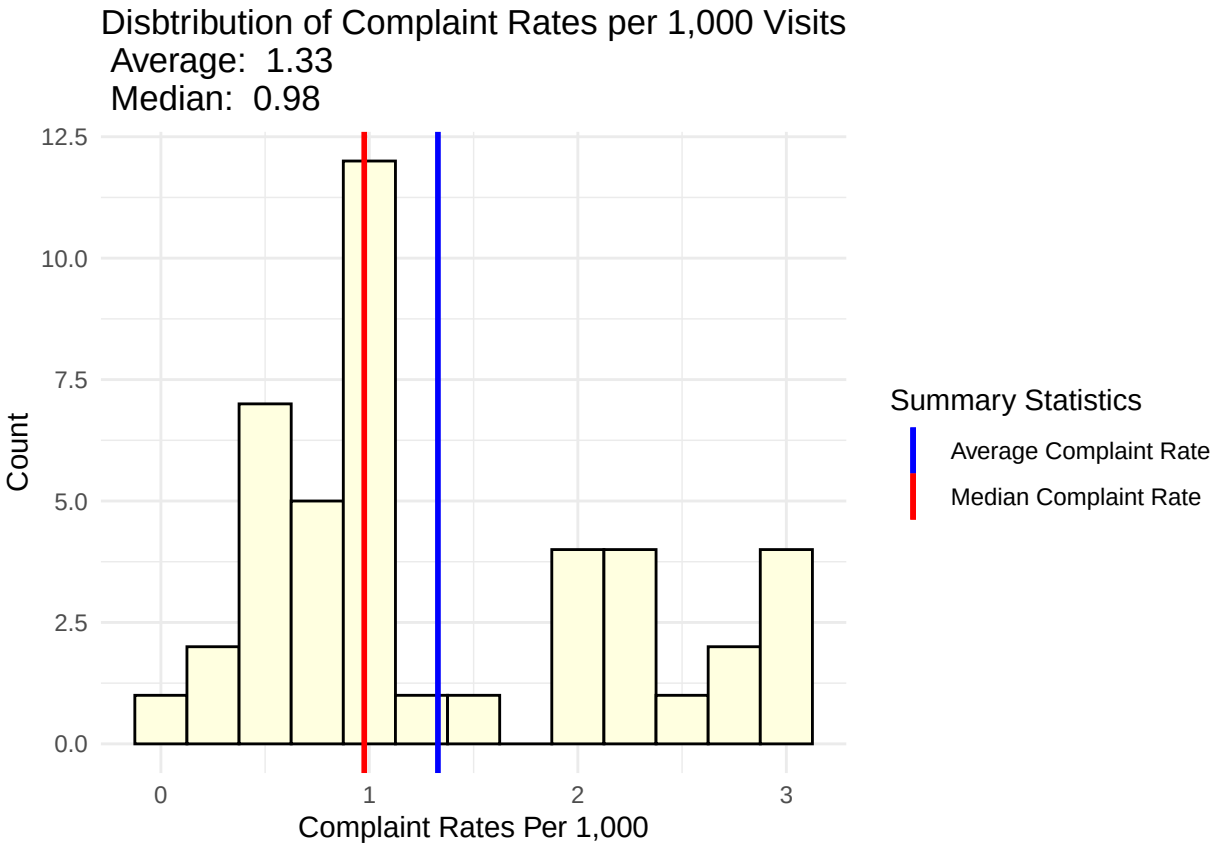
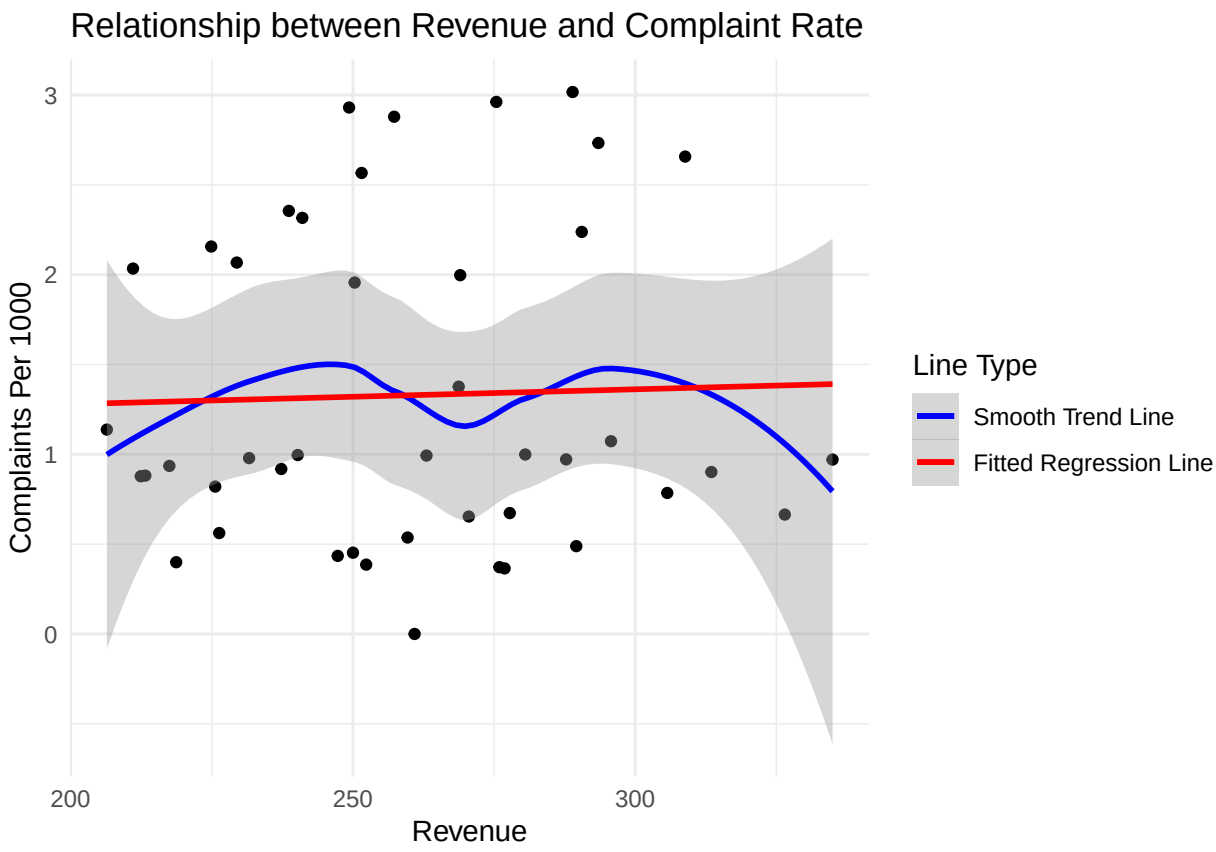


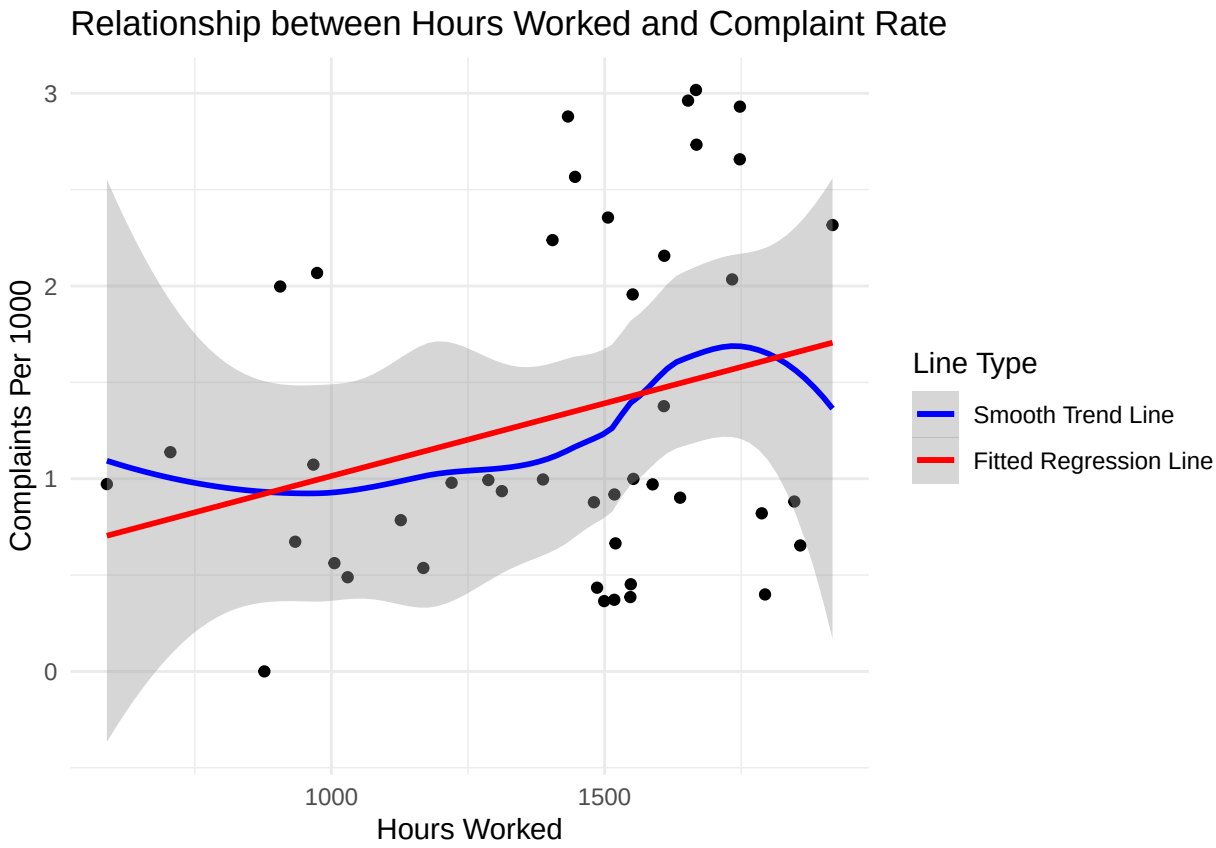
Table 1: Summary of Numeric Variables

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

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





We have categorical predictors also:

- Residency has two levels: Y, N with 54.55%, 45.45% class presence respectively
- Gender has two levels: F, M with 27.27%, 72.73% class presence respectively

Overall comments on variables

Model Assumptions

- one
- two

*three

Model Statement

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

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

Overall ANOVA

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:

Regression Coefficients

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
- Explain Coefficients

1- B

T-test for hours worked

- 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

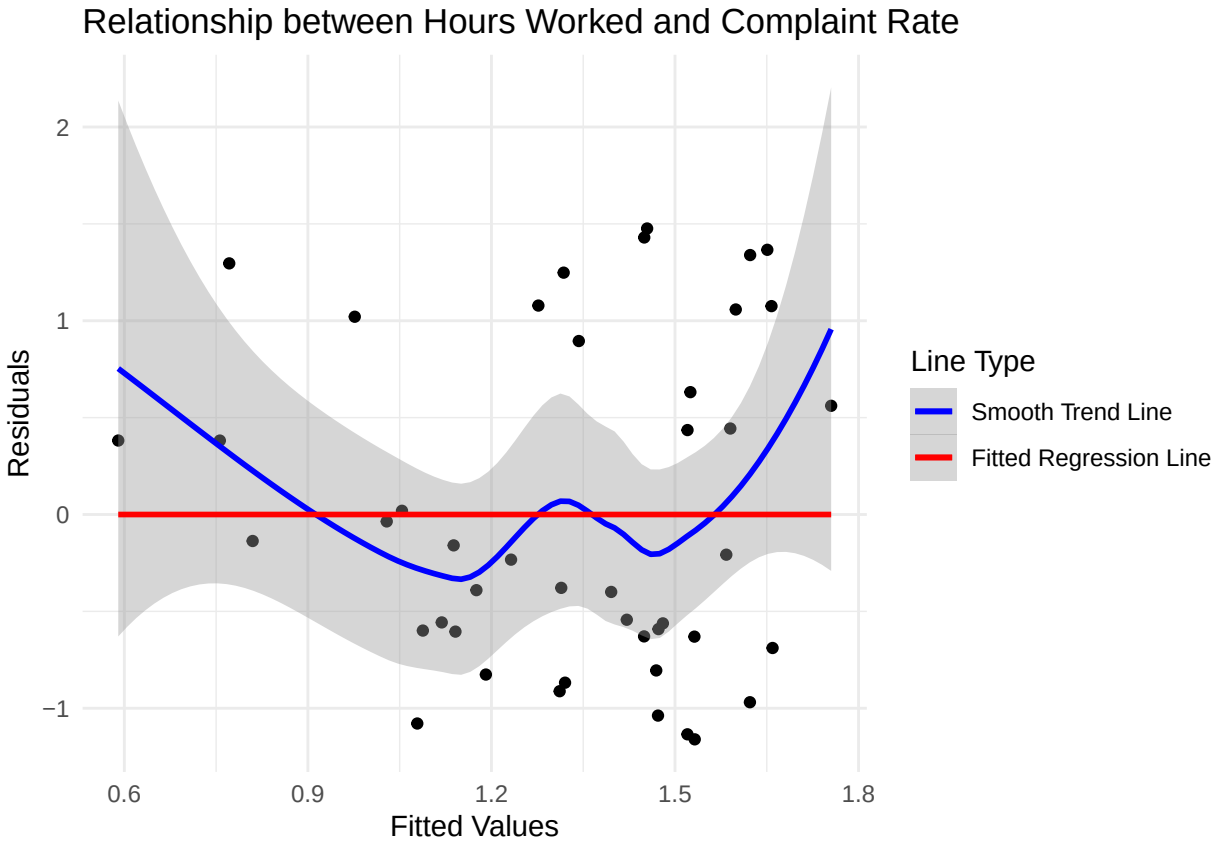
Interpretation of coefficient One additional Hour worked results in 0.000676 additional complaints on average. However, it makes more sense to say look at 100 hours, which is 0.0676

C.I.

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

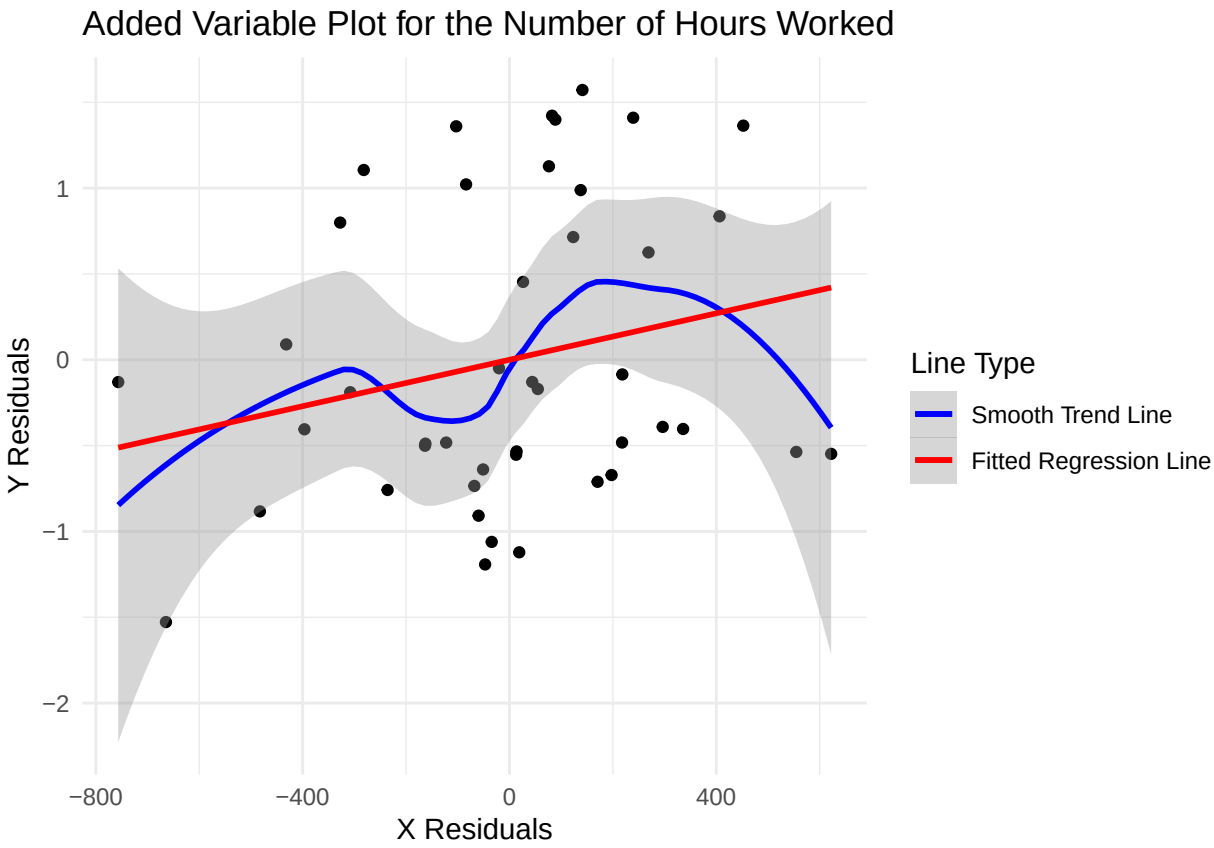
C.I. for the estimate 0.000676 with a 0.000461 standard error is (-0.000256, 0.001609)

1- C



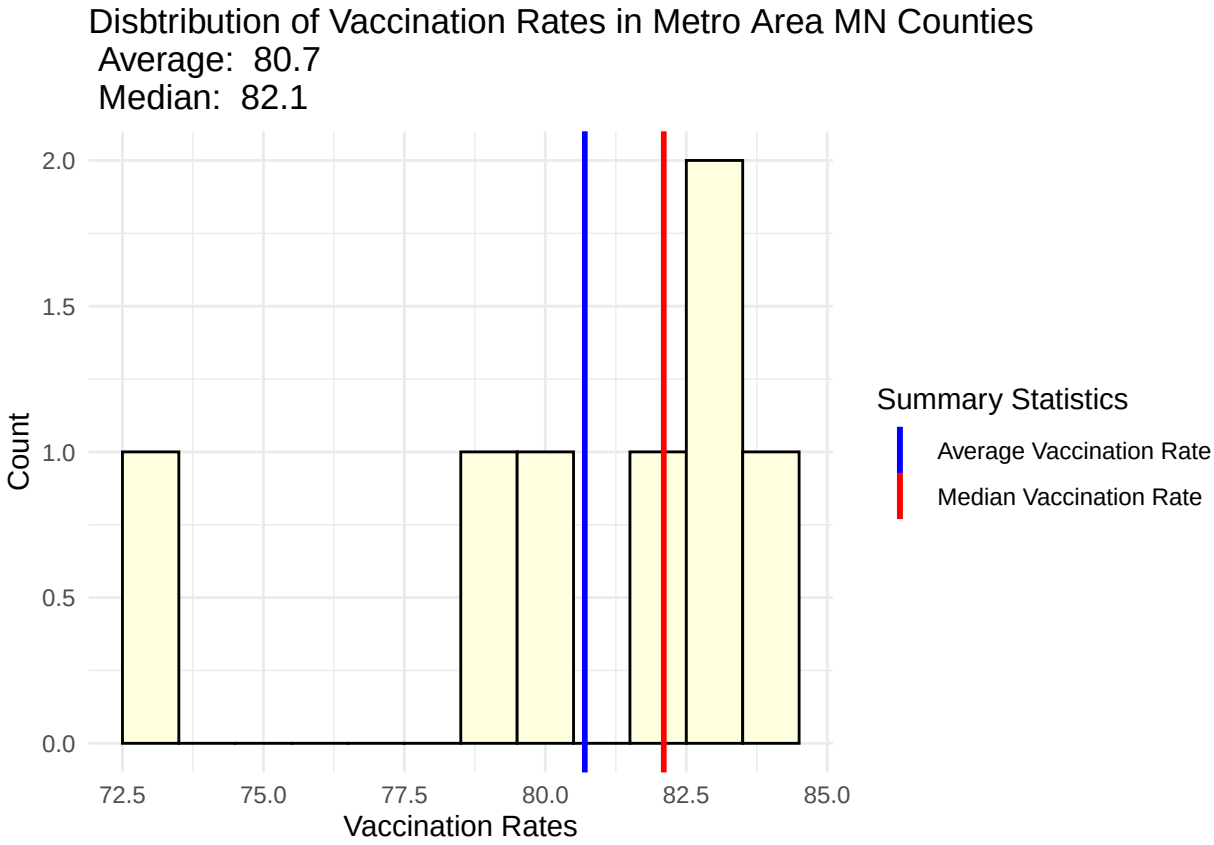
1- D

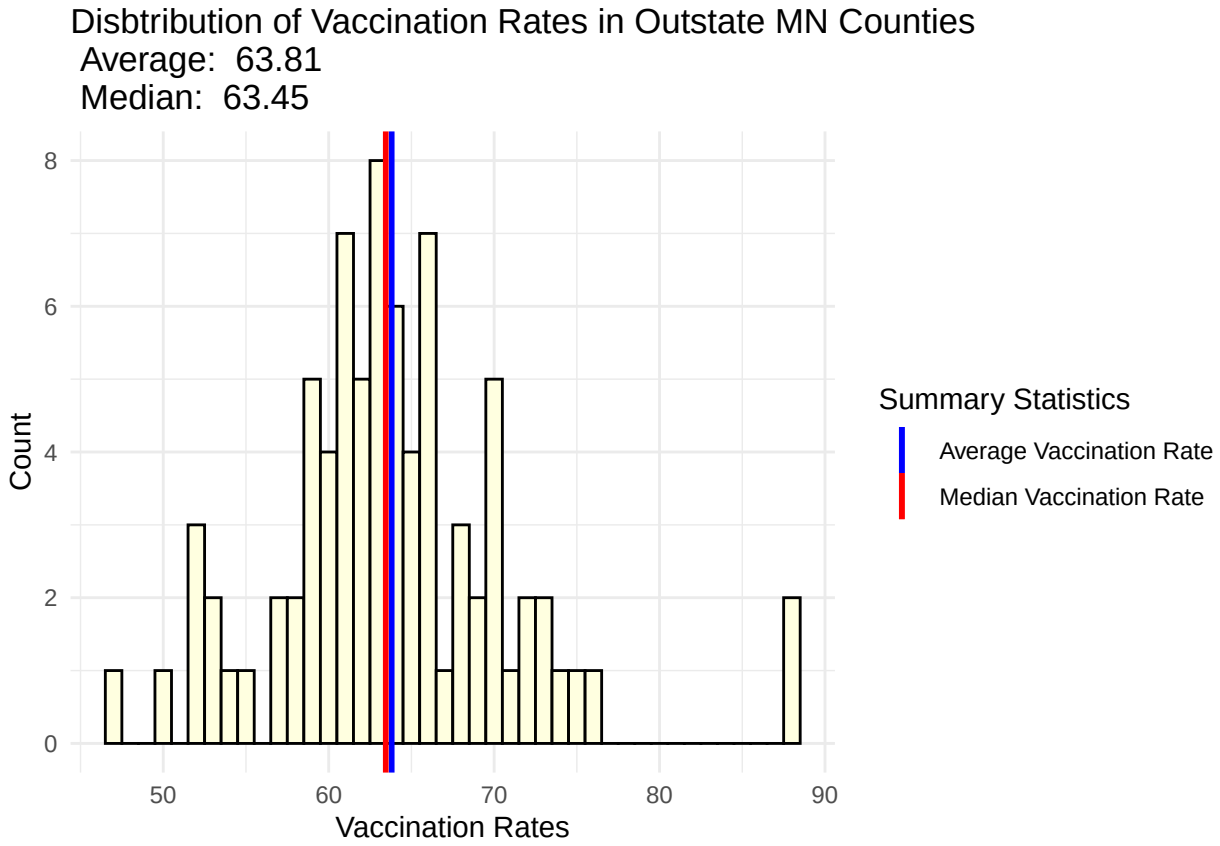
Effect plots needed here, find a nice package


Added Variable Plot for the Number of Hours Worked
Y Residuals
X Residuals
1
0
−1
−2
−800
−400
0
400
Line Type
Smooth Trend Line
Fitted Regression Line

Problem 2

2 - A


Disbtribution of Vaccination Rates in Metro Area MN Counties
Average: 80.7
Median: 82.1
Count
2.0
1.5
1.0
0.5
0.0
72.5
75.0
77.5
80.0
82.5
85.0
Vaccination Rates
Summary Statistics
Average Vaccination Rate
Median Vaccination Rate



Outlier: Olmsted, Cook 88.5, 87.9

Olmsted includes Rochester Cook is by the Canadian Border

Table 3: Vaccination Rates Summary by County Type

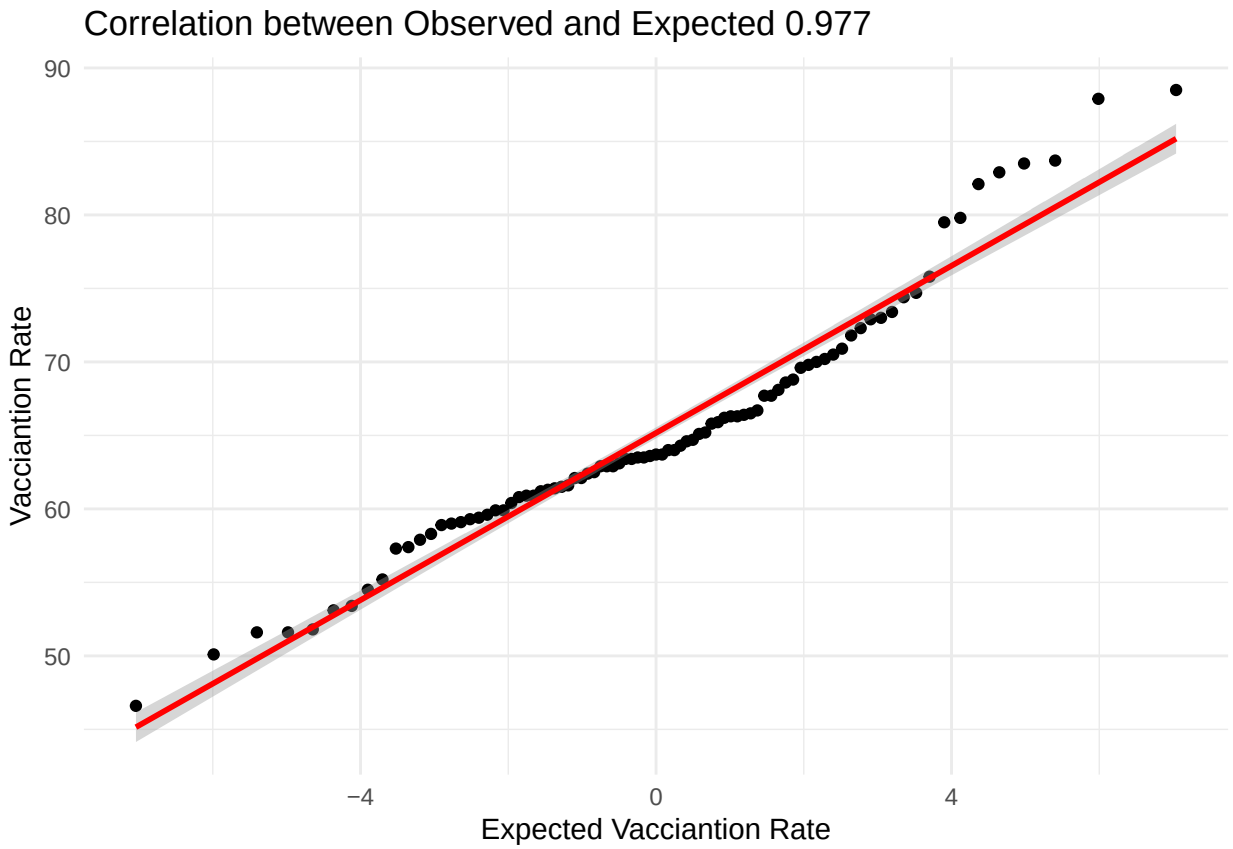
Type	N	Mean	Median	S.D.
Outstate	80	63.81	63.45	7.08
Metro	7	80.70	82.10	3.63

- **Normality of Vaccination 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:

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

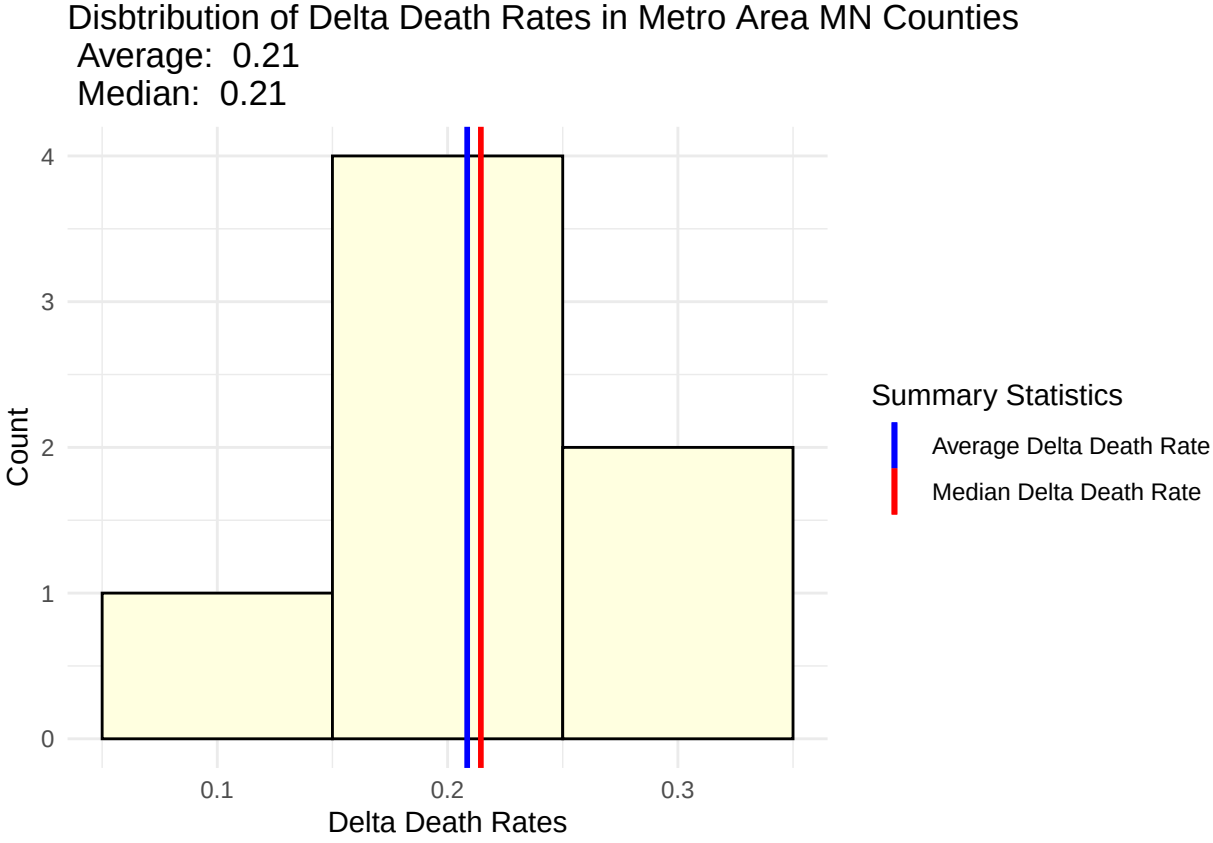


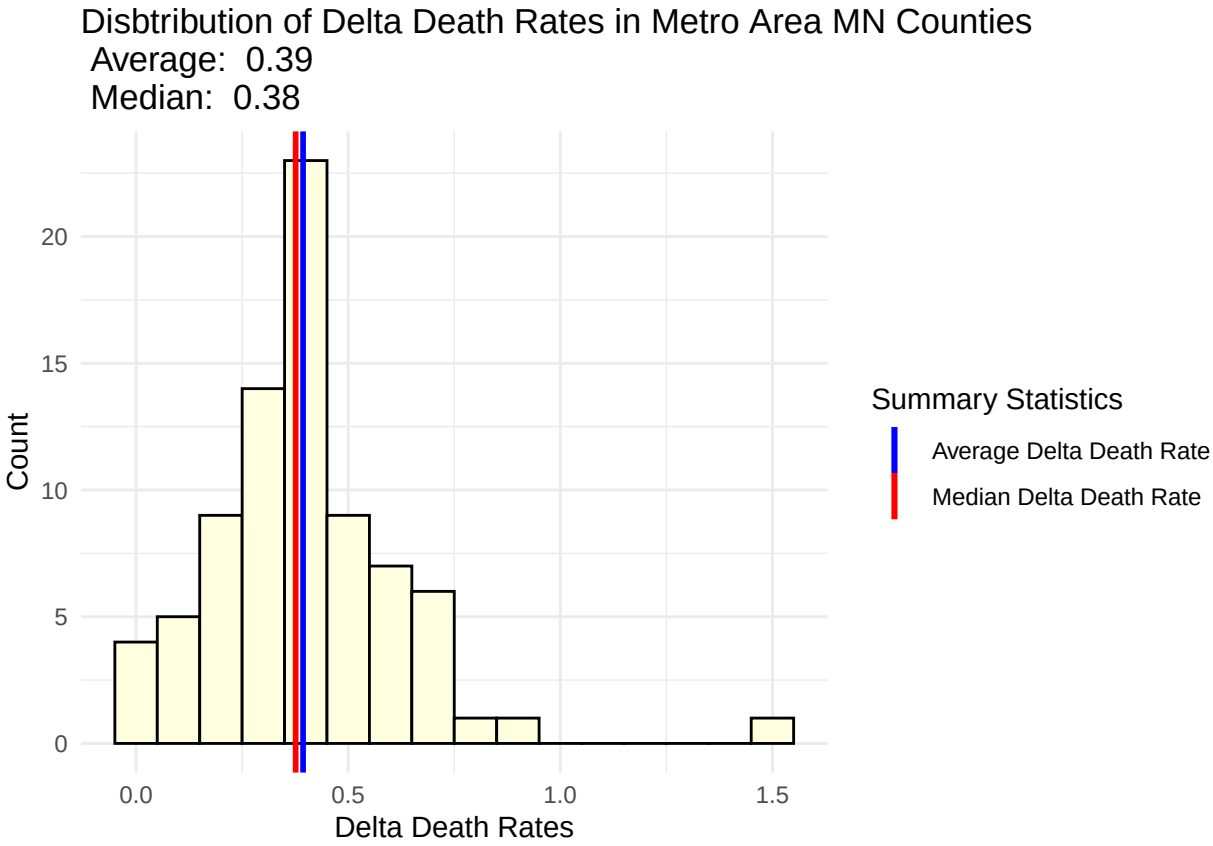
Use T test here. No heavy tails or skewed distributions So nothing should affect the results of the T test In such samples wilcoxon only has 95% of statistical power that the t test brings, so we will use the t test

Test results summary and interpretation:

- Null Hypothesis: $H_0 : Mean_{metro\ area} = Mean_{outstate}$
- Test statistic: $H_a : Mean_{metro\ area} \neq Mean_{outstate}$
- 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:

2 - B


Disbtribution of Delta Death Rates in Metro Area MN Counties
Average: 0.21
Median: 0.21
Count
0
1
2
3
4
0.1
0.2
0.3
Delta Death Rates
Summary Statistics
Average Delta Death Rate
Median Delta Death Rate



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: Vaccination Rates Summary by County Type

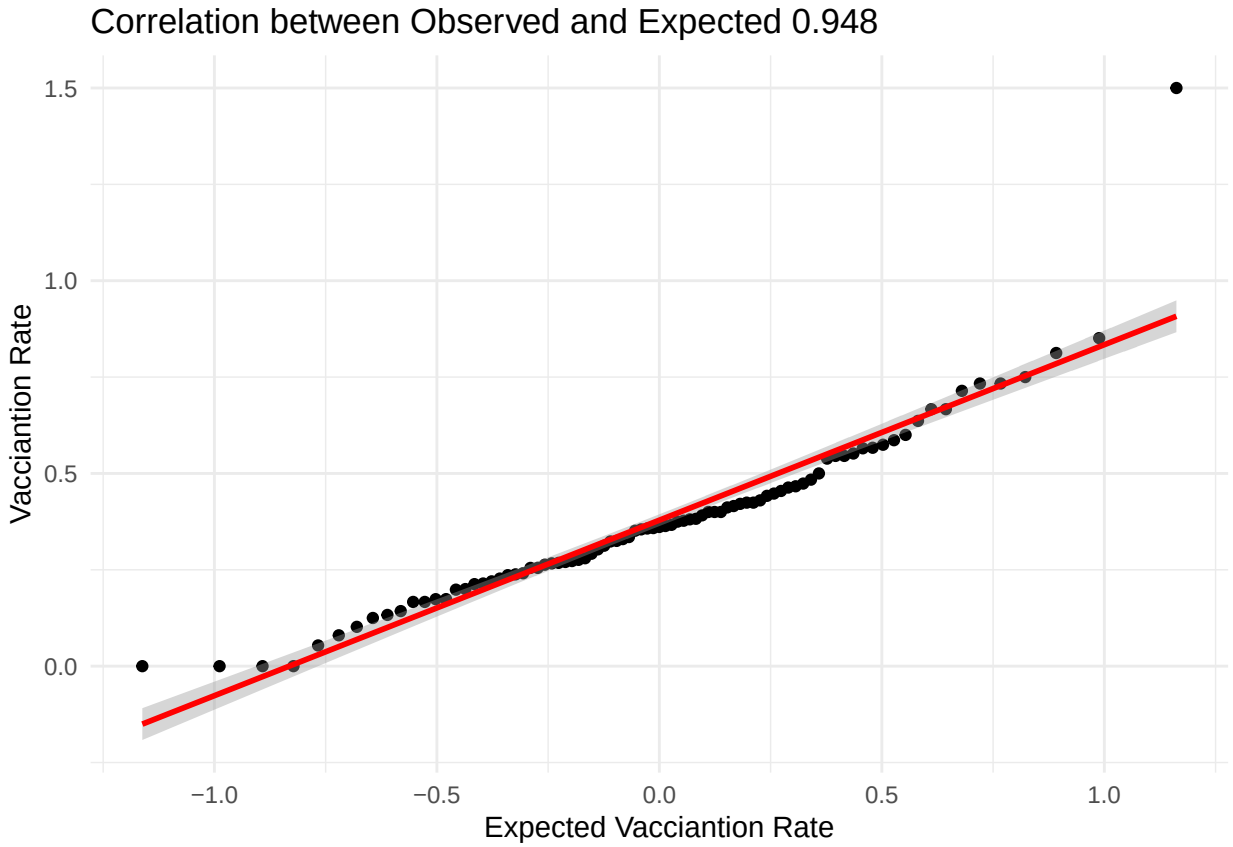
Type	N	Mean	Median	S.D.
Outstate	80	0.3936199	0.3761792	0.23
Metro	7	0.2085040	0.2144522	0.04

- **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: $H_0 : Mean_{metro\ area} = Mean_{outstate}$
- Test statistic: $H_a : Mean_{metro\ area} \neq Mean_{outstate}$
- 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

```
## Analysis of Variance Table
##
## Model 1: death_rate ~ 1
## Model 2: death_rate ~ v_rate + region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     86 4.3595
## 2     84 3.7854  2   0.57403 6.3689 0.002659 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 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)

2 - D