Correlation and Simple Linear Regression

We are often interested in studying the relationship among variables to determine whether they are associated with one another. When we think that changes in a variable *x* explain, or maybe even cause, changes in a second variable *y*, we call *x* an explanatory variable and *y* a response variable. If the form of the plot looks like a straight line, this indicates there may be a linear relationship between the two variables. The relationship is strong if all the data points approximately make up a straight line and weak if the points are widely scattered about the line.

The covariance and correlation are measures of the strength and direction of a linear relationship between two quantitative variables. A regression line is a mathematical model for describing a linear relationship between an explanatory variable, *x*, and a response variable, *y*. It can be used to predict the value of *y* for a given value of *x*.

Covariance, correlations and regression lines can all be computed using R.

A. Covariance and correlation

We can compute the covariance and correlation in R using the cov() and cor() functions.

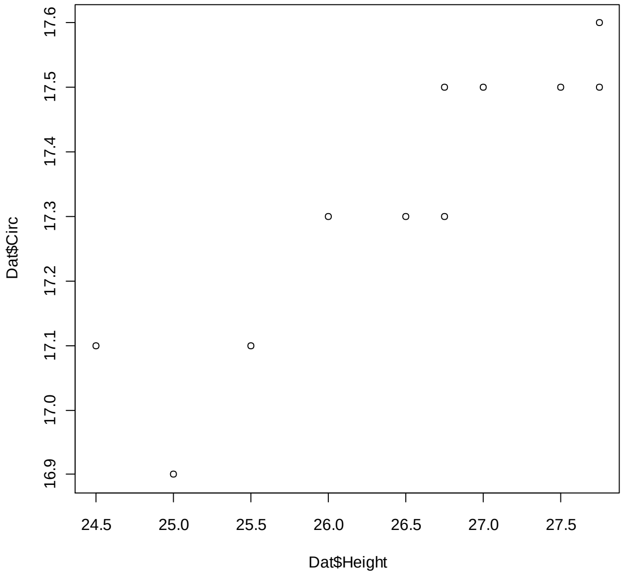
Ex. A pediatrician wants to study the relationship between a child’s height and their head circumference (both measured in inches). She selects a SRS of 11 three-year old children and obtains the following data. (See lecture notes for data)

Begin by reading in the data:

```
> Height = c(27.75, 24.5, 25.5, 26, 25, 27.75, 26.5, 27, 26.75, 26.75, 27.5)
> Circ = c(17.5, 17.1, 17.1,17.3, 16.9, 17.6, 17.3, 17.5, 17.3, 17.5, 17.5)
> Dat = data.frame(Height,Circ)
> attach(Dat)
> Dat
   Height Circ
1   27.75 17.5
2   24.50 17.1
3   25.50 17.1
…
8   27.00 17.5
9   26.75 17.3
10  26.75 17.5
11  27.50 17.5
```

To make a scatter plot of circumference against height type:

```
> plot(Dat$Height,Dat$Circ)
```



Studying the plot, there appears to be a linear relationship between the two variables.

This relationship can be quantified by computing the covariance and correlation between variables.

```
> cov(Dat)     # Covariance matrix
           Height        Circ
Height  1.1977273   0.21886364
Circ    0.2188636   0.04818182
```

From the output we see that the variance of Height and Circ is 1.198 and 0.048, respectively. The covariance between the two variables is 0.219 indicating a positive relationship.

```
> cor(Dat)               # Correlation matrix
           Height        Circ
Height  1.0000000    0.9110727
Circ    0.9110727    1.0000000
```

From the output we see that the correlation between Height and Circ is 0.911. Hence, the positive linear relationship between the variables is quite strong.

B. Simple Linear Regression

If there exists a strong linear relationship between two variables it is often of interest to model the relationship using a regression line. The main function for performing regression in R is lm(). It has many options that we will explore throughout the semester.

To perform simple linear regression we can use the command:

```
lm(response ~ explanatory)
```

Here the terms *response* and *explanatory* in the function should be replaced by the names of the response and explanatory variables, respectively, used in the analysis.

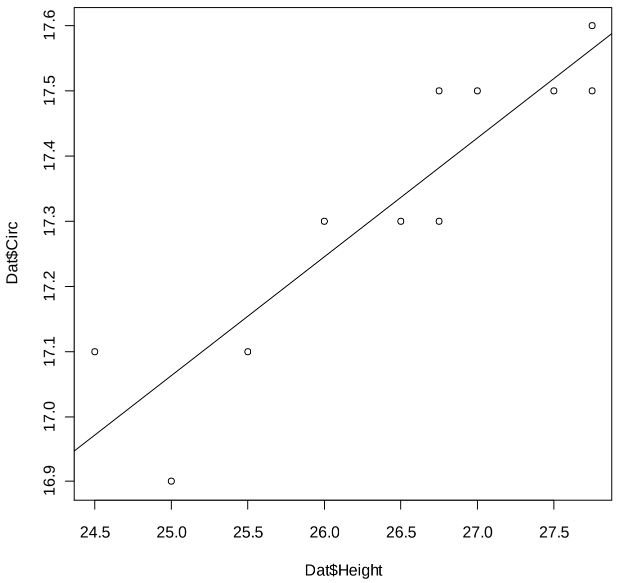
Ex. Fit a regression line that describes the relationship between Height and Circumference.

```
> results = lm(Circ ~ Height)
> results
Call:
lm(formula = Circ ~ Height)
Coefficients:
(Intercept)      Height
    12.4932      0.1827
```

The results indicate that the least squares regression line takes the form: yhat = 12.493 + 0.183x. Hence the model states that a one inch increase in height would lead to a 0.183 inch increase in head circumference.

To superimpose the regression line over the data first make a scatter plot of Circ against Height, and therefore overlay the regression line using the command abline(results). Here *results* contains all relevant information about the regression line.

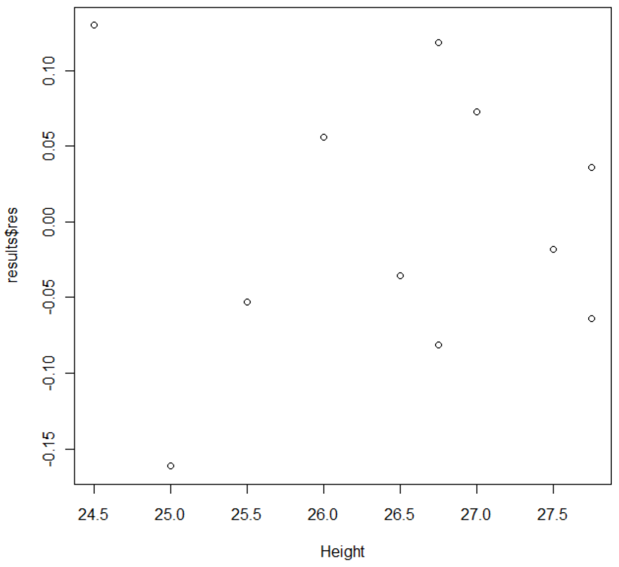
```
> plot(Height,Circ)
> abline(results)
```



The next step in our analysis is to verify all the relevant model assumptions needed for using the simple linear regression model. The residuals should be normally distributed with equal variance for every value of x. We can check the model assumptions by making appropriate plots of the residuals. Note that after fitting the model, the residuals are saved in the variable results$res.

We begin by plotting the residuals against the explanatory variable. The residuals should be randomly scattered about 0. The width should be equal throughout for the constant variance assumption to hold.

```
> plot(Height,results$res)
```

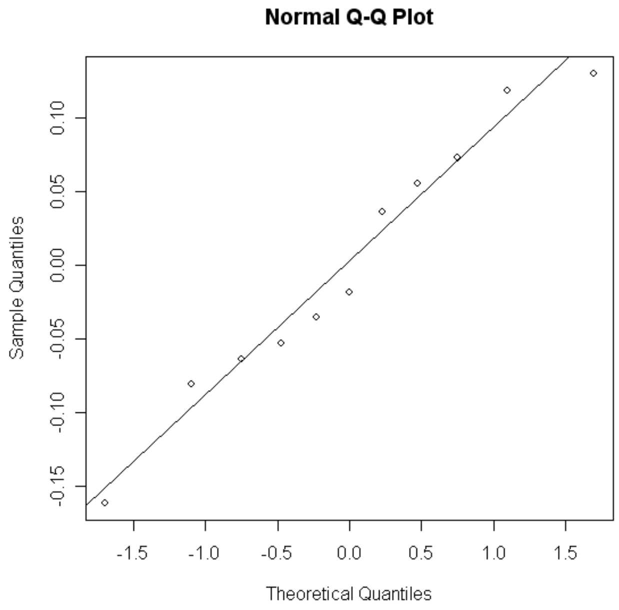


This plot shows no apparent pattern in the residuals indicating no clear violations of any model assumptions.

To check the normality assumption, make a QQ-plot using the command:

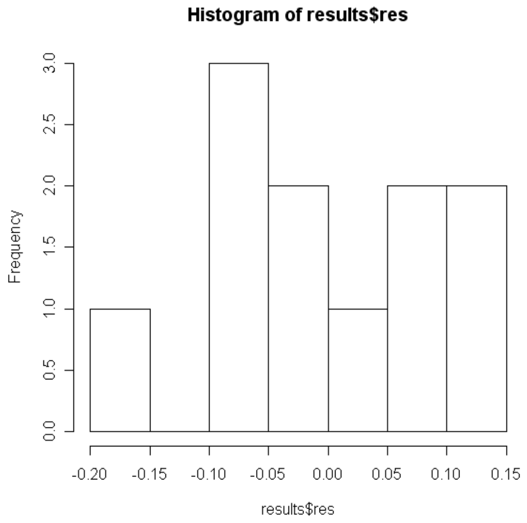
```
> qqnorm(results$res)
> qqline(results$res)
```

This makes a QQ-plot and overlays a straight line for comparison purposes.



To make a histogram of the residuals type:

```
> hist(results$res)
```



After verifying the assumptions, the next step is to perform inference. We want to construct tests and confidence intervals for the slope and intercept, confidence intervals for the mean response and prediction intervals for future observations.

To test whether the slope is significantly different from 0 we can use the function summary(results).

```
> summary(results)
Call:
lm(formula = Dat$Circ ~ Dat$Height)
Residuals:
     Min        1Q       Median       3Q         Max
-0.16148 -0.05842 -0.01831  0.06442  0.12989
Coefficients:
                  Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)    12.49317    0.72968   17.12    3.56e-08 ***
Dat$Height   0.18273     0.02756    6.63      9.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.09538 on 9 degrees of freedom
Multiple R-Squared: 0.8301,     Adjusted R-squared: 0.8112
F-statistic: 43.96 on 1 and 9 DF, p-value: 9.59e-05
```

From the output we see that we can clearly reject the null hypothesis of no linear relationship between height and head circumference (p=9.59e-05). The term ‘Residual standard error’ gives us an estimate of the standard deviation around the regression line, or s, which is equal to 0.09538.

To construct 95% confidence intervals for β_1 we can use the command confint(results).

```
> confint(results)
                      2.5 %     97.5 %
(Intercept) 10.8425070 14.1438307
Height        0.1203848   0.2450801
```

Finally, we may want to use our regression equation to predict future values of the response variable. The predicted value of head circumference for a child of a given height has two interpretations; it can either represent the mean circumference for all children whose height is x, or it can represent the predicted circumference for a randomly selected child whose height is x. The predicted value will be the same in both cases. However, the standard error will be larger in the second case due to the additional variation of individual responses about the mean.

The function predict() can be used to make both types of intervals. To make confidence intervals for the mean response use the option interval=”confidence”. To make a prediction interval use the option interval=”prediction”.

Ex. Obtain a 95% confidence intervals for the mean head circumference of children who are 25 inches tall.

```
> predict(results,data.frame(Height =25),interval="confidence")
          fit      lwr      upr
[1,] 17.06148 16.94987 17.17309
```

The confidence interval lies in the range (16.95, 17.17).

Ex. Obtain a 95% prediction intervals for a child who is 25 inches tall.

```
> predict(results,data.frame(Height =25),interval="prediction")
          fit      lwr      upr
[1,] 17.06148 16.81855 17.30441
```

The prediction interval lies in the range (16.82, 17.30). Note that it is slightly wider than the confidence interval.