Homework 2

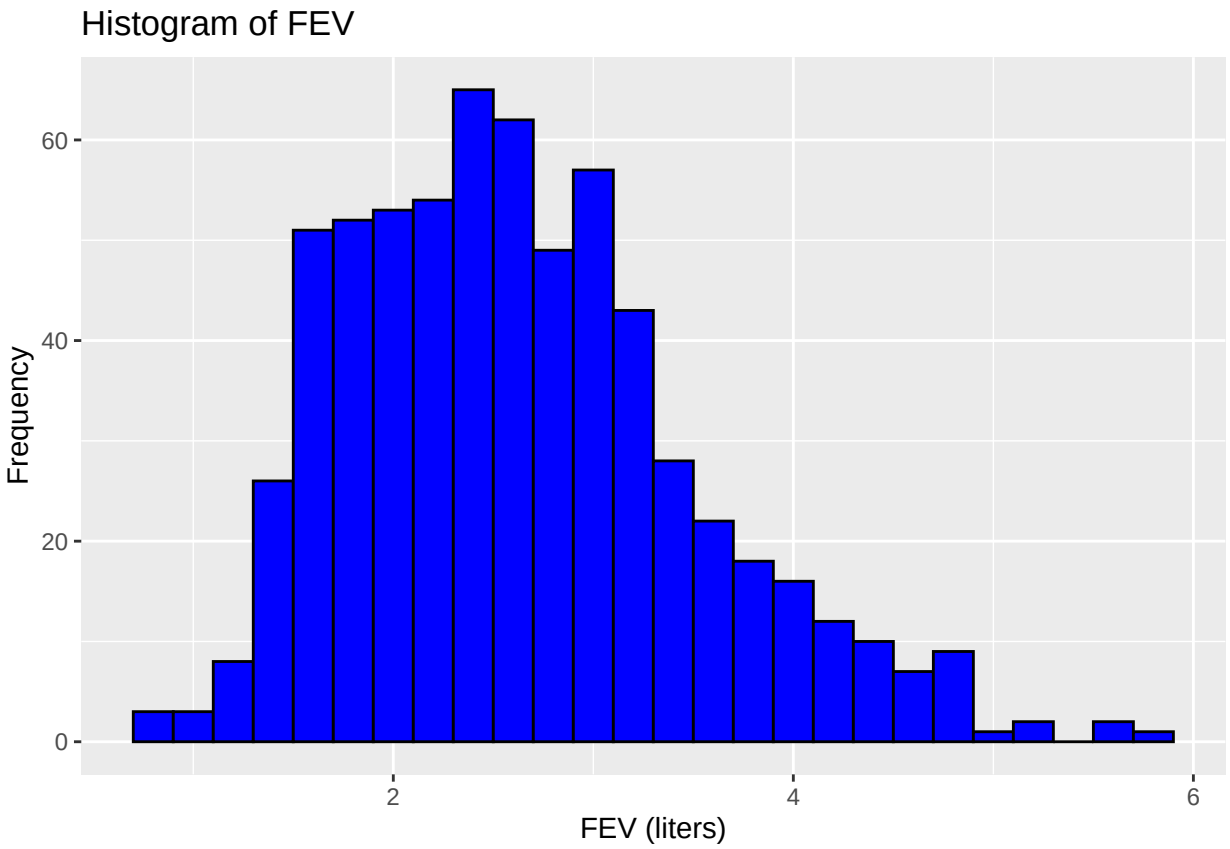
Jacob Thielemier

22 February 2024

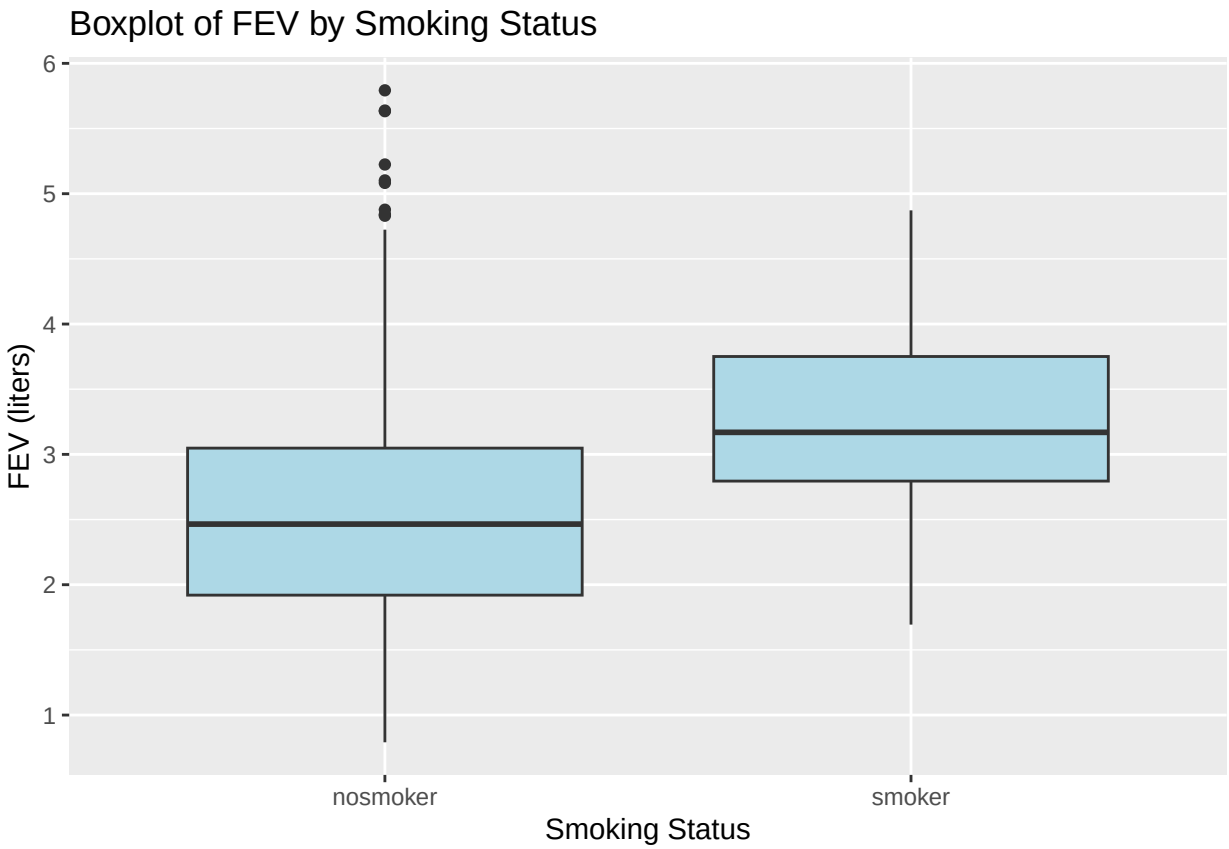
Question 1

(a)

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```



- The histogram shows that the data is appropriate for simple linear regression due to the large amount of the patients FEV value being grouped together. This right-skewed shape of the histogram shows the bulk of patients are grouped around the value of 2 with small amounts being much higher or much lower.



(b)

- The boxplot shows that the data is appropriate for simple linear regression due to the overlap of male and female patients FEV value. The boxplot shows that the mean value of the two groups is close as well as the majority of the values being between 2 - 3.5. There are more outlines for nonsmoker, but smoker looks right skewed. The boxplot does show counter intuitive information since traditionally smokers should have lower FEV.

(c)

```
##
## Call:
## lm(formula = fev ~ smoke, data = fev_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7751 -0.6339 -0.1021  0.4804  3.2269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
## smokesmoker  0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8412 on 652 degrees of freedom
## Multiple R-squared:  0.06023,    Adjusted R-squared:  0.05879
```

```
## F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

Table 1: FEV Regression Results

Coefficient Name	Point Estimate	Standard Error	P-value
Average Patient	2.566	0.035	<0.05
Smoker	0.710	0.109	<0.05

(d)

```
##                  2.5 %    97.5 %
## (Intercept) 2.4980831 2.6342021
## smokesmoker 0.4948346 0.9266033
```

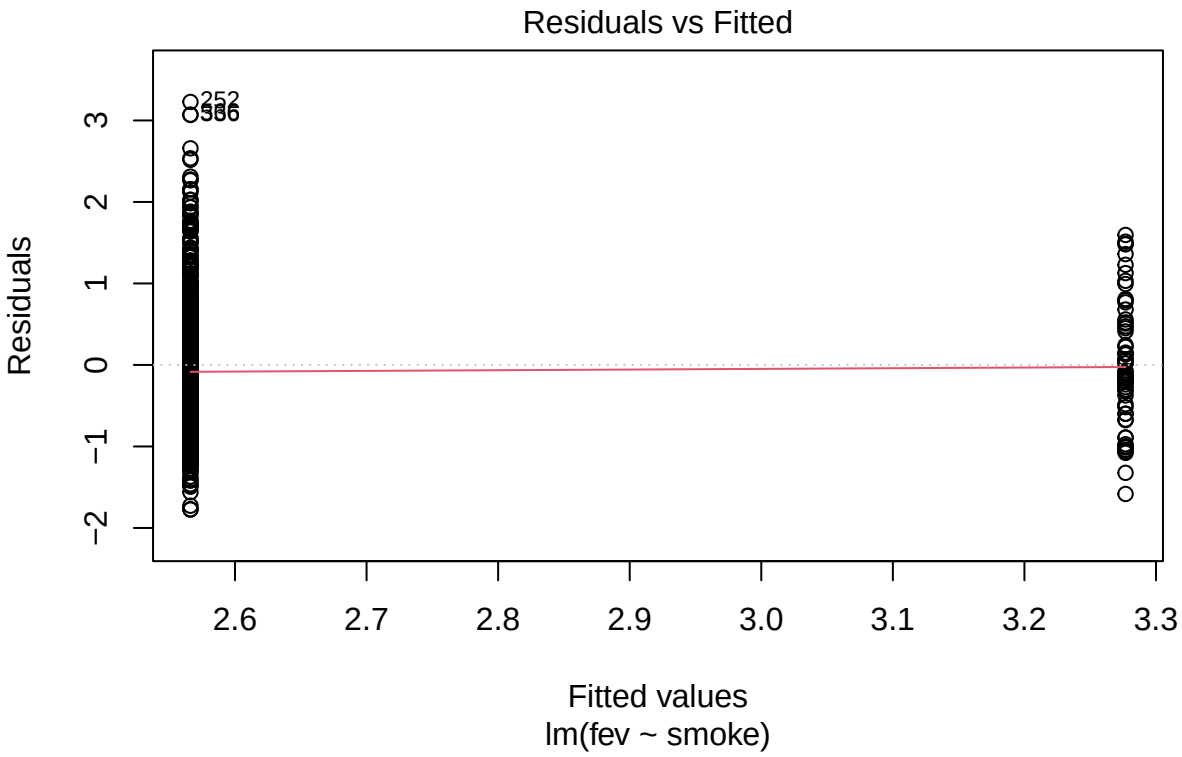
- The 95% confidence interval is from 0.495 to 0.926

(e)

- This C.I. of 0.495 - 0.926 means that if we were to repeat the study many times, 95% of the calculated confidence intervals from those studies would contain the true effect size. For this problem it gives a range within which we can reasonably expect the true effect of smoking on FEV to lie and it informs decisions or recommendations regarding smoking's impact on lung function.

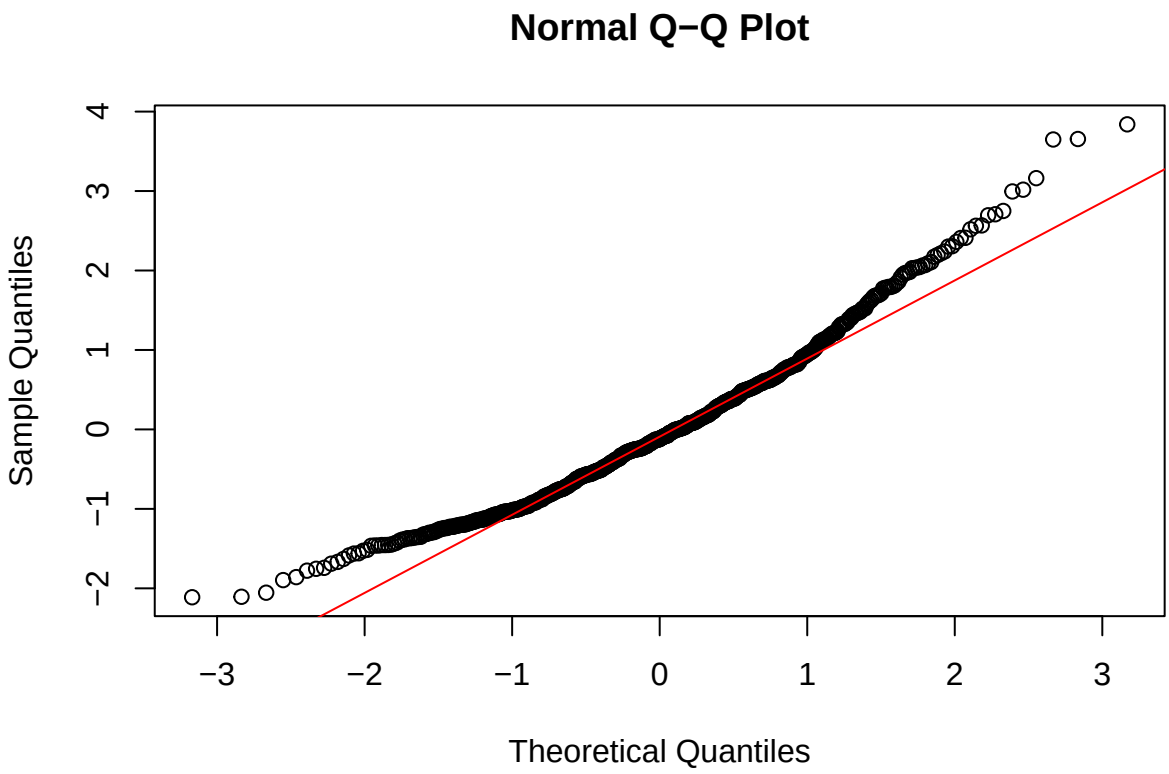
(f)

- 95% CI for the average FEV among nonsmokers: 2.498083 2.634202
- 95% CI for the average FEV among smokers: 2.992918 3.560805



(g)

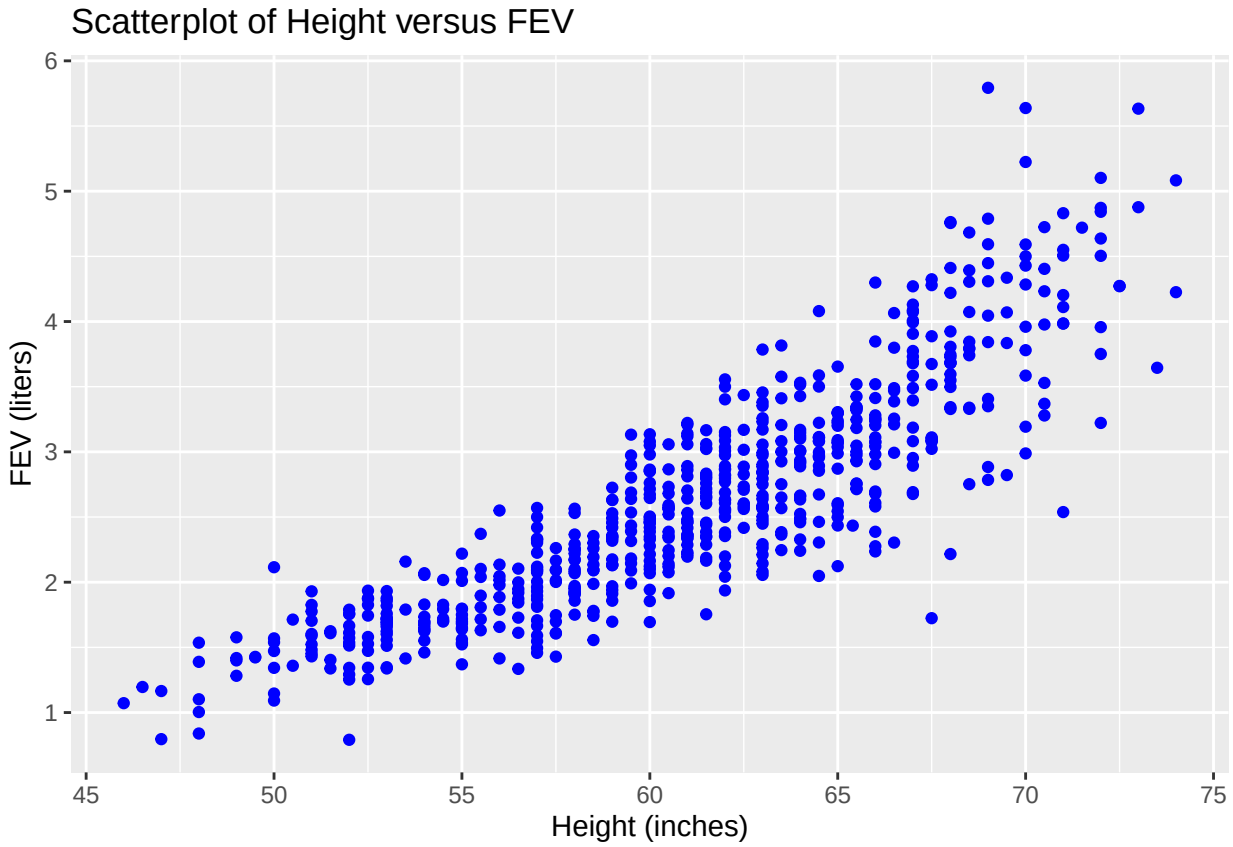
- The data appears to violate the assumption of linearity of linear regression. This is because there is no random scattering of points around the horizontal line at 0.



(h)

- The data appears to satisfy the assumptions of linear regression. This is becuase the points on the Q-Q Plot closely follow the reference line with minor deviations at both ends.

Question 2



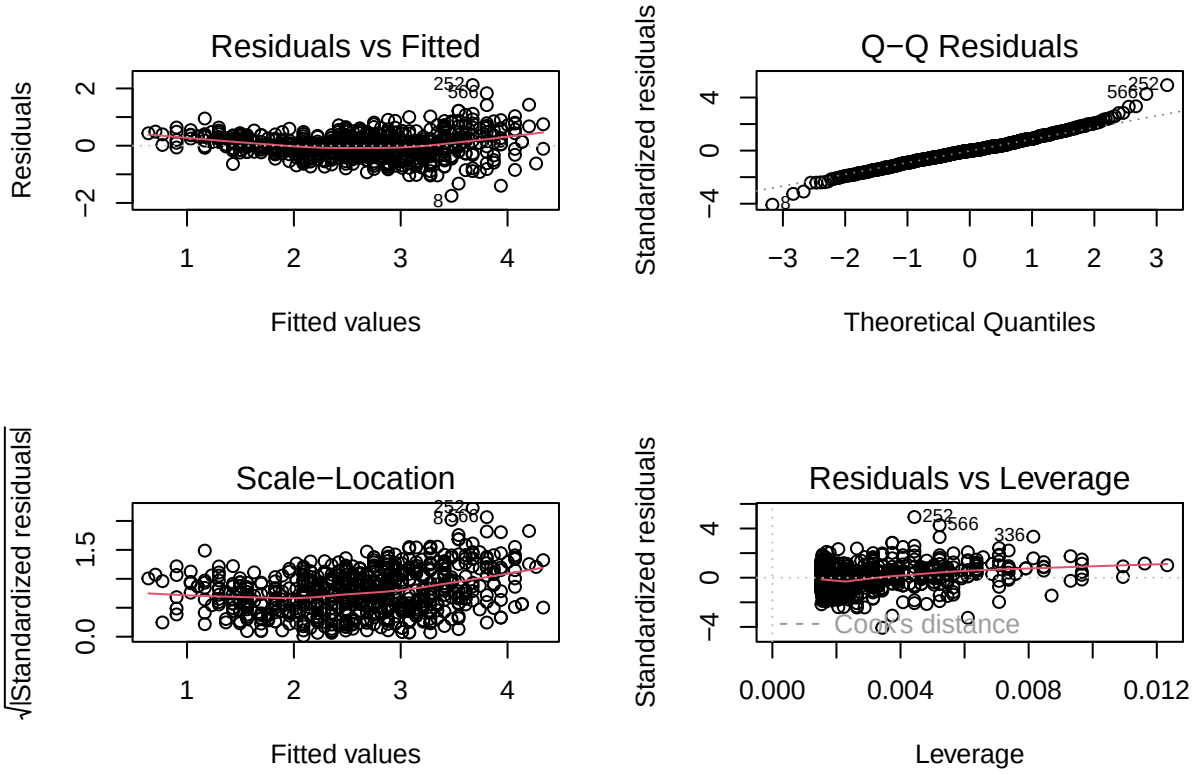
(a)

(b)

```
##
## Call:
## lm(formula = fev ~ height, data = fev_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75167 -0.26619 -0.00401  0.24474  2.11936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.432679   0.181460  -29.94   <2e-16 ***
## height       0.131976   0.002955   44.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4307 on 652 degrees of freedom
## Multiple R-squared:  0.7537, Adjusted R-squared:  0.7533
## F-statistic:  1995 on 1 and 652 DF,  p-value: < 2.2e-16
```

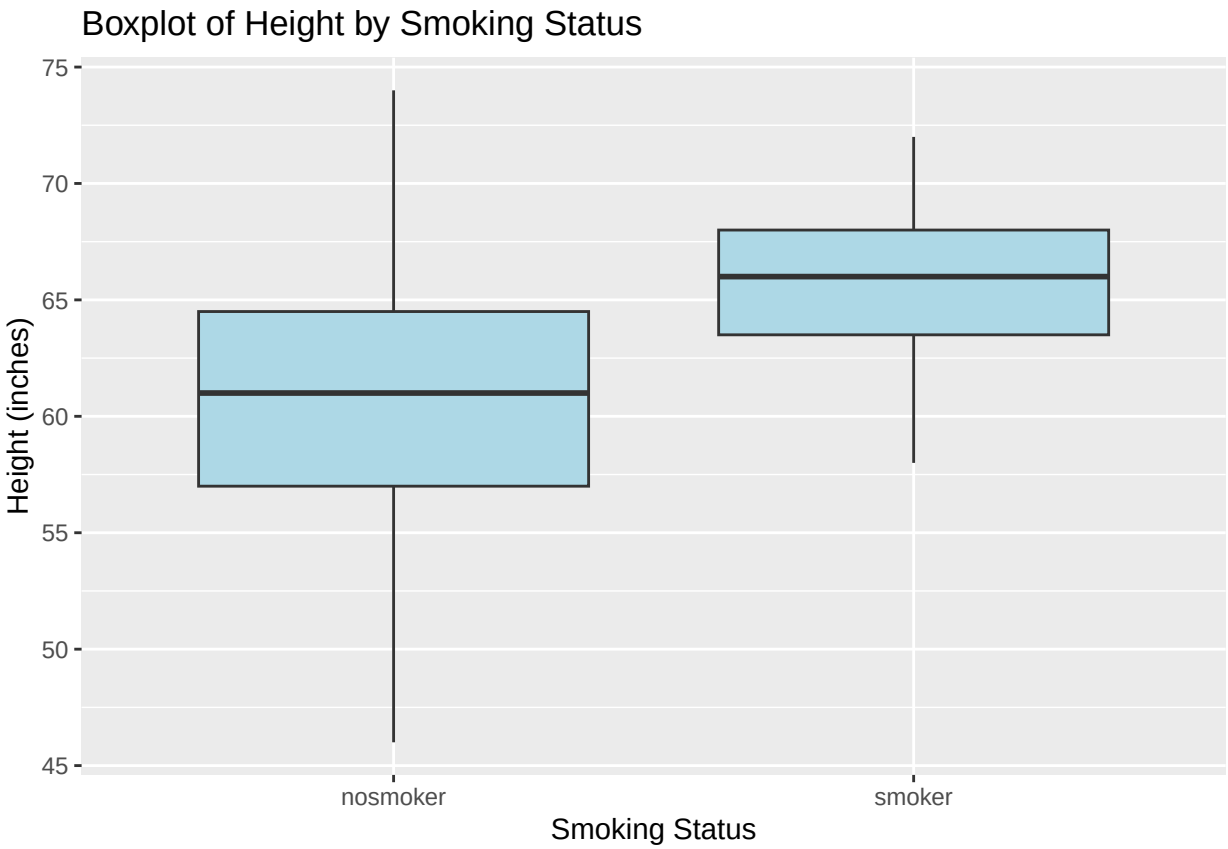
Table 2: FEV Regression Results of Height

Coefficient Name	Point Estimate	Standard Error	P-value
Average Patient	-5.433	0.181	<0.05
Height	0.132	0.003	<0.05



(c)

- The Residuals vs Fitted plot shows a < shape getting wider on the right hand side. This violates the assumption of constant variance.



(d)

- The box plot shows that smokers tend to be taller. Smokers have a higher average height around 66in and nonsmokers have an average height around 61in.

(e)

$$\text{FEV} = \beta_0 + \beta_1(\text{SmokingStatus}) + \beta_2(\text{Height}) + \varepsilon$$

```
##
## Call:
## lm(formula = fev ~ smoke + height, data = fev_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7505 -0.2660 -0.0041  0.2447  2.1207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.427620   0.187577 -28.935   <2e-16 ***
## smokesmoker  0.006319   0.058686   0.108    0.914
## height       0.131883   0.003081  42.808   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.431 on 651 degrees of freedom
## Multiple R-squared:  0.7537, Adjusted R-squared:  0.7529
## F-statistic: 995.9 on 2 and 651 DF,  p-value: < 2.2e-16
```

Table 3: FEV Regression Results of Smoking and Height

Coefficient Name	Point Estimate	Standard Error	P-value
Average Patient	-5.433	0.181	<0.05
Smoker	0.006	0.089	0.914
Height	0.132	0.003	<0.05

(f)

```
##                   2.5 %   97.5 %
## smokesmoker -0.1089173 0.121556
```

- The confidence interval is from -0.109 to 0.122

(g)

- The C.I. has gotten smaller and shifted lower than in question 1. This is due to adjusting for height when calculating the effect of smoking FEV. A negative value is concerning since FEV cannot be negative in people.

(h)

```
## Warning: package 'lmtest' was built under R version 4.3.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## 95% CI for the difference: -0.01138133 - 0.07350448
```

Question 3

(a)

- To prove $\hat{Y}_h$ is unbiased estimator for $E[Y|X = X_h]$ we need:

$$E[\hat{Y}_h] = E[\hat{\beta}_0 + \hat{\beta}_1 X_h] = E[Y|X = X_h]$$

- The expected values of the estimators:

$$E[\hat{\beta}_0] = \beta_0$$

$$E[\hat{\beta}_1] = \beta_1$$

- Then combine them:

$$E[\hat{Y}_h] = E[\hat{\beta}_0] + E[\hat{\beta}_1]X_h$$

- Showing the estimator is unbiased:

$$E[\hat{Y}_h] = \beta_0 + \beta_1 X_h$$

(b)

- Variance of $\hat{\beta}_1$:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- Variance of $\hat{\beta}_0$:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

- The covariance between $\bar{Y}$ and $\hat{\beta}_1$:

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = 0$$

- Due to independence and identically normally distributed.
- The variance of $\text{Var}(\hat{Y}_h)$:

$$\text{Var}(\hat{Y}_h) = Var[\hat{\beta}_0 + \hat{\beta}_1 X_h]$$

- Since the covariance is 0:

$$\text{Var}(\hat{Y}_h) = Var[\hat{\beta}_0] + Var[\hat{\beta}_1 X_h]$$

$$\text{Var}(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] + \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} X_h$$

$$\text{Var}(\hat{Y}_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

(c)

- The error term ϵ is normally distributed:

$$\epsilon \sim N(0, \sigma^2)$$

- The predicted value of $\hat{Y}_h$ at X_h is normally distributed:

$$E[\hat{Y}_h] = \beta_0 + \beta_1 X_h$$

- Using the variance of $\hat{Y}_h$ from the last question to calculate:

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h, \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\right)$$

(d)

- The variance of $\hat{Y}_h$ is:

$$\text{Var}(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

- As X_h and $\bar{X}$ increase the variance of $\hat{Y}_h$ also increases. This is because $(X_h - \bar{X})^2$ is a numerator in our formula.

(e)

- The numerator $\hat{Y}_h - E[Y|X = X_h]$ is the difference between the predicted value and the expected value of Y given $X = X_h$ which is assumed to be 0.
- The denominator is the standard error of the prediction $\hat{Y}_h$ which we estimate using the sample variance s^2
- The t-distribution is used instead of the normal distribution when estimating the variance from a sample rather than knowing the true population variance.

$$\frac{\hat{Y}_h - E[Y|X = X_h]}{\sqrt{s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}} \sim t_{n-2}$$

(f)

- The prediction interval:

$$PI = \hat{Y}_{\text{new}} \pm t_{\alpha/2,n-2} \times SE_{\text{pred}}$$

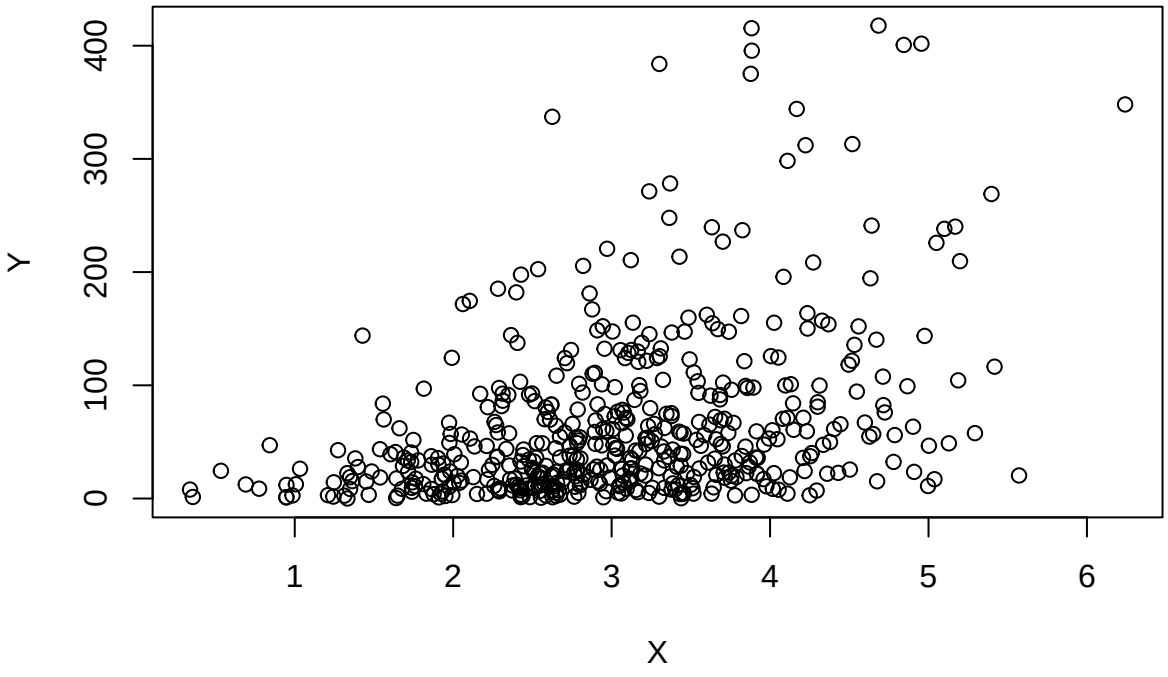
- The standard error of the prediction:

$$SE_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(X_{\text{new}} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

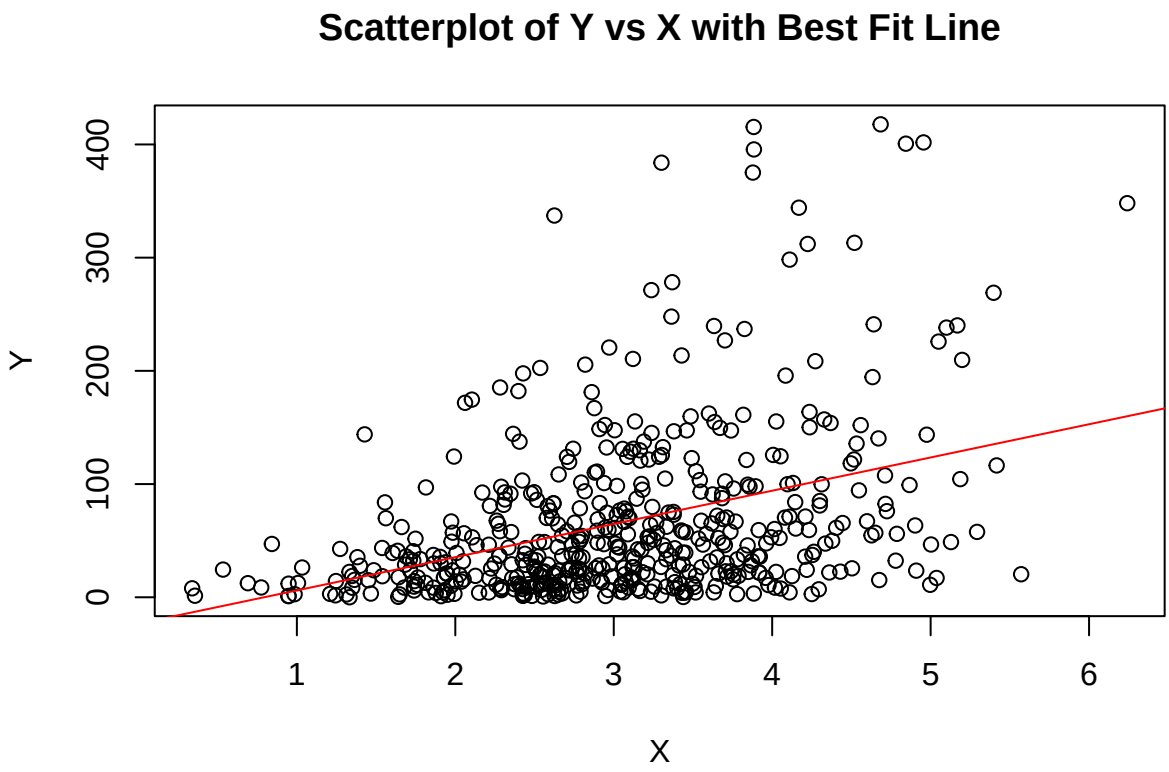
(g)

- The prediction interval is wider than the confidence interval for the mean of Y at a given X because it accounts for the additional variability associated with the individual outcome Y_{new} rather than the mean outcome.

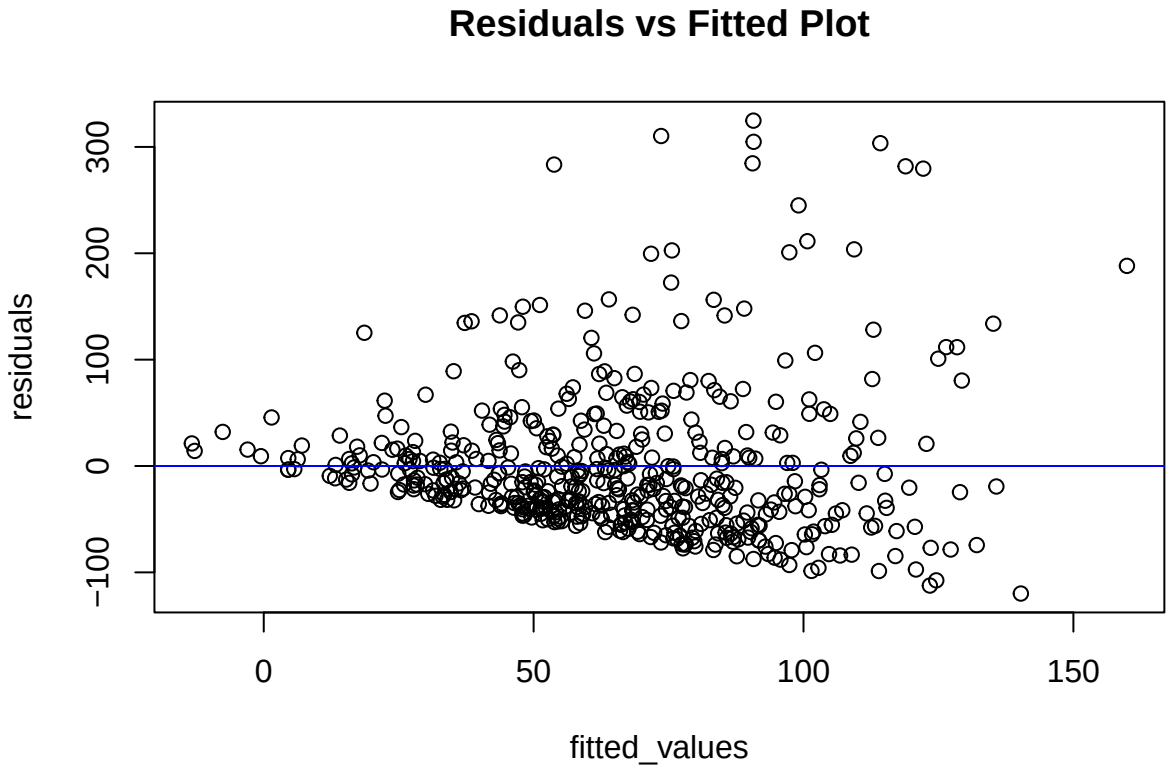
Question 4



(a)


Scatterplot of Y vs X with Best Fit Line
Y
X
0
100
200
300
400
1
2
3
4
5
6

(b)


Residuals vs Fitted Plot
residuals
fitted_values
−100
0
100
200
300
0
50
100
150

- The Scatterplot does show the relationship between X and Y to be linear, but since the points do not cluster symmetrically around the line the relationship may not be perfectly linear.
- The Residuals vs Fitted plot shows a $<$ shape getting wider on the right hand side. This violates the assumption of constant variance.

(c)

$$Y \sim \text{Exp}(\mu)$$

The variance of Y is:

$$\text{Var}(Y) = \mu^2$$

The variance of the transformed variable $T(Y)$ using Taylor's expansion is approximately:

$$\text{Var}(T(Y)) \approx (T'(\mu))^2 \text{Var}(Y)$$

For $T(Y)$ to stabilize the variance, $\text{Var}(T(Y))$ must be constant, so we set the derivative of μ to zero:

$$\frac{d}{d\mu}\text{Var}(T(Y)) = \frac{d}{d\mu}\left[(T'(\mu))^2 \mu^2\right] = 0$$

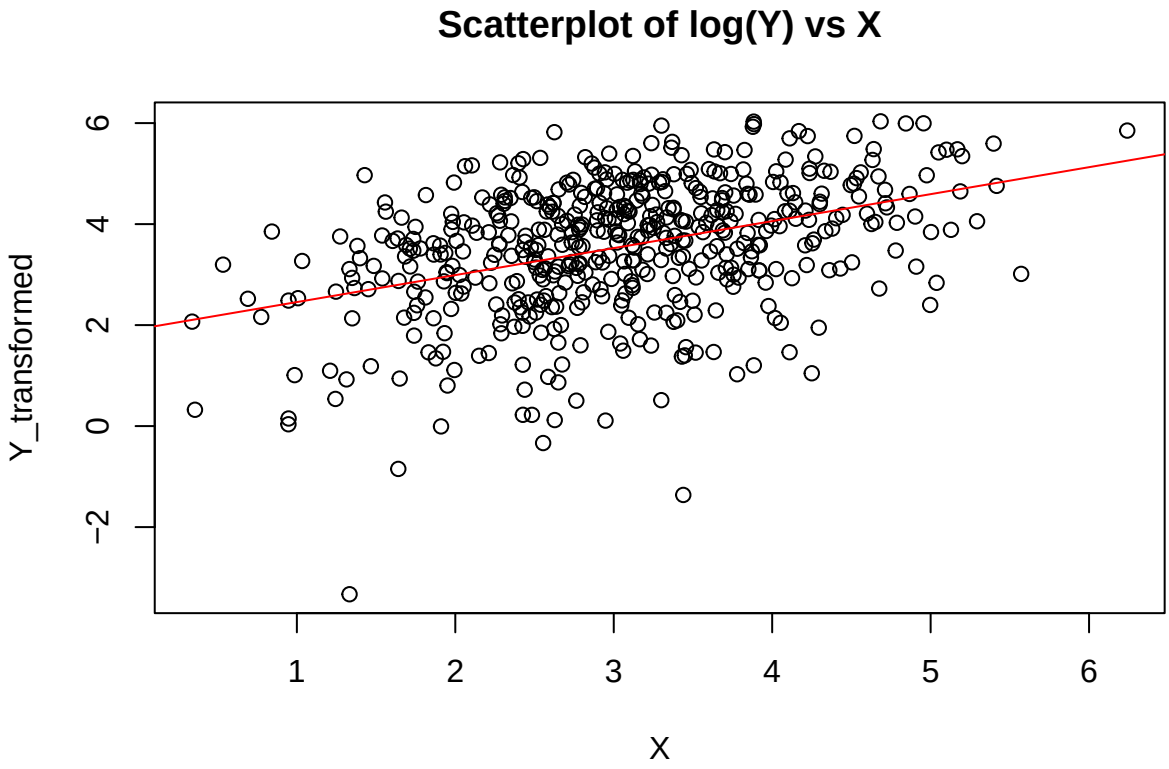
$T(Y) = \log(Y)$. The derivative of $T(Y)$ with respect to Y is:

$$T'(Y) = \frac{1}{Y}$$

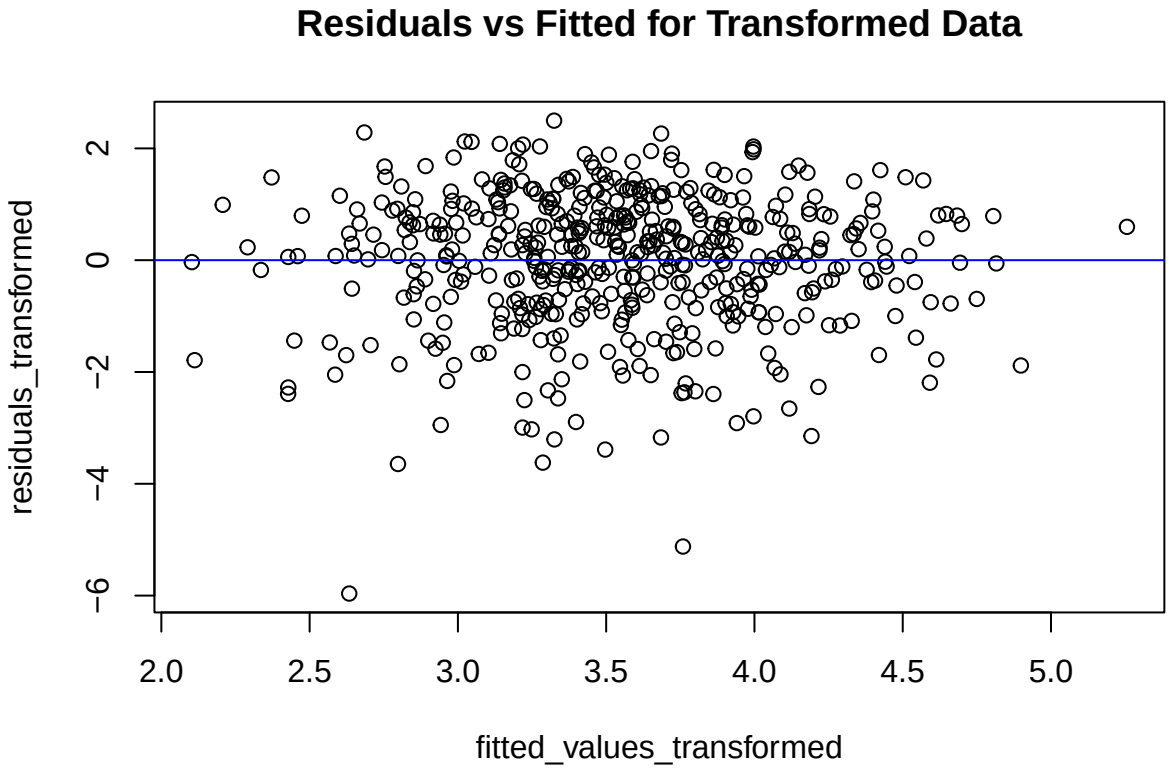
Substituting $T'(Y)$ and $\text{Var}(Y)$ into the equation:

$$\text{Var}(T(Y)) \approx \left(\frac{1}{\mu}\right)^2 \mu^2 = 1$$

This shows that the natural logarithm is a variance stabilizing transformation for the exponential distribution, as it makes the variance of $T(Y)$ constant.


Scatterplot of log(Y) vs X
Y_transformed
X
−2
0
2
4
6
1
2
3
4
5
6

(d)


Residuals vs Fitted for Transformed Data
residuals_transformed
fitted_values_transformed
−6
−4
−2
0
2
2.0
2.5
3.0
3.5
4.0
4.5
5.0

```
##
## Call:
## lm(formula = Y_transformed ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9633 -0.7109  0.1476  0.8262  2.4963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.92165    0.17677  10.871   <2e-16 ***
## X            0.53428    0.05548   9.631   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.205 on 498 degrees of freedom
## Multiple R-squared:  0.157,  Adjusted R-squared:  0.1553
## F-statistic: 92.75 on 1 and 498 DF,  p-value: < 2.2e-16
```

- The transformed data shows a more linear pattern on the Scatterplot. There is an observable slope without a discernible pattern.
- The Residuals vs Fitted plot is also more equal dispersed. This validates the constant variance assumption.

Question 5

(a)

- In multiple linear regression models, β_1 represents the expected change in the dependent variable Y for a change in the independent variable X_1 assuming all other independent variable remain constant.

(b)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Labels:

- $\mathbf{Y}$ is the response vector of size $n \times 1$, with each element representing the observed value of the dependent variable for each observation.
- $\mathbf{X}$ is the design matrix of size $n \times p$, which includes a column of ones for the intercept (β_0) and the values of the $p-1$ covariates for each observation.
- $\boldsymbol{\beta}$ is the coefficient vector of size $p \times 1$, containing the regression coefficients $\beta_0, \beta_1, \ldots, \beta_{p-1}$.
- $\boldsymbol{\varepsilon}$ is the error vector of size $n \times 1$, representing the random errors or residuals for each observation.

Dimensions:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

(c)

- Likelihood function:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right)$$

- Log-likelihood function:

$$\ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right)$$

$$\ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2$$

(d)

- Start with the log-likelihood from last question:

$$\ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2$$

- Partial derivative with respect to beta of the sum of squared residuals:

$$\frac{\partial}{\partial \boldsymbol{\beta}}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 = -2\sum_{i=1}^{n}\mathbf{X}_i(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})$$

- Partial derivative of the log-likelihood function:

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{X}_i(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})$$

- Partial derivative in matrix notation:

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

(e)

- Set the derivative equal to zero:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

- Expand the equation:

$$\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$$

- Isolate beta:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$$

- Multiply both sides by the inverse:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

(f)

- Start with the MLE of β which is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Var}(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2 I)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

(g)

- Expected value of Y given $X = X_h$:

$$E[Y|X = X_h] = \beta_0 + \beta_1 X_{h1} + \ldots + \beta_{p-1}X_{h,p-1}$$

- Variance of Y given $X = X_h$:

$$\text{Var}(Y|X = X_h) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^T(X_h - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^T(X_i - \bar{X})}\right)$$

- Distribution of Y given $X = X_h$:

$$Y|X = X_h \sim \mathcal{N}\left(E[Y|X = X_h], \text{Var}(Y|X = X_h)\right)$$

(h)

- Expected value of $\hat{Y}$:

$$E[\hat{Y}_h] = \mathbf{X}_h^T \boldsymbol{\beta}$$

- Variance of the estimated coefficients $\hat{\beta}$:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

- Variance of Y hat as a linear combination of the estimated coefficients $\hat{\beta}$:

$$\text{Var}(\hat{Y}_h) = \mathbf{X}_h^T \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{X}_h$$

- Substituting the variance of $\hat{\beta}$ into the variance of $\hat{Y}$:

$$\text{Var}(\hat{Y}_h) = \sigma^2 \mathbf{X}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_h$$

- Distribution of Y hat given X_h:

$$\hat{Y}_h | X_h \sim \mathcal{N}\left(\mathbf{X}_h^T \boldsymbol{\beta}, \sigma^2 \mathbf{X}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_h\right)$$