Regression, OLS,Model Quality ,Parameter Significance, Variable Importance

1. Regression

In many cases, researchers/business analysts are interested in determining relationships of variables, e.g. advertisement expenditures and sales.

We always have one variable that we would like to explain that is called Y or dependent variable, and we have a set of variables that should explain Y which we call X or explanatory or independent variables.

A regression analysis will always yield a line that explains the relationship between the explanatory variables and the dependent variable in the best possible way (in the case of one explanatory variable): $Y = \alpha + \beta * X$

However, since we work with observational data, there are also sources of errors that prevent our line to fit the data perfectly well such as measurement error or non-linear relationships or omitted variables.

We will work with linear regressions only. Hence, we assume that the relationship between the explanatory variables and the dependent variable is linear! This assumption will be somewhat relaxed later on.

So how do we find the line fitting our data best?

2. Ordinary least squares (OLS)

Ordinary least squares (OLS) generates a line such that the sum of the squared errors of our prediction of the dependent variable (Y) given the explanatory variable (X) in our sample is minimized.

Under certain assumptions, the OLS estimator is BLUE, i.e. the best, linear, unbiased estimator.

Another advantage is the easy interpretation of the results: What happens to Y if X increases by one unit, HOLDING ALL OTHER VARIABLES CONSTANT? This is basically given by the respective coefficient estimate of X.

3. Quality of our model

The most famous “quality” measure is the R^2. It basically tells us what share/ percentage in the variation of Y is explained by all X jointly. It ranges from 0 to 1. $R^2 = ESS/TSS$

where ESS is the explained sum of squares and TSS is the total sum of squares

If none of our X does have explanatory power to predict Y,R^2 will be equal to zero. If our explanatory variables are able to perfectly predict Y, R^2 will be equal to one.

An issue with R^2 is that it increases by adding more variables to the model, even though the new X have no/very little explanatory power. The estimator will always be able to establish some relationship between X and Y.

Solution: Adjusted R-squared which corrects for the number of parameters included in the model.

4. Significance of parameters

So far, we have only looked at the overall quality of our model since R^2 tells us what share of the variation in Y is explained by all X together. However, we would also like to know whether the relationship between each of the single X and Y is significantly different from zero.

We can use a t-test for this again. It is now calculated by: t = $(\hat{\beta} - 0)/se(\hat{\beta})$

Moreover, we can test the Nullhypothesis that ALL parameters are JOINTLY equal to zero. This can be done using the F-Test. $F = (ESS/(k-1))/(RSS/(n-k))$

where k is the number of parameters to be estimated n is the number of observations.

5. Importance of variables

The size of a coefficient estimate does not tell us much about the importance of an explanatory variable to explain our dependent variable since we measure the variables in different units.

Solution: Standardized coefficients. $\hat{\beta}stand = \hat{\beta} * sd(X)/sd(Y)$ where

$\hat{\beta}stand$ is the standardized coefficient estimate of $\hat{\beta}$

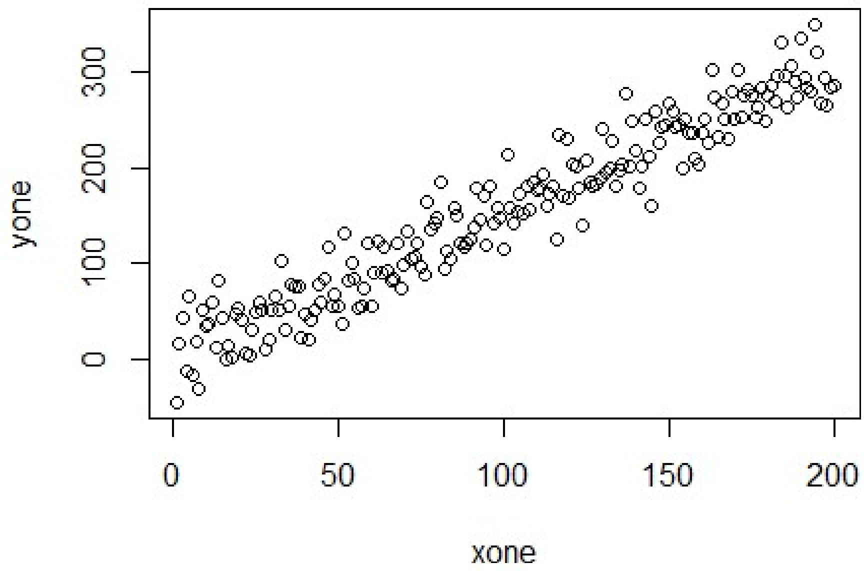
$sd(X)$ is the standard deviation of X

$sd(Y)$ is the standard deviation of Y

Theoretical Example

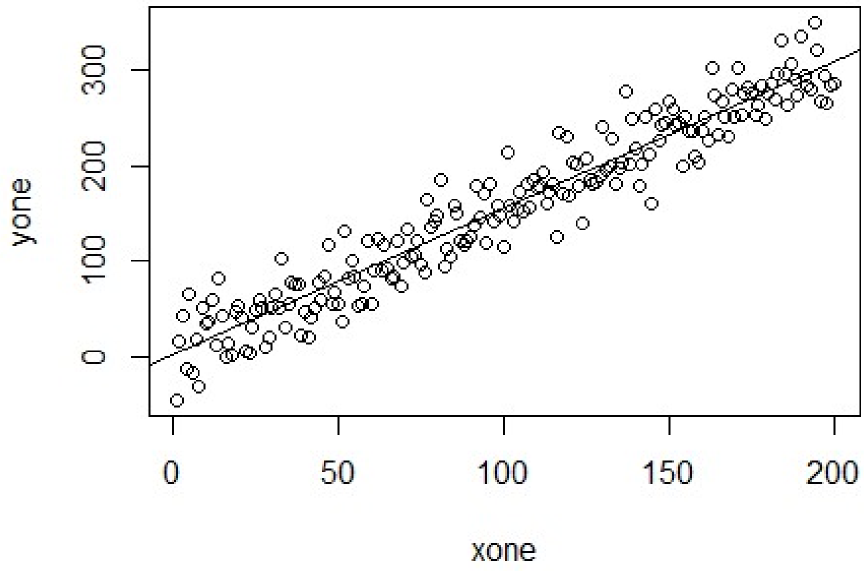
Let start with an illustration of OLS in a quite general way by evaluating a rather artificial example

```
n = 200
xone=seq(1:n)
yone = 1.5*xone + rnorm(n=n, mean = 3, sd=25)

plot(xone,yone)
```



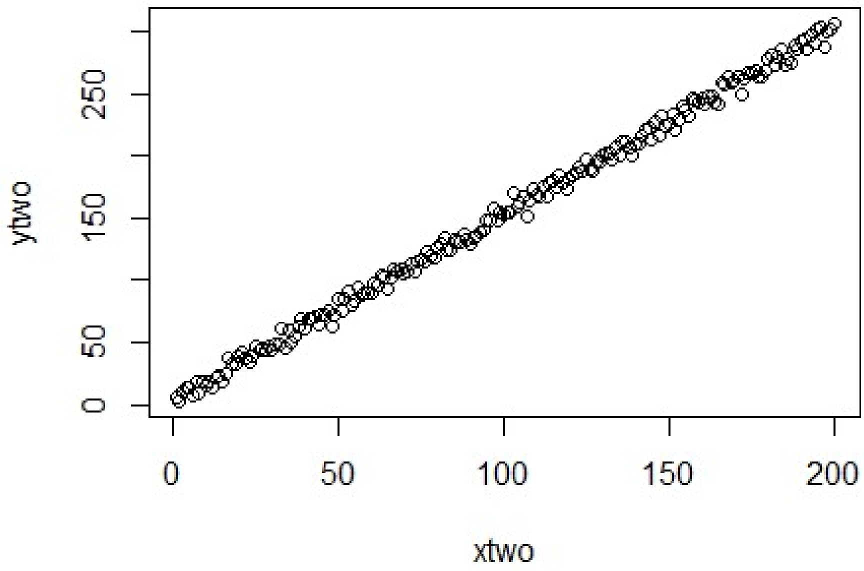

```
olsone = lm(yone~xone)
summary(olsone)

##
## Call:
## lm(formula = yone ~ xone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.439 -17.011  -3.201  16.345  65.895
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04406    3.64127   0.561    0.575
## xone         1.52494    0.03142  48.539   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.65 on 198 degrees of freedom
## Multiple R-squared:  0.9225, Adjusted R-squared:  0.9221
## F-statistic:  2356 on 1 and 198 DF,  p-value: < 2.2e-16

plot(xone,yone,abline(lm(yone~xone)))
```

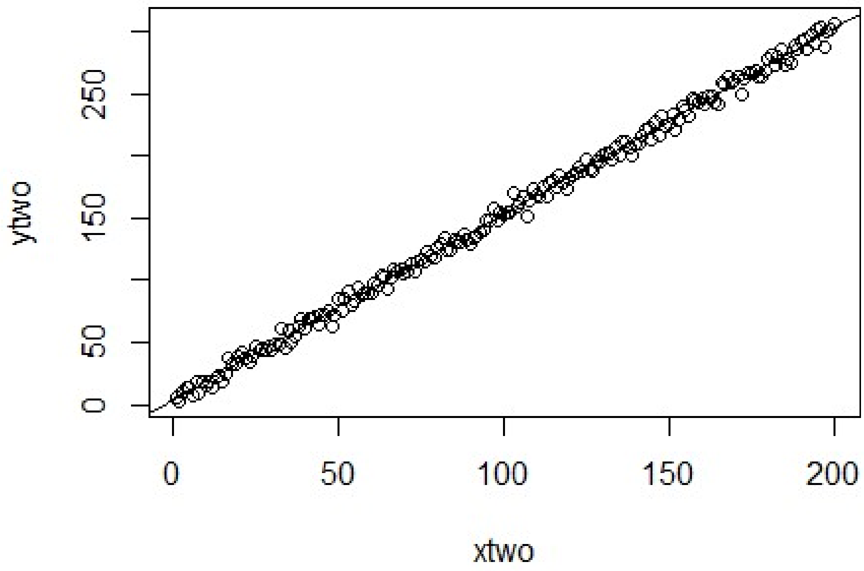


What happens, if we decrease the noise?

```
n = 200
xtwo=seq(1:n)
ytwo = 1.5*xtwo + rnorm(n=n, mean = 3, sd=5)

plot(xtwo,ytwo)
```

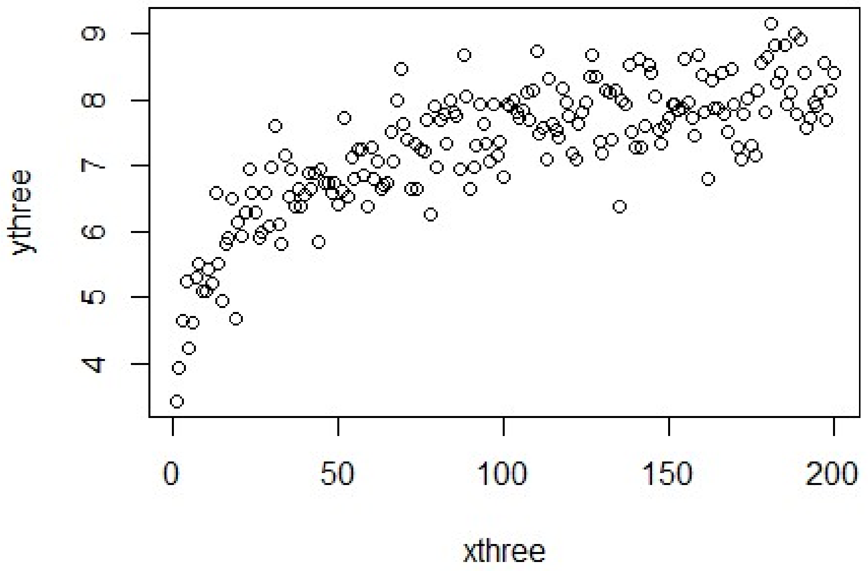



```
olstwo = lm(ytwo~xtwo)
summary(olstwo)

##
## Call:
## lm(formula = ytwo ~ xtwo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2871  -3.0705   0.2977   3.3078  12.1073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.326471   0.666865   4.988 1.33e-06 ***
## xtwo        1.498252   0.005754 260.401  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.698 on 198 degrees of freedom
## Multiple R-squared:  0.9971, Adjusted R-squared:  0.9971
## F-statistic: 6.781e+04 on 1 and 198 DF,  p-value: < 2.2e-16

plot(xtwo,ytwo,abline(lm(ytwo~xtwo)))
```

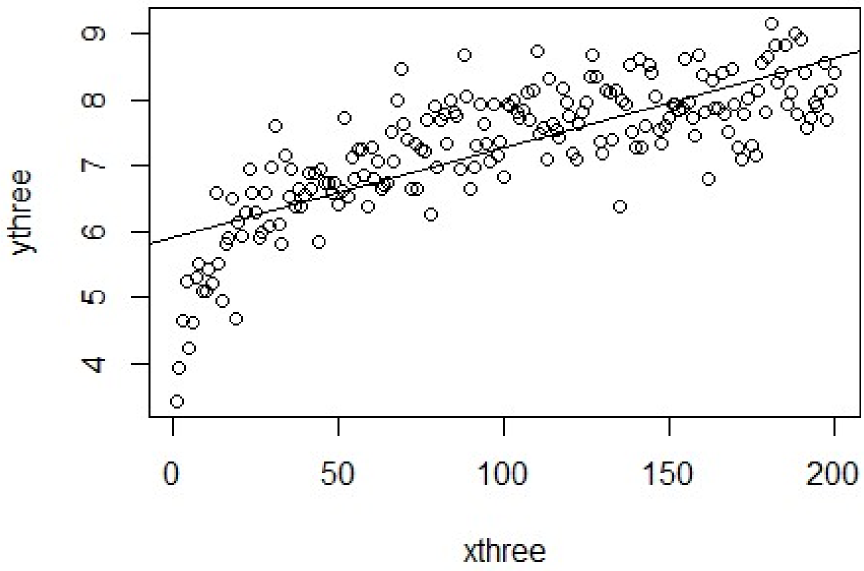


What happens, if the relationship was not linear?

```
xthree=seq(1:n)
ythree = log(xthree) + rnorm(n=n, mean = 3, sd=0.5)

plot(xthree,ythree)
```

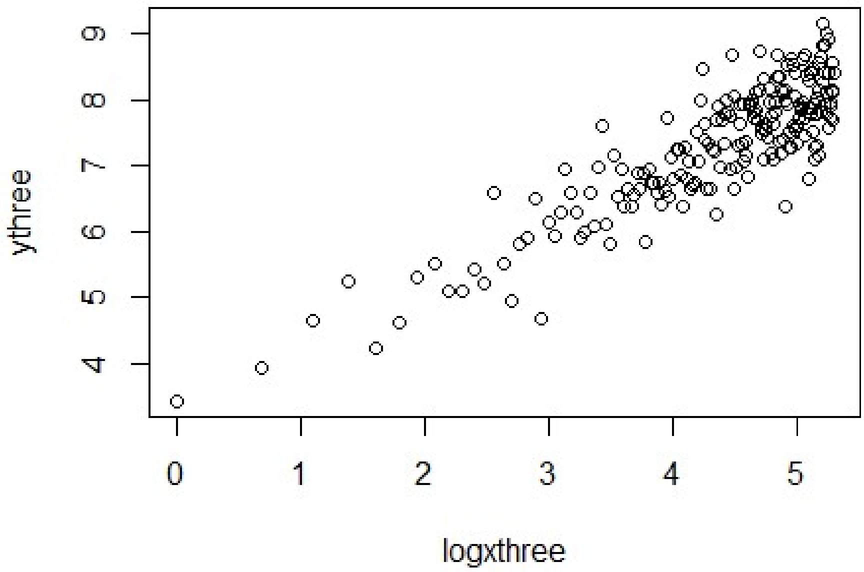



```
olsthree = lm(ythree~xthree)
summary(olsthree)

##
## Call:
## lm(formula = ythree ~ xthree)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50767 -0.34126  0.04646  0.43032  1.60078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.9242807  0.0905021   65.46   <2e-16 ***
## xthree      0.0136006  0.0007808   17.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6375 on 198 degrees of freedom
## Multiple R-squared:  0.6051, Adjusted R-squared:  0.6031
## F-statistic: 303.4 on 1 and 198 DF,  p-value: < 2.2e-16

plot(xthree,ythree,abline(lm(ythree~xthree)))
```

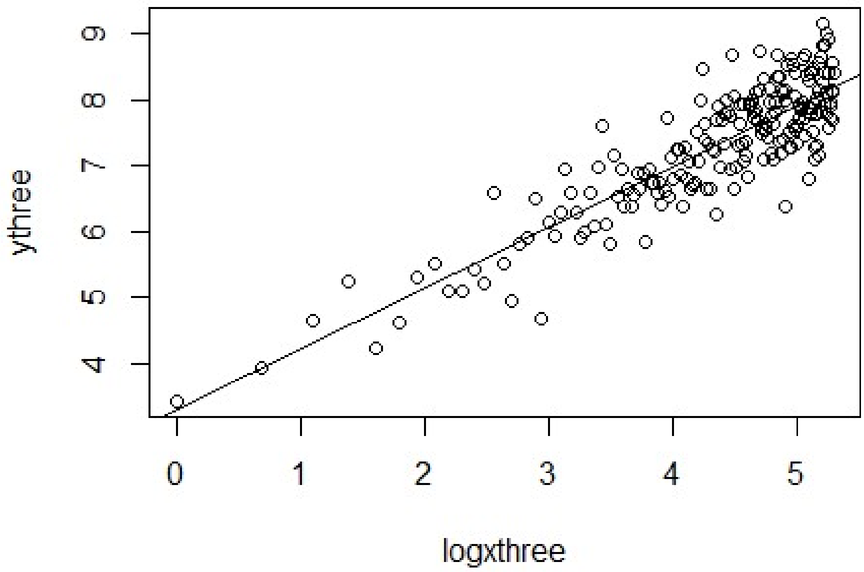


We can solve this by tranforming our X variable

```
logxthree = log(xthree)
plot(logxthree,ythree)
```




```
olsfour = lm(ythree~logxthree)
summary(olsfour)

##
## Call:
## lm(formula = ythree ~ logxthree)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46340 -0.32080 -0.03662  0.34615  1.25143
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2845     0.1609   20.41   <2e-16 ***
## logxthree     0.9283     0.0364   25.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4901 on 198 degrees of freedom
## Multiple R-squared:  0.7666, Adjusted R-squared:  0.7654
## F-statistic: 650.3 on 1 and 198 DF,  p-value: < 2.2e-16

plot(logxthree,ythree,abline(lm(ythree~logxthree)))
```



Real Example

Let's use real data: the mpg data set (publicly available) is shipped with ggplot2

The variables contained in the datset are:

manufacturer = manufacturer of the car

model = model

displ = engine displacement in liters

year = year of manufacturing

cyl = number of cylinders

trans = type of transmission (automatic/manual)

drv = drive type front, rear and 4-wheel

cty = city mileage in miles per gallon

hwy = highway mileage in miles per gallon

fl = fuel type

class = vehicle class (SUV etc.)

```r
library(ggplot2)
dat = mpg
hmile = lm(hwy~displ+year, dat=dat)
summary(hmile)
```

```
##
## Call:
## lm(formula = hwy ~ displ + year, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7616 -2.5187 -0.2899  1.8701 15.5852
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -276.15441  111.15444  -2.484  0.01369 *
## displ         -3.61099    0.19383 -18.630  < 2e-16 ***
## year           0.15579    0.05553   2.806  0.00545 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.78 on 231 degrees of freedom
## Multiple R-squared:  0.6004, Adjusted R-squared:  0.5969
## F-statistic: 173.5 on 2 and 231 DF,  p-value: < 2.2e-16
```

The TSS, RSS and ESS for our model are

```r
TSS = var(dat$hwy)*(nrow(dat)-1)
dat$errsq = (hmile$residuals)^2
RSS =sum(dat$errsq)
ESS=TSS-RSS
TSS
```

```
## [1] 8261.662
```

```r
RSS
```

```
## [1] 3301.333
```

```r
ESS
```

```
## [1] 4960.33
```

The R^2 can then be calculated manually as

```r
R_squared = ESS/TSS
R_squared
```

```
## [1] 0.6004034
```

whereas the $F - statistic$ is calculated manually as

```r
fishstat              =              (ESS/(length(coefficients(hmile))-1))/
  (RSS/(nrow(dat)-length(coefficients(hmile))))
fishstat

## [1] 173.5415
```

Of course, the above information together with any additional concerning the coefficients, their standard errors and the corresponding $p-values$ is already included in the collective table above. We realize

- The value of the $R^2 = 0.6$ is quite big, so the model explain about 60% of the variation of $Y = hwy$
- The coefficient of $disp$ is **significant**(not zero) at any level of significance(as $p - value << 1$)) whereas the coefficient of $year$ is **significant**(not zero) at a significance level of greater than 99%(as $p - value = 0.005 < 0.01$)
- As already expected, the two variables are also jointly significant at any level of significance (as $F - statistic = 173.5$ with $p - value << 1$)

However, the fact the coefficients of the respective variables are significant does not imply that the respective variable is also important! To examine that we should standardize **the coefficients** first

The standardized coefficient of $year$ becomes 0.118 from 0.156 whereas that of $displ$ decreases from -3.611 to -0.783. We observe that the the **variable** $displ$ **is not as important** as it would naively seem without standardizing the coefficients.