Introduction to Least Squares Regression

Chapter 6, Lab 1: Solutions

OpenIntro Biostatistics

Topics

- Fitting and interpreting a line
- Calculating a least squares line
- Checking assumptions with residual plots

The relationship between two numerical variables can be visualized using a scatterplot in the xy-plane. The *predictor* variable is plotted on the horizontal axis, while the *response* variable is plotted on the vertical axis. This lab introduces the idea of using a straight line, $y = b_0 + b_1 x$, where b_0 is the y-intercept and b_1 is the slope, to summarize data that exhibit an approximately linear relationship. The statistical model for least squares regression is also formally introduced, along with the residual plots used to assess the assumptions for linear regression.

The material in this lab corresponds to Section 6.1, 6.2, and 6.3.1 of *OpenIntro Biostatistics*.

Introduction

Least squares regression

The vertical distance between a point in the scatterplot and the predicted value on the regression line is the **residual** for the point. For an observation (x_i, y_i), where $\hat{y}_i$ is the predicted value according to the line $\hat{y} = b_0 + b_1 x$, the residual is the value $e_i = y_i - \hat{y}_i$.

The **least squares regression line** is the line which minimizes the sum of the squared residuals (SSE) for all the points in the plot; i.e., the regression line is the line that minimizes $e_1^2 + e_2^2 + ... + e_n^2$ for the n pairs of points in the dataset.[1]

For a general population of ordered pairs (x, y), the population regression model is

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is a normally distributed 'error term' with mean 0 and standard deviation σ.

The terms β_0 and β_1 are parameters with estimates b_0 and b_1. These estimates can be calculated from summary statistics: the sample means of x and y ($\overline{x}$ and $\overline{y}$), the sample standard deviations of x and y (s_x, s_y), and the correlation between x and y (r).

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \overline{y} - b_1\overline{x}$$

[1] SSE stands for "sum of squared errors" and refers to the sum of squared residuals.

Plots for checking assumptions

There are a variety of **residual plots** used to check the fit of a least squares line. The ones used in this textbook are scatterplots in which predicted values are on the x-axis and residual values on the y-axis. Residual plots are useful for checking the assumptions of linearity and constant variability.

To assess the normality of residuals, **normal probability plots** are used. These plots are also known as quantile-quantile plots, or Q-Q plots.

Background information

This lab uses data from the Prevention of REnal and Vascular END-stage Disease (PREVEND) study, which took place between 2003 and 2006 in the Netherlands. Clinical and demographic data for 4,095 individuals are stored in the `prevend` dataset in the `oibiostat` package.

As adults age, cognitive function declines over time; this is largely due to various cerebrovascular and neurodegenerative changes.

The Ruff Figural Fluency Test (RFFT) is one measure of cognitive function that provides information about cognitive abilities such as planning and the ability to switch between different tasks. Scores on the RFFT range from 0 to 175 points, where higher scores are indicative of better cognitive function.

The goal of this lab is to begin exploring the relationship between age and RFFT score in the `prevend` dataset.

Fitting and interpreting a line

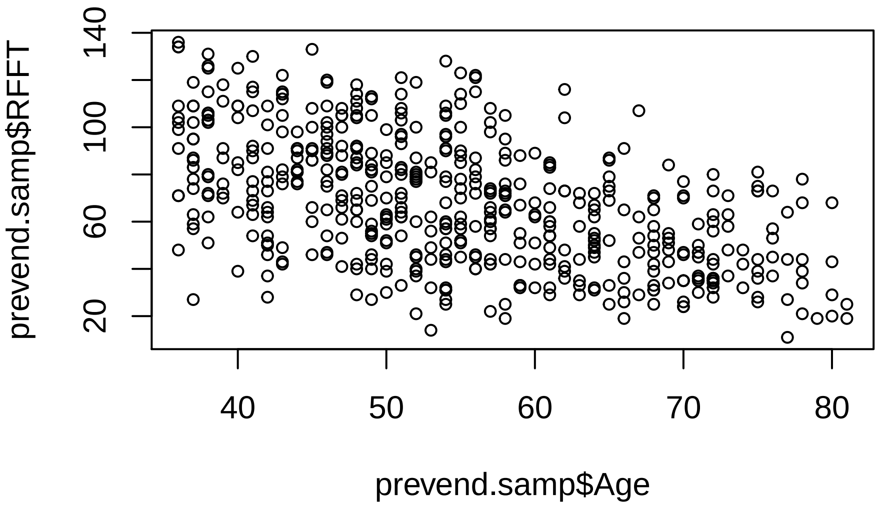
The questions in this lab will be based around data from a random sample of $n = 500$ individuals from the `prevend` dataset; the sample is stored as `prevend.samp` in the `oibiostat` package.

1. Run the following code chunk to load the `prevend.samp` dataset.

```r
#load the data
library(oibiostat)
data("prevend.samp")
```

2. Create a scatterplot of RFFT score (`RFFT`) and age in years (`Age`) in `prevend.samp`.

```r
#create a plot of RFFT versus age
plot(prevend.samp$RFFT ~ prevend.samp$Age, cex = 0.75)
```



3. Examine the plot and consider possible lines that are a reasonable approximation for the relationship in the plot.

   a) Consider the line $\hat{y} = -20 + 2x$.

      i. Using the code provided in the template, add the line to the plot. Does the line appear to be a good fit to the data?

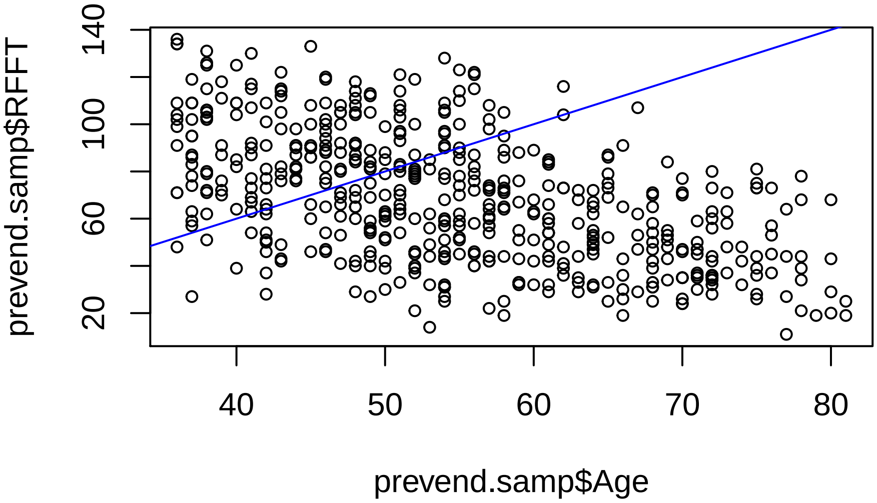
         No, the line does not appear to be a good fit to the data. The general trend in the data is a negative relationship, while the line has a positive slope.

      ii. Calculate the SSE, the sum of the squared residuals, for this line. Do you expect this SSE to be relatively low or relatively high? Explain your answer.

         Based on the answer in part i., the SSE should be relatively high, indicating a poor model fit and high amount of error (from large residuals) associated with the model. The SSE is 1,206,875.

```r
#enter line coefficients
b0 = -20
b1 = 2

#plot the data and line
plot(prevend.samp$RFFT ~ prevend.samp$Age, cex = 0.75)
abline(b0, b1, col = "blue")
```



```
#calculate sse
y = prevend.samp$RFFT
x = prevend.samp$Age

sse = sum((y - (b0 + b1*x))^2)
sse
```

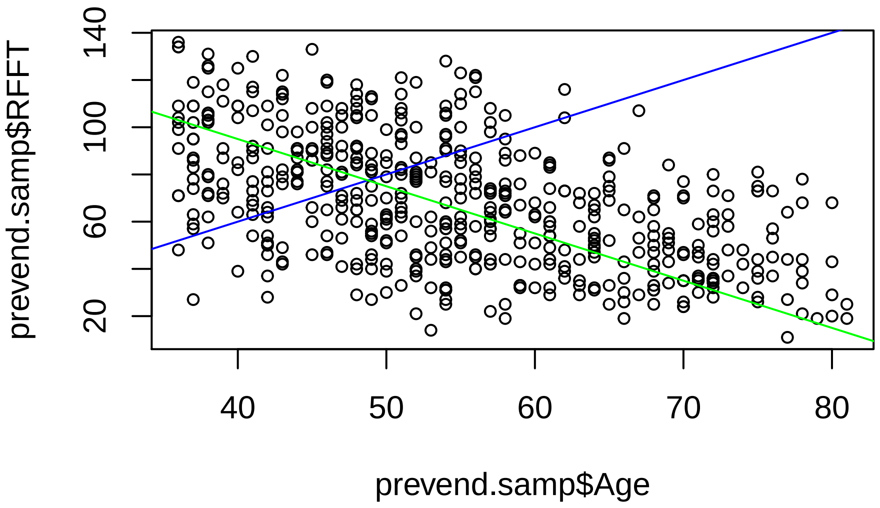
```
## [1] 1206875
```

b) From a visual inspection, determine a line that you think is a good fit to the data and add the line to the plot. Calculate the SSE and compare it to the SSE from the line in part a).

One potential reasonable line is $\hat{y} = 175 - 2x$. This line, shown in green, generally passes through the center of the downward sloping point cloud. The SSE of this line is 309,001, which is much lower than the SSE of the line from part a).

```
#enter line coefficients
b0.new = 175
b1.new = -2

#plot the data and lines
plot(prevend.samp$RFFT ~ prevend.samp$Age, cex = 0.75)
abline(b0, b1, col = "blue")
abline(b0.new, b1.new, col = "green")
```



```
#calculate sse
sse.new = sum((y - (b0.new + b1.new*x))^2)
sse.new
```

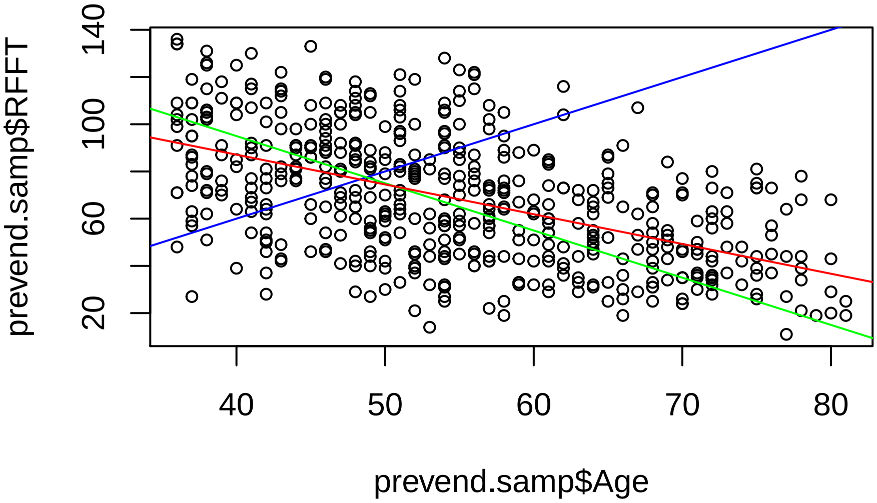
```
## [1] 309001
```

c) Consider the line $\hat{y} = 137.55 - 1.261x$. Add this line to the plot. Calculate the SSE and compare it to the SSE from the line in part b).

This line, shown in red on the plot, has SSE 26,771.9, which is lower than the SSE of the line from part b). In fact, any other line fit to the data must have SSE larger than this line—this is the least squares model, in which b_0 and b_1 have been chosen to minimize the sum of squared residuals.

```
#enter line coefficiens
b0.model = 137.55
b1.model = -1.261

#plot the data and lines
plot(prevend.samp$RFFT ~ prevend.samp$Age, cex = 0.75)
abline(b0, b1, col = "blue")
abline(b0.new, b1.new, col = "green")
abline(b0.model, b1.model, col = "red")
```

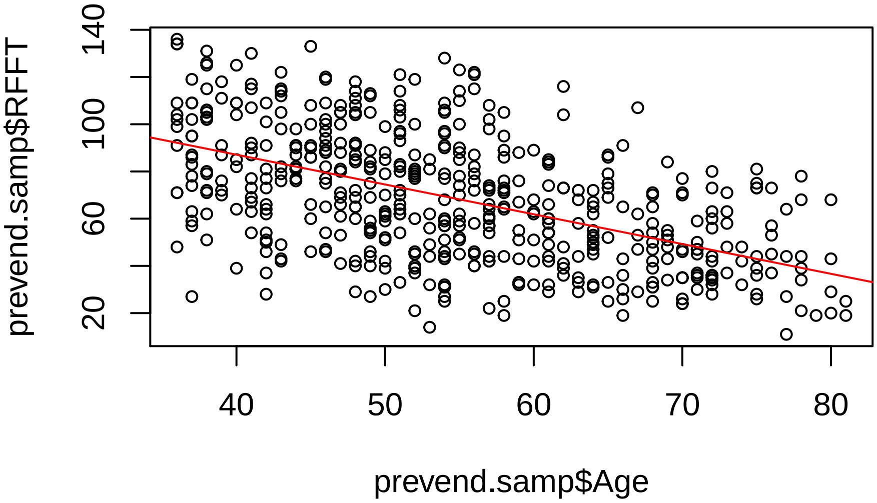


```
#calculate sse
sse.model = sum((y - (b0.model + b1.model*x))^2)
sse.model
```

```
## [1] 267771.9
```

4. Create a scatterplot of RFFT score (`RFFT`) and age in years (`Age`) in `prevend.samp`, then add a line of best fit.

```
#create a plot of RFFT versus age with line of best fit
plot(prevend.samp$RFFT ~ prevend.samp$Age, cex = 0.75)
abline(lm(prevend.samp$RFFT ~ prevend.samp$Age), col = "red")
```



```
#print the equation of the line
lm(prevend.samp$RFFT ~ prevend.samp$Age)
```

```
##
## Call:
## lm(formula = prevend.samp$RFFT ~ prevend.samp$Age)
##
## Coefficients:
##      (Intercept)  prevend.samp$Age
##          137.550            -1.261
```

a) What are the slope and intercept values of the line of best fit?

The slope is -1.261 and the intercept is 137.55.

b) Interpret the slope and intercept values in the context of the data; i.e., explain the linear model in terms that a non-statistician would understand. Comment on whether the intercept value has any interpretive meaning in this setting.

According to the model slope, an increase in age of 1 year is associated with an average decrease in RFFT score by 1.3 points. The model intercept suggests that the average RFFT score for an individual of age 0 is 137.55. The intercept value does not have interpretive meaning in this setting, since it is not reasonable to assess a newborn's cognitive function with a test like the RFFT.

c) Based on the linear model, how much does RFFT score differ, on average, between an individual who is 60 years old versus an individual who is 50 years old?

Based on the linear model, RFFT score differs, on average, by 1.261(10) = 12.61 points for individuals that have an age difference of 10 years; the younger individual is expected to have a higher RFFT score.

d) Write the equation of the least squares line in the form $\hat{y} = b_0 + b_1 x$. According to the linear model, what is the average RFFT score for an individual who is 70 years old?

The equation of the line is $\widehat{RFFT} = 137.55 - (1.261)(age)$. According to the model, the average RFFT score for an individual who is 70 years old is $137.55 - (1.261)(70) = 49.28$ points.

e) Is it valid to use the linear model to estimate RFFT score for an individual who is 20 years old? Explain your answer.

No, it is not valid to use the linear model to estimate RFFT score for an individual who is 20 years old. The data in the sample only extend from ages 36 to 81 and should not be used to predict scores for individuals outside that age range. It may well be that the relationship between RFFT and age is different for individuals in a different age group.

```r
summary(prevend.samp$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36.00   46.00   54.00   54.82   64.00   81.00
```

Checking assumptions with residual plots

There are four assumptions that must be met for a linear model to be considered reasonable: linearity, constant variability, independent observations, and normally distributed residuals.

Even though linearity and constant variability can be assessed from the scatterplot of y versus x, it is helpful to consult residual plots for a clearer view. Normality of residuals is best assessed through a normal probability plot; although skew can be visible from a histogram of the residuals, deviations from normality are more obvious on a normal probability plot.

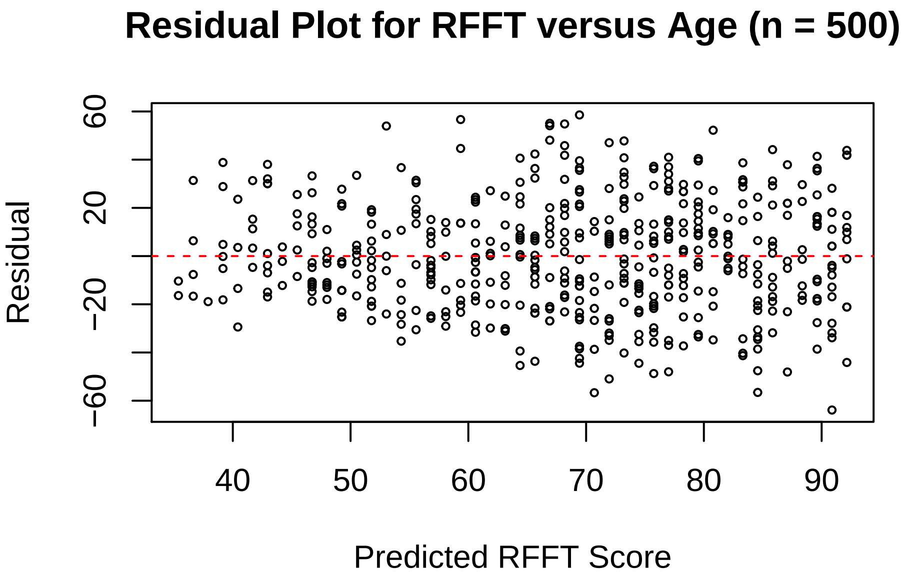
RFFT and age in the `prevend` *data*

5. Run the following chunk to create a residual plot where the residual values are plotted on the y-axis against predicted values from the model on the x-axis, using data in `prevend.samp`.

```r
#store the residuals from the linear model
prevend.residuals = residuals(lm(RFFT ~ Age, data = prevend.samp))

#store the predicted RFFT scores from the linear model
prevend.predicted = predict(lm(RFFT ~ Age, data = prevend.samp))

#create residual plot
plot(prevend.residuals ~ prevend.predicted,
 cex = 0.5,
 main = "Residual Plot for RFFT versus Age (n = 500)",
 xlab = "Predicted RFFT Score",
 ylab = "Residual")
abline(h = 0, col = "red", lty = 2)
```



a) When a linear model is a good fit for the data, the residuals should scatter around the horizontal line $y = 0$ with no apparent pattern. Does a linear model seem appropriate for these data?

Yes, a linear model seems appropriate; the residuals scatter about the horizontal line $y = 0$ with no apparent pattern. There is a roughly equal number of points above the line as below it, which indicates that the line goes through the center of the 'cloud' of data points.

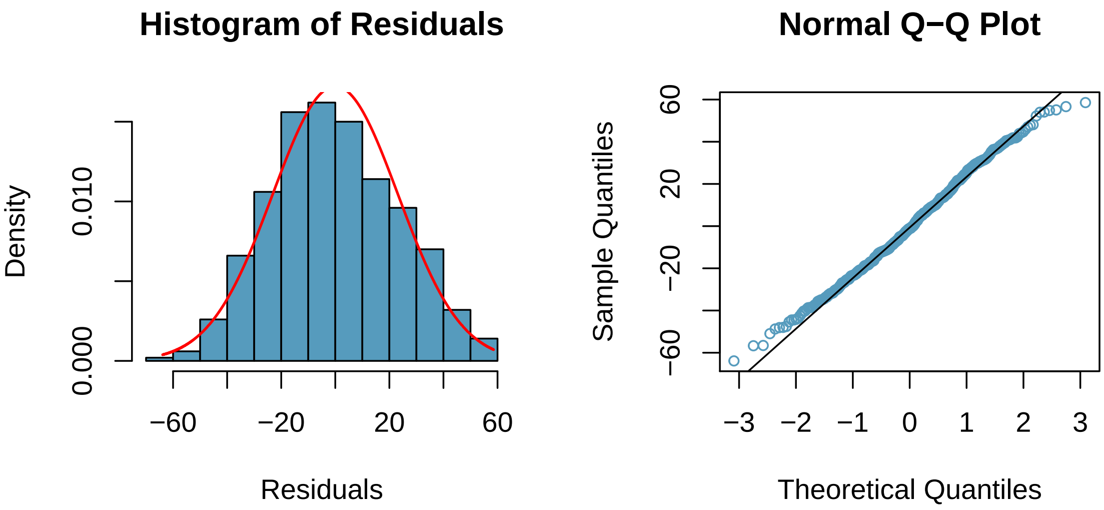
b) Does the variability of the residuals seem constant across the range of predicted RFFT scores?

The variability are generally constant across predicted RFFT scores between 60 to 90 points, but there seems to be less variability for lower predicted scores.

6. Run the code chunk shown in the template to create a normal probability plot of the residuals. For comparison purposes, the following figure shows a histogram of the residual values overlaid with a normal curve and the normal probability plot.

   Do the residuals appear to be normally distributed?

   Yes, the residuals follow a straight line on the Q-Q plot, with only slight deviations from normality in the tails. This is also visible in the histogram of the residuals with an overlaid normal curve.



7. Overall, does it seem that a least squares regression line is an appropriate model for estimating the relationship between cognitive function (as measured by RFFT score) and age?

   Overall, a least squares regression line seems appropriate. There is some nonconstant variability for lower predicted RFFT scores that suggest caution should be exercised when using the model to conduct inference about older individuals (since older individuals are predicted to have lower scores).[2] At this point, it is more important to note that data almost

[2] Inference for regression will be discussed in a later lab.

never perfectly satisfy model assumptions; the learning goal here is to understand the mechanics of assumptions checking and to be able to identify especially severe violations.

Clutch volume and body size in the `frog` *data*

The `frog` dataset in the `oibiostat` package contains observations from a study conducted on a frog species endemic to the Tibetan Plateau. Researchers collected measurements on egg clutches and female frogs found at breeding ponds across five study sites.

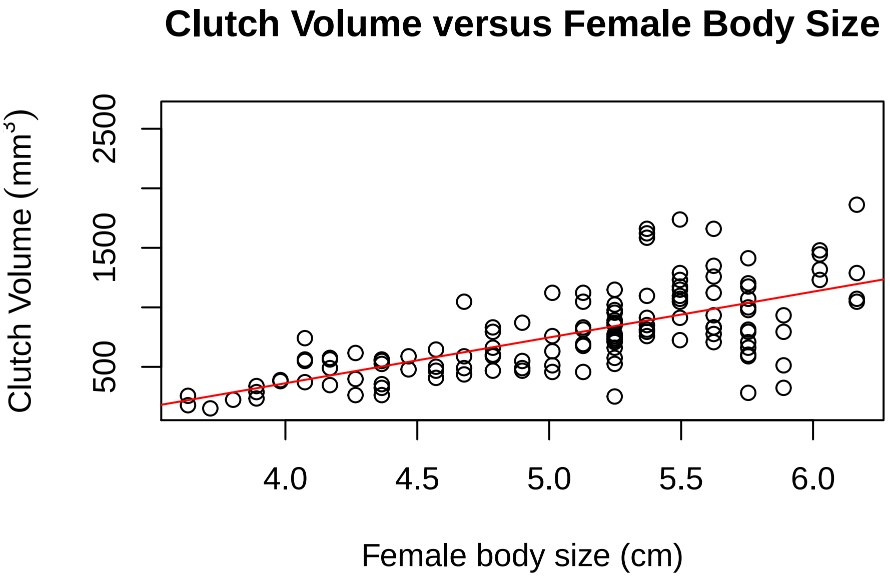
Previous research suggests that larger body size allows females to produce egg clutches with larger volumes. Frog embryos are surrounded by a gelatinous matrix that may protect developing embryos from temperature fluctuation or ultraviolet radiation; a larger matrix volume provides added protection. In the data, clutch volume (`clutch.volume`) is recorded in cubic millimeters and female body size (`body.size`) is measured as length in centimeters.

The following questions step through examining whether a linear regression model is appropriate for the relationship between female body size and clutch volume.

8. Create a scatterplot of clutch volume versus female body size and plot the least squares line.

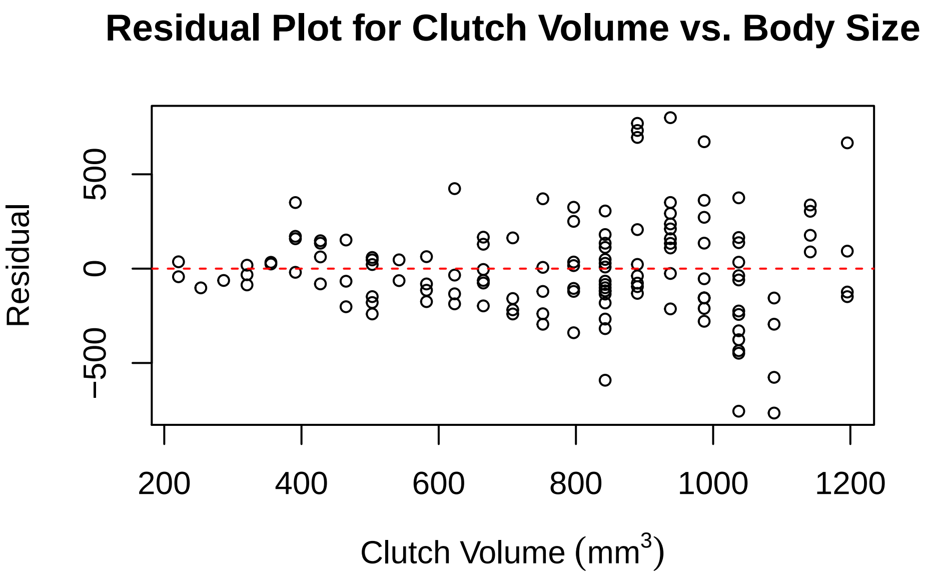
```
#load the data
data("frog")

#create a scatterplot and add a regression line
plot(frog$clutch.volume ~ frog$body.size,
 main = "Clutch Volume versus Female Body Size",
 xlab = "Female body size (cm)", ylab = expression("Clutch Volume" ~ (mm^3)))
abline(lm(frog$clutch.volume ~ frog$body.size), col = "red")
```



9. Create a residual plot where the residual values are plotted on the y-axis against predicted values from the model on the x-axis.

```r
#create residual plot
frog.model = lm(clutch.volume ~ body.size, data = frog)
plot(resid(frog.model) ~ predict(frog.model),
 main = "Residual Plot for Clutch Volume vs. Body Size",
 ylab = "Residual",
 xlab = expression("Clutch Volume" ~ (mm^3)),
 cex = 0.75)
abline(h = 0, col = "red", lty = 2)
```



a) Does the linearity assumption seem to be satisfied?

Linearity appears to be satisfied; there is no apparent pattern in the residuals and there is a roughly equal number of points above and below the $y = 0$ line.

b) Is the variability of the residuals constant across the range of predicted clutch volumes?

No; the residual plot makes it easier to see that the residuals are more variable for larger values of predicted clutch volume than smaller values.

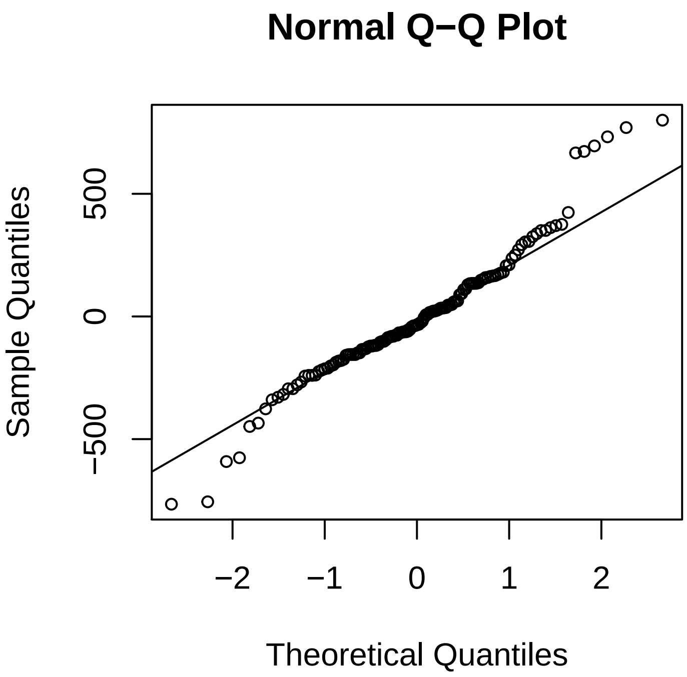
10. Assess whether it can be reasonably assumed that the observations are independent.

The data for body size and clutch volume are from frogs (and egg clutches) found at breeding ponds at five different study sites. In the absence of more specific biological information about the life cycle of this frog and its breeding habits, it seems reasonable to assume that the values of one pair (clutch volume and body size) provide no information about another pair. It would not be reasonable to assume independence if, for example, frogs found at a specific study site are more likely to be related to each other and that related frogs tend to have similar body sizes and/or clutch volumes.

11. Create a Q-Q plot and assess whether the residuals appear to be normally distributed.

    The majority of the residuals closely follow a normal distribution, but there exists some deviation from normality in both tails.

```
#create q-q plot
qqnorm(resid(frog.model), cex = 0.75)
qqline(resid(frog.model))
```



12. Evaluate whether a least squares regression line is an appropriate model for estimating the relationship between female body size and clutch volume.

    The least squares line seems to be a reasonable fit for the data such that predictions about average clutch volume can be made based on female body size. However, since the constant variability assumption is violated, caution is required when using the model to conduct inference; this concern will be discussed further once inference for regression is covered.

    The main point of showing the model with the `frog` data is to highlight the fact that data tend to be messy—the previous regression using the PREVEND data actually represent a case where data fit model assumptions surprisingly well. When assumptions are violated, the question of how to proceed is often a subtle one, requiring topics that would be covered in a specialized course on regression.