Tutorial 4

keywords: matrix, matrices, vector, column space, histogram, scatter plot, mean, standard deviation, regression, simple, conditional expectation, OLS estimator, prediction, R squared, residual, interpretation, intercept, slope

estimated reading time: 35 minutes

January 16, 2023

Question 1

Matrix multiplication

Show that post-multiplying a matrix by a vector produces a linear combination of the columns of the matrix weighted by the elements of the vector.

Background

If $\boldsymbol{M}$ and $\boldsymbol{N}$ are matrices/vectors then the matrix multiplication, $\boldsymbol{MN}$, is only computable if the number of columns in M equals to the number of rows in $\boldsymbol{N}$.

Suppose that $\boldsymbol{M}$ and $\boldsymbol{N}$ are given by,

$$\underset{2\times 2}{\boldsymbol{M}} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\underset{2\times 1}{\boldsymbol{N}} = \begin{bmatrix} e \\ f \end{bmatrix}$$

Since the number of columns in $\boldsymbol{M}$ equals to the number of rows in $\boldsymbol{N}$, $\boldsymbol{MN}$ is computable.

To compute $\boldsymbol{MN}$,

$$\underset{2\times 1}{\boldsymbol{MN}} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} ae + bf \\ ce + df \end{bmatrix}$$

Post-multiple matrix $\boldsymbol{X}$ by vector $\widehat{\boldsymbol{\beta}}$,

$$\underset{4\times 2}{\boldsymbol{X}} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$$

$$\underset{2\times 1}{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} 0.7 \\ 0.2 \end{bmatrix}$$

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}} = ?$$

Let $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

$$\underset{4\times 1}{\widehat{\boldsymbol{y}}} = \underset{4\times 1}{\boldsymbol{X}\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.7 \\ 0.2 \end{bmatrix}$$

Since the no. of columns in $\boldsymbol{X}$ equals to the no. of rows in $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{y}}$ is computable

$$\underset{4\times 1}{\widehat{\boldsymbol{y}}} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.7 \\ 0.2 \end{bmatrix} =$$

By factoring out the constants, 0.7 and 0.2, we can see that $\widehat{\boldsymbol{y}}$ is a linear combination of the columns of $\boldsymbol{X}$, weighted by 0.7 and 0.2,

$$\underset{4\times 1}{\widehat{\boldsymbol{y}}} =$$

Question 2

Generalising results from Question 1

Generalise the result from Question 1 for the case of n observations an $\boldsymbol{X}$ matrix of 3 columns and a $\widehat{\boldsymbol{\beta}}$ vector of 3 rows.

$$\underset{n\times 3}{\boldsymbol{X}} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

$$\underset{3\times 1}{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = ?$$

Again, we can show that the linear combination of the columns of $\boldsymbol{X}$ equals to the sum of each column of $\boldsymbol{X}$ weighted/multiplied by each corresponding element in $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned} \widehat{\boldsymbol{y}} = &\ 1^{st}\ column\ of\ \boldsymbol{X} \times 1^{st}\ element\ of\ \widehat{\boldsymbol{\beta}} \\ &+ 2^{nd}\ column\ of\ \boldsymbol{X} \times 2^{nd}\ element\ of\ \widehat{\boldsymbol{\beta}} \\ &+ 3^{rd}\ column\ of\ \boldsymbol{X} \times 3^{rd}\ element\ of\ \widehat{\boldsymbol{\beta}} \end{aligned}$$

$$\begin{aligned} \underset{n\times 1}{\widehat{\boldsymbol{y}}} &= \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} \\ &= \begin{bmatrix} x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + x_{13}\hat{\beta}_3 \\ x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + x_{23}\hat{\beta}_3 \\ \vdots \\ x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + x_{n3}\hat{\beta}_3 \end{bmatrix} \\ &= \begin{bmatrix} x_{11}\hat{\beta}_1 \\ x_{21}\hat{\beta}_1 \\ \vdots \\ x_{n1}\hat{\beta}_1 \end{bmatrix} + \begin{bmatrix} x_{12}\hat{\beta}_2 \\ x_{22}\hat{\beta}_2 \\ \vdots \\ x_{n2}\hat{\beta}_2 \end{bmatrix} + \begin{bmatrix} x_{13}\hat{\beta}_3 \\ x_{23}\hat{\beta}_3 \\ \vdots \\ x_{n3}\hat{\beta}_3 \end{bmatrix} \\ &= \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} \hat{\beta}_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} \hat{\beta}_2 + \begin{bmatrix} x_{13} \\ x_{23} \\ \vdots \\ x_{n3} \end{bmatrix} \hat{\beta}_3 \end{aligned}$$

Question 3

This question is based on question C4 of the textbook. It is based on data on monthly salary and other characteristics of a random sample of 935 individuals. These data are in the file $wage2.wf1$. We concentrate on $wage$ as the dependent variable and the IQ as the independent variable.

Descriptive analytics - "looking" at the data through summary measures

Obtain a histogram and summary statistics for the variables $wage$ and IQ and a scatter plot of $wage$ against IQ.

$wage$ - monthly earnings ($)

To obtain the histogram of the $wage$ in EViews,

1. *Double click on wage from the workfile*

2. *View → Descriptive Statistics & Tests → Histograms and Stats*

Figure 1: Histogram and descriptive statistics of monthly earnings ($).

As we can see from Figure, monthly earnings is positively skewed (right-tailed) with mean and median monthly earnings of $957.95 and $905.00 respectively.

IQ - IQ score

To obtain the histogram of the IQ in EViews,

1. *Double click on IQ from the workfile*

2. *View* $\rightarrow$ *Descriptive Statistics & Tests* $\rightarrow$ *Histograms and Stats*

Figure 2: Histogram and descriptive statistics of IQ score.

As we can see from Figure **??**, IQ score is very slightly negatively skewed (left-tailed) with mean and median IQ score of $101.28 and $102 respectively (almost symmetrical).

Theoretically, IQ score is normally distributed with the population parameters,

$$population\ mean = \mu = 100$$

$$population\ standard\ deviation = \sigma = 15$$

From the empirical rule, it follows that 68%, 95%, and 99.7% of individuals have an IQ score within 1, 2, and 3 standard deviations of the mean respectively,

$$68\% : IQ\ score\ [85, 115]$$

$$95\% : IQ\ score\ [70, 130]$$

$$99.7\% : IQ\ score\ [55, 145]$$

The sample mean and standard deviation are 'close' to their population counterpart,

$$sample\ mean = \hat{\mu} = \overline{IQ} = 101.28$$

$$sample\ standard\ deviation = \hat{\sigma} = 15.05$$

There is one outlier with an IQ score of 145 (theoretically, this individual is 3 population standard deviations above the population mean).

Scatter plot of *wage* against *IQ*

The dependent variable goes on the y-axis and the independent variable goes on the x-axis of a scatter plot. Since *wage* is the dependent variable it goes in the y-axis and *IQ* is the independent variable so it goes in the x-axis.

To obtain a scatter plot of *wage* against *IQ*,

1. *Quick* $\rightarrow$ *Graph* ...

$$2.\ Series\ List:\ iq\ wage$$
$$(x - variable\ first\ then\ y - variable)$$

$$3.\ Specific:\ Scatter \rightarrow Fit\ Lines:\ Regression\ Line$$

Figure 3: Scatter plot of monthly earnings ($) against IQ score.

Visual inspection of Figure reveals that there is a positive relationship between IQ score and monthly earnings, however, this relationship does not appear to be linear. We also observe that the variability of monthly earnings increases as IQ score increases.

The red line through the scatter plot is a regression line of the estimated model,

$$\widehat{wage} = \hat{\beta}_0 + \hat{\beta}_2 IQ$$

(a) Simple Regression Model, Estimation, Interpretation of Slope Coefficient and R^2

Estimate a simple regression model where a one-point increase in IQ score changes monthly earnings by a constant dollar amount.

Background

Simple Linear Regression Model

A *simple* linear regression model has only one independent variable,

$$y = \beta_0 + \beta_1 x_1 + u$$

With a sample of n observations, we can express this model in terms of each observation i,

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i \qquad i = 1, 2, \ldots, n$$

or more compactly, in matrix notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

where,

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \underset{n\times 2}{\boldsymbol{X}} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \quad \underset{2\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underset{n\times 1}{\boldsymbol{u}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

which gives,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} \beta_1 + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Multiple Linear Regression Model

A *multiple* linear regression model has more than one independent variable. For example, the following model contains 2 independent variables, x_1 and x_2,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

As with the simple case, the multiple regression model can also be expressed in terms of each observation i,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \qquad i = 1, 2, \ldots, n$$

or in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} \beta_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} \beta_2 + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

So a simple regression model where a one-point increase in IQ score changes wage by a constant amount is given by,

$$wage = \beta_0 + \beta_1 IQ + u$$

To estimate this model from the menu bar in EViews,

$$Quick \rightarrow Estimate\ Equation$$

$$Equation\ Estimation : wage\ c\ iq$$

To estimate the model from the **Command window**,

$$Command\ window : ls\ wage\ c\ iq$$

$$(press\ Enter\ to\ execute\ code)$$

To name (save) the estimated equation,

$$Name \rightarrow Name\ to\ identify\ object : eq01$$

$$(This\ names\ the\ equation\ \boldsymbol{eq01})$$

To save the residuals of this estimated model,

$$eq01 \rightarrow Proc \rightarrow Make\ Residual\ Series \cdots \rightarrow Name\ of\ resid\ series : uhat01$$

$$(This\ names\ the\ equation\ \boldsymbol{uhat01})$$

After clicking OK, the residuals should appears

Background

Residuals

Every estimated model comes with a set of residuals. The residuals from an estimated model is the difference between the actual value of y and $\hat{y}$ (predicted y) for each observation i,

$$\hat{u}_i = y_i - \hat{y}_i \qquad i = 1, 2, \ldots, n$$

$$\begin{aligned}
\hat{u}_1 &= y_1 - \hat{y}_1 \\
\hat{u}_2 &= y_1 - \hat{y}_2 \\
&\vdots \\
\hat{u}_n &= y_n - \hat{y}_n
\end{aligned}$$

For our estimated model of *wage* on a constant and *IQ* with a sample of 935 observations,

$$\widehat{wage}_i = 116.9916 + 8.3031 IQ_i \qquad i = 1, 2, \ldots, 935$$

the residuals are expressed as follows,

$$\hat{u}_i = wage_i - \widehat{wage}_i \qquad i = 1, 2, \ldots, 935$$

$$\begin{aligned}
\hat{u}_1 &= wage_1 - \widehat{wage}_1 \\
\hat{u}_2 &= wage_2 - \widehat{wage}_2 \\
&\vdots \\
\hat{u}_{935} &= wage_{935} - \widehat{wage}_{935}
\end{aligned}$$

Concretely,

$$\begin{aligned}
wage_1 - \widehat{wage}_1 &= wage_1 - (116.9916 + 8.3031 IQ_1) \\
&= 769 - (116.9916 + 8.3031 \times 93) \\
&= -120.18 \\
wage_2 - \widehat{wage}_2 &= wage_2 - (116.9916 + 8.3031 IQ_2) \\
&= 808 - (116.9916 + 8.3031 \times 119) \\
&= -297.05 \\
&\vdots \\
wage_{935} - \widehat{wage}_{935} &= wage_{935} - (116.9916 + 8.3031 IQ_{935}) \\
&= 1000 - (116.9916 + 8.3031 \times 107) \\
&= -5.42
\end{aligned}$$

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	116.9916	85.64153	1.366061	0.1722
IQ	8.303064	0.836395	9.927203	0.0000

R-squared	0.095535	Mean dependent var	957.9455
Adjusted R-squared	0.094566	S.D. dependent var	404.3608
S.E. of regression	384.7667	Akaike info criterion	14.74529
Sum squared resid	$1.38E+08$	Schwarz criterion	14.75564
Log likelihood	-6891.422	Hannan-Quinn criter.	14.74924
F-statistic	98.54936	Durbin-Watson stat	1.802114
Prob(F-statistic)	0.000000		

Table 1: Regression output of *wage* on a constant and IQ

When reporting the estimated model, we must not forget to include a 'hat' above the dependent variable and report the standard error of $\hat{\beta}$ in parenthesis underneath its corresponding estimated coefficient,

$$\widehat{wage} = \underset{(85.6415)}{116.9916} + \underset{(0.8364)}{8.3031} IQ$$

Background

Conditional Expectation

Our simple regression model,

$$wage = \beta_0 + \beta_1 IQ + u$$

can also be written in terms of the expectation of $wage$ conditional on IQ,

$$wage = E(wage|IQ) + u$$

since,

$$\begin{aligned}E(wage|IQ) &= E(\beta_0 + \beta_1 IQ + u|IQ)\\ &= E(\beta_0|IQ) + E(\beta_1 IQ|IQ) + E(u|IQ)\\ &= \beta_0 + \beta_1 IQ + 0\\ &= \beta_0 + \beta_1 IQ\end{aligned}$$

β_0 and β_1 are the true but unknown population parameters in our model, which we wish to estimate using our sample data set. To estimate the simple regression model of *wage* on a constant and IQ is to effectively estimate the expected *wage* conditional on IQ,

$$\begin{aligned}\widehat{wage} &= \hat{\beta}_0 + \hat{\beta}_1 IQ\\ E(\widehat{wage|IQ}) &= \hat{\beta}_0 + \hat{\beta}_1 IQ\end{aligned}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1 i.e. they are random variables which gives us estimates of β_0 and β_1 depending on the sample data set we feed it.

But how does $\hat{\beta}_0$ and $\hat{\beta}_1$ estimate β_0 and β_1?

<u>OLS estimator</u>

Estimating the expected *wage* conditional on IQ involves estimating the unknown population parameters β_0 and β_1 and we do so by using the OLS estimator. The OLS estimator 'finds' or 'chooses' a value for $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the sum of the squared differences between *wage* and $\widehat{wage}$ for each observation i in our sample,

$$\begin{aligned}&\sum_{i=1}^{n}(wage_i - \widehat{wage}_i)^2\\ =&\sum_{i=1}^{n}(wage_i - (\hat{\beta}_0 + \hat{\beta}_1 IQ_i))^2\\ =&\sum_{i=1}^{n}\hat{u}_i^2\end{aligned}$$

is as small as possible. Stated differently, the OLS estimator is a formula,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

which after given some sample data, will find the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimises the sum of squared residuals.

For a simple linear regression model like *wage*, the OLS formula for $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by,

$$\begin{aligned}\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x}\\ \hat{\beta}_1 &= \frac{\widehat{cov}(x_i, y_i)}{\widehat{var}(x_i)}\end{aligned}$$

Such that,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \begin{bmatrix}\overline{y} - \hat{\beta}_1\overline{x}\\ \frac{\widehat{cov}(x_i, y_i)}{\widehat{var}(x_i)}\end{bmatrix}$$

Find the predicted increase in *wage* for an increase in *IQ* of 15 points:

$$\widehat{wage} = 116.9916 + 8.3031IQ$$

The model predicts/estimates that a 1-point increase in IQ score, increases monthly earnings by \$8.30 ($\hat{\beta}_1 = 8.30$), on average. This implies that for a 15-point increase in IQ score, the model predicts monthly earnings to increase by 15 $\times$ \$8.3031 = \$124.55, on average.

Does *IQ* explain most of the variation in *wage*?

From Table **??**, $R^2 = 0.09554$. This means that the model explains 9.55% of the variability in *wage*. Since this model contains only one independent variable, *IQ*, this implies that *IQ* explains 9.55% of the variability in *wage* and 100-9.55%=90.45% is left unexplained by the model. (There is a lot of unexplained variation.)

What is the relationship between the R^2 of this regression and the sample correlation coefficient between *wage* and *IQ*?

$$R^2 = 0.0955$$

$$\widehat{corr}(wage, IQ) = 0.309088$$

$$(\widehat{corr}(wage, IQ))^2 = 0.309088^2 = 0.0955 = R^2$$

This relationship can only hold for a simple linear regression model.

(b) Interpretation of the intercept coefficient

What does the estimated intercept coefficient of,

$$\widehat{wage} = \underset{(85.6415)}{116.9916} + \underset{(0.8364)}{8.3031} IQ$$

mean?

$$\hat{\beta}_0 = 116.9916$$

The model predicts that an individual with an IQ score of 0 will earn \$116.99 each month, on average. This interpretation is not meaningful because:

1. Not possible for living person to have an IQ score of 0.
2. We estimated our model with a sample of individuals with an IQ score between 50 (minimum) and 145 (maximum). Predictions based on values of the independent variable(s) outside of the range of values in the sample used to estimate the model should be avoided. Since $IQ = 0$ is outside this range, this interpretation is not meaningful.

Run a regression of $wage$ on a constant and $(IQ - 100)$ and name this equation $eq02$ in EViews.

$$wage = \alpha_0 + \alpha_1(IQ - 100) + u$$

$$Quick \rightarrow Estimate\ Equation$$

$$Equation\ Estimation : wage\ c\ iq - 100$$

$$3.\ Name \rightarrow Name\ to\ identify\ object : eq02$$

To save residuals of this estimated model,

$$eq02 \rightarrow Proc \rightarrow Make\ Residual\ Series \cdots \rightarrow Name\ of\ resid\ series : uhat02$$

After clicking OK, the residuals should appears (if not, double-click $uhat02$ from the workfile to see the residuals),

Dependent Variable: WAGE
Method: Least Squares
Date: 07/13/17 Time: 18:28
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	947.2980	12.62885	75.01066	0.0000
IQ-100	8.303064	0.836395	9.927203	0.0000

R-squared	0.095535	Mean dependent var	957.9455
Adjusted R-squared	0.094566	S.D. dependent var	404.3608
S.E. of regression	384.7667	Akaike info criterion	14.74529
Sum squared resid	$1.38E+08$	Schwarz criterion	14.75564
Log likelihood	-6891.422	Hannan-Quinn criter.	14.74924
F-statistic	98.54936	Durbin-Watson stat	1.802114
Prob(F-statistic)	0.000000		

Table 2: Regression output of $wage$ on a constant and $(IQ - 100)$

When reporting the estimated model, we must not forget to include a 'hat' above the dependent variable and $se(\hat{\beta}_j)$ underneath $\hat{\beta}_j$ in parenthesis,

$$\widehat{wage} = \underset{(12.6289)}{947.2980} + \underset{(0.8364)}{8.3031}(IQ - 100)$$

<u>Similarities between both estimated models:</u>

- Estimated slope coefficient is the same for both estimated models, $\hat{\beta}_1 = \hat{\alpha}_1 = 8.3031$.

- R^2 statistic is the same for both estimated models, $R^2_{eq01} = R^2_{eq02} = 0.0955$.

- The OLS residuals are the same for both estimated models ($Highlight\ uhat01$ & $uhat02 \rightarrow Open \rightarrow As\ Group$),

Differences between both estimated models:

- Estimated intercept coefficient is different for both estimated models, $\hat{\beta}_0 = 116.9916 \neq \hat{\alpha}_0 = 947.2980$.

- Interpretation of estimated intercept of $eq02$, $\hat{\alpha}_0 = 947.2980$, is meaningful:

  The model predicts that a person with an IQ score of 100 (when IQ-100=0, IQ=100), will earn, on average, $947.30 per month.

(c) What have we learnt?

Subtracting 100 (or any constant) from the independent variable IQ only changes the estimated intercept coefficient. Why?

Firstly, consider a graph with both regression lines,

($Discuss\ in\ class$)

If instead, we subtracted IQ by 1 rather than 100, the regression line will shift to the left by 1 unit (thinking about the data points plotted in a scatter plot which form the first regression line. If these data points take one step to the left, the regression line will also take one step to the left.) so the estimated intercept should increase by 1 unit of the estimated slope i.e. $1 \times \hat{\beta}_1$,

$$(Discuss\ in\ class)$$

If we now subtract IQ by 100, the regression line shifts to the left by 100 units and estimated intercept should increase by 100 units of the estimated slope i.e $100 \times \hat{\beta}_1$,

$$(Discuss\ in\ class)$$

Question 4

Dummy variables & linear independence of matrix columns

EViews workfile: $wage1tute4.wf1$

Suppose that each observation in our sample belongs into one of categories, $female$ or $male$.

The dummy variable $female$ is binary and equals to 1 if the individual is female and 0 otherwise,

$$female =$$

The dummy variable $male$ is binary and equals to 1 if the individual is male and 0 otherwise,

$$male =$$

Consider a regression of $wage$ on these two dummy variables without a constant,

$$wage = \beta_1 female + \beta_2 male + u$$

(a) Sketch the $\mathbf{X}$ matrix for this regression and assume that n_1 observations are females, n_2 observations are males and that our sample size is given by $n_1 + n_2 = n$

For cross-sectional data ($wage1tute4.wf1$ contains cross-sectional data), rearranging the order of observations does not change statistical results.

For ease of notation, we arrange the first n_1 observations to be females and the last n_2 observations to be males. This gives us the following $\mathbf{X}$ matrix,

$$\underset{n\times 2}{\mathbf{X}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\\vdots\\1\end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix}0\\0\\\vdots\\0\end{matrix}}\right\} n_2 \end{matrix}$$

The 1^{st} and 2^{nd} column of the $\mathbf{X}$ matrix represents the $female$ and $male$ dummy variable respectively.

Are the columns of $\mathbf{X}$ linearly independent?

Background

Linear independence of columns of matrix

The columns of a matrix are linearly independent if the linear combination of the columns of the matrix equals to 0 only when each column is weight (multiplied) by 0. (Another way to think about this is if the columns of a matrix can form a linearly dependent set, then the columns of the matrix are linearly dependent.)

If we denote the 1^{st}, 2^{nd} through to the r^{th} column of matrix $\boldsymbol{X}$ as, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ and $\boldsymbol{x_r}$ respectively, then the columns of $\mathbf{X}$ will be linearly independent if,

$$\boldsymbol{x_1}a_1 + \boldsymbol{x_2}a_2 + \cdots + \boldsymbol{x_r}a_r = 0$$

$$\underline{\textit{only when}}$$

$$a_1 = a_2 = \cdots = a_r = 0$$

If the columns of $\mathbf{X}$ are not linearly independent, then $\mathbf{X}'\mathbf{X}$ will not be invertible, $\cancel{(\mathbf{X}'\mathbf{X})^{-1}}$ and the OLS estimator,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

cannot be computed.

Since the linear combination of the columns of $\mathbf{X}$ equals to 0 only when each column is weighted (multiplied) by 0,

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} a_1 + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} a_2 = 0 \qquad \underline{\textit{only when}}\ a_1 = a_2 = 0$$

the columns of $\mathbf{X}$ are linearly independent $\therefore$ the OLS estimator can be calculated.

(b) Use the OLS formula $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and derive the OLS estimator in this case.

Let $y = wage$

$$\boldsymbol{X}'\boldsymbol{X} = \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix}} \underset{n\times 2}{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}} = \begin{bmatrix} \sum_{i=1}^{n_1} 1 & 0 \\ 0 & \sum_{i=1}^{n_2} 1 \end{bmatrix} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}$$

$$\therefore (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \dfrac{1}{n_1} & 0 \\ 0 & \dfrac{1}{n_2} \end{bmatrix}$$

$$(\textit{for the inverse of a diagonal matrix, simply take inverse of diagonal elements})$$

$$\boldsymbol{X}'\boldsymbol{y} = \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix}} \underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ y_{n_1+2} \\ \vdots \\ y_{n_1+n_2} \end{bmatrix}} = \underset{2\times 1}{\begin{bmatrix} y_1 + y_2 + \cdots + y_{n_1} \\ y_{n_1+1} + y_{n_1+2} + \cdots + y_{n_1+n_2} \end{bmatrix}}$$

$$= \begin{bmatrix} \sum_{i=1}^{n_1} y_i \\ \sum_{i=n_1+1}^{n_1+n_2} y_i \end{bmatrix}$$

$$\therefore \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \dfrac{1}{n_1} & 0 \\ 0 & \dfrac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n_1} y_i \\ \sum_{i=n_1+1}^{n_1+n_2} y_i \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n_1} \times \sum_{i=1}^{n_1} y_i + 0 \times \sum_{i=n_1+1}^{n_1+n_2} y_i \\ 0 \times \sum_{i=1}^{n_1} y_i + \dfrac{1}{n_2} \times \sum_{i=n_1+1}^{n_1+n_2} y_i \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1}{n_1} \sum_{i=1}^{n_1} y_i \\ \dfrac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} y_i \end{bmatrix} = \begin{bmatrix} \overline{y}_{female} \\ \overline{y}_{male} \end{bmatrix} = \begin{bmatrix} \overline{wage}_{female} \\ \overline{wage}_{male} \end{bmatrix}$$

Comment on the result.

We find that by regressing the dependent variable on dummy variables without a constant term, the OLS estimator is the sample mean of the dependent variable for each group,

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \overline{y}_{female} \\ \overline{y}_{male} \end{bmatrix}$$

Verify results by creating the dummy variable $male$ in $wage1tute3.wf1$ and running a regression of $wage$ on $female$ and $male$ without a constant.

$$wage = \beta_1 female + \beta_2 male + u$$

$wage1tute3.wf1$ contains the $female$ dummy variable,

Since the $male$ dummy variable is a function of the $female$ dummy variable,

$$male_i = 1 - female_i \qquad i = 1, 2, \ldots, n$$

To create the $male$ dummy variable in EViews,

$$Quick \rightarrow Generate\ Series$$

$$Enter\ Equation : male = 1 - female$$

To estimate $wage$ on $female$ and $male$ without a constant,

$$wage = \beta_1 female + \beta_2 male + u$$

$$Quick \rightarrow Estimate\ Equation$$

$$Equation\ Estimation : wage\ female\ male$$

Dependent Variable: WAGE
Method: Least Squares
Date: 07/16/17 Time: 17:08
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
FEMALE	4.587659	0.218983	20.94980	0.0000
MALE	7.099489	0.210008	33.80578	0.0000

R-squared	0.115667	Mean dependent var	5.896103
Adjusted R-squared	0.113979	S.D. dependent var	3.693086
S.E. of regression	3.476254	Akaike info criterion	5.333582
Sum squared resid	6332.194	Schwarz criterion	5.349800
Log likelihood	−1400.732	Hannan-Quinn criter.	5.339932
Durbin-Watson stat	1.817601		

Table 3: Regression output of $wage$ on $female$ and $male$

When reporting the estimated model, we must not forget to include a 'hat' above the dependent variable and $se(\hat{\beta}_j)$ underneath $\hat{\beta}_j$ in parenthesis,

$$\widehat{wage} = \underset{(0.2190)}{4.5877} female + \underset{(0.2100)}{7.0995} male$$

To obtain the sample mean of $wage$ for females and males in EViews,

$$\overline{wage}_{female} = ?$$

$$\overline{wage}_{male} = ?$$

$$Double\ click\ on\ wage \rightarrow View \rightarrow Descriptive\ Stats\ \&\ Tests \rightarrow Stats\ by\ Classification$$

$$Series/Group\ for\ classify : female$$

Descriptive Statistics for WAGE
Categorized by values of FEMALE
Date: 07/16/17 Time: 17:34
Sample: 1 526
Included observations: 526

FEMALE	Mean	Std. Dev.	Obs.
0	7.099489	4.160858	274
1	4.587659	2.529363	252
All	5.896103	3.693086	526

Table 4: Sample mean and standard deviation of female and male *wage*

$$\begin{aligned}\overline{wage}_{female} &= 4.5877 = \hat{\beta}_1 \\ \overline{wage}_{male} &= 7.0995 = \hat{\beta}_2\end{aligned}$$

(c) Suppose we have a constant in addition to the two dummy variables *female* and *male*.

Write down the $\boldsymbol{X}$ matrix for this case.

$$\underset{n\times 3}{\boldsymbol{X}} = \begin{array}{c} \left.\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}\right. \end{array} \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} n_1 \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} n_2 \end{array}$$

Are the columns of $\boldsymbol{X}$ linearly independent?

The columns of $\boldsymbol{X}$ will be linearly independent if,

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} a_1 + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} a_2 + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} a_3 = 0 \qquad \underline{\textit{only when}}\ a_1 = a_2 = a_3 = 0$$

However, there are linear combinations of the columns of $\boldsymbol{X}$ that equal to 0 when $a_1 \neq 0, a_2 \neq 0, a_3 \neq 0$. For example, when, $a_1 = 1, a_2 = -1$ and $a_3 = -1$

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = 0 \qquad a_1 = 1, a_2 = -1, a_3 = -1$$

Therefore, the columns of $\boldsymbol{X}$ are not linearly independent. This implies that $\boldsymbol{X}'\boldsymbol{X}$ will not be invertible so the OLS estimator cannot be computed. This is also called a problem of perfect collinearity i.e. an exact linear relationship among the independent variables.

What is the dimension of the column space of $\boldsymbol{X}$?

Consider an example in which we have only 3 rows in our $\boldsymbol{X}$ matrix (3 observations in our sample) and let the first 2 observations be female and the last observation be male,

$$\underset{3\times 3}{\boldsymbol{X}} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$(2 \; females \; and \; 1 \; male)$$

Each column of the $\boldsymbol{X}$ matrix plotted on a 3-dimensional (x, y, z) axes, would be vectors pointed to the coordinates,

$$(1, 1, 1)$$

$$(1, 1, 0)$$

$$(0, 0, 1)$$

from the origin.

Since the column space of a matrix is characterised by the columns of the matrix, we find that although we have 3 columns, it only forms a 2-dimensional column space. All 3 column vectors lie on this column space, but we only need 2 columns and not 3 (it can be any 2 of the 3 columns) to obtain this column space i.e. one of the columns is redundant given the other two.

If a regression has a constant, only add one dummy variable for an attribute that has two categories (such as male, female). Explain.

For a regression with a constant, we cannot include both the $female$ AND $male$ dummy variable because it will cause the columns of the matrix to be linearly dependent (problem of perfect collinearity) and $\boldsymbol{X}'\boldsymbol{X}$ will not be invertible, so the OLS estimator,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

cannot be computed.