MLR – Exercise 3 (Optional)

In the following we focus on a small data set with n=5 observations so that it will not be too cumbersome to do the math by hand. Data is given below.

i	X_{1i}	X_{2i}	X_{3i}	Y_i
1	3	6	2	6
2	5	3	-1	10
3	7	1	3	15
4	4	5	-6	7
5	1	10	2	2

Univariate regressions

Plot Y against X_1, X_2 and X_3 respectively. If You could pick one X-variable to model Y in the linear regression $Y = \beta_0 + \beta_1 \cdot X_?$ which would it be?

Regress Y against X_1, against X_2 and finally against X_3. Report the three sets of parameters. Is the conclusion that some variables aren't important in the description of Y? Ask the opinion of other students!

For a little help on the math see the next slides.

A little help for $X=X_1$

$\hat{\beta}=(X'X)^{-1}(X'Y)$

First allow for an intercept β_0 adding a column of ones to the X:

$$X = \begin{bmatrix} 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 4 \\ 1 & 1 \end{bmatrix}$$

Calculate X'X:

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 5 & 7 & 4 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \sum_i 1 & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

and X'Y:

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 5 & 7 & 4 & 1 \end{bmatrix} * \begin{bmatrix} 6 \\ 10 \\ 15 \\ 7 \\ 2 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

A little help in general

Invert X'X:

$$(X'X)^{-1} = \begin{bmatrix} \left(1-\sum_i x_i\right) \Big/ \left(n-\sum_i x_i\right) & -\sum_i x_i \Big/ \left(n-\sum_i x_i\right) \\ -\sum_i x_i \Big/ \left(\sum_i x_i^2 - \sum_i x_i\right) & \left(1-\sum_i x_i\right) \Big/ \left(\sum_i x_i^2 - \sum_i x_i\right) \end{bmatrix}$$

If original X had been centered then $\sum X_i = 0$ making the inverse

$$(X'X)^{-1} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum_i x_i^2 \end{bmatrix}$$

which makes finding the parameters simple

$$\hat{\beta} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum_i x_i^2 \end{bmatrix} * \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} = \begin{bmatrix} \sum_i y_i \Big/ n \\ \sum_i x_i y_i \Big/ \sum_i x_i^2 \end{bmatrix}$$

Implications of centering

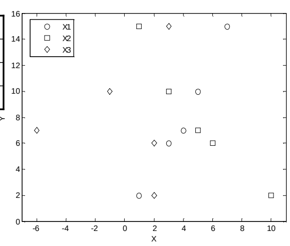
Centering of X doesn't affect β_1 (why?) and the intercept becomes $\bar{Y}$. Try convince yourself that without centering of X the intercept will be $\beta_1 \bar{X}$ lower.

Do the calculus by hand and check with your calculator! Are you (still) able to invert a matrix like the X'X?

Solution

	β_0	β_1	R^2
$Y=f(X_1)$	-0.6	2.15	0.9835
$Y=f(X_2)$	14.96	-1.39	0.9473
$Y=f(X_3)$	8	0.167	0.0160

- On its own X_1 seems to describe Y best.
- Alone X_3 doesn't seem useful but it might be valuable combined with other X's!



Multivariate regressions

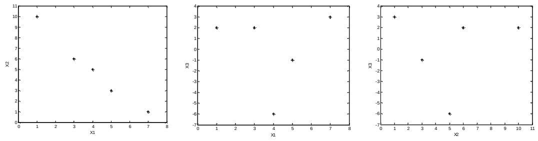
Plot X_1 vs. X_2, X_1 vs. X_3 and X_2 vs. X_3. Do the variables seem to correlate? What are the implications?

Regress Y against X_1 and X_2 simultaneously. Report the parameter estimates and calculate the correlation matrix!

Regress Y against X_1 and X_3 simultaneously. Report the parameter estimates and calculate the correlation matrix!

Compare the multivariate coefficients to the univariate ones. Are there coefficients that don't change? Is there a connection to the correlation matrices? What's the explanation? Which regression performed best and why? Was it surprising?

Solution



	β_0	β_1	β_2	β_3	R^2	Comments
$Y=f(X_1, X_2)$	-6.1	2.9	0.5	-	0.986	
$Y=f(X_1, X_3)$	-0.6	2.15	-	0.167	0.999	These are the univariate slopes!

Correlation	Y	X_1	X_2
Y	1	0.9917	-0.973
X_1	0.9917	1	-0.989
X_2	-0.973	-0.989	1

Correlation	Y	X_1	X_3
Y	1	0.9917	0.1263
X_1	0.9917	1	0
X_3	0.1263	0	1

Solution

In the multivariate regression using X_1 and X_3 these variables are uncorrelated and the regression coefficients (the slopes) are the same as in the two univariate regressions for Y against X_1 and X_3 respectively.

The $Y=f(X_1, X_3)$ regression performs the best. X_3 explains a great deal of the residuals from $Y=f(X_1)$ while X_2 does not.