

# Regression with $\mathbf{R}$

Miguel Palhoto Rodrigues\*

August, 2005

## Abstract

This document aims to explain how to use  $\mathbf{R}$  matrix capacity in the context of regression analysis.

**Keywords:** R, Matrices, Regression.

**JEL Classification:** A2, C8.

## 1 The Model

We will build a model that will try to explained the consumption of petrol in thousands of tons (endogenous variable), based of the number of automobiles in thousands of units and the number of foreign automobiles in circulation in thousands of units (exogenous variables):

- $Y$  - Consumption of petrol in thousands of tons;
- $X_2$  - Number of automobiles in thousands of units;
- $X_3$  - Number of foreign automobiles in circulation in thousands of units

The model could be:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\epsilon}_i, \quad i = 1, 2, \dots, N \quad (1)$$

In a matrix form:

$$\mathbf{Y} = \hat{\beta}\mathbf{X} + \hat{\epsilon} \quad (2)$$

$$\text{Where } \mathbf{Y} = \begin{bmatrix} 136 \\ 144 \\ 145 \\ \vdots \\ 553 \\ 613 \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 91 & 11 \\ 1 & 105 & 13 \\ 1 & 109 & 17 \\ \vdots & \vdots & \vdots \\ 1 & 510 & 220 \\ 1 & 575 & 271 \end{bmatrix} \text{ and } \hat{\epsilon} = \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \hat{\epsilon}_3 \\ \vdots \\ \hat{\epsilon}_{17} \\ \hat{\epsilon}_{18} \end{bmatrix}$$

---

\*miguel@lpmotores.pt



```
X 60.4599
X2 0.7563
X3 0.4135
```

$$\hat{\beta} = \begin{bmatrix} 60.4599 \\ 0.7563 \\ 0.4135 \end{bmatrix}$$

The estimated model:

$$\hat{y}_1 = 60.459 + 0.7563x_{i2} + 0.4135x_{i3} \quad (4)$$

## 4 The estimation of Variance and Residual Standard Error

Variance is given by this formula,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N - k} \quad (5)$$

Where  $N$  is the number of observations and  $K$  is the number of exogenous variable plus the independent term, 3.

Residual standard error is square of formula (5),

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N - k}} \quad (6)$$

```
> (N <- length(Y))
[1] 18
> K <- 3
```

$\hat{\epsilon}$  is the error,

$$\hat{\epsilon}_i = y_i - \hat{y}_i \quad (7)$$

```
> Ye <- beta[1]*X[,1]+beta[2]*X[,2]+beta[3]*X[,3]
> erro <- Y-Ye
```

The sum of squared of error  $\sum_{i=1}^{18} \hat{\epsilon}_i^2 = 1084.42$

```
> sum(erro^2)
[1] 1084.42
```

The result of equation (5) is:

$$\frac{1084.42}{18 - 3} = 72.29$$

```

> sigma2 <- sum(erro^2)/(N-K)
> sigma2
[1] 72.3
> sigma <- sqrt(sigma2)
> sigma
[1] 8.502627

```

The Residual standard error, formula (6), on 15 degrees of freedom is:

$$\hat{\sigma} = 8.503$$

## 5 Standard Error

The standard error is the diagonal of the matrix of variance and covariance:

$$Var[\hat{\beta}] = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \quad (8)$$

```

> var.beta <- sigma2*XX
> var.beta
      X      X2      X3
X  67.6512 -0.66586  1.19755
X2 -0.6659  0.00766 -0.01451
X3  1.1976 -0.01451  0.02819

```

The diagonal of var.beta matrix:

```

> diag(var.beta)
      X      X2      X3
67.65119  0.00766  0.02819

```

We can get the standard error by calculating the squared root,

```

> sqrt(diag(var.beta))
      X      X2      X3
8.22503  0.08752  0.16789

```

Now we can write the model with the standard error,

$$\hat{y}_1 = \underset{(8.225)}{60.459} + \underset{(0.088)}{0.7563}x_{i2} + \underset{(0.168)}{0.4135}x_{i3} \quad (9)$$

## 6 Coefficient of Determination (R-squared)

The calculation of coefficient of determination,

$$R^2 = 1 - \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (10)$$

```
> R2 <- 1-sum(erro^2)/sum((Y-mean(Y))^2)
> R2
[1] 0.997
```

The calculation of the adjusted coefficient of determination (adjusted R-squared),

$$\bar{R}^2 = 1 - (1 - R^2) \frac{N - 1}{N - k} \quad (11)$$

```
> R22 <- 1-(1-R2)*(18-1)/(18-3)
> R22
[1] 0.9965
```

## 7 Testing the null hypotheses: there is no systematic relationship between exogenous variables ( $X_1$ and $X_2$ ) and endogenous variable ( $Y$ ).

Testing the null:

$$H_0 : \beta_2 = 0 \\ \beta_3 = 0$$

Putting the hypotheses test in a form of matrix,

$$H_0 : \mathbf{R}\mathbf{B} = \mathbf{r} \quad (12)$$

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \text{ and } \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

With the F-statistics we calculate the critical region,

$$F = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}]^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})/q}{\frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N-K}} \quad (13)$$

Where  $q$  is the number of restrictions of null hypotheses, 2.

```

> R <- matrix(c(0,1,0,0,0,1),2,3,byrow=T)
> R
      [,1] [,2] [,3]
[1,]    0    1    0
[2,]    0    0    1
> r <- matrix(c(0,0),2,byrow=T)
> r
      [,1]
[1,]    0
[2,]    0
> q <- 2
> (t(R%*%beta-r)%*%solve(R%*%XX%*%t(R))%*%(R%*%beta-r)/q)/(sum(erro^2)/(N-K))
      x
x 2450.172
> qf(0.95,2,15)
[1] 3.68232

```

Because 2450 is in critical region  $[368, +\infty[$ , we don't accept the null hypotheses, all variables are significant.

## 8 lm

We could obtain the same result using the function `lm`.

```

> summary(lm(Y~X[,2]+X[,3]))

Call:
lm(formula = Y ~ X[, 2] + X[, 3])

Residuals:
    Min       1Q   Median       3Q      Max
-13.221  -4.854  -1.596   4.576  15.834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.45995    8.22503   7.351 2.40e-06 ***
X[, 2]        0.75635    0.08752   8.642 3.29e-07 ***
X[, 3]        0.41349    0.16789   2.463  0.0264 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.503 on 15 degrees of freedom
Multiple R-Squared:  0.9969,    Adjusted R-squared:  0.9965
F-statistic:  2450 on 2 and 15 DF,  p-value: < 2.2e-16

```

As you can see we obtain the same result as using matrix to make the calculus of:

- Estimation of  $\beta$
- Standard Error
- Residual Standard Error
- R-Squared and Adjusted R-Squared
- F-statistics for testing the null

We also obtain with the function `lm` the probability of error type I: at 0.1% we reject the null for the intercept and  $X_2$ , at 5% we reject the null for  $X_3$ , all variables are significant.

## References

- [1] ARAI, M. A Brief Guide to R for Beginners in Econometrics. Tech. rep., Stockholm University, 2004.
- [2] CRIBARI-NETO, F., AND ZARCOS, S. R: Yet Another Econometric Programming Environment. *Journal of Applied Econometrics* 14 (1999), 319–329.
- [3] FOX, J. *An R and S-Plus Companion to Applied Regression*, first ed. Sage Publications Ltd., 6 Bonhill Street, London EC2A 4PU, United Kingdom, 2002.
- [4] R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.
- [5] RACINE, J., AND HYNDMAN, R. J. Using R to Teach Econometrics. Tech. rep., Monash University, Australia, 2001.