CSCI E-106:Assignment 4

Problem 1

Refer to the CDI data set. The number of active physicians in a CDI is the dependent variable (Y). The number of hospital beds is the independent variable (X). Build a regression model to predict Y (Total=40 points, each part is 5 points).

```
suppressPackageStartupMessages({
  library(ggplot2)
  library(lattice)
  library(caret)
  library(ggfortify)
  library(dplyr)
  library(car)
})
cdi_data <- read.csv('/Users/shreyabajpai/CSCI E-106 - Data Modeling/CSCI E-106 Assignment 4/CDI Data.c
```

a-) Create train and test data sets. Use 70% of the data for train data set and use remaining data (30% of data) for test data set (use set.seed(1023)).

```
set.seed(1023)

DataSplit<-createDataPartition(y = cdi_data$Number.of.active.physicians, p = 0.7, list = FALSE)

training_data <- cdi_data[DataSplit,]
testing_data<-cdi_data[-DataSplit,]
```

b-) Test whether there is linear association between number of active physicians and hospital beds. Using a t test with alpha=0.05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

First, we state the hypotheses:

$H_o \; : \beta_1 = 0$ There is no linear association between hospital beds and number of active physicians.

$H_a : \beta_1 \neq 0$ There is a linear association between hospital beds and number of active physicians.

The decision rule is as follows:

1. If $|t^*| \leq t_{\alpha/2,n-2}$, then we fail to reject H_0, concluding that there is no evidence of a linear association between hospital beds and number of active physicians.
2. If $|t^*| > t_{\alpha/2,n-2}$, then we reject the null hypothesis and conclude H_a, which indicates that there is evidence of a linear association hospital beds and number of active physicians.

```
# Fit regression model of CDI training data
fit_cdi_data <- lm(Number.of.active.physicians ~ Number.of.hospital.beds, data=training_data)

# Extract slope coefficient and its standard error
beta1_hat <- coef(summary(fit_cdi_data))["Number.of.hospital.beds", "Estimate"]  # Slope coefficient
se_beta1 <- coef(summary(fit_cdi_data))["Number.of.hospital.beds", "Std. Error"] # Standard error of th

# Compute the t-statistic for the slope
```

```r
t_stat_cdi <- beta1_hat / se_beta1

# Determine df for the t-test
df_cdi <- nrow(training_data) - 2

# Find the critical t-value at alpha = 0.01
alpha <- 0.05
crit_val_cdi <- qt(1 - alpha/2, df_cdi)

# Calculate p-value
p_val_cdi <- 2 * pt(-abs(t_stat_cdi), df_cdi)

# Output
cat("Test Statistic (t*):", t_stat_cdi, "\n")
```

```
## Test Statistic (t*): 54.95605
```

```r
cat("Critical t-value (at alpha = 0.05):", crit_val_cdi, "\n")
```

```
## Critical t-value (at alpha = 0.05): 1.967721
```

```r
cat("p-value:", p_val_cdi, "\n")
```

```
## p-value: 6.548999e-161
```

Given $|t^*| > t_{\alpha/2,n-2}$ where $54.956051 > 1.9677213$ with a significant p-value $6.5489989 \times 10^{-161} < \alpha = 0.05$, we can reject the null hypothesis and conclude that there is a linear association between hospital beds and number of active physicians.

c-) Set up the ANOVA table for the regression models for the independent variable. How much percent of the variation is explained by the independent variable?

```r
# ANOVA
anova_cdi_dta <- anova(fit_cdi_data)
print(anova_cdi_dta)
```

```
## Analysis of Variance Table
##
## Response: Number.of.active.physicians
##                          Df     Sum Sq    Mean Sq F value    Pr(>F)
## Number.of.hospital.beds   1 1095188129 1095188129  3020.2 < 2.2e-16 ***
## Residuals               307  111325862     362625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Multiple_R_squared_NHB<-summary(fit_cdi_data)$r.square
print(Multiple_R_squared_NHB)
```

```
## [1] 0.9077293
```

As per the ANOVA output and the R^2, $\approx 90.8\%$ of the variability in Number of Active Physicians is explained by Number of Hospital Beds with a significant p-value of $< 2.2\text{e-}16$.

d-) Use Generalized linear test approach to test the significance of the independent variable.

```r
# Reduced Model
fit_redu_cdi_data <- lm(Number.of.active.physicians ~ 1, data=training_data)

# Perform General Linear Test of the two models
```

```r
glt_tst <- anova(fit_redu_cdi_data,fit_cdi_data)

print(glt_tst)
```

```
## Analysis of Variance Table
##
## Model 1: Number.of.active.physicians ~ 1
## Model 2: Number.of.active.physicians ~ Number.of.hospital.beds
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    308 1206513991
## 2    307  111325862  1 1095188129 3020.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Multiple_R_squared<-summary(fit_cdi_data)$r.square
print(Multiple_R_squared)
```

```
## [1] 0.9077293
```

When no predictor is used (Reduced Model - Model 1), the total variability is 1206513991. However, adding the Number of Hospital Beds as a predictor (Full Model - Model 2) reduces the unexplained variability to 111325862. The difference in Sum of Squares between the Reduced Model and the model with the predictor is 1095188129, amounting Model 2 (Full Model) explaining ≈ 90.8% of the variability in Number of Active Physicians. The F-statistic 3020.2 supported by a p-value less than 2.2e-16 allows us to reject the null hypothesis that the number of hospital beds has no effect on the number of active physicians and instead adopt the view that a statistically significant relationship between these two variables exists.

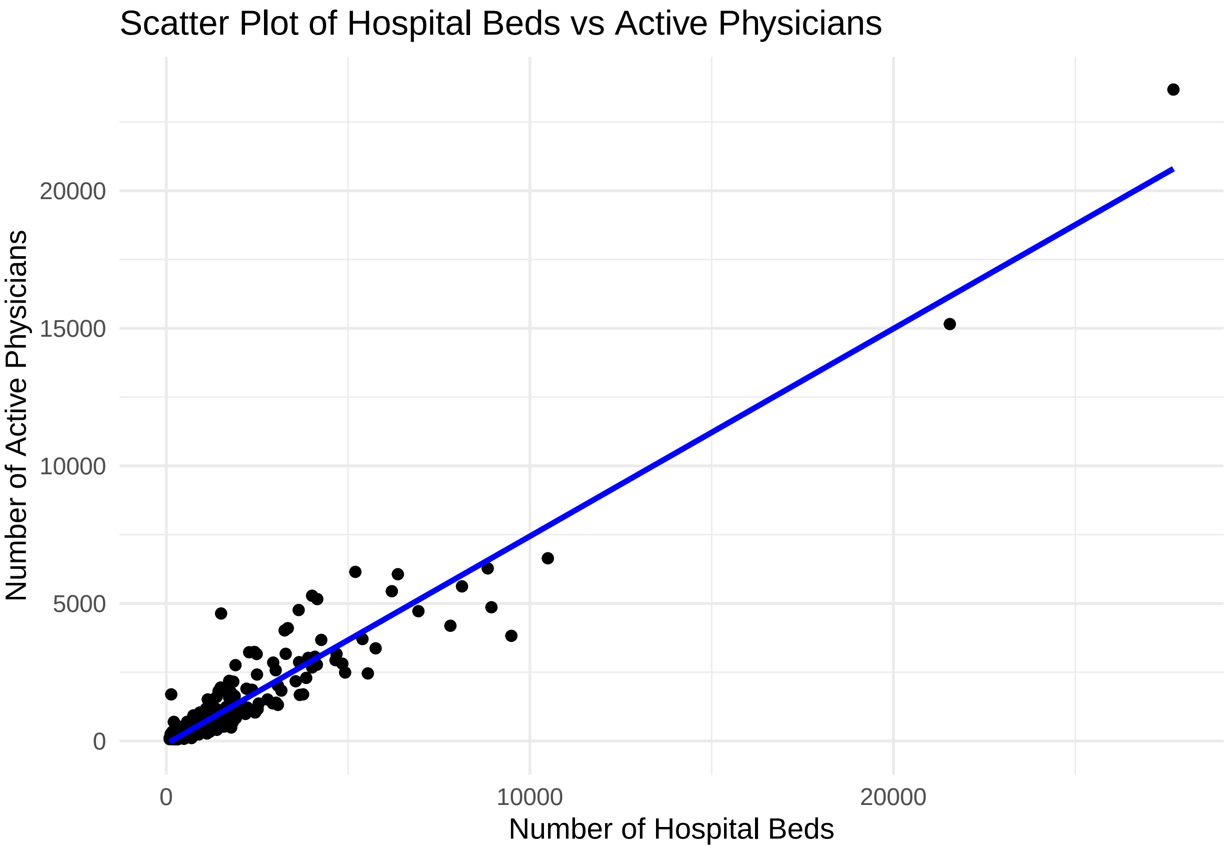
e-) Is the linearity assumption appropriate? Use graphs.

Linear regression has the following assumptions:
(1) The residuals (errors) are identically and independently distributed.
(2) The residuals (errors) are normally distributed, with mean 0 and equal variance (homoscedasticity).

When we plot the linear regression model, we eyeball a linear regression line that supports the linear regression nature of this relationship; however, we know this is not the way to validate the linearity assumption.

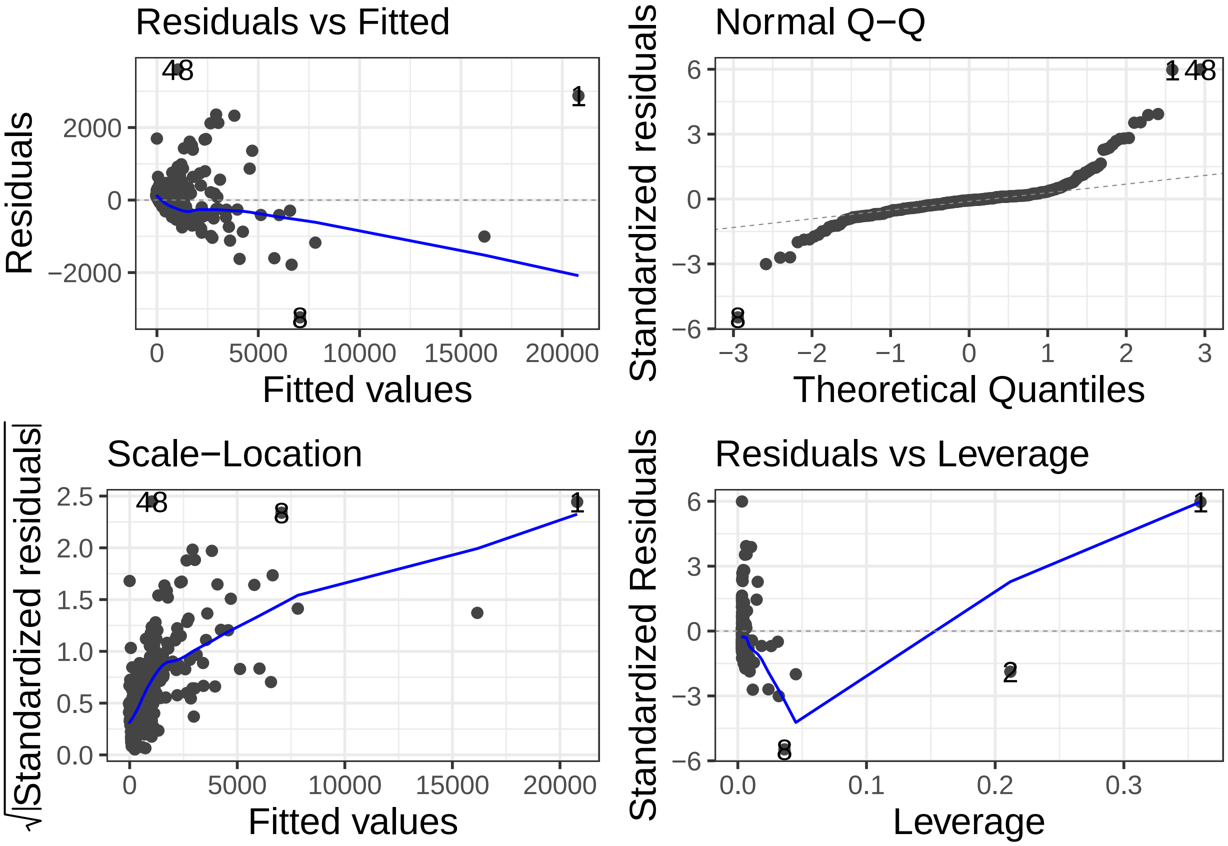
```r
ggplot(training_data, aes(x = Number.of.hospital.beds, y = Number.of.active.physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of Hospital Beds vs Active Physicians",
       x = "Number of Hospital Beds",
       y = "Number of Active Physicians") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



It is when we plot the residuals of our model that we can identify if they are normally distributed and assess if linearity assumptions of the model are met.

```r
autoplot(fit_cdi_data) + theme_bw() + theme(
    text = element_text(size = 12),
    axis.title = element_text(size = 14),
    legend.position = "bottom"
  )
```



The diagnostic plots above show residuals in four different ways:

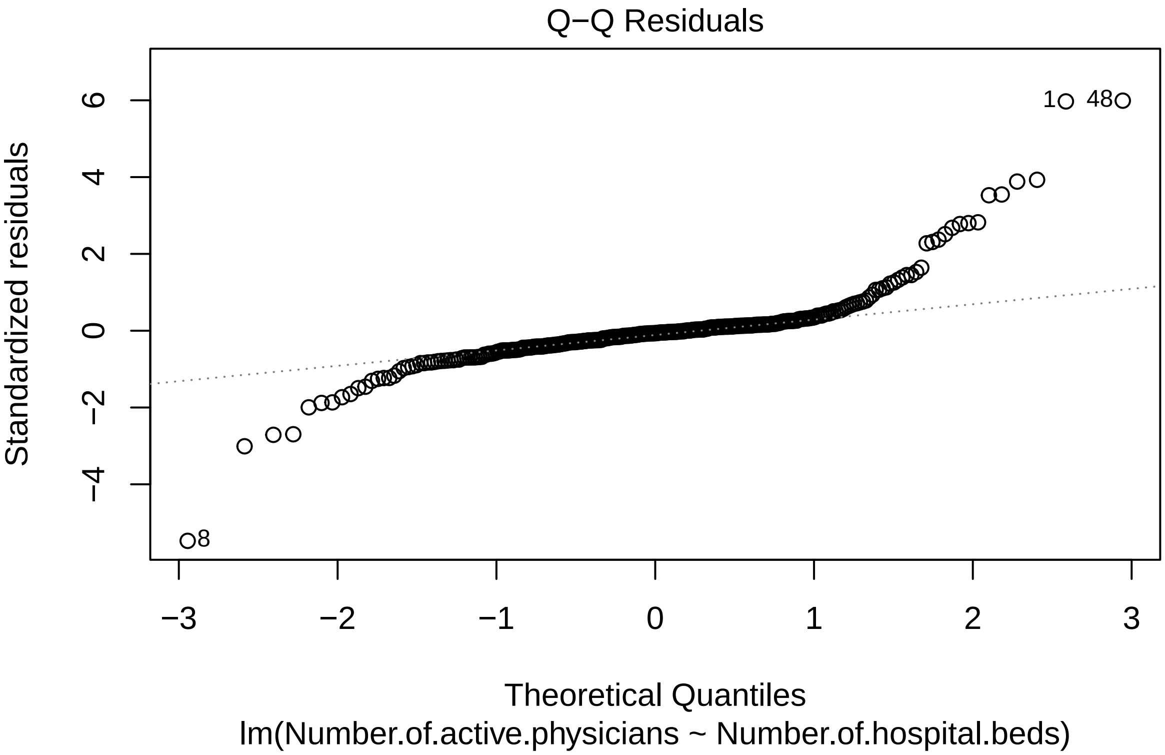
1. **Residuals vs Fitted** This plot checks whether the assumption of a linear relationship holds in the model. Ideally, residuals should be scattered randomly around zero, showing a balanced fit. In this case, while the residuals cluster tightly near zero at lower fitted values, the slight downward curve hints at a non-linear relationship. The fanning out of residuals as fitted values increase points to heteroscedasticity, meaning that error variance grows with higher fitted values. The downward trend suggests the model tends to overestimate the number of active physicians for higher fitted values and underestimate them for lower ones, indicating the model struggles to capture the relationship properly, likely due to non-linearity. This can lead to biased predictions, especially at the extremes.
2. **Normal QQ** This plot evaluates whether the residuals follow a normal distribution, with the ideal scenario being that all points align closely with the diagonal line. However, the heavy tails and pronounced potential outliers at both ends in this plot show a clear departure from normality. These extreme values indicate that the residuals deviate significantly from what's expected under normal distribution, suggesting the presence of outliers. This non-normality can skew the model's predictions and lead to less reliable parameter estimates, as it violates a core assumption of linear regression.
3. **Scale-Location** This plot helps assess whether the residuals have constant variance, a key assumption in linear regression. Ideally, the residuals should be scattered randomly around a horizontal line, but here, the upward curve suggests that the residual variance increases as the fitted values rise. This increasing variance violates the assumption of constant error variance, which can lead to inefficiencies in the model's predictions and affect its reliability for future data.
4. **Residuals vs Leverage** The plot is crucial for identifying influential data points that could disproportionately impact the model. Ideally, residuals should be randomly scattered, but here we see a checkmark-shaped curve with the bottom point of the "V" occuring at ≈ (0.05, -4.5) corresponding to the X and Y value of this graph, following by a exponential rise in standardized residuals (the long extension in the check-mark shape) as leverage increases. This pattern suggests that data points with higher leverage—those further from the center of the data—are exerting a stronger influence on the model's fit. The checkmark shape indicates that as leverage increases, residuals follow a predictable trend, highlighting potential outliers that may be

distorting the regression results. These high-leverage points could be pulling the model in their direction, making it less accurate overall.

Based on the analysis above of the graphs, we see that linearity assumptions are not met and we need to dive deeper into the data to see whether a linear relationship is the best means to explain the effect of Number of Hospital Beds on the Number of Active Physicians.

f-) Prepare a normal probability plot. Analyze your plot and summarize your findings.

```r
plot(fit_cdi_data, which=2)
```

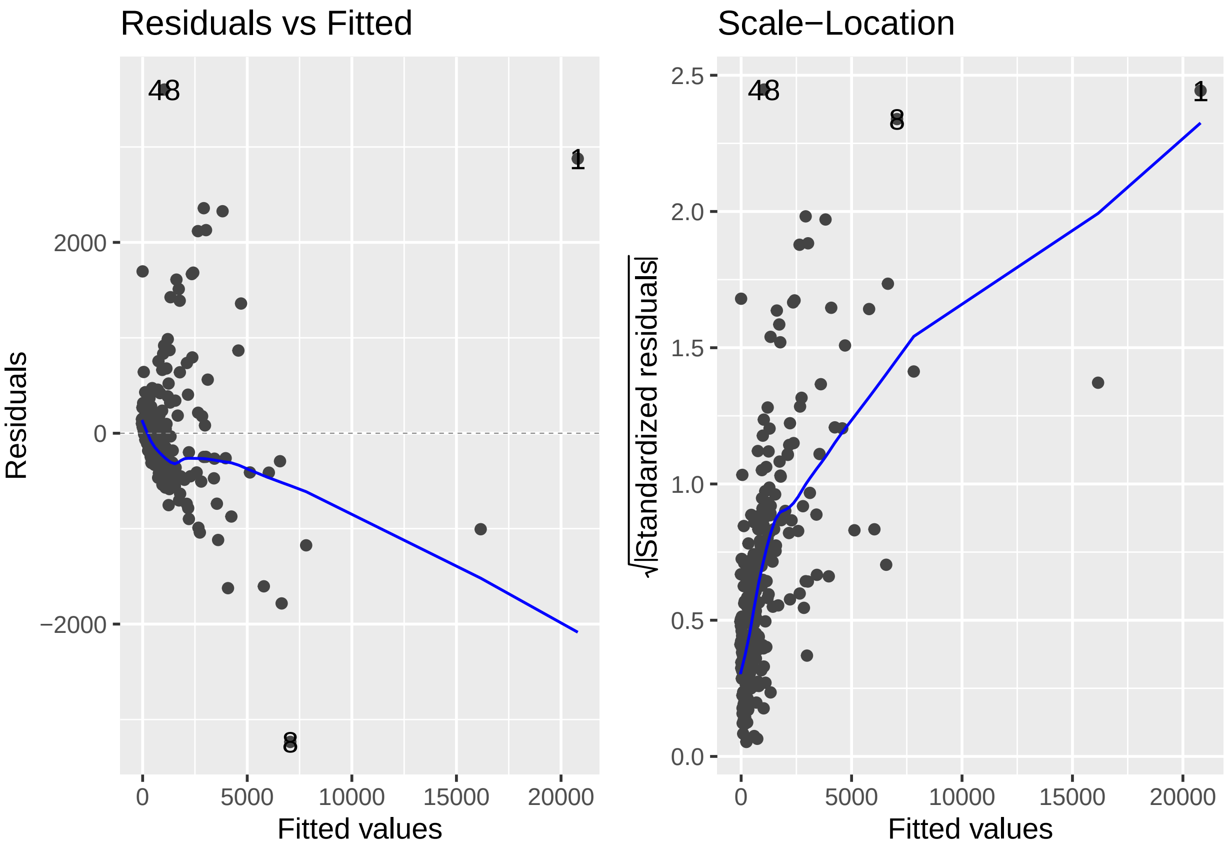


The dotted gray line represents the theoretical normal distribution, and our residuals seem to deviate, specifically in the tails, from this line, indicating the presence of outliers or non-normality (8, 1, 48). One of the key assumptions of linear regression is that the residuals should be normally distributed. Heavy tails imply that this assumption may be violated, which can impact the reliability of hypothesis tests and confidence intervals associated with the regression coefficients.

g-) Are the error variances constant? Use graphs. Perform Brown-Forsythe Test, write down Null and Alternative Hypotheses and conclusion of the test?

The error variances are not constant. Before we dive into the test, we leverage the *Residuals vs Fitted Plot* and *Scale-Location Plot* to visually depict this.

```r
# Use autoplot() to display the "Residuals vs Fitted" and "Scale-Location" plots
autoplot(fit_cdi_data, which = c(1, 3))
```



As mentioned before, the ‘Residuals vs Fitted’ plot depicts the residuals as densely packed near zero at low fitted values, with a subtle downward curvature suggesting a potential non-linear relationship. The ‘Scale-Location’ plot depicts an upward curve indicating an increasing variance with higher fitted values, reinforcing the presence of heteroscedasticity and violating the assumption of constant error variance. We can leverage the Brown-Forsythe Test to identify which data points are producing potential outliers.

```
# Extract residuals
ei <- residuals(fit_cdi_data)

# Create data frame
bf_data <- data.frame(
  Number.of.active.physicians = training_data$Number.of.active.physicians,
  Number.of.hospital.beds = training_data$Number.of.hospital.beds,
  ei = ei)

# Calculate the median of hospital beds
median_hospital_beds <- median(training_data$Number.of.hospital.beds)

# Create hospital_beds_group variable based on median
bf_data$hospital_beds_group <- as.factor(ifelse(bf_data$Number.of.hospital.beds < median_hospital_beds,

# Perform Levene's test using the residuals and hospital_beds_group based on the median
leveneTest(ei ~ hospital_beds_group, data = bf_data, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  48.591 1.935e-11 ***
##       307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, we state the hypotheses:
$H_o \ : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ The error variances are constant (homoscedasticity).

$H_a : \sigma_i^2 \neq \sigma_j^2$ The error variances are not constant (heteroscedasticity), where at least one group variance is not equal.

The decision rule is as follows:

1. If $p - value \geq \alpha$, then we fail to reject H_0, concluding the error variances are constant.

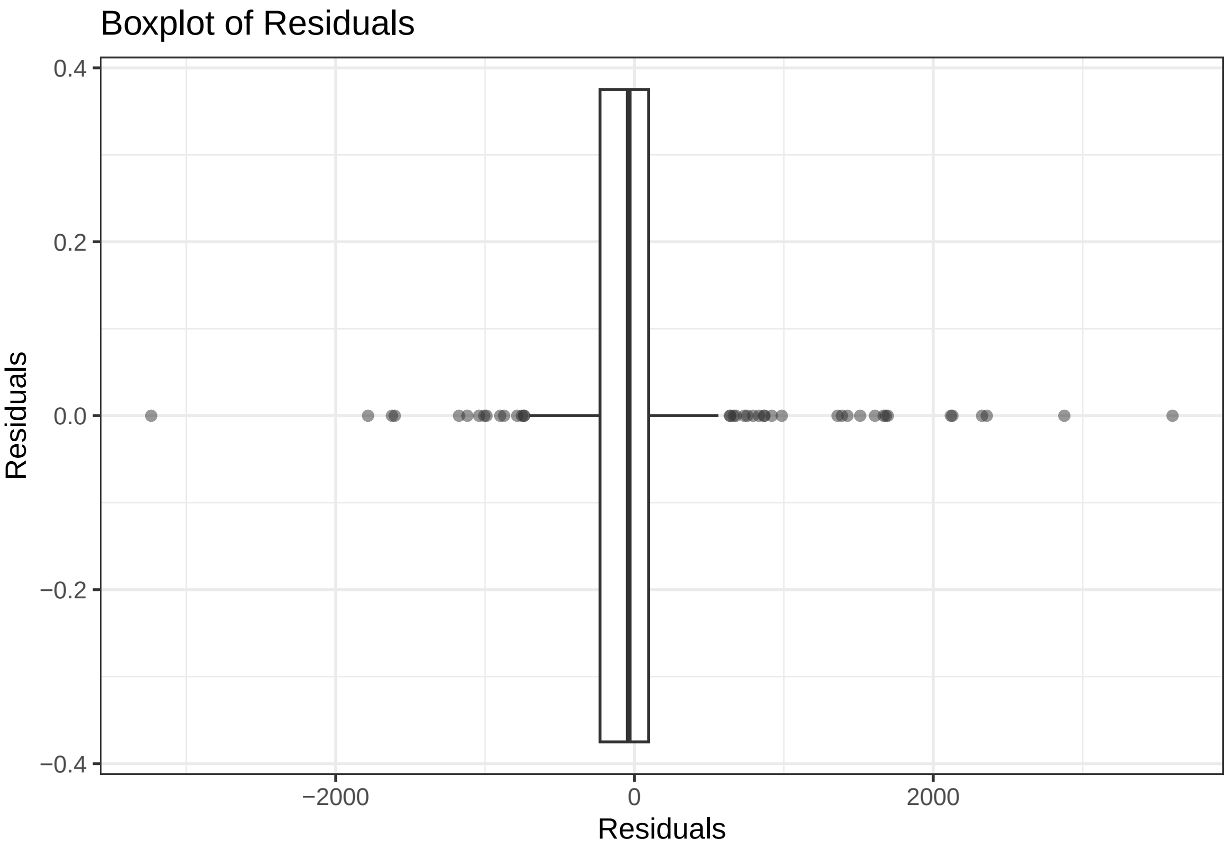
2. If $p - value < \alpha$, then we reject H_0 and conclude H_a, concluding the error variances are not constant.

The Brown-Forsythe test is a useful tool for assessing whether different groups have equal variances, also known as homogeneity of variances. One of its strengths is that it remains reliable even when the data doesn't follow a normal distribution. In our analysis, we obtained an F statistic of 48.591 and a p-value of 1.935e-11. Because the p-value is well below the typical significance level of 0.05, we are led to reject H_0 and conclude that the error variances are not constant, indicating the presence of heteroscedasticity in our data. In other words, the variability of the residuals changes across different levels of the fitted values, rather than remaining consistent.

h-) Are there any outliers in the dataset based on the errors?

Until now, we used Box Plots to visualize outliers; however, we know that residuals provide vast insight into the performance of a regression model, and into outliers as well. To depict the potential of outliers, we will calculate and leverage Standardized Residuals.

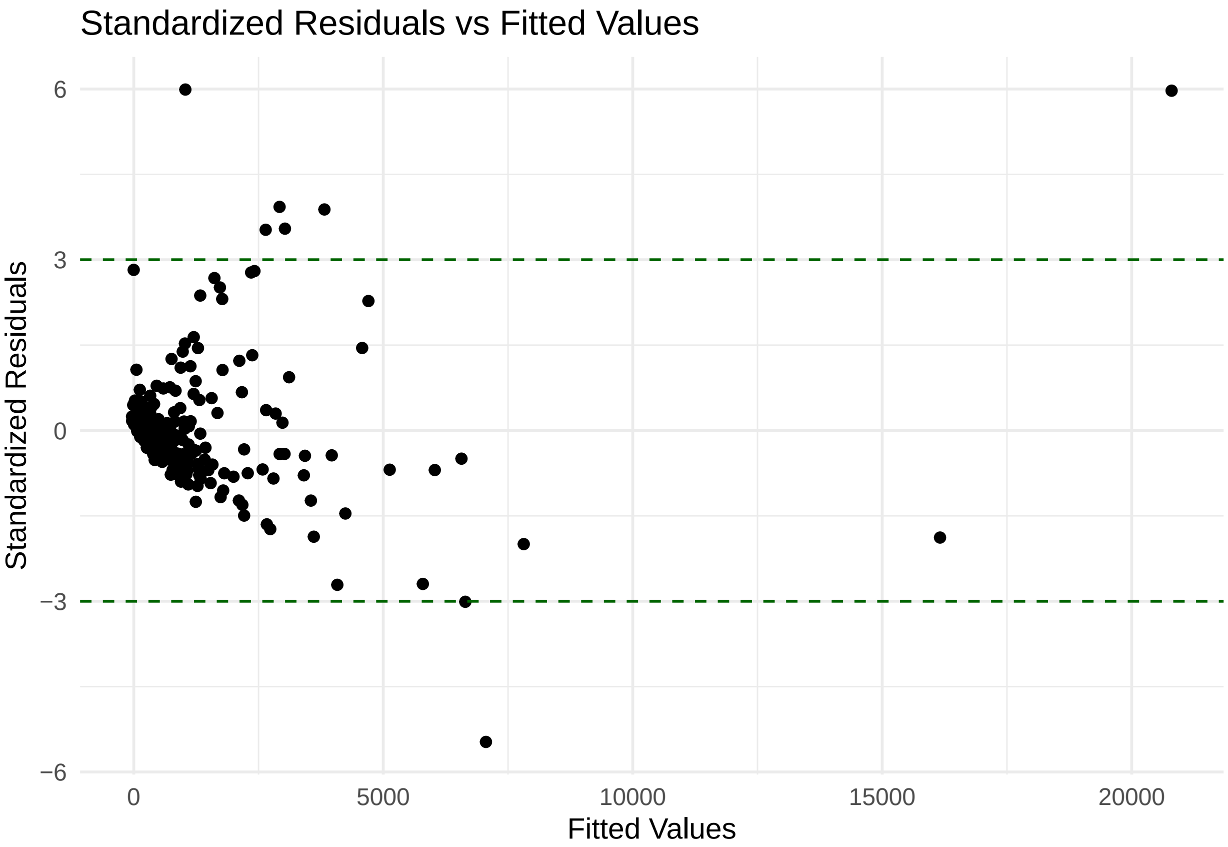
```r
training_data$Residuals <- residuals(fit_cdi_data)
training_data$Standardized_Residuals <- rstandard(fit_cdi_data)

# Boxplot of Residuals
ggplot(training_data, aes(x = Residuals)) +
  geom_boxplot(outlier.alpha = 0.5) +
  theme_bw() +
  labs(title = "Boxplot of Residuals", y = "Residuals")
```



```
# Option 1: Identify Outliers Using Standardized Residuals
std_residuals <- rstandard(fit_cdi_data)
outliers <- training_data %>% filter(abs(Standardized_Residuals) > 3)
print(outliers)
```

```
##   Identification.number        County State Land.area Total.population
## 1                     1   Los_Angeles    CA      4060          8863164
## 2                     6         Kings    NY        71          2300664
## 3                     8         Wayne    MI       614          2111687
## 4                    12          King    WA      2126          1507319
## 5                    16     Middlesex    MA       824          1398468
## 6                    19        Nassau    NY       287          1287348
## 7                    48    Montgomery    MD       495           757027
## 8                    53 San_Francisco    CA        47           723959
##   Percent.of.population.aged.18.34 Percent.of.population.65.or.older
## 1                             32.1                               9.7
## 2                             28.3                              12.4
## 3                             27.4                              12.5
## 4                             30.1                              11.1
## 5                             31.7                              12.5
## 6                             25.7                              14.2
## 7                             28.6                              10.2
## 8                             32.2                              14.5
##   Number.of.active.physicians Number.of.hospital.beds Total.serious.crimes
## 1                       23677                   27700               688936
## 2                        4861                    8942               680966
## 3                        3823                    9490               193978
## 4                        5280                    4009               124959
## 5                        5158                    4152                35825
## 6                        6147                    5200                43203
## 7                        4635                    1507                34754
## 8                        4761                    3640                71234
##   Percent.high.school.graduates Percent.bachelor.s.degrees
## 1                          70.0                       22.3
## 2                          63.7                       16.6
## 3                          70.0                       13.7
## 4                          88.2                       32.8
## 5                          84.3                       35.4
## 6                          84.2                       30.0
## 7                          90.6                       49.9
## 8                          78.0                       35.0
##   Percent.below.poverty.level Percent.unemployment Per.capita.income
## 1                        11.6                  8.0             20786
## 2                        19.5                  9.5             16803
## 3                        16.9                 10.0             17461
## 4                         5.0                  4.6             23779
## 5                         4.2                  7.3             25312
## 6                         2.5                  5.1             31679
## 7                         2.7                  3.3             30081
## 8                         9.7                  5.6             28532
##   Total.personal.income Geographic.region Residuals Standardized_Residuals
## 1                184230                 4  2877.210               5.970354
## 2                 38658                 1 -1783.322              -3.009757
## 3                 36872                 2 -3234.863              -5.471963
## 4                 35843                 4  2358.299               3.928939
## 5                 35398                 1  2128.386               3.546582
## 6                 40782                 1  2326.527               3.883474
## 7                 22772                 3  3601.399               5.990269
## 8                 20656                 4  2117.760               3.526632
```

```
ggplot(data = data.frame(Fitted = fit_cdi_data$fitted.values, Std_Residuals = std_residuals),
       aes(x = Fitted, y = Std_Residuals)) +
  geom_point() +
  geom_hline(yintercept = 3, linetype = "dashed", color = "darkgreen") +
  geom_hline(yintercept = -3, linetype = "dashed", color = "darkgreen") +
  labs(title = "Standardized Residuals vs Fitted Values",
       y = "Standardized Residuals",
       x = "Fitted Values") +
  theme_minimal()
```



Standardized residuals are the residuals of a regression model scaled by their estimated standard deviation. A common rule of thumb is that standardized residuals greater than 3 or less than -3 indicate potential outliers. This threshold corresponds to approximately 99.7% of values in a normal distribution (due to the empirical rule). Residuals falling outside this range suggest that the observation is significantly different from what the model predicts. Outliers can have a large influence on the regression results, potentially skewing estimates of coefficients and predictions. By examining standardized residuals, we can identify which data points may be exerting undue influence on the regression model. We see from the dataset output and the visual depiction of the graph that there are 8 potential outliers (Identification.number(s) 1, 6, 8, 12, 16, 19, 48, 53) identified by this analysis.

Problem 2

Repeat every item included in the first question by using total personal income as the independent variable (X) instead of the total number of hospital beds. Again the number of active physicians in a CDI is the dependent variable (Y) (Total 40 points).

b-) Test whether there is linear association between number of active physicians and total personal income. Using a t test with alpha=0.05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

First, we state the hypotheses:

$H_o\ : \beta_1 = 0$ There is no linear association between personal income and number of active physicians.

$H_a : \beta_1 \neq 0$ There is a linear association between personal income and number of active physicians.

The decision rule is as follows:

1. If $|t^*| \leq t_{\alpha/2,n-2}$, then we fail to reject H_0, concluding that there is no evidence of a linear association between personal income and number of active physicians.

2. If $|t^*| > t_{\alpha/2,n-2}$, then we reject the null hypothesis and conclude H_a, which indicates that there is evidence of a linear association between personal income and number of active physicians.

```
# Fit regression model of CDI training data
fit_cdi_tpi <- lm(Number.of.active.physicians ~ Total.personal.income, data=training_data, na.action = n

# Extract slope coefficient and its standard error
beta1_hat <- coef(summary(fit_cdi_tpi))["Total.personal.income", "Estimate"]  # Slope coefficient
se_beta1 <- coef(summary(fit_cdi_tpi))["Total.personal.income", "Std. Error"] # Standard error of the s

# Compute the t-statistic for the slope
t_stat_cdi_tpi <- beta1_hat / se_beta1

# Determine df for the t-test
df_cdi <- nrow(training_data) - 2

# Find the critical t-value at alpha = 0.01
alpha <- 0.05
crit_val_cdi_tpi <- qt(1 - alpha/2, df_cdi)

# Calculate p-value
p_val_cdi_tpi <- 2 * pt(-abs(t_stat_cdi_tpi), df_cdi)

# Output
cat("Test Statistic (t*):", t_stat_cdi_tpi, "\n")
```

```
## Test Statistic (t*): 56.555
```

```
cat("Critical t-value (at alpha = 0.05):", crit_val_cdi_tpi, "\n")
```

```
## Critical t-value (at alpha = 0.05): 1.967721
```

```
cat("p-value:", p_val_cdi_tpi, "\n")
```

```
## p-value: 2.162765e-164
```

Given $|t^*| > t_{\alpha/2,n-2}$ where $56.5550015 > 1.9677213$ with a significant p-value $2.1627646 \times 10^{-164} < \alpha = 0.05$, we can reject the null hypothesis and conclude that there is evidence of a linear association between personal income and number of active physicians.

c-) Set up the ANOVA table for the regression models for the independent variable. How much percent of the variation is explained by the independent variable?

```
# ANOVA
anova_cdi_tpi <- anova(fit_cdi_tpi)
print(anova_cdi_tpi)
```

```
## Analysis of Variance Table
##
## Response: Number.of.active.physicians
##                        Df     Sum Sq    Mean Sq F value    Pr(>F)
## Total.personal.income   1 1100850560 1100850560  3198.5 < 2.2e-16 ***
## Residuals             307  105663431     344181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Multiple_R_squared_TPI<-summary(fit_cdi_tpi)$r.square
print(Multiple_R_squared_TPI)
```

```
## [1] 0.9124225
```

$\approx 91.2\%$ of the variability in Number of Active Physicians is explained by personal income.

d-) Use Generalized linear test approach to test the significance of the independent variable.

```r
# Reduced Model
fit_redu_cdi_data <- lm(Number.of.active.physicians ~ 1, data=training_data)

glt_tpi_tst <- anova(fit_redu_cdi_data,fit_cdi_tpi)
print(glt_tpi_tst)
```

```
## Analysis of Variance Table
##
## Model 1: Number.of.active.physicians ~ 1
## Model 2: Number.of.active.physicians ~ Total.personal.income
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    308 1206513991
## 2    307  105663431  1 1100850560 3198.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Multiple_R_squared_TPI<-summary(fit_cdi_tpi)$r.square
print(Multiple_R_squared_TPI)
```

```
## [1] 0.9124225
```

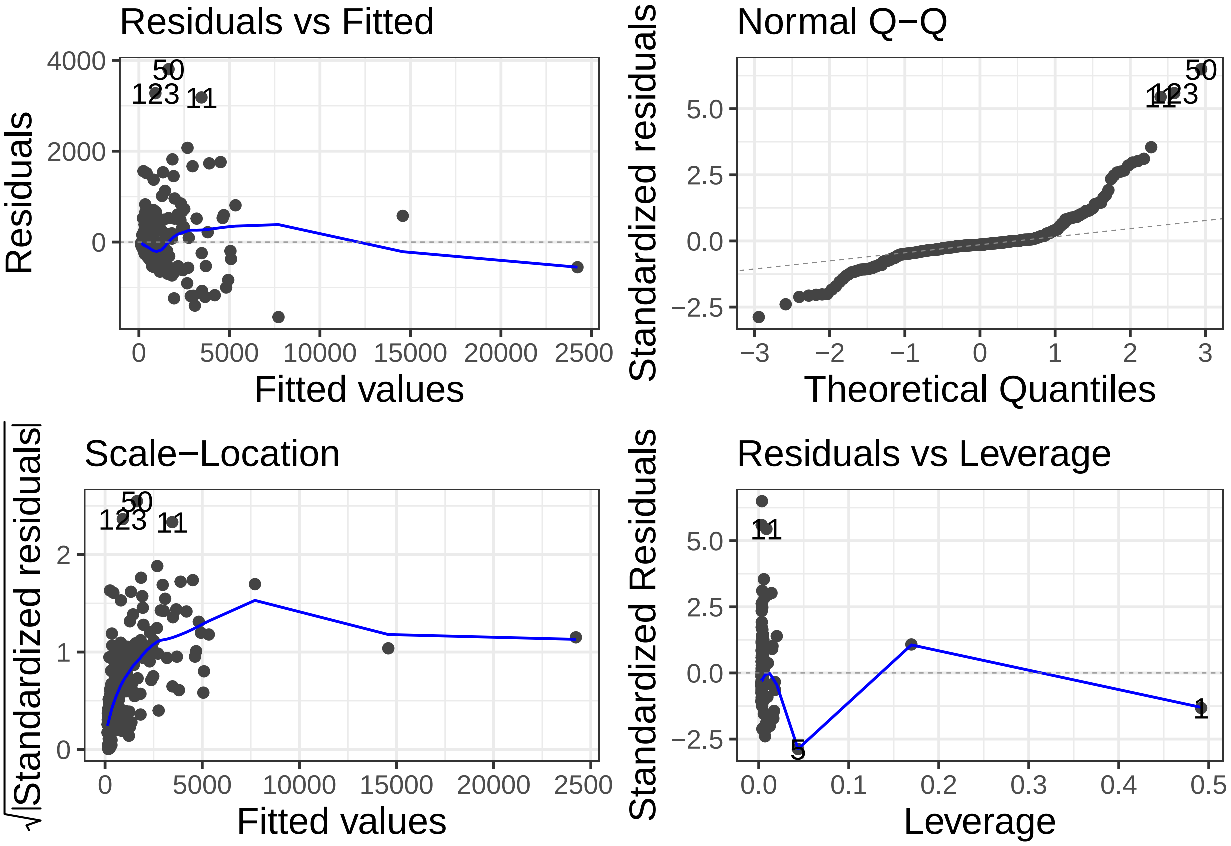
When no predictor variable is used (Reduced Model), the total variability is 1206513991. However, adding Total Personal Income as a predictor (Full Model) reduces the unexplained variability to 105663431. The difference in Sum of Squares between the Reduced Model (Model 1) and the Independent Variable Model (Model 2) is 1100850560, amounting to the full model explaining $\approx 91\%$ of the variability in Number of Active Physicians. The F^*-statistic of 3198.5, with a p-value of less than 2.2e-16 (less than $\alpha = 0.05$), shows strong evidence that Total Personal Income has a significant effect on the number of active physicians.

e-) Is the linearity assumption appropriate? Use graphs.

Linear regression has the following assumptions:
(1) The residuals (errors) are identically and independent distributed
(2) The residuals (errors) are normally distributed, with mean 0 and equal variance (homoscedasticity).

When we plot the residuals of our model, we can identify if they are normally distributed and assess if linearity assumptions of the model are met.

```r
library(ggfortify)
autoplot(fit_cdi_tpi) + theme_bw() + theme(
    text = element_text(size = 12),
    axis.title = element_text(size = 14),
    legend.position = "bottom"
  )
```



The diagnostic plots above show residuals in four different ways:

1. **Residuals vs Fitted** At first glance, the residuals seem randomly scattered around the zero line, which generally indicates a decent fit. However, there's a subtle but noticeable pattern: the residuals show an upward curve before dipping downward as fitted values increase. This suggests potential non-linearity in the model. Additionally, the residuals start to fan out at higher fitted values, hinting at heteroscedasticity—meaning the variance of the residuals isn't constant across the range of predicted values. This could imply the model struggles to maintain consistent accuracy as the fitted values grow.

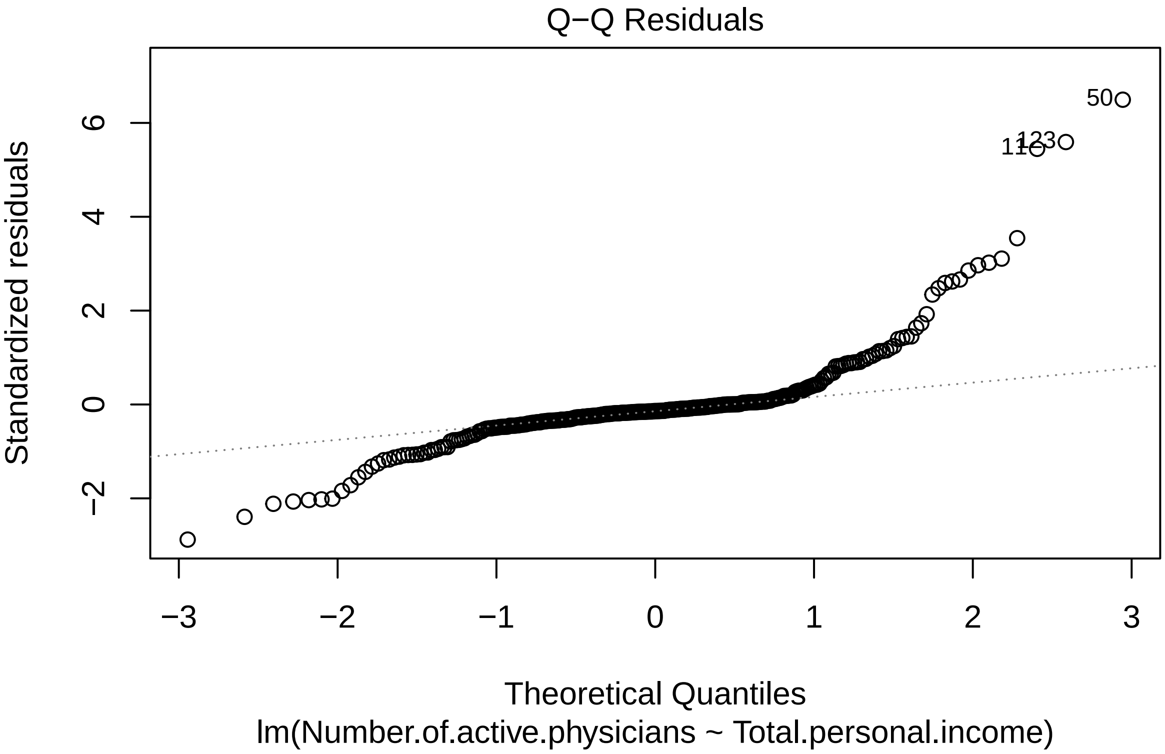
2. **Normal QQ** Although the majority of residuals cluster around the diagonal line, significant deviations emerge at both tails, indicating the presence of heavy tails. These deviations suggest the likelihood of outliers and point towards potential violations of the normality assumption in the residuals. Consequently, these extreme values require closer examination to validate the reliability of our results.

3. **Scale-Location** Ideally, we would expect to see a horizontal scatter of points, indicating a consistent spread. However, the plot reveals an upward trajectory starting from the origin, peaking around a fitted value of 7500, and then tapering off downwards as it approaches a fitted value just below 25000. This pattern suggests that the variability of the residuals increases with higher fitted values, reinforcing that the error variance is not constant across different levels of the predictor variable.

4. **Residuals vs Leverage** Initially, most points cluster around the horizontal line, indicating low leverage, but a few observations at the higher end stand out as potentially influential. The "check mark" shape of the plot reveals a pattern where residuals are more negative for lower-leverage points and more positive for those with higher leverage. Points such as 1, 5 and 11 are flagged due to their elevated leverage levels, warranting further investigation.

f-) Prepare a normal probability plot. Analyze your plot and summarize your findings.

```r
plot(fit_cdi_tpi, which=2)
```

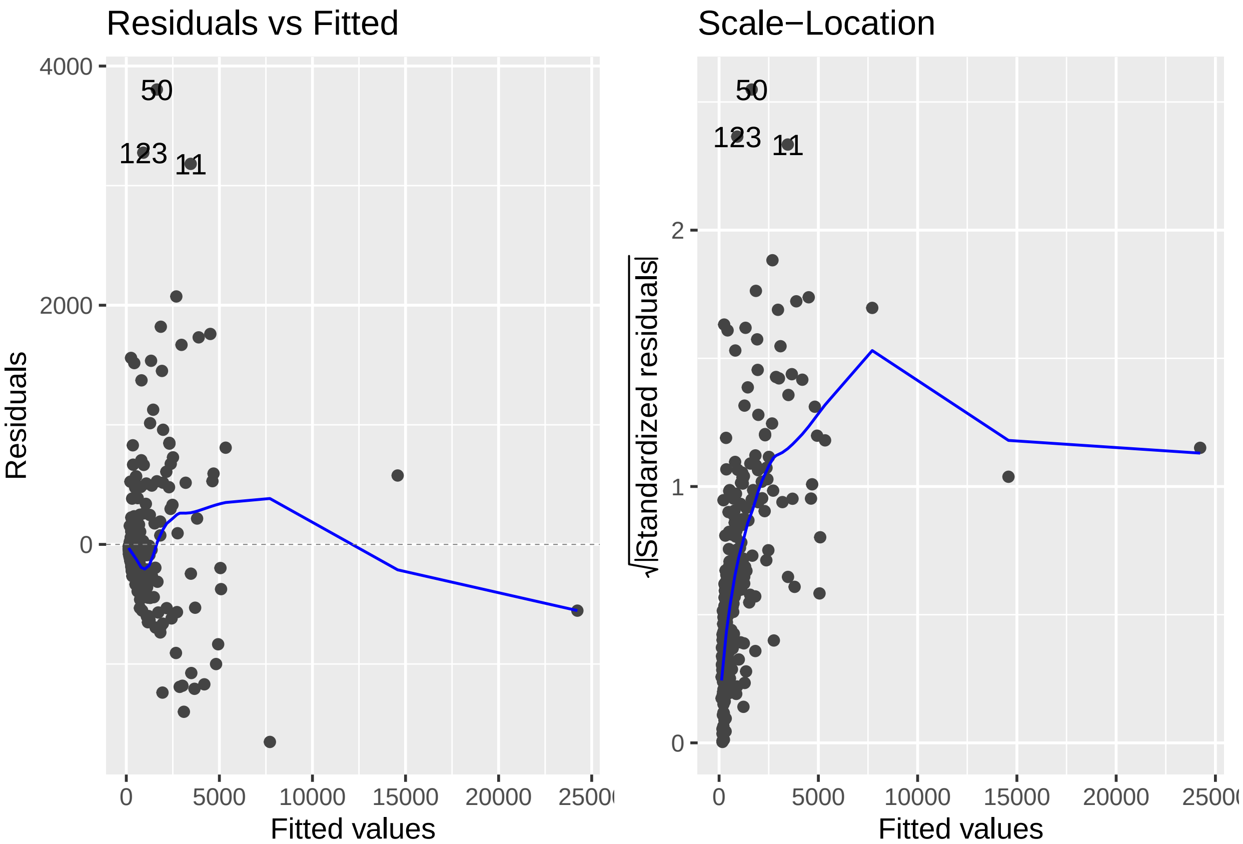


The dotted gray line represents the theoretical normal distribution, and our residuals seem to deviate, particulary in the tails, from this line, indicating the presence of outliers or non-normality. One of the key assumptions of linear regression is that the residuals should be normally distributed. Heavy tails imply that this assumption may be violated, which can impact the reliability of hypothesis tests and confidence intervals associated with the regression coefficients.

g-) Are the error variances constant? Use graphs. Perform Brown-Forsythe Test, write down Null and Alternative Hypotheses and conclusion of the test?

As analyzed above, the error variances are not constant. Before performing the Brown-Forsythe Test, we leverage the *Residuals vs Fitted Plot* and *Scale-Location Plot* to visually depict this.

```r
# Use autoplot() to display the "Residuals vs Fitted" and "Scale-Location" plots
autoplot(fit_cdi_tpi, which = c(1, 3))
```



As mentioned before, the 'Residuals vs Fitted' plot depicts the residuals as densely packed near zero at low fitted values, with a subtle curved, downward pattern suggesting a potential non-linear relationship. The'Scale-Location' plot depicts an upward curve indicating an increasing variance from the origin then tapering off as fitted values increase, reinforcing the presence of heteroscedasticity and violating the assumption of constant error variance. We can leverage the Brown-Forsythe Test to identify which data points are producing potential outliers.

```
# Extract residuals
ei <- residuals(fit_cdi_tpi)

# Create data frame
bf_data <- data.frame(
  Number.of.active.physicians = training_data$Number.of.active.physicians,
  Number.of.hospital.beds = training_data$Number.of.hospital.beds,
  ei = ei)

# Calculate the median of hospital beds
median_X <- median(training_data$Number.of.hospital.beds)

# Create group variable based on median
bf_data$group <- as.factor(ifelse(bf_data$Number.of.hospital.beds < median_X, 1, 2))

# Perform Levene's test using the residuals and group based on the median
leveneTest(ei ~ group, data = bf_data, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1   70.35 1.83e-15 ***
##       307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, we state the hypotheses:
$H_o \ : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ The error variances are constant (homoscedasticity).

$H_a : \sigma_i^2 \neq \sigma_j^2$ The error variances are not constant (heteroscedasticity), where at least one group variance is not equal.

The decision rule is as follows:

1. If $p - value \geq \alpha$, then we fail to reject H_0, concluding the error variances are constant.

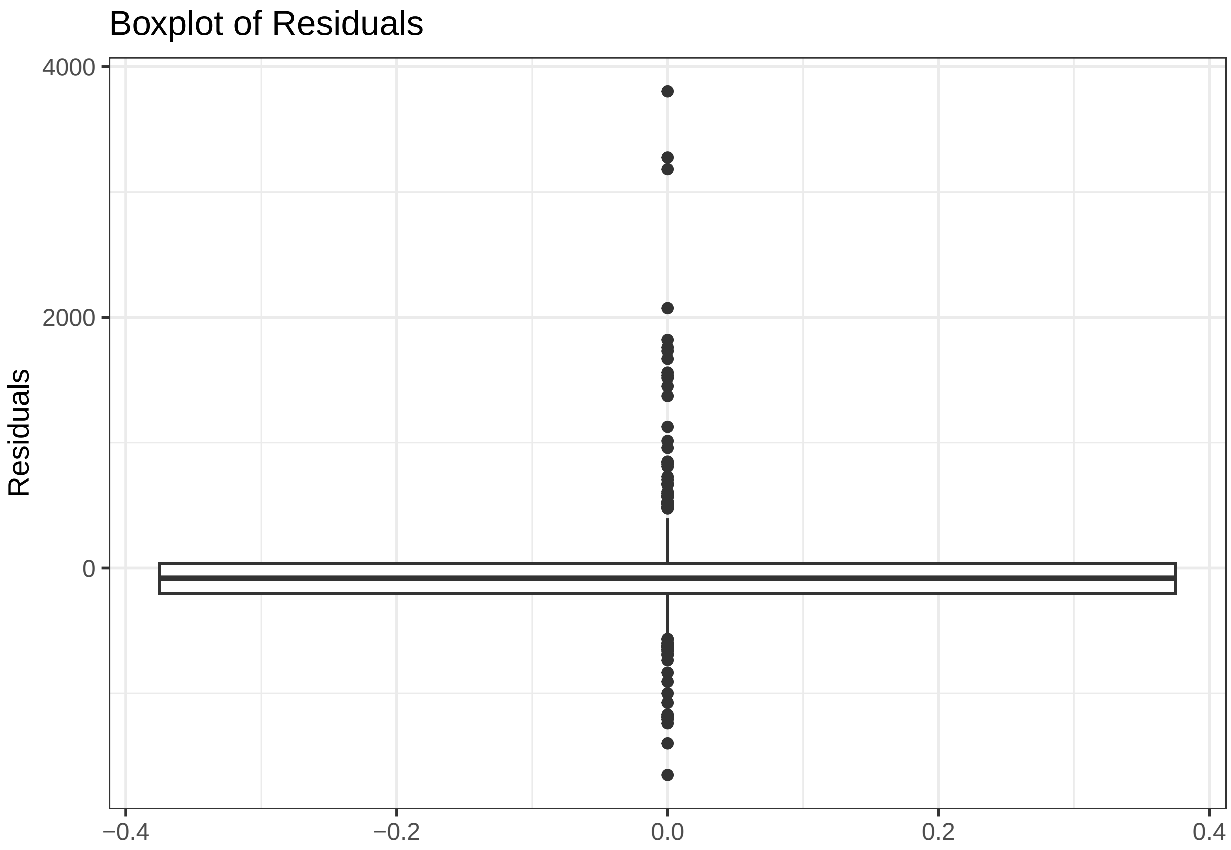
2. If $p - value < \alpha$, then we reject H_0 and conclude H_a, concluding the error variances are not constant.

The Brown-Forsythe test is a useful tool for assessing whether different groups have equal variances, also known as homogeneity of variances. One of its strengths is that it remains reliable even when the data doesn't follow a normal distribution. In our analysis, we obtained an F statistic of 70.35 and a p-value of 1.83e-15. Because the p-value is well below the typical significance level of 0.05, we are led to reject H_0 and conclude that the error variances are not constant, indicating the presence of heteroscedasticity in our data.

h-) Are there any outliers in the dataset based on the errors?

Until now, we used Box Plots to visualize outliers; however, we know that residuals provide vast insight into the performance of a regression model, and into outliers as well. To depict the potential of outliers, we will calculate and leverage Standardized Residuals.

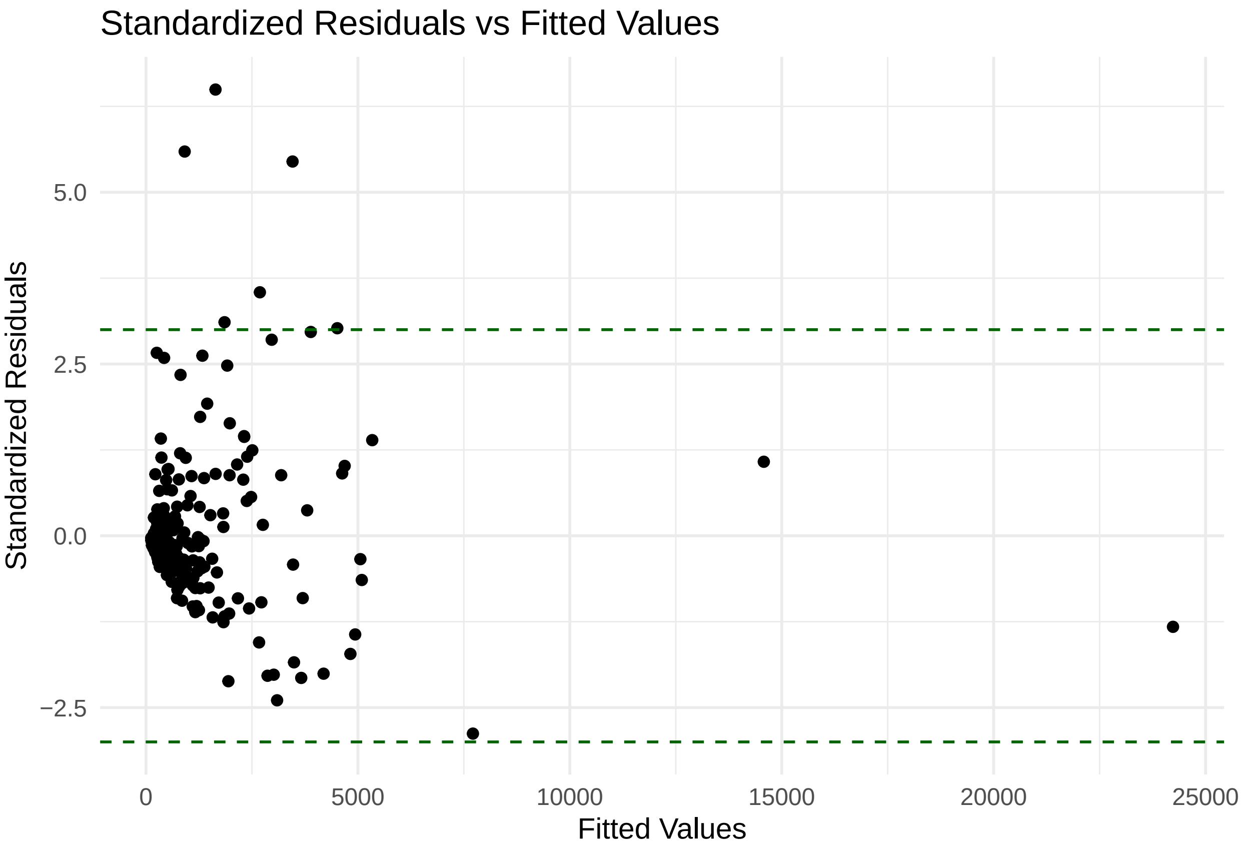
```r
training_data$Residuals <- residuals(fit_cdi_tpi)
training_data$Standardized_Residuals <- rstandard(fit_cdi_tpi)

# Boxplot of Residuals
ggplot(training_data, aes(y = Residuals)) +
  geom_boxplot() +
  theme_bw() +
  labs(title = "Boxplot of Residuals", y = "Residuals")
```



```
# Option 1: Identify Outliers Using Standardized Residuals
std_residuals <- rstandard(fit_cdi_tpi)
outliers <- training_data %>% filter(abs(Standardized_Residuals) > 3)
print(outliers)
```

```
##   Identification.number               County State Land.area Total.population
## 1                     9                 Dade    FL      1945          1937094
## 2                    11         Philadelphia    PA       135          1585577
## 3                    50       Baltimore_City    MD        81           736014
## 4                    53        San_Francisco    CA        47           723959
## 5                    73 District_of_Columbia    DC        61           606900
## 6                   123        St._Louis_City    MO        62           396685
##   Percent.of.population.aged.18.34 Percent.of.population.65.or.older
## 1                             27.1                              13.9
## 2                             29.1                              15.2
## 3                             30.0                              13.7
## 4                             32.2                              14.5
## 5                             33.6                              12.8
## 6                             28.7                              16.6
##   Number.of.active.physicians Number.of.hospital.beds Total.serious.crimes
## 1                        6274                    8840               244725
## 2                        6641                   10494               109148
## 3                        5444                    6203                87355
## 4                        4761                    3640                71234
## 5                        3674                    4262                64393
## 6                        4189                    7814                64103
##   Percent.high.school.graduates Percent.bachelor.s.degrees
## 1                          65.0                       18.8
## 2                          64.3                       15.2
## 3                          60.7                       15.5
## 4                          78.0                       35.0
## 5                          73.1                       33.3
## 6                          62.8                       15.3
##   Percent.below.poverty.level Percent.unemployment Per.capita.income
## 1                        14.2                  8.7             17823
## 2                        16.1                  8.0             16721
## 3                        17.8                  9.4             17263
## 4                         9.7                  5.6             28532
## 5                        13.3                  7.7             23603
## 6                        20.6                  9.0             18113
##   Total.personal.income Geographic.region Residuals Standardized_Residuals
## 1                 34525                 3  1759.738               3.021016
## 2                 26512                 1  3182.086               5.447262
## 3                 12706                 3  3803.396               6.494600
## 4                 20656                 4  2073.347               3.544189
## 5                 14325                 3  1820.167               3.108503
## 6                  7185                 2  3275.536               5.592379
```

```r
ggplot(data = data.frame(Fitted = fit_cdi_tpi$fitted.values, Std_Residuals = std_residuals),
       aes(x = Fitted, y = Std_Residuals)) +
  geom_point() +
  geom_hline(yintercept = 3, linetype = "dashed", color = "darkgreen") +
  geom_hline(yintercept = -3, linetype = "dashed", color = "darkgreen") +
  labs(title = "Standardized Residuals vs Fitted Values",
       y = "Standardized Residuals",
       x = "Fitted Values") +
  theme_minimal()
```



Standardized residuals are the residuals of a regression model scaled by their estimated standard deviation. A common rule of thumb is that standardized residuals greater than 3 or less than -3 indicate potential outliers. This threshold corresponds to approximately 99.7% of values in a normal distribution (due to the empirical rule). Residuals falling outside this range suggest that the observation is significantly different from what the model predicts. Outliers can have a large influence on the regression results, potentially skewing estimates of coefficients and predictions. By examining standardized residuals, we can identify which data points may be exerting undue influence on the regression model. We see from the dataset output and the visual depiction of the graph that there are 6 potential outliers (Identification.number(s) 9, 11, 50, 53, 73, 123) identified by this analysis.

Problem 3

Use the test data set to compare the models performances in question I and II as well as considering your answers from part b-) to h-). Which model would you pick and why? (Total 20 points).

```
# Extract predicted error stats for Train Data for No of Active Physicians and Hospital Beds
PredictedTrainNHB<-predict(fit_cdi_data,training_data)
ModelTrainNHB<-data.frame(obs = training_data$Number.of.active.physicians, pred=PredictedTrainNHB)

# Extract predicted error stats for Test Data for No of Active Physicians and Hospital Beds
PredictedTestNHB <- predict(fit_cdi_data, newdata=testing_data)
ModelTestNHB <- data.frame(obs= testing_data$Number.of.active.physicians, pred=PredictedTestNHB)

rbind(defaultSummary(ModelTrainNHB), defaultSummary(ModelTestNHB))
```

```
##          RMSE  Rsquared      MAE
## [1,] 600.2315 0.9077293 337.3255
## [2,] 435.8014 0.8838889 274.1029
```

Model 1: Number of Hospital Beds on Active Physicians

When examining the performance of Model 1, which predicts the number of active physicians based on the number of hospital beds, we observe a relatively small decline in R-squared from 0.908 in the training dataset to 0.884 in the testing dataset. This minimal drop indicates that the model generalizes well to unseen data. A slight decrease in R-squared is expected, and the fact that it remains strong suggests the model is not significantly overfitting.

Interestingly, both the RMSE and MAE decrease when moving from the training to the test dataset, suggesting that the model performs slightly better on the test data. This consistency in performance metrics reinforces the robustness of Model 1.

```r
PredictedTrainTPI<-predict(fit_cdi_tpi,training_data)
ModelTrainTPI<-data.frame(obs = training_data$Number.of.active.physicians, pred=PredictedTrainTPI)

PredictedTestTPI <- predict(fit_cdi_tpi, newdata=testing_data)
ModelTestTPI <- data.frame(obs= testing_data$Number.of.active.physicians, pred=PredictedTestTPI)

rbind(defaultSummary(ModelTrainTPI), defaultSummary(ModelTestTPI))
```

```
##          RMSE  Rsquared      MAE
## [1,] 584.7673 0.9124225 331.2821
## [2,] 528.5132 0.8153974 294.5912
```

Model 2: Total Personal Income on Active Physicians

In the case of Model 2, which predicts the number of active physicians based on total personal income, we observe a slightly larger decline in R-squared, dropping from 0.912 in the training dataset to 0.815 in the testing dataset. While this still indicates that the model generalizes adequately, the decrease in R-squared is more pronounced than in Model 1, suggesting some loss of explained variance.

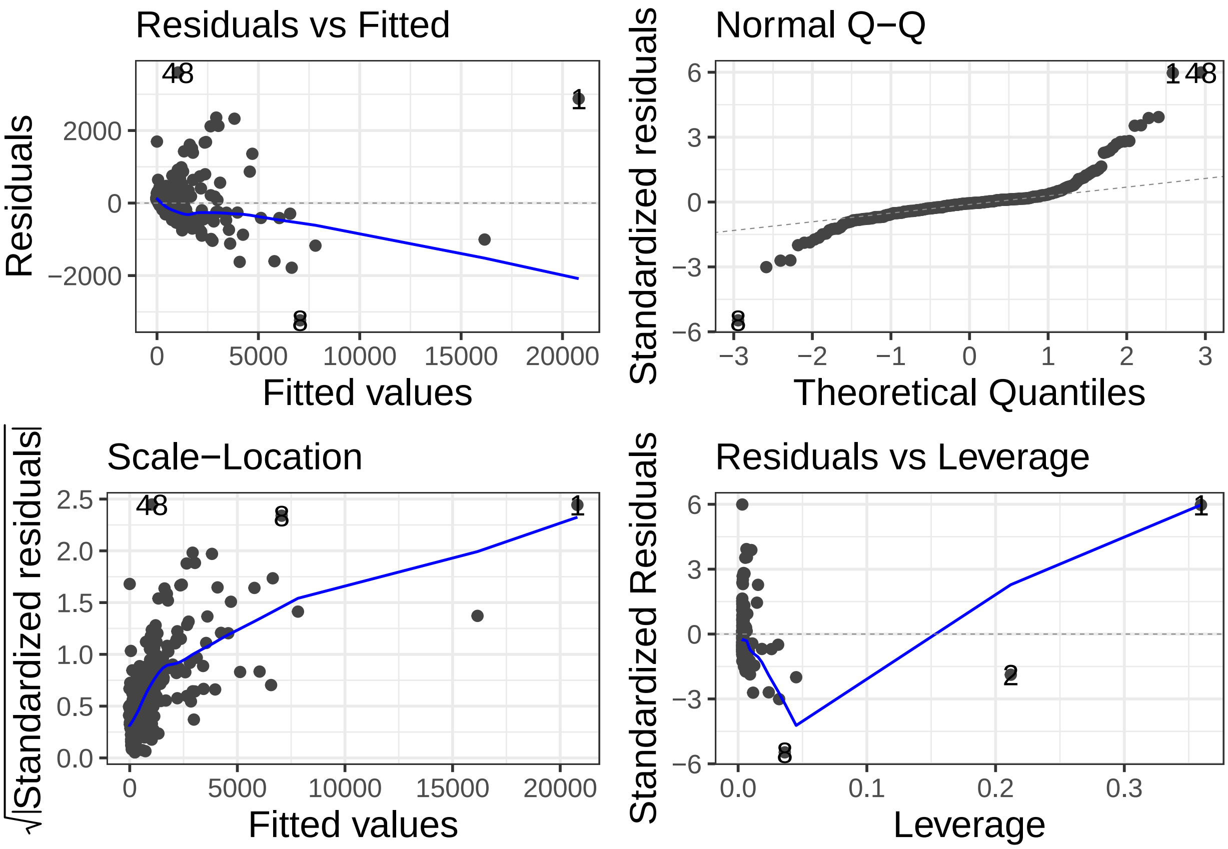
Despite this decline, the RMSE and MAE for the test dataset are lower than those for the training dataset, indicating that Model 2 performs relatively better on unseen data. This suggests good generalization; however, the substantial drop in R-squared raises concerns about potential overfitting, especially given the lower fit to the test dataset.

Comparative Analysis of Models

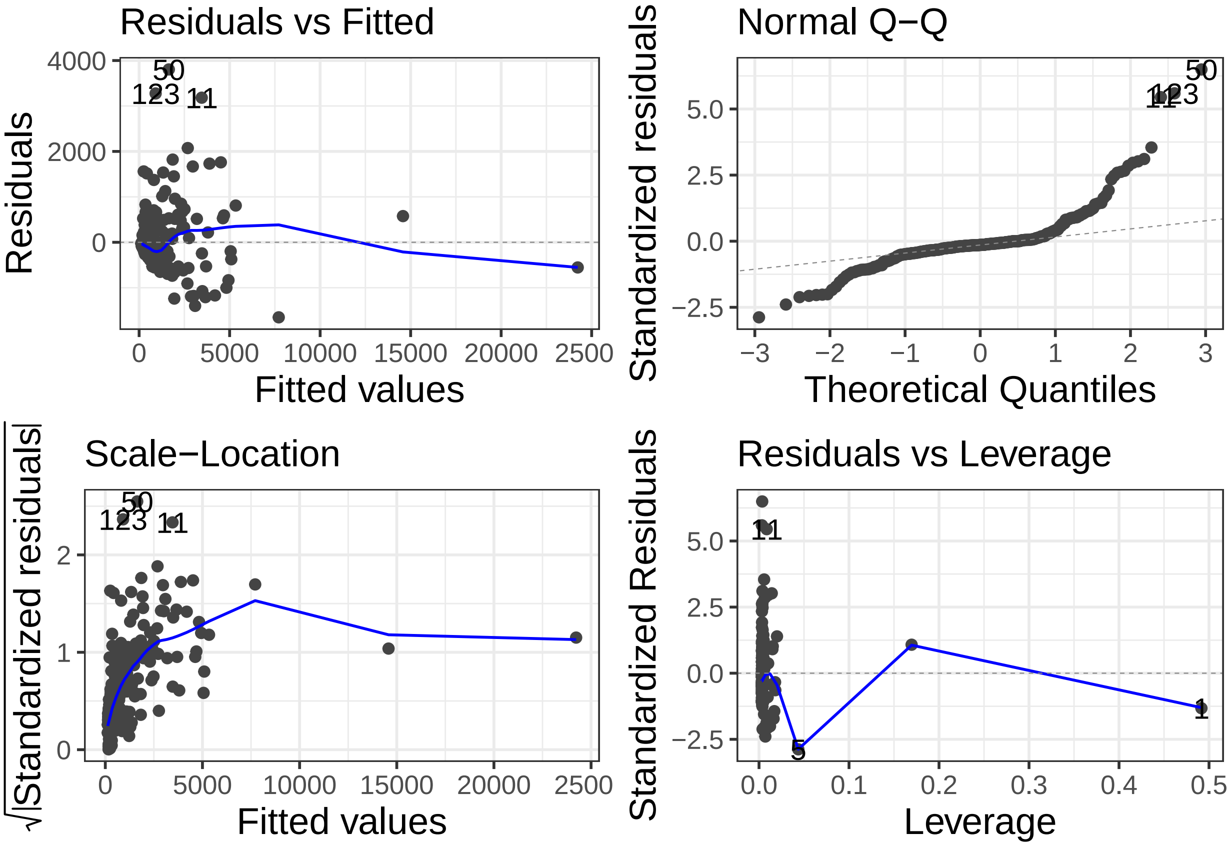
When comparing the performance of both models on their respective training and testing datasets, we find that Model 2 exhibits superior performance on the training data, with lower RMSE and MAE and a higher R-squared value. This suggests that Model 2 fits the training data more accurately. In contrast, Model 1 shows better performance on the test data, indicated by lower RMSE and MAE and a higher R-squared value, suggesting it generalizes more effectively to unseen data and may be less prone to overfitting than Model 2.

To further assess model performance, we will analyze the residual plots and identify any outliers. This will help us determine whether Model 2 is overfitting the training data or if Model 1's better generalization comes at the expense of influential outliers or increased heteroscedasticity. Having already plotted the residuals for both models, we can now compare their performances side by side.

```r
# Model 1 - NHB
autoplot(fit_cdi_data) + theme_bw() + theme(
    text = element_text(size = 12),
    axis.title = element_text(size = 14),
    legend.position = "bottom"
  )
```



```r
# Model 2 - TPI
autoplot(fit_cdi_tpi) +
  theme_bw() +
  theme(
    text = element_text(size = 12),
    axis.title = element_text(size = 14),
    legend.position = "bottom")
```



1. **Residuals vs. Fitted** In Model 1, the distribution of residuals appears somewhat random, but there is a notable increase in variation at higher fitted values. This suggests a potential non-linear relationship, indicated by a slight downward curvature in the residual plot. In contrast, Model 2 exhibits tighter clusters of residuals with less variation across the fitted values. However, a non-random pattern persists, with a trend of decreasing residuals as fitted values increase, further hinting at possible non-linearity.

2. **Normal Q-Q Plot** The Normal Q-Q Plot for Model 1 displays a significant deviation from the normal line at both ends, indicating that the residuals may not be normally distributed. This deviation suggests the model may struggle to capture extreme values or that the dataset contains outliers. Although Model 2 also shows some deviation at the extremes, the points align more closely with the normal line, particularly in the central portion of the distribution. This indicates that the residuals for Model 2 are better approximated to normality compared to Model 1.

3. **Scale Location Plot** In the Scale-Location Plot for Model 1, the residuals demonstrate an increasing spread with higher fitted values, indicative of heteroscedasticity—where the variance of residuals is not constant—thus violating regression assumptions. While the residual spread in Model 2 also increases with fitted values, the trend is less pronounced than in Model 1. Nonetheless, heteroscedasticity persists in both models.

4. **Residuals vs Leverage Plot** Model 1 exhibits several points with high leverage, which could significantly influence the model's results. These high-leverage points are positioned farther from the main cluster, suggesting that some data points exert considerable influence on the regression line. In Model 2, there are fewer points with high leverage compared to Model 1, though the influence of specific outliers (e.g., points 11, 5) remains apparent. Overall, Model 2 seems to manage influential points more effectively.

Conclusion

Model 1 displays more signs of non-linearity and heteroscedasticity, indicating that it may not be the best fit for the data. In contrast, Model 2 shows improvements in terms of residual normality, homoscedasticity, and leverage, despite still facing challenges with influential points and residual spread at higher fitted values.

When examining the outlier analysis for both models using the training data, Model 2 identified 6 outliers beyond the 3/-3 boundary according to standardized residuals, and 21 potential outliers according to Cook's Distance. In comparison, Model 1 identified 8 outliers beyond the 3/-3 boundary and 19 potential outliers based on Cook's Distance.

Considering the overall model performance, along with the residual plots, I would favor Model 2 due to its lower variance in residuals and higher R^2 value (91.2% compared to Model 1's 90.8%), as indicated by the General Linear Test results. Both models, however, exhibit significant inconsistencies in error variances, as confirmed by the Brown-Forsythe test, and are subject to the presence of outliers, as confirmed by the Standardized Residuals analysis above.