Prediction of the U.S. Population Growth

Eugene Monforte

May 8, 2025

Abstract:

Our goal for this project is determine the best model to predict the growth of the US population. Our data consisted of a variable Year (x) that spanned from 1790 to 1980 and was used to predict Population (Y) in millions. By taking in account World War 2 (WW2), we were able to create an indicator variable to add to our best fitting model. We determine our 3rd model was the best fit by way of AIC and MSE. The following model was a quadratic regression model with the WW2 indicator variable with the following equation:

$$\hat{Y} = 4.971 - .07447 \cdot x_1 + .006526 \cdot x_1^2 - 8.641 \cdot x_2$$

Where $\hat{Y}$ = **Population** is the predicted U.S. population (in millions),

x_1= **Year** is the **rescaled year** (e.g., 1790 = 0, 1800 = 10, ..., 1980 = 190),

And x_2 = indicator variable for **WW2** (1=WW2 took place, 0=WW2 did not take place).

This model predicted that in the year 2020, the population would be 333.084 million. Comparing this to the actual population in 2020, we found there was only a -.48% percent error.

Introduction:

In this project, we are tasked with finding the best-fitting model to predict the population, where the population is in millions. The dataset given contains two variables: Population (y) and Year (x), where Population is our dependent variable and Year is our independent variable. The variable Population ranges from 3.93 million to 226.55 million. The variable Year starts at 1790 and ends in 1980. We will rescale the Year variable to start at 0 and go up in increments of 10 for each Year after (example: 1790 =0, 1800 = 10, 1810 = 20, …etc.). We are also given the fact that World War 2 occurred during 1940 to 1950. This known fact would affect population growth, so should be taken into account when finding a Regression Model. We will attempt to find a best fitting model and then use it to predict the population in 2020.

Methods:

We will start by fitting a Simple Linear Regression (SLR) model. We then find descriptive statistics and assess the adequacy of the model using different metrics such as overall F-test for the model, t-test for regression coefficient(s), adjusted R^2, MSE and residual analysis.

The SLR model is used as a baseline model for the two (or possibly three) alternative models we will find. The alternative models should outperform the SLR model using Akaike Information Criterion (AIC) and Mean Square Error (MSE) metrics.

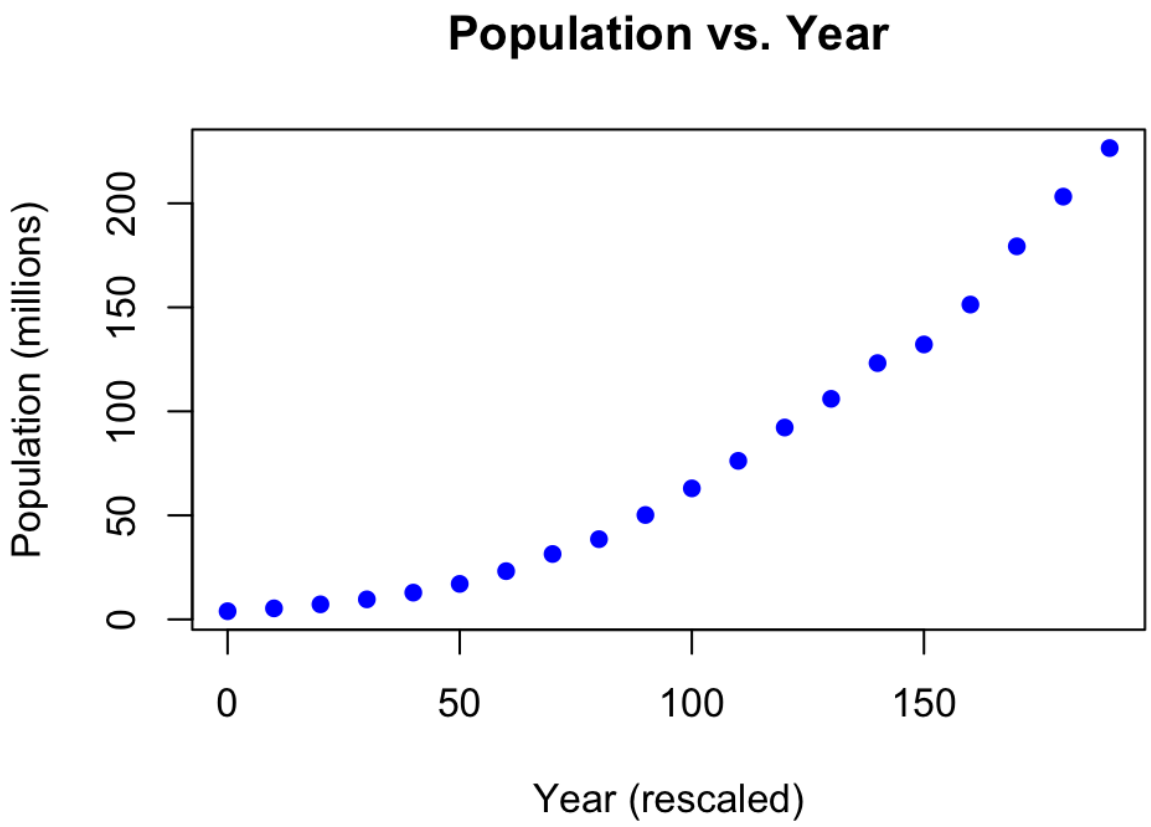
The "BEST" model selected at the end should satisfy the following criteria:

1. Assumptions of normality and homoscedasticity are satisfied
2. Given the time-dependent nature of this dataset, it is reasonable to expect autocorrelation in the error terms. Consequently, the assumption of independence will not be evaluated.
3. All regression coefficients are statistically significant at the 5% significance level
4. The model has the lowest AIC
5. The model has the smallest MSE (close to one, if possible)

In the end, we will use our "BEST" model to predict the population size for the year 2020.

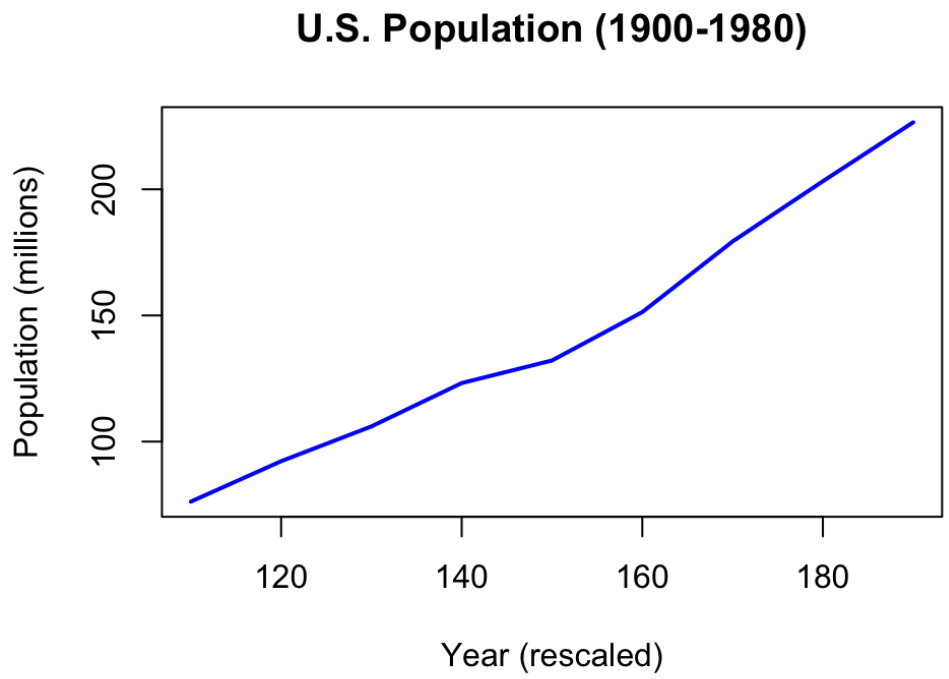
Preliminary Data Analysis:

Before fitting our first model, we seek to understand the data better. We fit a scatterplot of the data and get the following,



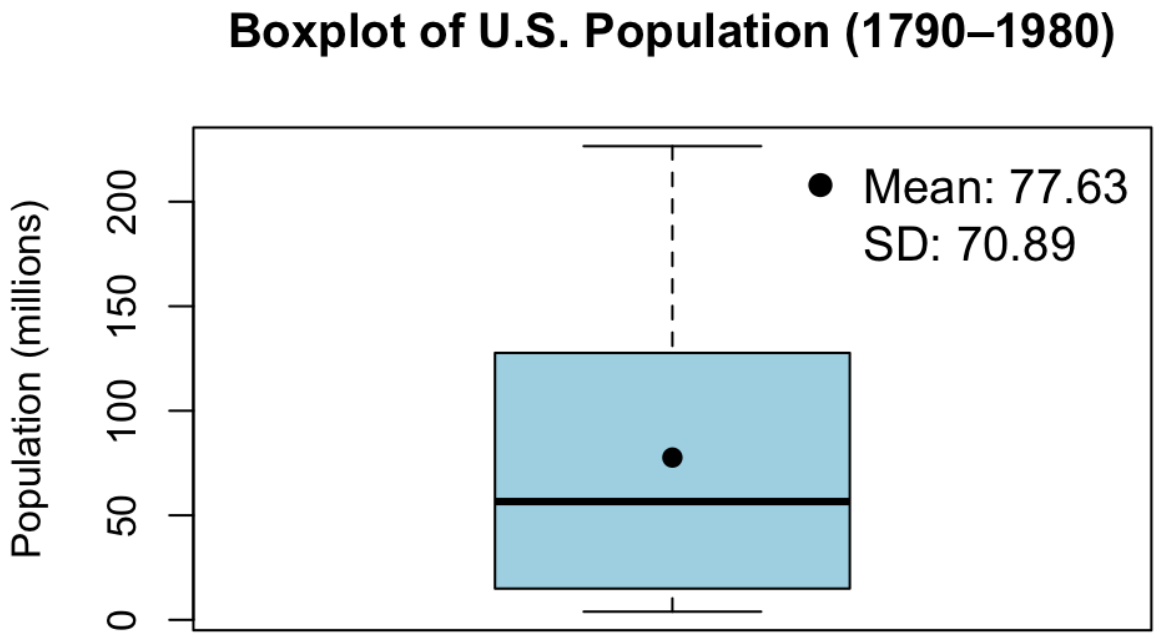
The scatterplot shows a nonlinear relationship between Year and Population. We notice that in the earlier years, population seems to grow a little slower, before really increasing as time increases.

We investigate the years where WW2 took place. To do this, we created a line graph and zoomed in on the years where WW2 took place, years 150 (1940) and 160 (1950) in our rescaled data.



We notice that there is indeed a change in the increasing population from 150 to 160. So, considering WW2 for later modeling could be important.

Next, we make a boxplot and find the summary statistics of our Population variable.



Summary Statistics for All Observations:

Min.	1st Qu.	Median	3rd Qu.	Max	Mean	SD
3.93	16.02	56.59	125.44	226.55	77.63	70.89

Analyzing the box plot, we see that although the data has a wide range of values, we do not have any outliers in our data. Although we see a strong growth pattern in our data as shown in our scatterplot, it is not statistically inconsistent. We also see that our data is right skewed, where we have small values for population for the lower years, but the few values from the later years bring the distribution upward. The wide IQR suggests that we have a large amount of variability, especially in the upper half of the boxplot.

Results:

Model 1: Simple Linear Regression

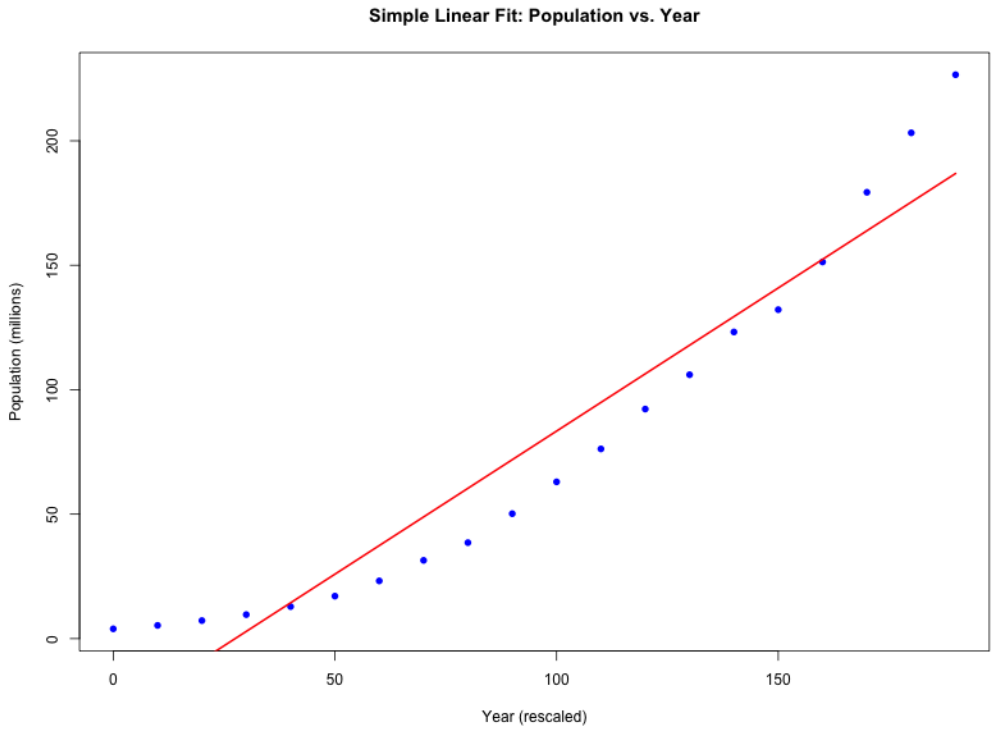
We start by fitting a baseline model with the rescaled Year as a predictor variable for the Population. We get the following equation:

$$\hat{Y} = -31.611 + 1.1499 \cdot x_1$$

Where $\hat{Y}$ = **Population** is the predicted U.S. population (in millions)

And x_1= **Year** is the **rescaled year** (e.g., 1790 = 0, 1800 = 10, ..., 1980 = 190).

Creating a scatterplot and then graphing the regression line on top, we can see how well the regression line predicts our data. We get the following plot:



From the following scatterplot, we can see that the regression line is able to capture the overall trend of the data, but it is not able to capture the curvature of the data. The fit of the line is not bad but can be improved with our second model. There is a nonlinear relationship between Year and Population, so a nonlinear model may improve this.

Continuing our analysis of the SLR, we perform an Overall F-test.

$$H_0: All\ regression\ coefficients\ (except\ the\ intercept)\ are\ equal\ to\ 0\ (\beta_1 = 0)$$

$$H_A: All\ regression\ coefficients\ (except\ the\ intercept)\ are\ not\ equal\ to\ 0\ (\beta_1 \neq 0)$$

Results for the Overall F-Test

P-value
2.318e-11

We reject H_0, and conclude that all regression coefficients are not equal to 0. So, our model explains a statistically significant portion of the variation in Population.

Next, we perform a t-test for the regression coefficient.

$$H_0: Year\ has\ no\ effect\ on\ Population\ (\beta_1 = 0)$$

$$H_A: Year\ has\ a\ significant\ effect\ on\ Population (\beta_1 \neq 0)$$

Results for t-test for β_1

P-value
2.32e-11

We reject H_0, and conclude that Year has a significant effect on Population.

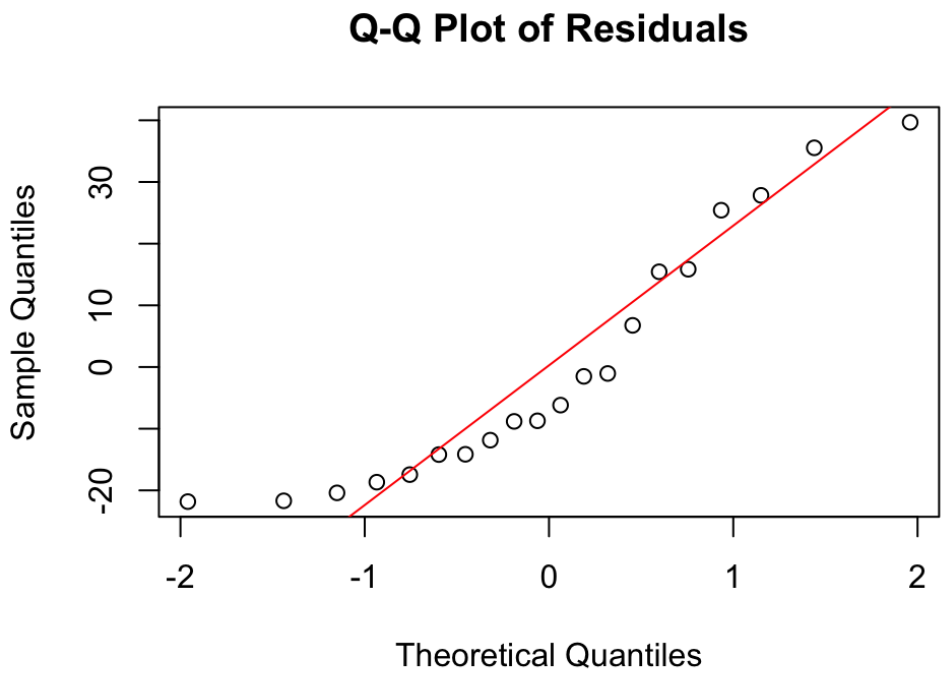
Additionally, we get the following:

$Adjusted\ R^2$	MSE	AIC
.9166	419.43	181.42

The $Adjusted\ R^2$ says that 91.66% of our variation is explained. A 419.4304 is relatively high since we want the number close to 1.

To check the validity of our results, we check normality and homoscedasticity.

Looking at normality, we first try to visualize it.



The Q-Q plot of Residuals shows deviation from the tail, specifically in the lower tail. This is a sign our data may be non-normal.

We formally test this doing the Shapiro-wilk test.

$$H_0: Residuals\ are\ normally\ distributed$$

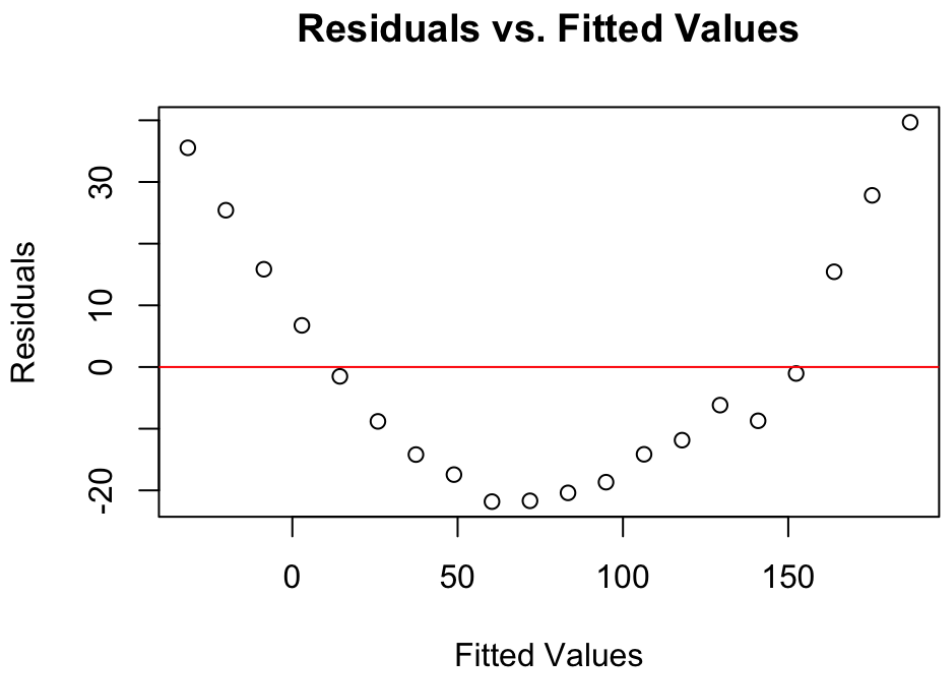
$$H_A: Residuals\ are\ not\ normally\ distributed$$

Results for the Shapiro-Wilk test

P-value
.02443

Since p-value < α=.05, we reject the null and conclude that the residuals are not normally distributed.

Next, we test for homoscedasticity.



The residuals vs. fitted values show a clear systematic pattern. This is signs that constant error variance is violated. We formally test this using the Breusch-Pagan test.

$$H_0: The\ residuals\ have\ constant\ variance\ (homoscedasticity)$$

$$H_A: The\ residuals\ have\ non-constant\ variance\ (heteroscedasticity)$$

Results for the Breusch-Pagan test

P-value
.7609

Since p-value > α=.05, we fail to reject the null. Despite the systematic pattern we saw, the residuals have constant variance (homoscedasticity) by the Breusch-Pagan test.

Concluding the analysis of model 1, although the model explained a high amount of variance because of the high $adjusted\ R^2$, the MSE is still a large number and ultimately the model is invalid because of the failed assumption of normality. This model is inadequate for this data.

Model 2: Quadratic Regression Model

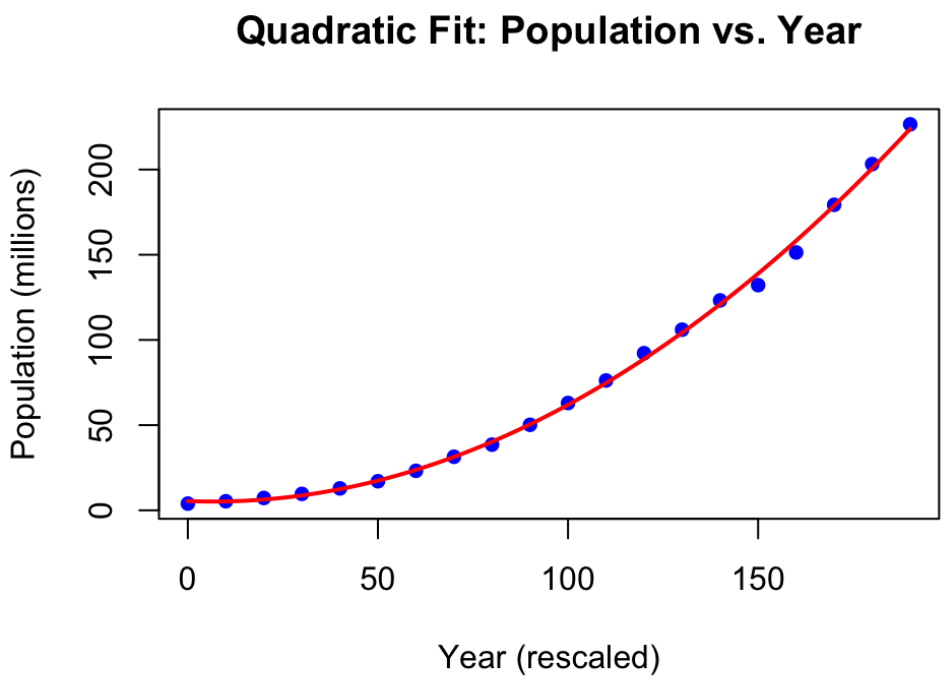
Due to seeing the nature of our data with our earlier scatterplot, we will now fit a quadratic regression model. Fitting the model we get:

$$\hat{Y} = 5.4203 - .0845 \cdot x_1 + .0065 \cdot x_1^2$$

Where $\hat{Y}$ = **Population** is the predicted U.S. population (in millions)

And x_1= **Year** is the **rescaled year** (e.g., 1790 = 0, 1800 = 10, ..., 1980 = 190).

We use our model with the scatterplot to see how well our model predicted the scatterplot. We get the following plot:



The model fits the data very well. We now see the regression curve, match the curve of the scatterplot which we didn't see before in the Simple Linear Model. We notice there are two points that are a little further from the line than the other points. These points appear to be the WW2 years that we saw in our preliminary analysis. This may be important for our next model if needed.

Proceeding with our analysis of model 2, we first perform an overall F-test.

H_0: *All regression coefficients (except the intercept) are equal to* 0 ($\beta_1 = \beta_2 = 0$)

H_A: *All regression coefficients (except the intercept) are not equal to* 0 ($\beta_1 \neq \beta_2 \neq 0$)

Results for the Overall F-Test

P-value
<2.2e-16

We reject H_0, and conclude that all regression coefficients are not equal to 0. So, our model explains a statistically significant portion of the variation in Population.

Next, we perform a t-test for the regression coefficients.

$$H_0: Year\ has\ no\ effect\ on\ Population\ (\beta_1 = 0)$$

$$H_A: Year\ has\ a\ significant\ effect\ on\ Population (\beta_1 \neq 0)$$

Results for t-test for β_1

P-value
0.06169

We fail to reject H_0 and conclude that Year has no significant effect on Population.

We perform another t-test for the regression coefficient for Year^2.

$$H_0: \text{Year}^2\ has\ no\ effect\ on\ Population\ (\beta_2 = 0)$$

$$H_A: \text{Year}^2 has\ a\ significant\ effect\ on\ Population (\beta_2 \neq 0)$$

Results for t-test for β_2

P-value
3.11e-16

We conclude that Year^2 has a statistically significant effect on population.

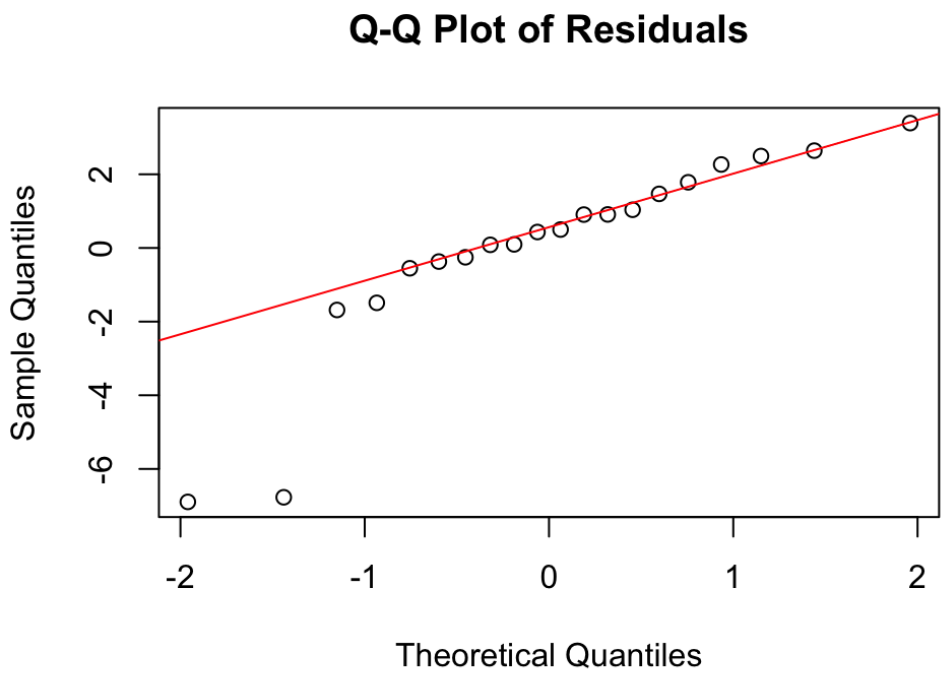
Additionally, we get the following from the summary statistics.

$Adjusted\ R^2$	MSE	AIC
.9984	8.08	103.29

The $Adjusted\ R^2$ says that 99.84% of our variation is explained. Nearly all the variance is explained. An 8.076964 MSE is a lot lower than we previously saw with the Simple Linear Regression Model.

Now we will perform residual diagnostics to determine the validity of the model.

Looking at the Q-Q plot we get the following,



We see that the residuals seem to follow the line very well except to points on the lower tail of the line. These two points could show that our data is still non-normal.

We formally test this doing the Shapiro-Wilk test.

$$H_0: Residuals\ are\ normally\ distributed$$

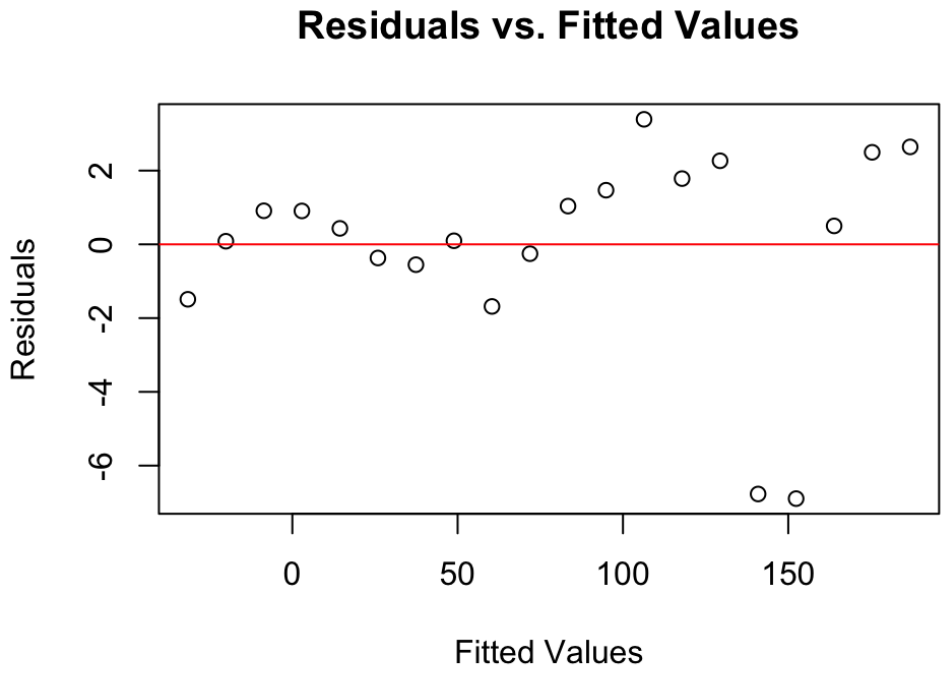
$$H_A: Residuals\ are\ not\ normally\ distributed$$

Results for the Shapiro-Wilk test

P-value
0.001765

Thus, we conclude that our data is not normally distributed.

Next, we test for homoscedasticity.



There seems to be a pattern. However, we formally test this using the Breusch-Pagan test.

$$H_0: The\ residuals\ have\ constant\ variance\ (homoscedasticity)$$

$$H_A: The\ residuals\ have\ non-constant\ variance\ (heteroscedasticity)$$

Results for Breusch-Pagan Test

P-value
.1252

Despite seeing what looks like a patter, we conclude that the residuals have constant variance by the Breusch-Pagan test.

Overall, our model 2 was improved from model 1. Our $adjusted\ R^2$ increased and our MSE was reduced by a lot. We can also compare AIC and see that model 2 has a lower AIC. Although model 2 is a better fit, it is still invalid since model 2 showed non-normality because of the Shapiro-Wilk test.

Model 3: Quadratic Regression Model with Indicator Variable for WW2

The next model, we seek to consider that WW2 occurred during the years 1940 (x=150) to 1950 (x=160). We feature engineer a new variable called WW2 where x=150 and x=160 is 1, to

indicate WW2 occurred during those years, and the rest of the entries are 0. In our preliminary analysis, we saw some changes to the growth rate occur during that time with the line plot and scatterplot.

Fitting the model we get the following regression equation:

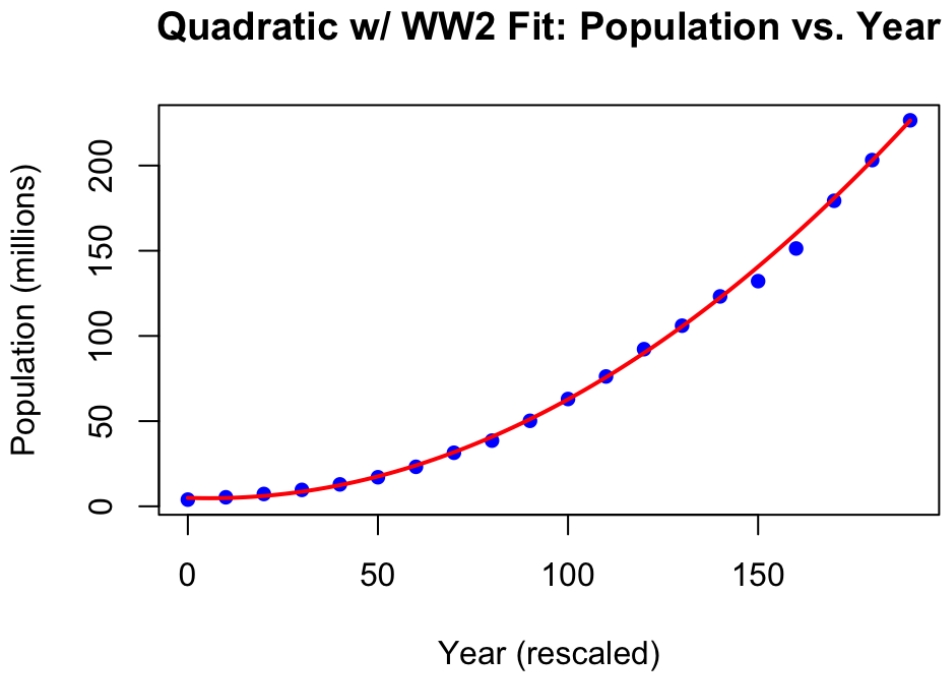
$$\hat{Y} = 4.971 - .07447 \cdot x_1 + .006526 \cdot x_1^2 - 8.641 \cdot x_2$$

Where $\hat{Y}$ = **Population** is the predicted U.S. population (in millions),

x_1= **Year** is the **rescaled year** (e.g., 1790 = 0, 1800 = 10, ..., 1980 = 190),

And x_2 = indicator variable for **WW2** (1=WW2 took place, 0=WW2 did not take place).

We use model 3 to see how well the model predicts our data.



Notice that the observed values are well predicted by our model.

We proceed now with an Overall F-test.

H_0: *All regression coefficients (except the intercept) are equal to* 0 ($\beta_1 = \beta_2 = \beta_3 = 0$)

H_A: *All regression coefficients (except the intercept) are not equal to* 0 ($\beta_1 \neq \beta_2 \neq \beta_3 \neq 0$)

Results for the Overall F-Test

P-value
<2.2e-16

We reject the null hypothesis and conclude that all regression coefficients are not equal to 0. So, our model explains a statistically significant portion of the variation in Population.

We proceed by doing t-tests for each regression coefficient.

$$H_0: Year\ has\ no\ effect\ on\ Population\ (\beta_1 = 0)$$

$$H_A: Year\ has\ a\ significant\ effect\ on\ Population (\beta_1 \neq 0)$$

Results for t-test for $\boldsymbol{\beta_1}$

P-value
0.000317

Thus, Year has a significant effect on Population.

Now doing it for β_2,

$$H_0: Year^2\ has\ no\ effect\ on\ Population\ (\beta_2 = 0)$$

$$H_A: Year^2\ has\ a\ significant\ effect\ on\ Population (\beta_2 \neq 0)$$

Results for t-test for $\boldsymbol{\beta_2}$

P-value
<2e-16

Thus, we reject the null hypothesis and conclude that $Year^2$ has a significant effect on Population.

Now doing it for β_3,

$$H_0: WW2\ has\ no\ effect\ on\ Population\ (\beta_3 = 0)$$

$$H_A: WW2\ has\ a\ significant\ effect\ on\ Population\ (\beta_3 \neq 0)$$

Results for t-test for β_3

P-value
3.09e-08

Thus, we conclude that WW2 has a significant effect on Population.

We determined that all regression coefficients have statistically significant effects on Population.

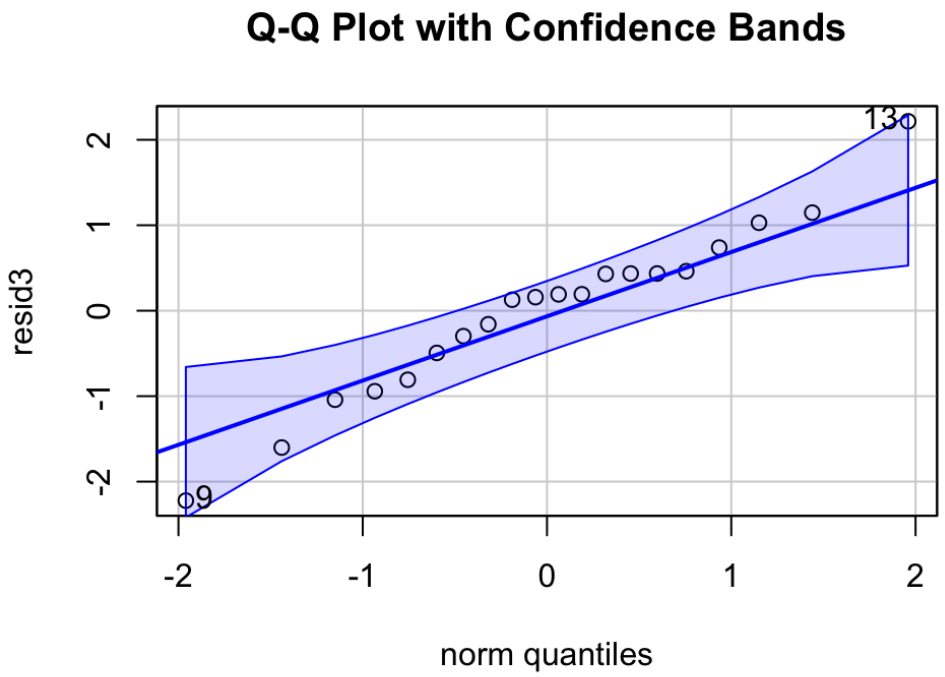
We can also obtain the following values from the summary statistics:

$Adjusted\ R^2$	MSE	AIC
.9998	1.201216	65.95306

We notice that the $Adjusted\ R^2$ shows that nearly all the variance is explained. The MSE value is also very close to 1. The AIC also improved from the last two models.

Now we will perform residual diagnostics on the model.

Looking at the Q-Q plot, we get:



The residuals seem to lie close to the line and inside the confidence bands. There are two points, 9 and 13, that seem to be close to outside or on the border of the confidence bands. We will do the Shapiro-Wilk to formally test for normality.

$$H_0: Residuals\ are\ normally\ distributed$$

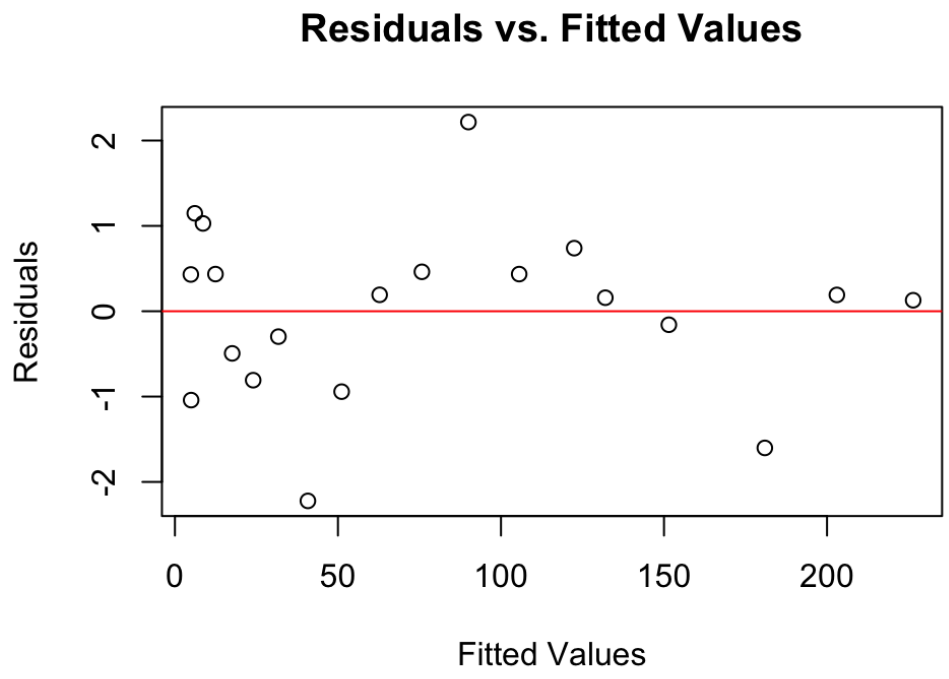
$$H_A: Residuals\ are\ not\ normally\ distributed$$

Results for the Shapiro-Wilk test

P-value
.8094

Thus, we fail to reject H_0, the residuals are normally distributed.

Looking at the residuals vs. fitted values, we get the following plot:



There does not seem to be any pattern. We further investigate with the Breusch-Pagan Test.

$$H_0: The\ residuals\ have\ constant\ variance\ (homoscedasticity)$$

$$H_A: The\ residuals\ have\ non-constant\ variance\ (heteroscedasticity)$$

Results for Breusch-Pagan Test

P-value
.7609

Thus, we conclude that our residuals have constant variance, homoscedasticity.

After analyzing model 3, we seem to improve AIC, MSE and $Adjusted\ R^2$. We also find that model 3 passes the assumption of normality and constant variance.

Conclusion:

Comparing the three models using AIC and MSE, we form the following table:

Model	AIC	MSE
Model 1: **Simple Linear Regression**	419.43	181.42
Model 2: **Quadratic Regression Model**	103.29	8.08
Model 3: **Quadratic Regression Model with Indicator Variable for WW2**	65.95	1.20

By comparing the models with AIC and MSE, model 3 is the best model since it performs the best under both metrics. Model 3 was also the only model that passed both assumptions of normality and constant variance, making it the only statistically valid model.

We will use model 3, with the following regression equation:

$$\hat{Y} = 4.971 - .07447 \cdot x_1 + .006526 \cdot x_1^2 - 8.641 \cdot x_2$$

to predict the Population in the year 2020.

The model predicted that the Population in the United States would be 333.084 million in 2020. Using Google, the actual population reported was 331.5 million in 2020. Therefore, the percent error from our model is -.48%. So, our model overestimated the actual population by .48% percent.

Source Code and Outputs

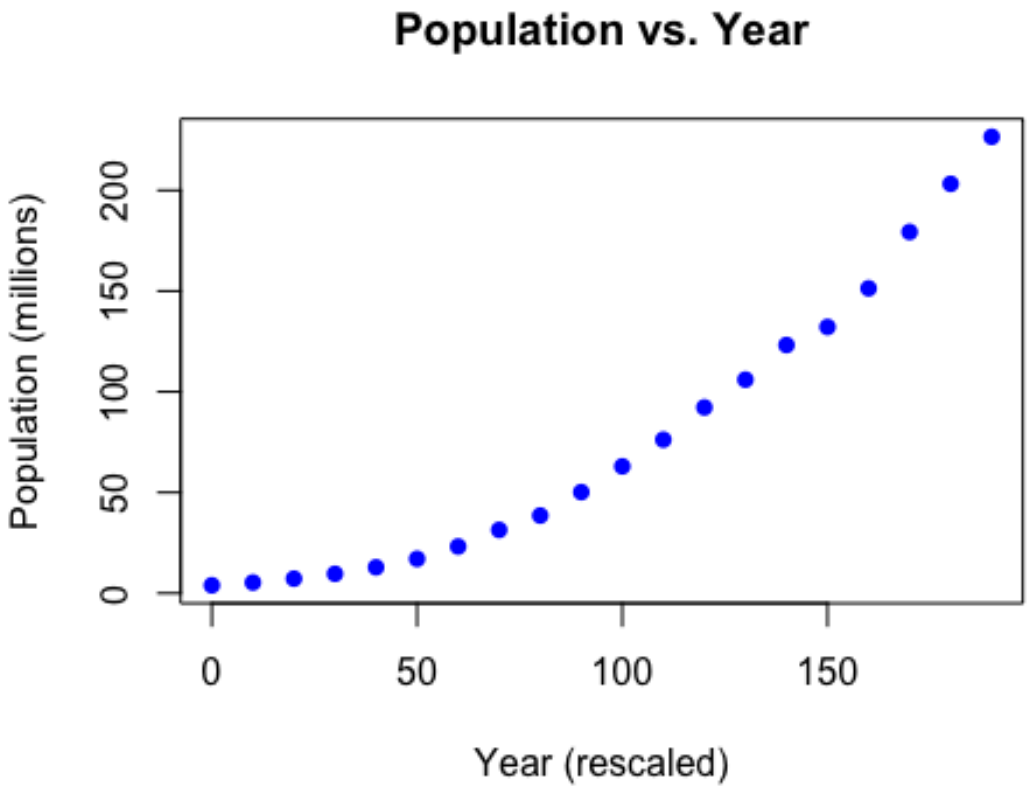
```
library(here)

## here() starts at /Users/emonforte/Downloads/Spring2025/Sta 5900/RStudio_Sp
ring2025

uspop <- read.csv(here("USpop.csv"))
```

Re-scale Year variable

```
uspop$Year = uspop$Year - 1790
```

#preliminary Data Analysis

```
plot(uspop$Year, uspop$Population,
     main = "Population vs. Year",
     xlab = "Year (rescaled)", ylab = "Population (millions)",
     pch = 16, col = "blue")
```




```
png("lineplot.png", width = 800, height = 600)
plot(uspop$Year, uspop$Population, type="l", main="U.S. Population Over Time"
, xlab="Year", ylab="Population (millions)")
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
png("lineplot_sub.png", width = 800, height = 600)
subset_data <- subset(uspop, Year >= 110 & Year <= 190)

plot(subset_data$Year, subset_data$Population,
     type = "l",
     main = "U.S. Population (1900-1980)",
     xlab = "Year (rescaled)", ylab = "Population (millions)",
     col = "blue", lwd = 2)
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
# Calculate summary stats
mean_val <- mean(uspop$Population)
sd_val <- sd(uspop$Population)

png("Population_box.png", width = 800, height = 600)
# Create the boxplot
boxplot(uspop$Population,
        main = "Boxplot of U.S. Population (1790-1980)",
        ylab = "Population (millions)",
        col = "lightblue")

# Add mean as a red dot
points(1, mean_val, col = "black", pch = 19)

legend("topright",
       legend = c(
         paste("Mean:", round(mean_val, 2)),
         paste("SD:", round(sd_val, 2))
       ),
       pch = c(19, NA),
       col = c("black", NA),
       bty = "n",
       cex = 1.2)

dev.off()
```

```
## quartz_off_screen
##                 2
```

```
summary(uspop$Population)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.93   16.02   56.59   77.63  125.44  226.55
```

#model 1 Running analysis on the SLR model #Fitting model

```r
attach(uspop)
mod1 <- lm(Population~ Year, uspop)

png("mod1_scatter.png", width = 800, height = 600)
# Scatterplot of original data
plot(uspop$Year, uspop$Population,
     main = "Simple Linear Fit: Population vs. Year",
     xlab = "Year (rescaled)", ylab = "Population (millions)",
     pch = 16, col = "blue")

# Add the quadratic regression curve
curve(predict(mod1, newdata = data.frame(Year = x)),
      from = min(uspop$Year), to = max(uspop$Year),
      col = "red", lwd = 2, add = TRUE)

dev.off()
```

```
## quartz_off_screen
##                 2
```

Scatterplot shows a nonlinear relationship between year and population. A tranformation on x could be better Possible quadratic independent variable may be more suitable.

#summary statistics

```r
summary(mod1)
```

```
##
## Call:
## lm(formula = Population ~ Year, data = uspop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.824 -15.007  -7.449  15.545  39.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.6111     8.8243  -3.582  0.00213 **
## Year          1.1499     0.0794  14.482 2.32e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.48 on 18 degrees of freedom
## Multiple R-squared:  0.921,  Adjusted R-squared:  0.9166
## F-statistic: 209.7 on 1 and 18 DF,  p-value: 2.318e-11
```

Overall F-test: p-val 2.318e-11 Reject H0

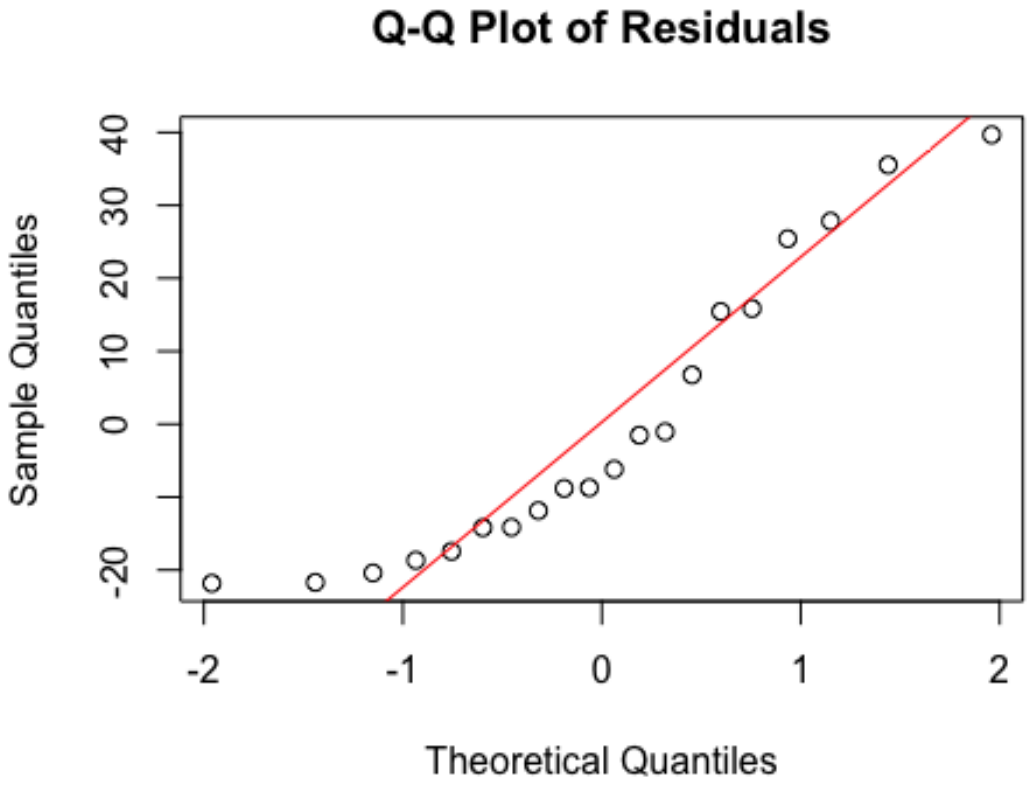
T-test for the Regression Coefficient: p-val 2.32e-11 Reject H0

Adjusted R^2 = .9166

MSE = RSE^2 = (20.48)^2 = 419.4304

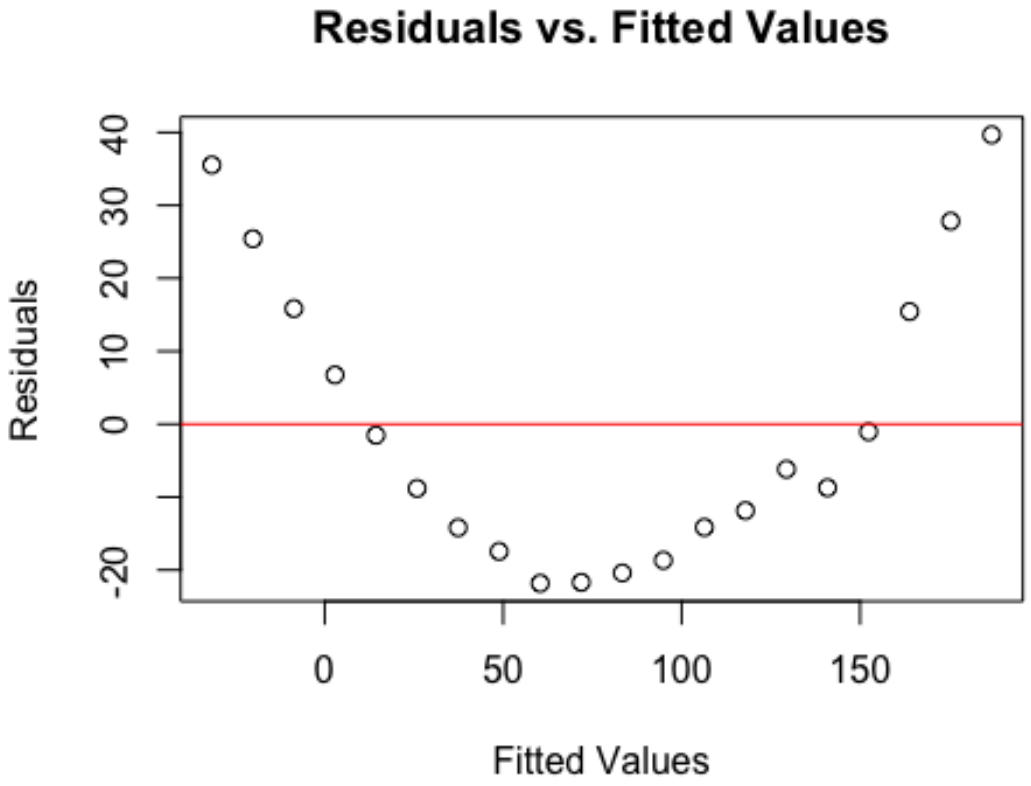
#Residual Analysis

```
resid1 = mod1$residuals
fitted1 = mod1$fitted.values

#qqplot
qqnorm(resid1,
       main = "Q-Q Plot of Residuals")
qqline(resid1,
        col = "red")
```




```
#normality
shapiro.test(resid1)

##
##  Shapiro-Wilk normality test
##
## data:  resid1
## W = 0.88773, p-value = 0.02443

#constant error variance
plot(fitted1, resid1, main = "Residuals vs. Fitted Values",
```

```
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

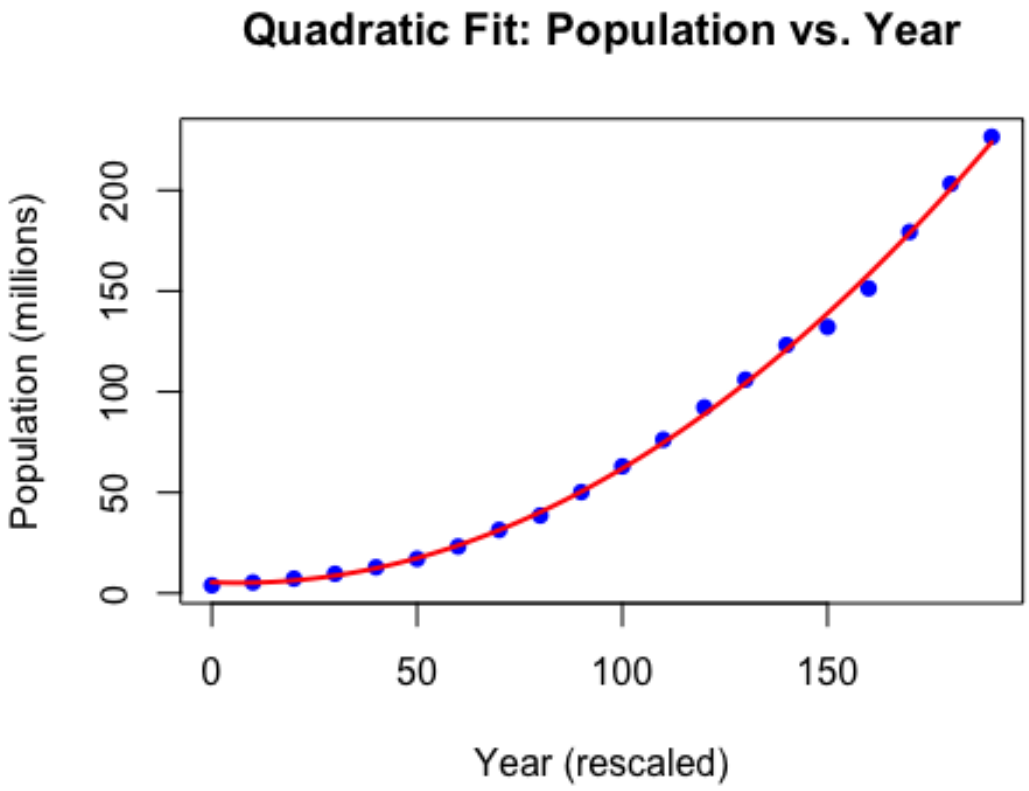



```
bptest(mod1)

##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 0.092625, df = 1, p-value = 0.7609
```

#model 2

Analyzing a quadratic regression model #Fitting a model

```
mod2 = lm(Population ~ Year + I(Year^2))

#scatterplot

# Scatterplot of original data
plot(uspop$Year, uspop$Population,
     main = "Quadratic Fit: Population vs. Year",
     xlab = "Year (rescaled)", ylab = "Population (millions)",
     pch = 16, col = "blue")

# Add the quadratic regression curve
curve(predict(mod2, newdata = data.frame(Year = x)),
      from = min(uspop$Year), to = max(uspop$Year),
      col = "red", lwd = 2, add = TRUE)
```



#Summary Statistics

```
summary(mod2)

##
## Call:
## lm(formula = Population ~ Year + I(Year^2))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8952 -0.4152  0.4664  1.5470  3.3904
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.4202532  1.7304335   3.132  0.00607 **
## Year        -0.0844468  0.0422148  -2.000  0.06169 .
## I(Year^2)    0.0064967  0.0002145  30.291 3.11e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.842 on 17 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9984
## F-statistic:  5903 on 2 and 17 DF,  p-value: < 2.2e-16
```

Overall F-test: p-val=2.2e-16 Reject H0

T-Test

I(Year^2) p-val = 3.11e-16 Reject H0

Year p-val = 0.06169 Fail to reject

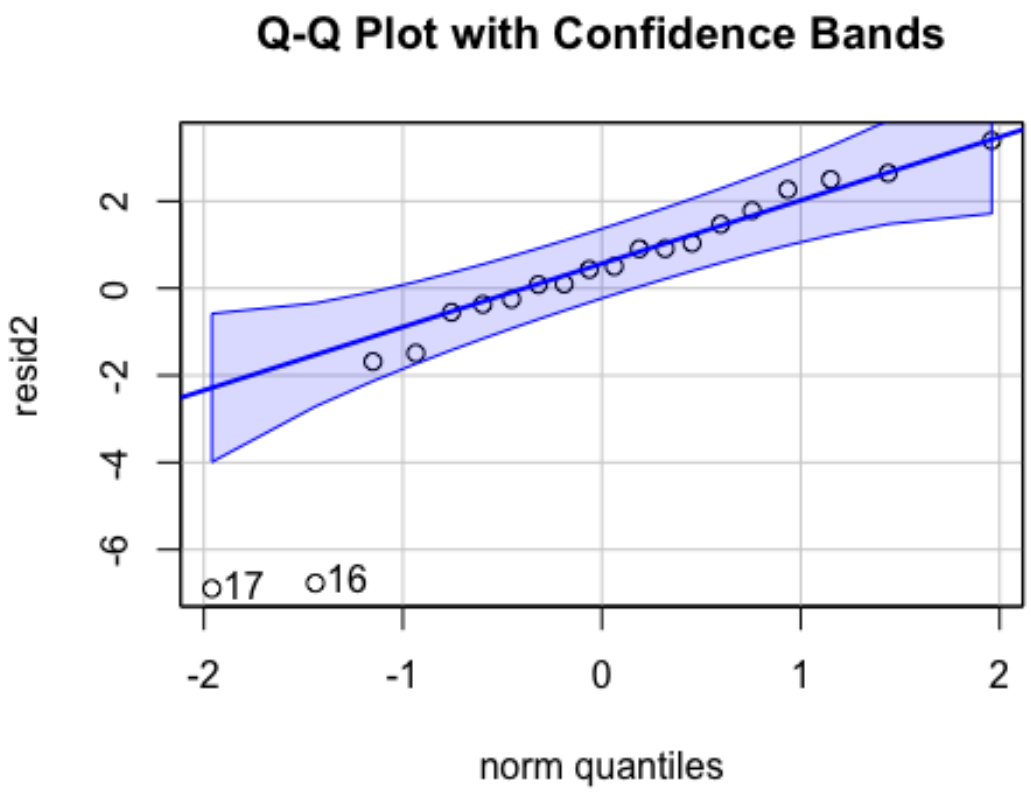
A t-test for the quadratic coefficient indicates a highly significant contribution ($p < 0.001$), while the linear term is marginally insignificant ($p = 0.062$). However, both terms are retained to preserve the polynomial structure of the model.

Adjusted R^2 = .9984

MSE = RSE^2 = 8.076964

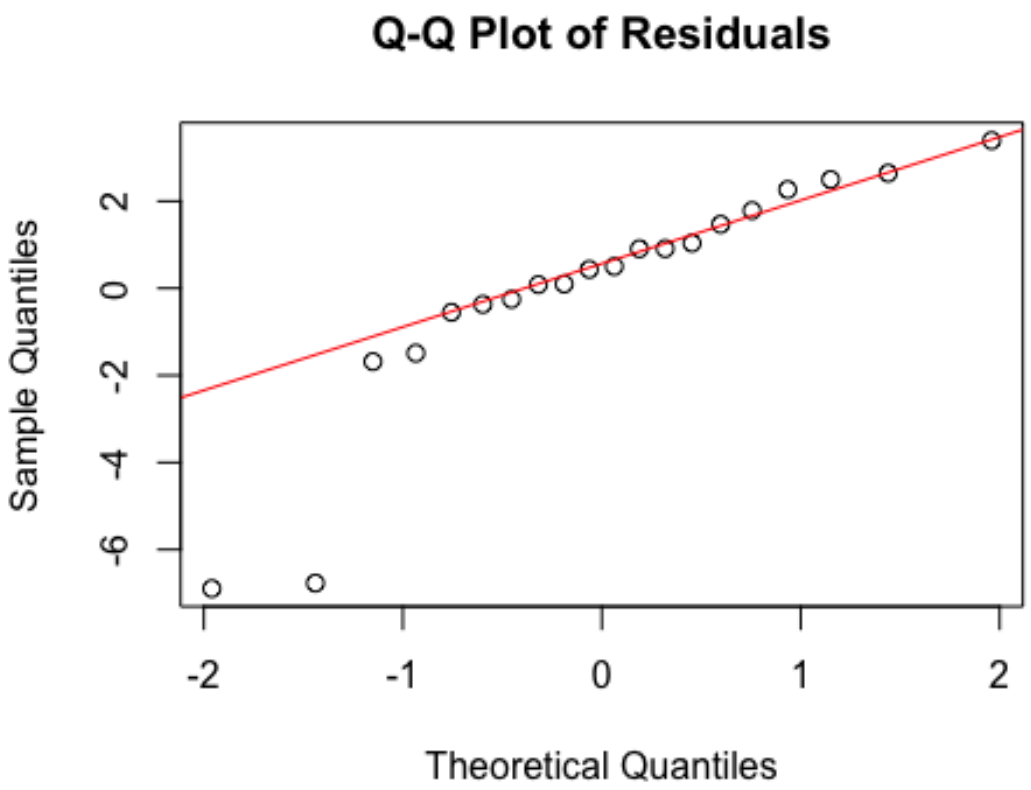
#Residual Analysis

```
resid2 = mod2$residuals
fitted2 = mod2$fitted.values

library(car)

## Loading required package: carData

qqPlot(resid2,
       main = "Q-Q Plot with Confidence Bands",
       envelope = 0.95)
```

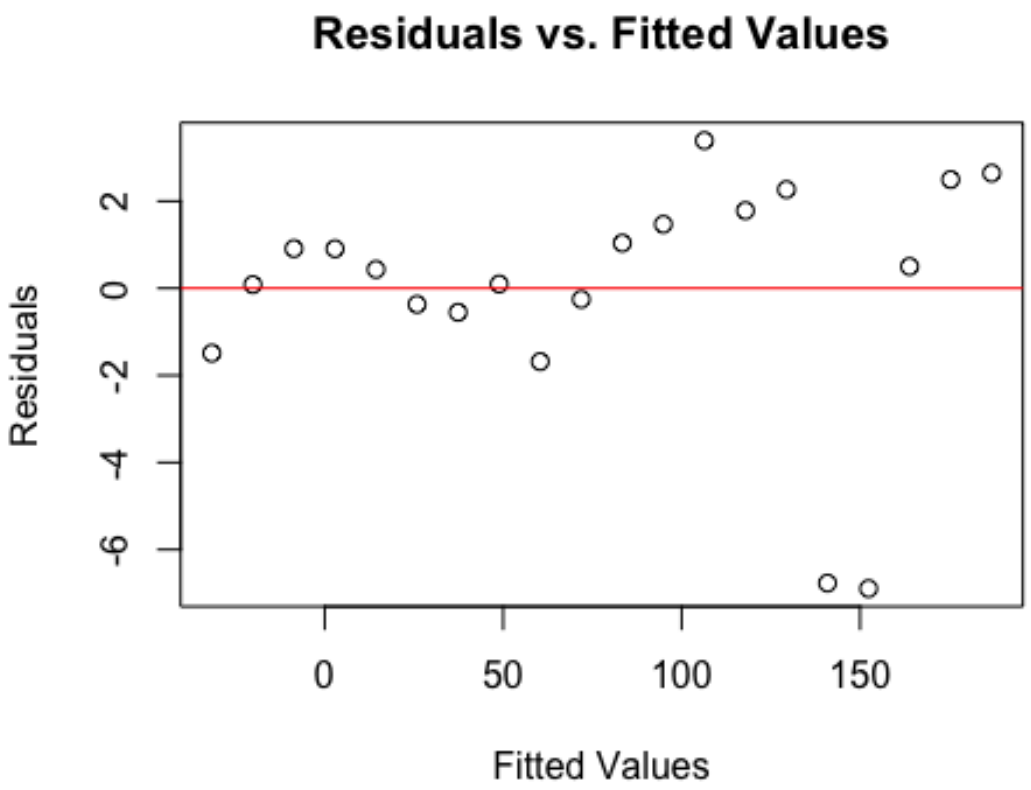



```
## [1] 17 16
```

```
#qqplot
qqnorm(resid2,
       main = "Q-Q Plot of Residuals")
qqline(resid2,
        col = "red")
```




```
#normality
shapiro.test(resid2)

##
##  Shapiro-Wilk normality test
##
## data:  resid2
## W = 0.82034, p-value = 0.001765

#constant error variance
plot(fitted1, resid2, main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```




```
bptest(mod2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 4.1561, df = 2, p-value = 0.1252
```

Shapiro Wilk Reject H0, not normal

Breusch Pagan Fail to reject

#model 3 Quadratic Model using WW2 fact

Creating Indicator variable for WW2

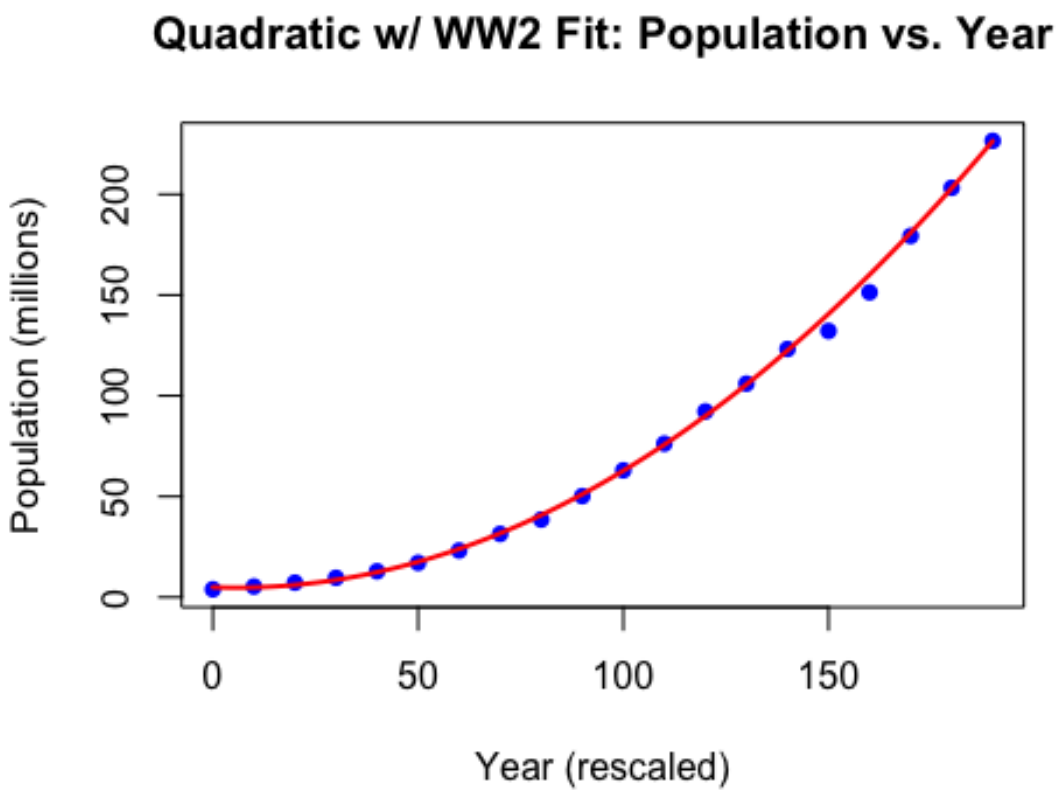
```
uspop$WW2 <- ifelse(uspop$Year %in% c(150, 160), 1, 0)
```

#Fitting model

```
attach(uspop)
```

```
## The following objects are masked from uspop (pos = 7):
##
##     Population, Year
```

```r
mod3 = lm(Population ~ Year + I(Year^2) + WW2)

# Scatterplot of original data
plot(uspop$Year, uspop$Population,
     main = "Quadratic w/ WW2 Fit: Population vs. Year",
     xlab = "Year (rescaled)", ylab = "Population (millions)",
     pch = 16, col = "blue")

# Add the cubic regression curve
curve(predict(mod3,
              newdata = data.frame(
                Year = x,
                WW2 = ifelse(x %in% c(150, 160), 1, 0)  # mimic your original
WW2 definition
              )),
      from = min(uspop$Year), to = max(uspop$Year),
      col = "red", lwd = 2, add = TRUE)
```



#Summary Statistics

```r
summary(mod3)

##
## Call:
```

```
## lm(formula = Population ~ Year + I(Year^2) + WW2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2223 -0.5724  0.1748  0.4419  2.2162
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.971e+00  6.688e-01   7.434 1.42e-06 ***
## Year        -7.447e-02  1.631e-02  -4.566 0.000317 ***
## I(Year^2)    6.526e-03  8.275e-05  78.863  < 2e-16 ***
## WW2         -8.641e+00  8.714e-01  -9.917 3.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 16 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 2.65e+04 on 3 and 16 DF,  p-value: < 2.2e-16
```

Overall F-test p-val= 2.2e-16 Reject H0

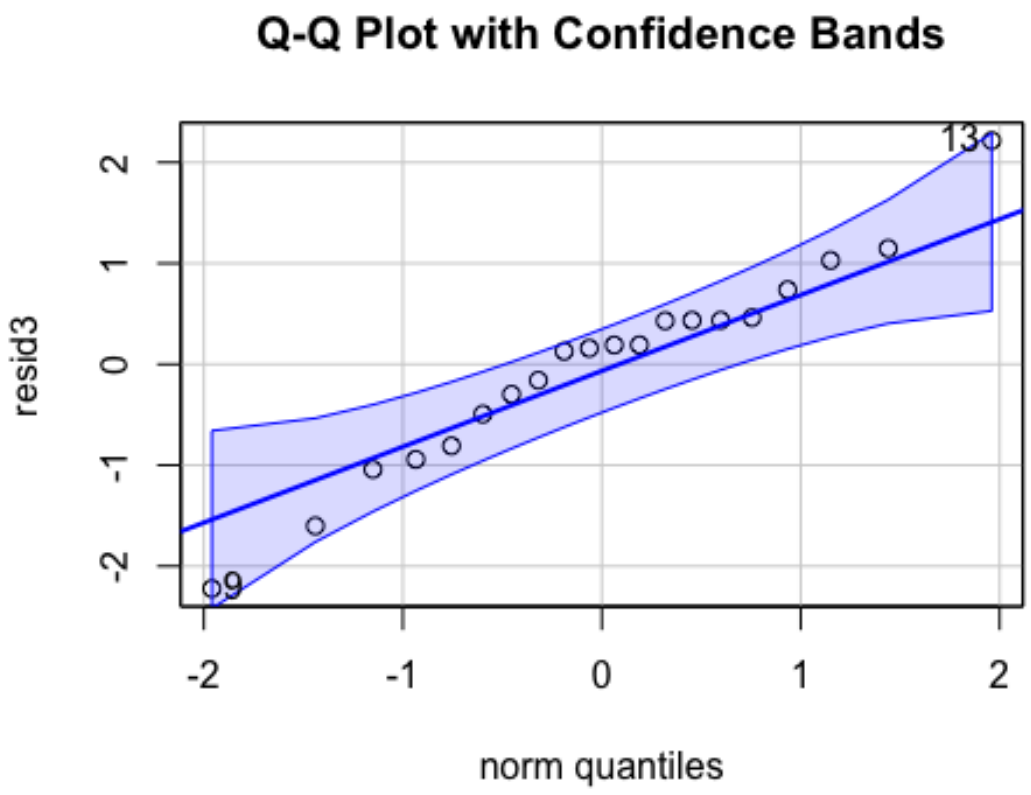
T-test for regression coefficients All p-val < .05

Adjusted R^2 = .9998

MSE = RSE^2 = 1.201216

#Residual Analysis

```
resid3 = mod3$residuals
fitted3 = mod3$fitted.values

qqPlot(resid3,
       main = "Q-Q Plot with Confidence Bands",
       envelope = 0.95)
```

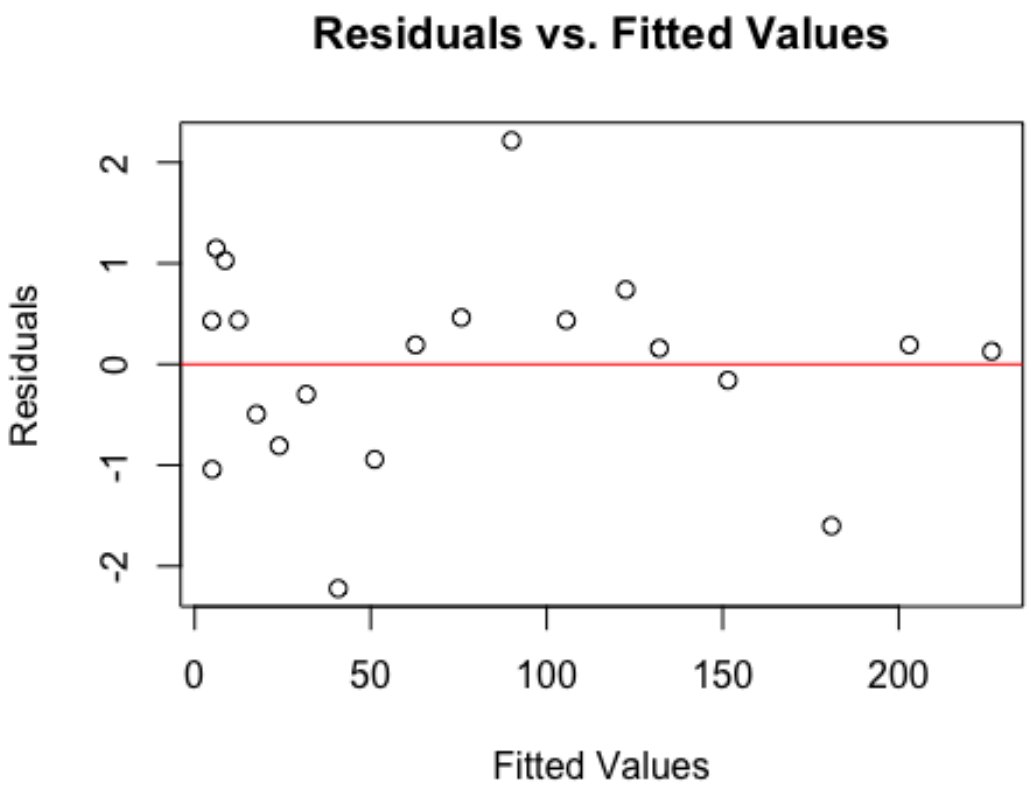



```
## [1]  9 13
```

```
shapiro.test(resid3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid3
## W = 0.97264, p-value = 0.8094
```

```
#constant error variance
plot(fitted3, resid3, main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```




```
bptest(mod1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 0.092625, df = 1, p-value = 0.7609
```

Shapiro-Wilk p-val = .8094 fail to reject H0, Data is normal

Breusch Pagan P-val = .7609 Fail to reject H0, Constant error variance

#Comparing AIC

```
AIC(mod1, mod2,mod3)
```

```
##      df       AIC
## mod1  3 181.42168
## mod2  4 103.28501
## mod3  5  65.95306
```

#Predict Population for 2020

```
predicted = predict(mod3, newdata = data.frame(
  Year = 230,
```

```
  WW2 = 0  # 2020 is not during WWII
))

actual = 331.5
percent_error <- (actual - predicted) / actual * 100
```

My model predicted the population would be 333.084 million Percent error is -.48%. Meaning our model had a slight overestimation of -.48% of the actual population in 2020.