Simple Linear Regression (SLR) and Multiple Linear Regression (MLR)

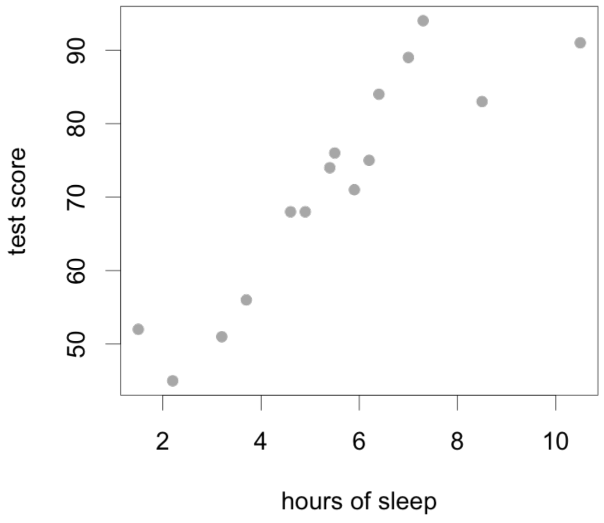
2018-02-01
Andres Kroker

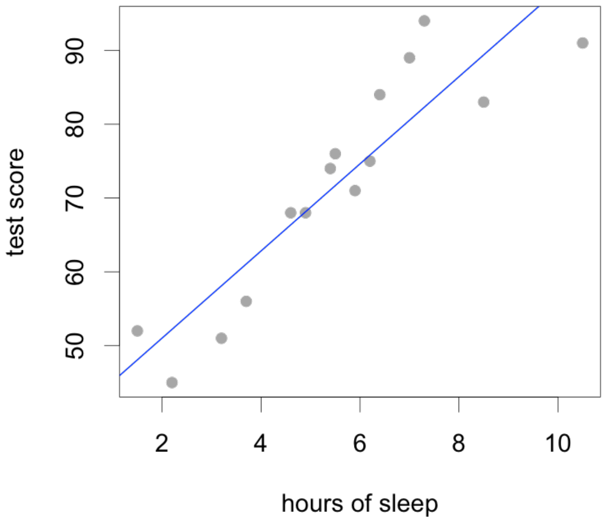
Let's assume we have the following data:

- Test scores
- Hours of sleep the night before
- The liters of coffee drunk the morning of the exam

Score [%]	Sleep [h]	Coffee [l]
68	4.6	0.3
83	8.5	0.1
51	3.2	0.12
76	5.5	0.32
89	7.0	0.14
75	6.2	0.19
84	6.4	0.23
71	5.9	0.28
52	1.5	0.22
91	10.5	0.13
94	7.3	0.22
74	5.4	0.31
68	4.9	0.08
56	3.7	0.23
45	2.2	0.1

First, let's learn if **sleep** and test **score** are related. Here is what the data looks like. Naturally, we (or at least I) want to place a (regression) line through it.





What are regression models?

The general equation for a line is:

$$y = mx + c$$

- y and x are variables
- m (slope) and c (intercept) are coefficients

The regression line is described in the same way:

$$\hat{Y}_i = \hat{b}_1 X_i + \hat{b}_0$$

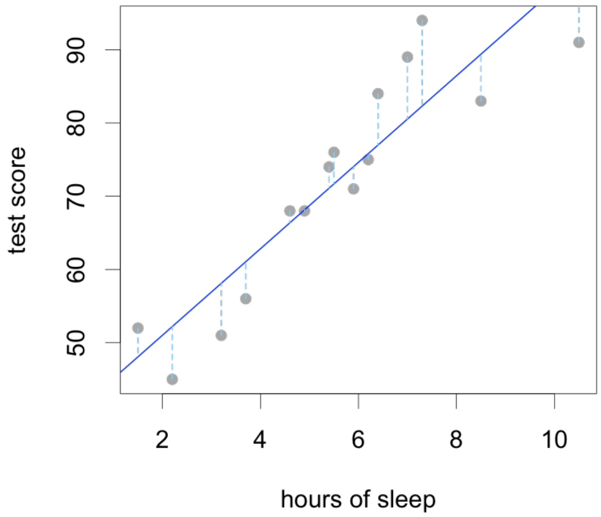
- X_i is the value of the predictor variable of the i$^{\text{th}}$ observation
- Y_i is the value of the response variable corresponding to the i$^{\text{th}}$ observation
- $\hat{Y}_i$ is the estimate of Y_i
- $\hat{b}_1$ is the estimate of slope (change in $\hat{Y}$if X increased by 1 unit)
- $\hat{b}_0$ is the estimate of intercept (value of $\hat{Y}$if $X \ = 0$)

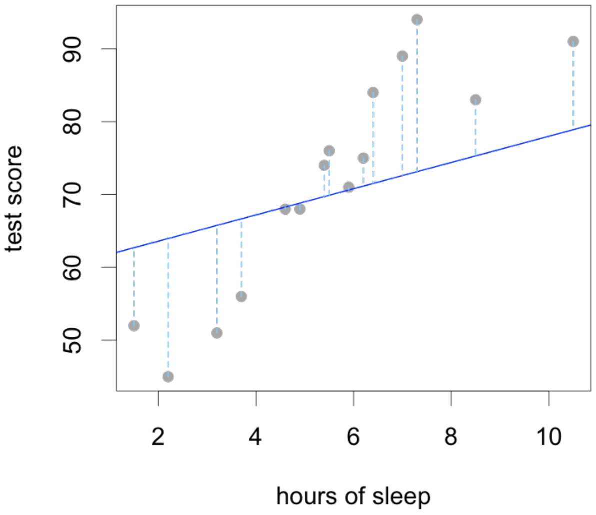
However, clearly no all the data points fall onto the line. Thus, the model we are trying to fit to the data needs to account for this deviation from response variable and it's estimate

$$\text{Residuals: } \varepsilon_i = Y_i - \mu_{i|X}$$
$$\text{Model: } Y_i = \ b_1 X_i + b_0 + \varepsilon_i$$

We do not know the above population parameters, but we can estimate them

$$\text{Estimated residuals: } \epsilon_i = Y_i - \hat{Y}_i$$
$$\text{Estimated/fitted model: } \hat{Y}_i = \ \hat{b}_1 X_i + \hat{b}_0 \text{ that minimized } SS_R = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$





$$\hat{b}_1 = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}$$

…or you can let R (SPSS/STATA) do the math for you…

Here is what R returns to us:

```
> lm(score ~ sleep, data=data)

Call:
lm(formula = score ~ sleep, data = data)

Coefficients:
(Intercept)        sleep
     39.221        5.902
```

This regression model can easily be extended to more than 1 predictor variable. In our case let's add **coffee** consumption to the model:

$$\text{Model: } Y_i = b_2X_{i.2} + b_1X_{i.1} + b_0 + \varepsilon_i$$
$$\text{Estimated/fitted model: } \hat{Y}_i = \hat{b}_2X_{i.2} + \hat{b}_1X_{i.1} + \hat{b}_0$$

Estimating the coefficients just became a more complicated unless you are comfortable with matrices. So, let's just skip the math and go straight to R:

```
> lm(score ~ sleep + coffee, data=data)

Call:
lm(formula = score ~ sleep + coffee, data = data)

Coefficients:
(Intercept)        sleep       coffee
     31.523        6.038       35.088
```

Quantifying the fit of a regression model

We now estimated our coefficients! But we don't know if the model fits the data well…this is where the R^2-value comes in handy.

We want to get a feeling for how much of the variance in the data is explained by the model…which sounds oddly familiar from our discussion on ANOVA. So, let's recap:

$$SS_T = SS_M + SS_R$$

we know that:

$$SS_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

and

$$SS_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

So, we can calculate the model sum of squares as follows

$$SS_M = SS_T - SS_R$$

And, therefore, we can calculate R^2as follows

$$R^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_R}{SS_T}$$

Let’s take a look at how our model is performing: $R^2 = 0.857$

```
> model2 = lm(score ~ sleep + coffee, data = data)
> summary(model2)

Call:
lm(formula = score ~ sleep + coffee, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-8.483 -3.705 -1.005  3.892 10.681 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.5235     6.1744   5.106 0.000259 ***
sleep         6.0379     0.7151   8.444 2.15e-06 ***
coffee       35.0883    20.4473   1.716 0.111834    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.258 on 12 degrees of freedom
Multiple R-squared:  0.857,	Adjusted R-squared:  0.8332 
F-statistic: 35.96 on 2 and 12 DF,  p-value: 8.554e-06
```

Hypothesis tests for regression models: The whole model

It’s nice to look at R^2, but we need to test if the data is adequately described by our model or not. Does the relationship we are modeling exists, or is there no relationship between outcome and predictor variables:

H_o: $Y_i = b_0 + \varepsilon_i$
H_1: $Y_i = b_2 X_{i.2} + b_1 X_{i.1} + b_0 + \varepsilon_i$

When we compare 2 models we run an F-test

$$SS_T = SS_M + SS_R$$
$$SS_M = SS_T - SS_R$$

$$MS_M = \frac{SS_M}{df_m} \quad \text{, with } df_m = p$$

where p is the number of predictors in the alterative hypothesis model (in this case 2)

$$MS_R = \frac{SS_R}{df_R} \quad \text{, with } df_R = n - p - 1$$

where n is the number of observations in our sample and p is the number of predictors in the alternative hypothesis model (in this case 12)

Like before, our F-statistic is: $F = \frac{MS_M}{MS_R}$

```
> summary(model2)

Call:
lm(formula = score ~ sleep + coffee, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-8.483 -3.705 -1.005  3.892 10.681 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.5235     6.1744   5.106 0.000259 ***
sleep         6.0379     0.7151   8.444 2.15e-06 ***
coffee       35.0883    20.4473   1.716 0.111834    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.258 on 12 degrees of freedom
Multiple R-squared:  0.857,	Adjusted R-squared:  0.8332 
F-statistic: 35.96 on 2 and 12 DF,  p-value: 8.554e-06
```

Hypothesis tests for regression models: individual coefficients

Assuming the model performed better than chances, it is time to look at the individual coefficients. Are they really different from zero? Depending on the units of your data this can be difficult to assess by good old fashioned eye-balling it.

So, for each coefficient b_x we have a hypothesis test with the following hypotheses

H_o: $b_x = 0$
H_1: $b_x \neq 0$

This looks awfully similar to a two-sided t-test. Which is good. Because this means you already know the test. Like before we can calculate the t-statistics using our estimates as follows:

$$t = \frac{\hat{b}_x - 0}{SE(\hat{b}_x)} = \frac{\hat{b}_x}{SE(\hat{b}_x)}$$

where $df = n - p - 1$ and $SE(\hat{b}_x) = [complicated]$. Thankfully, R (and all the other stats tools) calculate that for us.

```
> summary(model2)

Call:
lm(formula = score ~ sleep + coffee, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-8.483 -3.705 -1.005  3.892 10.681 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.5235     6.1744   5.106 0.000259 ***
sleep         6.0379     0.7151   8.444 2.15e-06 ***
coffee       35.0883    20.4473   1.716 0.111834    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.258 on 12 degrees of freedom
Multiple R-squared:  0.857,	Adjusted R-squared:  0.8332 
F-statistic: 35.96 on 2 and 12 DF,  p-value: 8.554e-06
```

…Looks like you can drink all the coffee you want before an exam, but can’t make up for lack of sleep to improve your test score. Thankfully this example is fabricated.

We can use the estimates $\hat{b}_x$ and their standard error $SE(\hat{b}_x)$ to calculate confidence intervals:

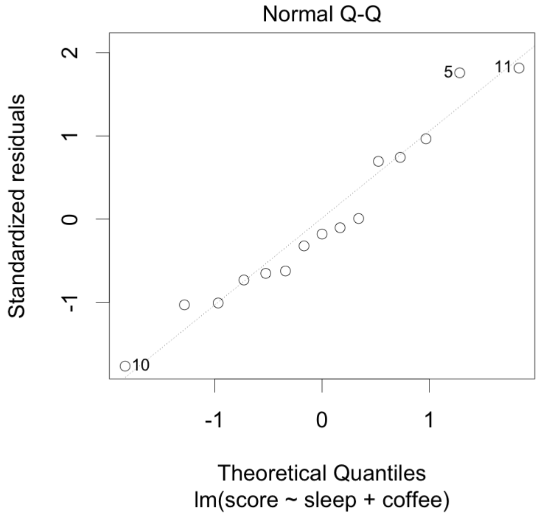
$$CI(\hat{b}_x) = \hat{b}_x \pm t_{df,\alpha/2} * SE(\hat{b}_x)$$

Here is how it’s done in R:

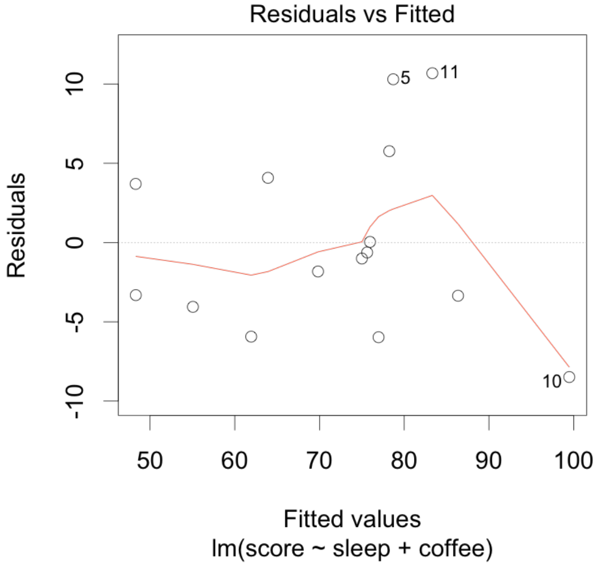
```
> confint(model2)
                2.5 %    97.5 %
(Intercept) 18.070582 44.976368
sleep        4.479831  7.595908
coffee      -9.462589 79.639221
```

Assumptions of regression and how to check them

1. Normality: residuals are normally distributed
    a. Histogram
    b. Shapiro-Wilk test
    c. QQ-plot



2. Linearity: there is a linear relationship between outcome and predictor variables
    a. Plot data and look at it (SLR)
    b. Plot residuals vs fitted values and check curvature



3. Homogeneity of variance: all residuals come from the same normal distribution
    a. Plot (standardized) residuals against fitted values and check distribution of residuals

How to compare two NESTED models

H_o: $Y = A + B$
H_1: $Y = A + B + C + D$

Comparing 2 nested models → F-test

$$SS_T = SS_{M0} + SS_{R0}$$
$$SS_T = SS_{M1} + SS_{R1}$$

Thus, the model sums of squares are

$$SS_{M0} = SS_T - SS_{R0}$$
$$SS_{M1} = SS_T - SS_{R1}$$

We want to know if there is a difference between the two models

$$\Delta SS_M = SS_{M1} - SS_{M0} = SS_{R0} - SS_{R1} = \sum_{i=1}^{n} (\hat{Y}_i^{(1)} - \hat{Y}_1^{(0)})^2$$

$$MS_{\Delta M} = \frac{\Delta SS_M}{df_{\Delta M}}$$

with

$$df_{\Delta M} = p_1 - p_0$$

were p_0 is the number of predictors model H$_0$ and p_1 is the number of predictors of model H$_1$

$$MS_R = \frac{SS_R}{df_R}$$

with

$$df_R = n - p_1 - 1$$

Thus, our F-statistic is

$$F = \frac{MS_{\Delta M}}{MS_R}$$

This is how it looks in R

```
> anova(model1, model2)
Analysis of Variance Table

Model 1: score ~ sleep
Model 2: score ~ sleep + coffee
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     13 585.32
2     12 469.99  1    115.33 2.9448 0.1118
```