I focus on the methods: hypothesis testing and Gaussian Response model. I begins with hypothesis testing for Binomial models, building on our previous analyses of the subject.sex variate. This will allow us to more formally test whether the proportion of people stopped who identify as female is similar to that proportion in the broader population. I will then extend the analyses of the third file named *interval estimation*, where explored the location of traffic stops in relation to the geographic centres of each city. Whereas in the third file named *interval estimation* we explored this using confidence intervals, we will now deepen our understanding using hypothesis testing. Our third analysis extends our investigation into the relationship between the latitude and longitude of traffic stops, which I began all the way back in the first file named *Subject age variate*. Whereas then I were limited to inspecting scatterplots and sample correlation, then I am now able to employ the extremely powerful technique of linear regression! Note that Analysis 3 is worth almost half the marks in this assignment, so do plan your time accordingly! I'll finish off with a final analysis that explores subject age and its relationship with subject sex, race, and/or the make of vehicle the subject was stopped in. This will build on analyses conducted in file1(*Subject age variate*) and 2(*interval estimation*), and you'll get to decide which analysis you're most interested in extending!

I will analyze data from Chicago and test the null hypothesis H_0: θ_c = 51.3%

The observed value of the test statistic is [57.767], and the resulting p-value is [2.953193e-14]. The p-value was calculated using the [chi-squared distribution & Binomial distribution].

```
# test statistic:
lambda <- (-2*log((theta0/thetahat)^(n*thetahat)*((1 - theta0)/(1 - thetahat))^(n - n*thetahat)))
lambda
# p-value:
w <- pchisq(lambda, df = 1)
pvalue <- 1 - w
pvalue
```

Based on our results from Analysis before. I conclude [there is very strong evidence against H0 based on the observed data, the traffic stop by female is equal to 51.3%. Since 2.953193e^(-14) is much smaller than 0.001, which is very small.].

I will analyze data from Chicago below:

To test H_0: μ = [41.8781].

I calculatd the observed value of the test statistic using $\frac{|mean-41.8781|}{sd/\sqrt{n}}$, where mean is sample mean, sd is sample standard deviation, n is sample size. The value of the test statistic for my sample is 7.609.

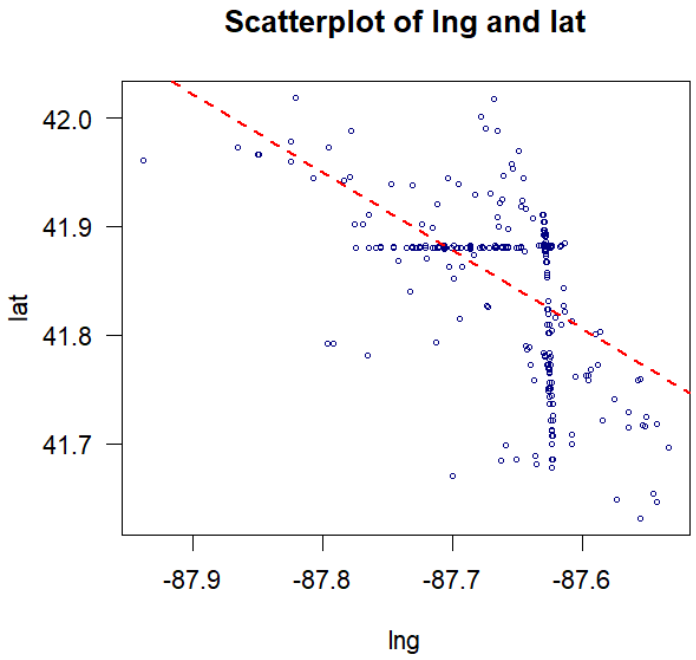
p-value = $P(D \geq d) = 2 * (1 - P(T \leq d))$, and the resulting p-value is 0.

Based on the results of Analysis 2b, I conclude there is very strong evidence against null hypothesis based on the observed data(almost no evidence). Since 0 is much smaller than 0.001, which is very small.

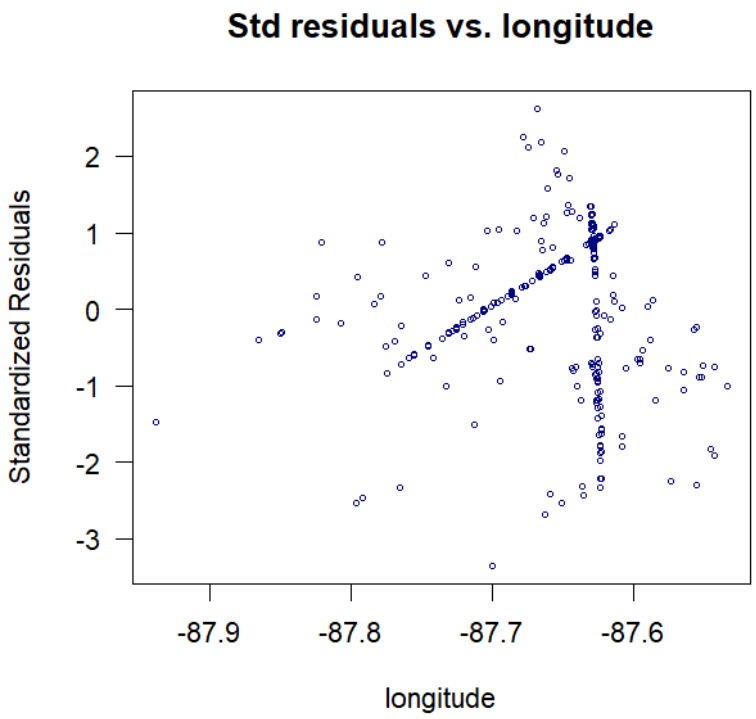
The least squares estimate of α is [-21.392], with 95% confidence interval [-29.937, -12.847]. The least squares estimate of β is [-0.721], with 95% confidence interval [-0.818, -0.623]. The estimate of sigma is [0.062].

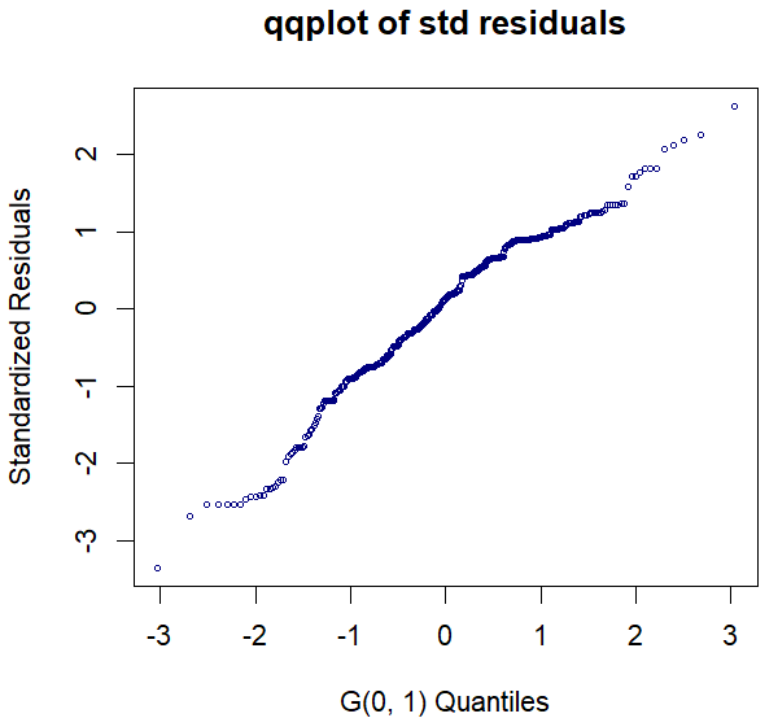
In the context of this study, α represents [the mean latitude in the study population of traffic stops in Chicago with a longitude=0].

Scatterplot:(3e)



Residual and Q-Q Plots: (3f)





The linear model assumes the variance does not depend on longitude, there is a linear relationship between the response and explanatory variates, and the standardized residuals have a Gaussian distribution: $Yi \sim G(-21.392 - 0.721xi , \sigma)$. If these hold, I would expect to see the data points fit reasonably along a straight line: $y=-21.392 - 0.721x$ and spread out reasonably equally in the scatterplot. For my sample, I observe the points don't lie in a straight line, which means data doesn't fit Ill in the linear regression model.

In the scatterplot illustrating standardized residuals versus the natural logarithm of traffic stops in Chicago, a noticeable dispersion of standardized residuals is observed within the longitude range of -87.8 to -87.6. In contrast, in other longitudes, these residuals tend to cluster around -1, suggesting a potential lack of constant variance.

In the qq-plot for standardized residuals, a discernible pattern emerges where the residuals consistently fall below the straight line beyond the 1st quantile of the G(0, 1) distribution. Additionally, within the interval of -1.5 to -0.6, the standardized residuals generally exceed the expectations set by the straight line. These observed patterns cast doubt on the adequacy of the Gaussian model to accurately capture the behavior of the standardized residuals.]. Overall, the linear model [does not] seem suitable for my sample.

An estimate of the value of lat for a future traffic stop that occurs at a longitude of 100 degrees west is value, with 95% prediction interval 49.543, 51.960.

The p-value of a test of H_0: $\beta = 0$ is smaller than $2 * 10^{(-16)}$. This was calculated using t distribution.

Based on the results of Analysis 3i, I conclude there is very strong evidence against the null hypothesis that there is no linear relationship between latitude and longitude of traffic stops in Chicago, Since $2 * 10^{(-16)}$ is much smaller than 0.001, which is very small.

I will be comparing subject.sex in Chicago below.

Sample Statistic	female	male
Size	137	280
Mean	3.458	3.541
Median	3.466	3.497
Standard deviation	0.325	0.373

To test H_0: $\mu_0 = \mu_1$ we use a [unpaired] test because [the samples are independent, the observations in sex are not related on the observations in age.]. The observed value of the test statistic is calculated by $d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{m_2}}}$, where y1 is the sample mean of female =3.458, y2 is the sample mean of male =3.541, n1 is the sample size of female =137, n2 is the sample size of male = 280, sp is the pooled standard deviation estimate (calculated by

```
sqrt(((femalenum - 1) * (sd(chifemale$subject.age.log))^2 +
        (malenum - 1) * (sd(chimale$subject.age.log))^2)/(nrow(chicago) - 2))
```

]. The value of the test statistic for my sample is [0.358]. To calculate the p-value we explanation of process of calculating p-value, and the resulting p-value is [0.270].

The results in Analysis 4c rely on the following assumptions: [

1. Gaussian model holds for both male and female in Chicago
2. The variance(σ) for both male and female in Chicago are the same].

Based on the results of Analysis 4c, I conclude [there is no evidence against null hypothesis based on the observed data, since 0.27 is larger than 0.10].