First, download the datasets as follows

```
> f = "http://www.math.csi.cuny.edu/verzani/R/regression-inference.R"
> source(url(f))
```

Suppose we want to investigate a linear relationship between `x`, a predictor, and `y` a response variable. The basics of regression involved the following commands.

```
> plot(y ~ x)                  # scatterplot
> res = lm(y~x)                # fit a linear model
                               # y = b_0 + b_1 x + error
> abline(res)                  # add regression line
> summary(res)                 # show a detailed summary
> coef(res)                    # the regression coefficients
> resid(res)                   # the residuals
```

1 Is the regression model correct?

In this project, we want to develop the skills to investigate if a linear model is a good fit to the data, and then make inferences about the estimates.

The basic simple regression model is that the data (x_i, y_i) are related by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where ε_i are normal with mean 0 and unknown variance σ^2. Alternately, this means that conditionally on the value of x_i the y_i are normal with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.

To see if the model is appropriate for the data we should have

- Data which shows a linear trend
- Residuals which should show no trend
- Residuals which are approximately normally distributed

To check if the data shows a linear trend a scatterplot will suffice.

The residuals are the values

$$y_i - \hat{y}_i.$$

As they sum to 0 they are not independent. As well, they do not have a common variance. However, for our purposes we will assume they are – if the model is true! Thus to check if the errors, ε_i, are i.i.d. normal with mean 0 and variance σ^2 we should check the following

independence There should be no apparent trend in the data when you plot the fitted (the $\hat{y}_i$) versus the residuals. This plot is done with

```
> plot(fitted(res),resid(res))
```

If there is a trend, then the model isn't right.

Common variance The same plot can show if the variances are the same. If you see a spreading of the points, then the variances are not the same. This can indicate a need to transform the variables first.

Normality You can check normality with a quantile plot of the residuals

```
> qqnorm( resid(res) )
```

For example, suppose we have the data

```
> x = c(1,1,2,2,3)
> y = c(1,2,1,2,3)
```

The scatterplot is made with

```
> plot(y ~ x)                          # or plot(x,y)
```

The residuals plots are made with

```
> res = lm(y ~ x)
> plot(fitted(res), resid(res))
```

This shows no trends so the assumption of independence and common variance seems okay.

Finally, a normal plot is done with

```
> qqnorm(resid(res))
```

This data set is so small it is hard to assess, but there is no reason to doubt the normality.

These plots (and two others) are made when the `plot()` command is called on the results of `lm()`.

1.1 Problems

For the following, you will first need to `attach()` the data set, or use the optional argument to `plot()` and `lm()`, `data = dataframe`.

1.1 The dataset `twins` contains 3 variables from a twin study. The twins were separated at birth with one in a foster home, the other in their parents. For each pair the IQ scores are given and the socioeconomic strata of the biological parents.

Investigate whether the relationship between the IQ's of the Foster and Biological twin is linear. Check all of the above and report on the answers.

1.2 The dataset `deflection` contains two variables – the deflection on a wire for a given load. Check if the linear model seems appropriate using the graphs above. What do you conclude?

1.3 The dataset `mhr` has simulated data on the maximum heart rate for various ages. Check to see if a linear model is appropriate. As both the variable and the dataframe have the same name, here you must use the `data = mhr` argument as in

```
> plot(mhr ~ age, data = mhr)
```

1.4 The dataset `too.young` has two variables. For each male age, people were surveyed for the youngest possible female age for a romantic partner. Is a linear model appropriate for this data? Explain.

2 Regression inference

If you are satisfied that the regression model is correct, then one can draw statistical inferences about the predictors $\hat{\beta}_0$ and $\hat{\beta}_1$ as well as the variance estiamate. For the $\hat{\beta}$'s the fact that under the assumption that the model is correct, the sampling distribution of

$$\frac{\hat{\beta} - \beta}{\mathsf{SE}(\hat{\beta})}$$

is a t-distribution with $n-2$ degres of freedom.

To find the standard errors, these formulas can be used. ($\hat{\sigma}^2$ is $(1/(n-2))\sum(y_i - \hat{y}_i)^2$.)

$$\mathsf{SE}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i - \bar{x})^2}},$$

$$\mathsf{SE}(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}.$$

Or, these can be found in the output of the `summary()` command.

With these we can do a signficance test. For example. Suppose our data is given below. And we want to test if $\beta_1 = 1.$ against a two-sided alternative. We do so as follows. First the data

```
> x = c(1,1,2,2,3)
> y = c(1,2,1,2,3)
```

A plot shows that the line with slope 1 is reasonable, but the output of the linear model shows it is not the correct estimate

```
> plot(x,y)
> res = lm(y ~ x)
> summary(res)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5
-0.2857  0.7143 -0.9286  0.0714  0.4286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.643      0.862    0.75     0.51
x              0.643      0.442    1.45     0.24

Residual standard error: 0.74 on 3 degrees of freedom
Multiple R-Squared: 0.413,      Adjusted R-squared: 0.218
F-statistic: 2.11 on 1 and 3 DF,  p-value: 0.242
```

Notice, the estimated value, $\hat{\beta}_1 = .643$, the standard error is 0.442 and there are $3 = n-2$ degrees of freedom. How to do a test

$$H_0 : \beta_1 = 1, \quad H_A : \beta_1 \neq 1?$$

The test statistic is

$$\frac{\hat{\beta}_1 - \beta_1}{\mathsf{SE}}$$

which has observed value

```
> SE = .442
> obs = (0.643 - 1)/SE
> obs
[1] -0.8077
```

A two-sided *p*-value is given by

```
> 2*pt(obs, df = 5 - 2)
[1] 0.4784
```

So the null hypothesis is not rejected.

2.1 For the heart rate data, do a significance test of the hypotheses

$$H_0 : \beta_0 = 220, \quad H_A : \beta_0 < 220.$$

2.2 For the heart rate data, `mrh`, do a significance test of the hypotheses

$$H_0 : \beta_1 = -1, \quad H_A : \beta_1 \neq -1.$$

2.3 For the twin data, Use `Foster` for a predictor and `Biological` for the response. Do a significance test of the hypotheses

$$H_0 : \beta_1 = 1, \quad H_A : \beta_1 \neq 1.$$