Simple Linear Regression

Blake Pappas

11/16/2021

Example: Tomatoes

a. Find a simple linear regression line that can predict the tomato height at measurement 2 (height2) using the pH measurement. Report the slope and intercept of the regression line.

```
tomato <- read.csv("tomato.csv")

lm_tomato <- lm(ph ~ height1, data = tomato)
summary(lm_tomato)
```

```
##
## Call:
## lm(formula = ph ~ height1, data = tomato)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67125 -0.55614  0.06756  0.84947  1.50532
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.8580     3.1874  -1.210   0.2607
## height1       0.8521     0.3230   2.638   0.0298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.114 on 8 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.3984
## F-statistic: 6.959 on 1 and 8 DF,  p-value: 0.02981
```

Answer: Please see above for the for the simple linear regression line (SLR) that can predict the tomato height. Based on the output, the slope of the regression is 1.1737 and the intercept is -11.6973.

b. Use your answer from part (a) to predict the height at measurement 2 for a tomato plant whose soil pH is 5.1.

```
newX <- 5.1
beta0_hat <- as.numeric(lm_tomato$coefficients[1])
beta1_hat <- as.numeric(lm_tomato$coefficients[2])
predictedY <-  beta0_hat + beta1_hat * newX
predictedY
```

```
## [1] 0.487516
```

Answer: Using the regression equation, the predicted height for a tomato plant whose soil pH is 5.1 is approximately 0.487516.

c. Find a 95% confidence interval for the slope of the regression line.

```
n <- nrow(tomato)
tstar <- qt(0.975, df = n - 2)
tomato_coefficients <- summary(lm_tomato)$coefficients

beta1_hat <- tomato_coefficients[2, 1]
se_beta1_hat <- tomato_coefficients[2, 2]

lower_slope <- beta1_hat - tstar * se_beta1_hat
upper_slope <- beta1_hat + tstar * se_beta1_hat

c(lower_slope, upper_slope)
```

```
## [1] 0.1072357 1.5968781
```

Answer: The 95% confidence interval for the slope of the regression line is (0.1072357, 1.5968781).

Example: Expenditures

The data set `2015\_revenue\_expenditures.csv` contains records of government revenue and expenditures for 150 of the largest cities in the U.S in 2015. Among the variables measured are `Intergovt revenue`, which records the total revenue in the cities coming from state and federal government, and expenditures in various spending categories.

a. Find the simple linear regression equation to predict expenditures in education services (`Education Services Expend.`) from `Intergovt revenue`. Give the estimated regression equation.

```
revex <- read.csv("2015_revenue_expenditures.csv")

lm_revex <- lm(Education_Services_Expend ~ Intergovt_Revenue, data = revex)
summary(lm_revex)
```

```
##
## Call:
## lm(formula = Education_Services_Expend ~ Intergovt_Revenue, data = revex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1518.81  -247.79    41.61   239.42  1130.21
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       807.36295   80.62925   10.01   <2e-16 ***
## Intergovt_Revenue   0.48871    0.03718   13.14   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 419.1 on 148 degrees of freedom
## Multiple R-squared:  0.5386, Adjusted R-squared:  0.5355
## F-statistic: 172.8 on 1 and 148 DF,  p-value: < 2.2e-16
```

Answer: The estimated regression equation to predict expenditures in education services is Education_Services_Expend = 0.48871 x Intergovt_Revenue + 807.36295.

b. Perform a test of the hypotheses

$$H_0 : \beta_1 = 0; \quad H_A : \beta_1 > 0.$$

Use $\alpha = 0.01$. Report the test statistic, p-value, decision, and conclusion in the context of the problem.

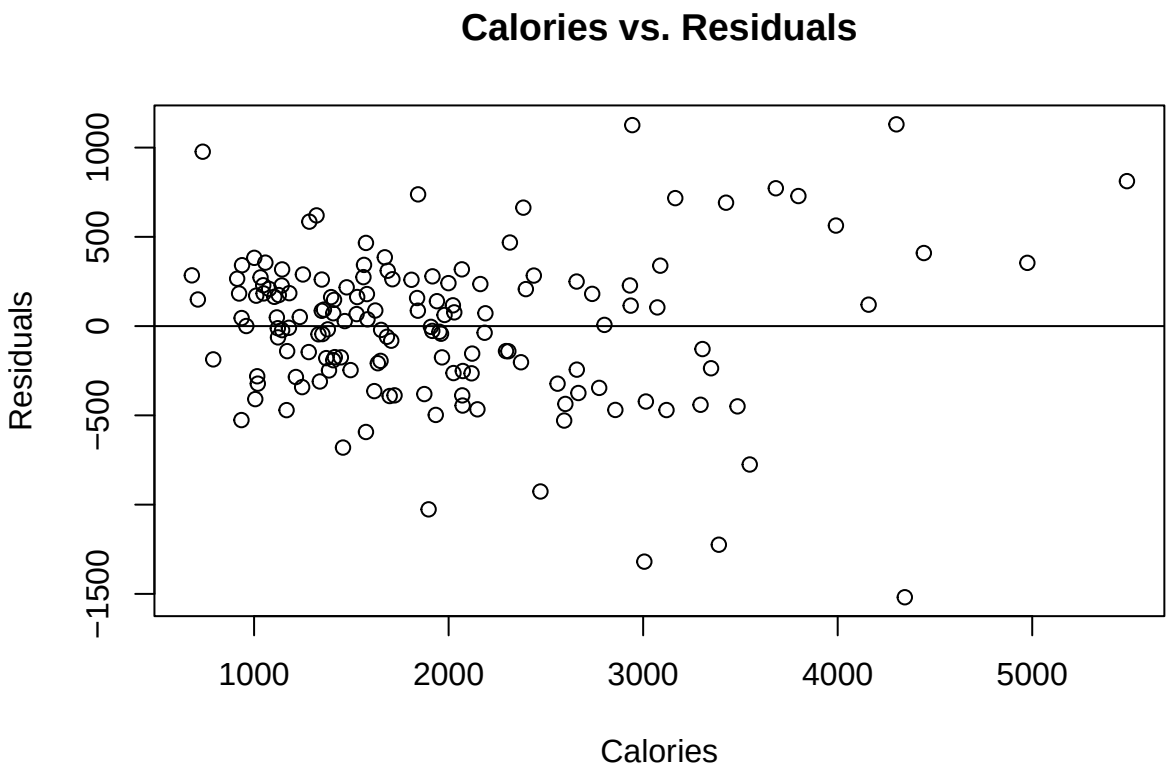
```
n <- nrow(revex)
tstar <- qt(0.995, df = n - 2)

revex_coefficients <- summary(lm_revex)$coefficients

pt(tstar, df = n - 2, lower.tail = TRUE)
```

```
## [1] 0.995
```

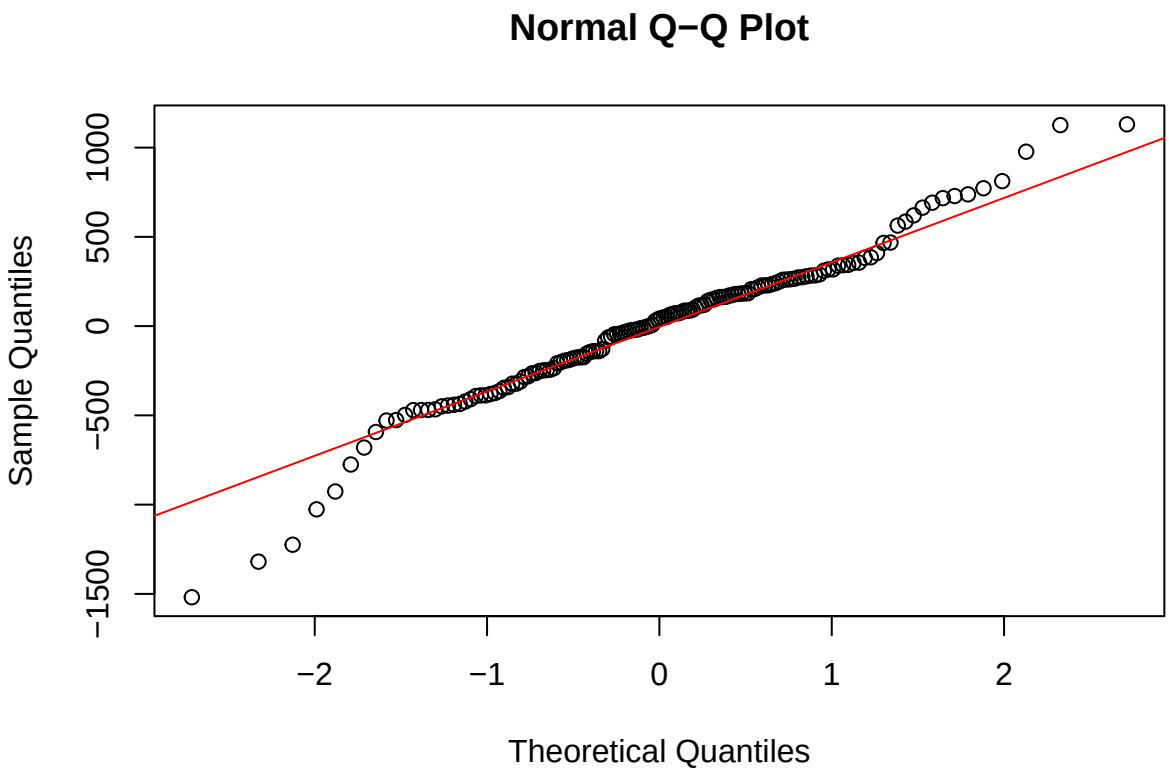
Answer: The test statistic is 2.609456 and the p-value is 0.995. Therefore, we fail to reject null hypothesis. Sufficient evidence does not exist to indicate that the slope is greater than 0.

c. Create a plot of the x variable vs residuals. Also, create a normal quantile plot of the residuals. Explain whether the modeling assumptions of normality and equal variances appear to be reasonable for these data.

```
lm_revex <- lm(Education_Services_Expend ~ Intergovt_Revenue, data = revex)

revex_resid <- residuals(lm_revex)
revex_fitted <- fitted(lm_revex)

plot(revex$Intergovt_Revenue, revex_resid, main = 'Calories vs. Residuals',
     xlab = 'Calories', ylab = 'Residuals')

abline(h = 0)
```




```
qqnorm(revex_resid)
qqline(revex_resid, col = 'red')
```



Answer: See above for the plot of the x variable vs residuals, as well as the normal quantile plot of the residuals. Based on these plots, the modeling assumptions of normality and equal variances appear to be reasonable. Regarding equal variances, the distribution of residuals appears to be random. Regarding normality, the pattern of the plot is very close to the line.

d. Find the observation in the data set for VA: Chesapeake. What is the observed value of y_i, the response variable, for this observation?

```
chesapeake <- c(subset(revex, revex$Label == 'VA: Chesapeake'))
chesapeake$Education_Services_Expend
```

```
## [1] 1951
```

Answer: The observed value for Education_Services_Expend for VA: Chesapeake is 1951.

e. What value does the estimated regression line from part (a) predict for this observation?

Answer: The regression line predicts the value for this observation to be approximately 1691.44.

f. Calculate its residual by subtracting the predicted value from the observed value. ($\widehat{\epsilon}_i = y_i - \widehat{y}_i$). This answer can be compared to the value automatically calculated by R using the `residuals` function.

```r
observed <- chesapeake$Education_Services_Expend
predicted <- 0.48871 * 1809 + 807.36295

residual <- observed - predicted
residual
```

```
## [1] 259.5607
```

Answer: The residual is 259.56.

Example: Body Fat

The data set `BodyFat.csv` contains body fat percentage and several other measurements for a sample of 253 men. Direct measurement body fat percentage can be a cumbersome process, so it might be useful to find a different measurement that is an accurate predictor of body fat.

a. Find a regression line that predicts `BODYFAT` from `WEIGHT`. Write the estimated regression equation and the R^2 for this model.

```r
body_fat <- read.csv("BodyFat.csv")

lm_body_fat <- lm(BODYFAT ~ WEIGHT, data = body_fat)
summary(lm_body_fat)
```

```
##
## Call:
## lm(formula = BODYFAT ~ WEIGHT, data = body_fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.434  -4.315   0.079   4.540  19.681
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.99515    2.38906  -4.184 3.97e-05 ***
## WEIGHT       0.16171    0.01318  12.273  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.135 on 250 degrees of freedom
## Multiple R-squared:  0.376,  Adjusted R-squared:  0.3735
## F-statistic: 150.6 on 1 and 250 DF,  p-value: < 2.2e-16
```

Answer: The regression equation for predicting `BODYFAT` from `WEIGHT` is BODYFAT = 0.16171 x WEIGHT - 9.99515. The r-squared for this model is 0.376.

b. Now find a regression line that predicts `BODYFAT` from `ABDOMEN` (the abdomen circumference). Write the estimated regression equation and the R^2 for this model.

```r
lm_body_fat <- lm(BODYFAT ~ ABDOMEN, data = body_fat)
summary(lm_body_fat)
```

```
##
## Call:
## lm(formula = BODYFAT ~ ABDOMEN, data = body_fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6257  -3.4672   0.0111   3.1415  11.9754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.19661    2.46229  -14.29   <2e-16 ***
## ABDOMEN       0.58489    0.02643   22.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.514 on 250 degrees of freedom
## Multiple R-squared:  0.6621, Adjusted R-squared:  0.6608
## F-statistic: 489.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

Answer: The regression equation for predicting `BODYFAT` from `ABDOMEN` is BODYFAT = 0.58489 x ABDOMEN - 35.19661. The r-squared for this model is 0.6621.

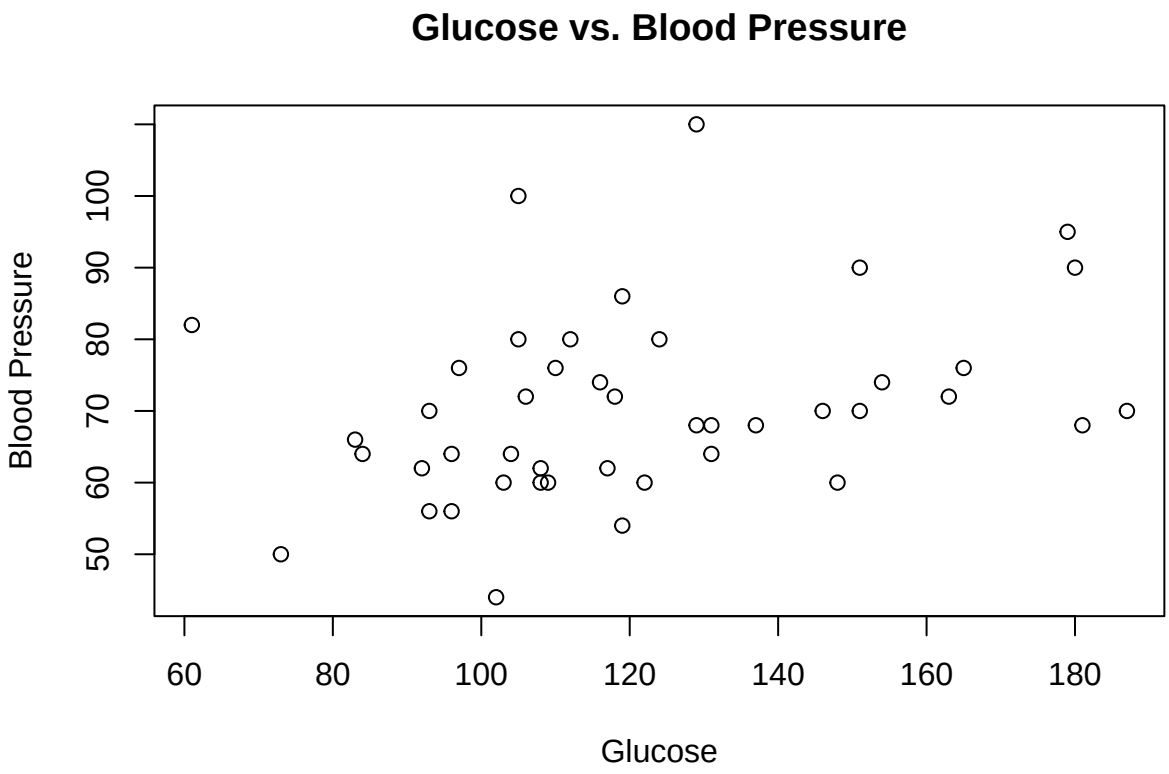
c. Based on the R^2 values, is weight or abdomen circumference a stronger predictor of of body fat? Which variable accounts for a greater proportion of the variability in body fat measurements?

Answer: Based on the r-squared values, abdomen circumference is a stronger predictor of of body fat. It has a proportion of 66.21%, compared to WEIGHT, which only has a proportion of 37.6%.

Example: Diabetes

a. Provide a descriptive plot and summary statistic that describes the relationship between glucose (glu) and diastolic blood pressure (bp). Does there appear to be a strong association between the two variables?

```r
diabetes <- read.csv("diabetes_sm.csv")

plot(diabetes$glu, diabetes$bp, main = 'Glucose vs. Blood Pressure', xlab = 'Glucose', ylab = 'Blood Pr
```




```
cor(diabetes$glu, diabetes$bp)
```

```
## [1] 0.3216093
```

Answer: See above for the descriptive plot and summary statistic describing the relationship between glucose and diastolic blood pressure. Based on the two, there does not appear to be a strong association between the two variables.

b. Fit a simple linear regression line to predict glucose (glu) using blood pressure (bp). Report the slope and intercept of the estimate regression line. Provide an interpretation of the value of the slope.

```
lm_diabetes <- lm(glu ~ bp, data = diabetes)
summary(lm_diabetes)
```

```
##
## Call:
## lm(formula = glu ~ bp, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.733 -18.269  -4.995  15.921  66.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  69.2087    24.0566   2.877  0.00629 **
## bp            0.7381     0.3353   2.201  0.03327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.76 on 42 degrees of freedom
## Multiple R-squared:  0.1034, Adjusted R-squared:  0.08209
## F-statistic: 4.845 on 1 and 42 DF,  p-value: 0.03327
```

Answer: The regression equation for predicting glucose using blood pressure is GLUCOSE = 0.7381 x BLOOD_PRESSURE + 68.2087. The slope of the equation is 0.7381 and the intercept is 69.2087. Glucose levels are expected to increase by 0.7381 if blood pressure rises by one.

c. Make 90% confidence intervals for the intercept and slope of the regression line.

```
# Intercept
diabetes_coefficients <- summary(lm_diabetes)$coefficients
n <- nrow(diabetes)
tstar <- qt(0.95, df = n - 2)

beta0_hat <- diabetes_coefficients[1, 1]
se_beta0_hat <- diabetes_coefficients[1, 2]

lower <- beta0_hat - tstar * se_beta0_hat
upper <- beta0_hat + tstar * se_beta0_hat

c(lower, upper)
```

```
## [1]  28.74661 109.67080
```

```
# Slope
beta1_hat <- diabetes_coefficients[2, 1]
se_beta1_hat <- diabetes_coefficients[2, 2]

lower_slope <- beta1_hat - tstar * se_beta1_hat
upper_slope <- beta1_hat + tstar * se_beta1_hat

c(lower_slope, upper_slope)
```

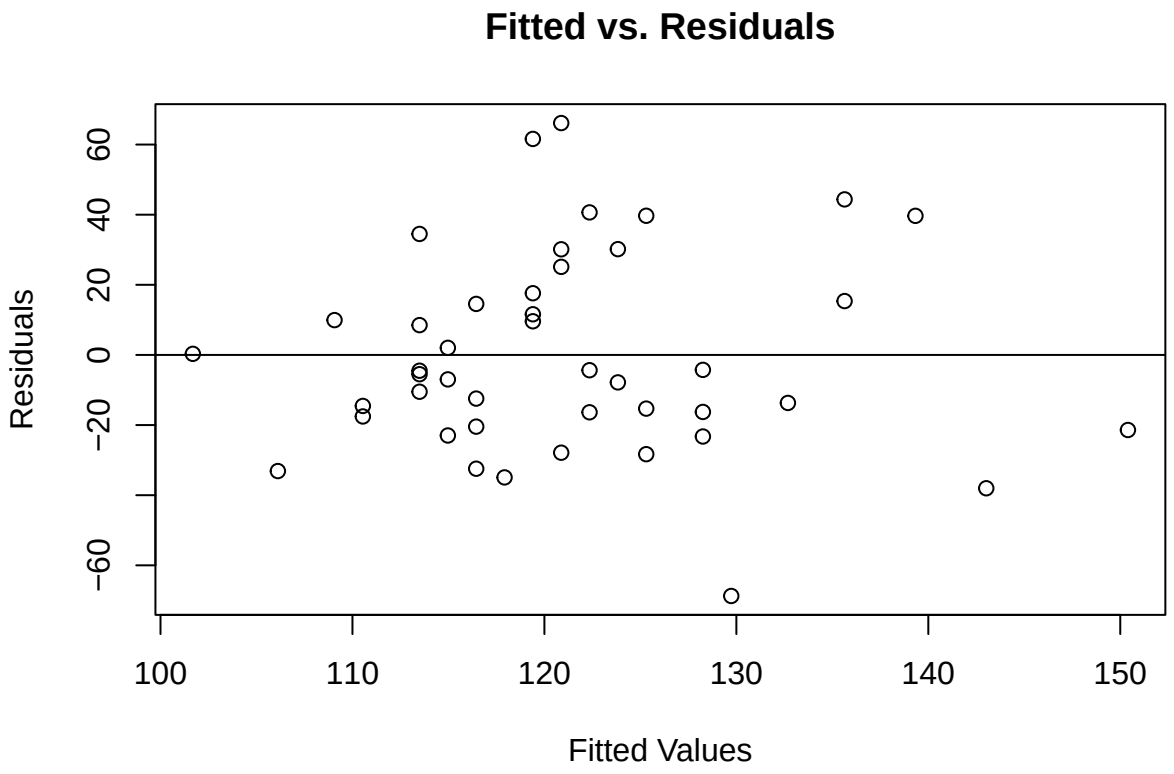
```
## [1] 0.1741167 1.3020939
```

Answer: See above for the 90% confidence intervals for the intercept and slope. They are (28.74661, 109.67080) for the intercept and (0.1741167, 1.3020939) for the slope.

d. Make a plot of the fitted values vs. the residuals. Do the assumptions of equal variances and linearity appear to be reasonable?

```
diabetes_resid <- residuals(lm_diabetes)
diabetes_fitted <- fitted(lm_diabetes)
```

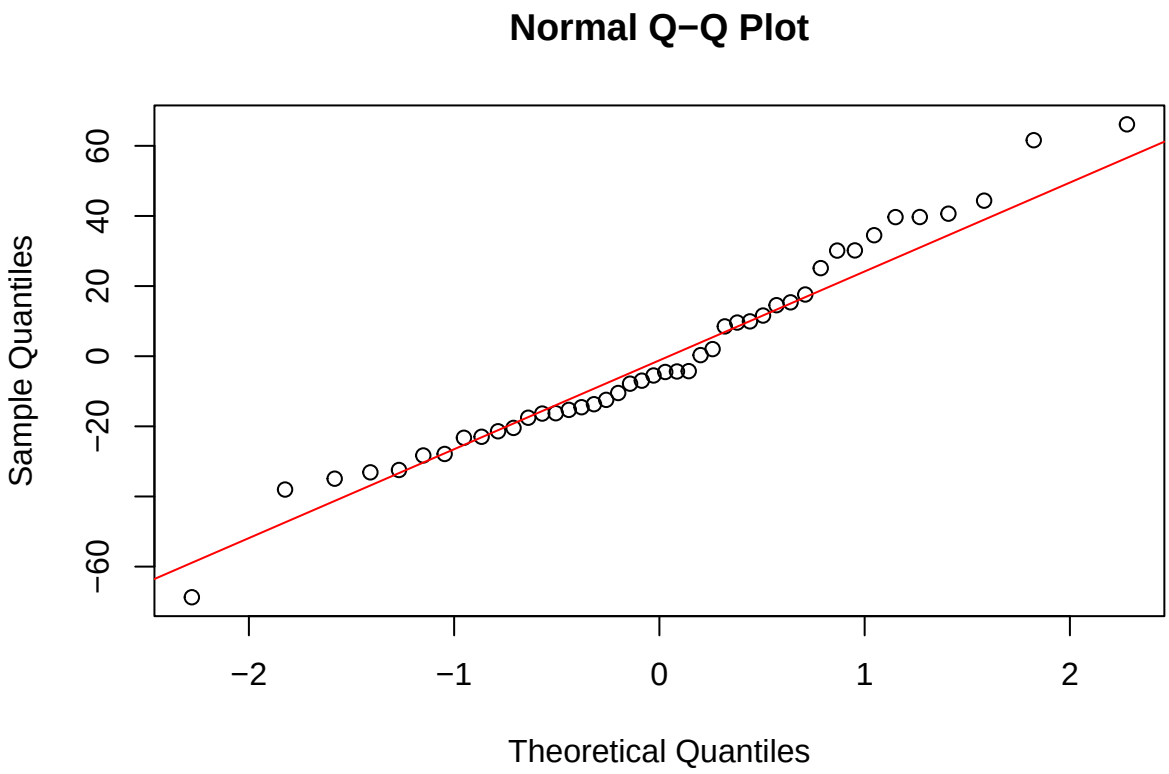
```r
plot(diabetes_fitted, diabetes_resid , main = 'Fitted vs. Residuals',
     xlab = 'Fitted Values', ylab = 'Residuals')

abline(h = 0)
```



Answer: See above for the plot of the fitted values vs. the residuals. Based on the plot, the assumptions of equal variances and linearity do appear to be reasonable.

e. Make a normal quantile plot of the residuals. Does the assumption of normality appear to be reasonable?

```r
qqnorm(diabetes_resid)
qqline(diabetes_resid, col = 'red')
```



Answer: See above for the normal quantile plot of the residuals. Based on the plot, the assumption of normality does indeed appear to be reasonable.