Quantitative Methods for Economics

Tutorial 10

Katherine Eyal

TUTORIAL 10

11 October 2010

ECO3021S

Part A: Problems

1. Consider the following regression output:

$$\begin{array}{ccccc} Q & = & 300 & - & 5 \quad P \\ (se) & & (60) & & (0.5) \\ (t) & & (5) & & (-10) \end{array}$$

where Q is the quantity of cheese demanded in the Waterfront Pick 'n Pay (measured in kg per day) and P is the price of cheese (measured in R/kg)

Explain how the regression output will change if:

   (a) Q is measured as the number of tons (i.e. 1000 kg) of cheese demanded per week (a shopping week has seven days)

   (b) P is measured in cents per kg

   (c) Q is measured as the number of tons (i.e. 1000 kg) of cheese demanded per week (a shopping week has seven days) and P is measured in cents per kg

2. Assume the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + u_i$$

where Y is personal consumption expenditure, X_1 is personal income, and X_2 is personal wealth.

   (a) The term $(X_{1i}X_{2i})$ is known as the *interaction term*. What is meant by this expression?

   (b) Show that the marginal propensity to consume, holding wealth constant, in this model is $\beta_1 + \beta_3 X_{2i}$.

   (c) Explain how you would test whether the marginal propensity to consume is significantly different from zero.

   (d) How would you test the hypothesis that the marginal propensity to consume is independent of the wealth of the consumer?

3. For a sample of firms in the chemical industry, the following equation was obtained by OLS (standard errors in parentheses):

$$\begin{aligned} \widehat{rdintens} &= \underset{(0.429)}{2.613} + \underset{(0.00014)}{0.00030}\, sales - \underset{(0.0000000037)}{0.0000000070}\, sales^2 \\ n &= 32, \quad R^2 = 0.1484 \end{aligned}$$

where *rdintens* denotes research and development (R&D) expenditure as a percentage of sales and *sales* denotes annual sales in millions of Rands.

(a) At what point does the marginal effect of *sales* on *rdintens* become negative?

(b) Would you keep the quadratic term in the model? Explain.

(c) Define *salesbil* as sales measured in billions of Rands: $salesbil = sales/1,000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the R-squared. (*Hint*: Note that $salesbil^2 = sales^2/(1,000)^2$.)

(d) For the purpose of reporting the results, which equation do you prefer?

Part B: Computer Exercises

1. The data set NBASAL.DTA contains salary information and career statistics for 269 players in the National Basketball Association (NBA).

(a) Estimate a model relating points-per-game (*points*) to years in the league (*exper*), *age* and years played in college (*coll*). Include a quadratic in *exper*; the other variables should appear in level form. Interpret your results.

(b) Holding college years and age fixed, at what value of experience does the next year of experience actually reduce points-per-game? Does this make sense?

(c) Why do you think *coll* has a negative and statistically significant coefficient? (*Hint*: NBA players can be drafted before finishing their college careers and even directly out of high school.)

(d) Add a quadratic in *age* to the equation. Is it needed? What does this appear to imply about the effects of age, once experience and college years are controlled for?

(e) Now regress log(*wage*) on *points*, *exper*, $exper^2$, *age*, and *coll*. Interpret your results in full. (Use the command: `gen expersq = exper^2` to create the $exper^2$ variable.)

(f) Find the predicted value of log(*wage*), when *points* = 10, *exper* = 5, *age* = 27 and *coll* = 4. Using the methods in Section 6.4 of Wooldridge, find the predicted value of *wage* at the same values of the explanatory variables.

(g) Test whether *age* and *coll* are jointly significant in the regression from part (e). What does this imply about whether age and college years have separate effects on wage, once productivity and seniority are accounted for?

2. Consider the data provided in MARKSANALYSIS.DTA. The file consists of variables that were used to investigate the determinants of performance in UCT's standard first-year course in microeconomics.

   The matric subjects are weighted as follows:

Points	*Requirements*
10	A at Cambridge System A levels
9	B at Cambridge system A levels
8	A at HG or C at Cambridge System A levels
7	B at HG or D at Cambridge System A levels
6	A at SG or C at HG or E at Cambridge System A levels
5	B at SG or D at HG or F at Cambridge System A levels

   To calculate the total number of entry points, the points received for maths and English are doubled. For UCT entry purposes only English and maths and the best four other subjects are considered. Thus a South African student who gets six As at HG level will get 64 entry points. Students who follow the Cambridge system can in principle obtain more entry points.

   The TOTALMARK consists of 72.5 per cent multiple-choice questions (MCQTOT) and 27.5 per cent essays (EXAMLQ).

   The interpretation of the data series is as follows:

   AFRIKAANSMARK: Entry points received for matric Afrikaans, where Afrikaans first language is fully weighted and Afrikaans second language is the entry point less 25 per cent;

   AFRSCHOOL: African school (e.g. student from Malawi, Mauritius, Zimbabwe, etc.);

   AGE: Age of student at start of academic year (28 Feb. 2002, continuous scale);

   ATTENDANCE: Number of monitored lectures attended by student (out of 6);

   BLACK: Dummy variable: 1 if black, zero otherwise;

   COLOURED: Dummy variable: 1 if coloured, zero otherwise;

   DACT: Dummy variable: 1 if passed Accounting at matric level; zero otherwise;

   DADM: Dummy variable: 1 if passed Additional Maths at matric level; zero otherwise;

DAFL: Dummy variable: 1 if passed an African language (e.g. Xhosa, Zulu, etc.) at matric level (1st, 2nd or 3rd language); zero otherwise;

DART: Dummy variable: 1 if passed Art at matric level; zero otherwise;

DBEC: Dummy variable: 1 if passed Business Economics at matric level; zero otherwise;

DBIO: Dummy variable: 1 if passed Biology at matric level; zero otherwise;

DCST: Dummy variable: 1 if passed Computer Studies at matric level; zero otherwise;

DECS: Dummy variable: 1 if passed Economics at matric level; zero otherwise;

DEGM: Dummy variable: 1 if passed English first language at matric level; zero otherwise;

DEGS: Dummy variable: 1 if passed English second language at matric level; zero otherwise;

DGEO: Dummy variable: 1 if passed Geography at matric level; zero otherwise;

DHIS: Dummy variable: 1 if passed History at matric level; zero otherwise;

DLNA: Dummy variable: 1 if passed a non-African language other than English, e.g. German, Chinese, or French at matric level; zero otherwise;

DMUS: Dummy variable: 1 if passed Music at matric level; zero otherwise;

DPSC: Dummy variable: 1 if passed Physical Science at matric level; zero otherwise;

ENGLISH: Dummy variable: 1 if English home language; zero otherwise;

ENGMARK: Entry points received for matric English, where English first language is fully weighted and English second language is the entry point less 25 per cent;

EXAMLQ: Mark obtained for the essays in the exam;

FEMALE: Dummy variable: 1 if female; zero if male;

INDIAN: Dummy variable: 1 if Indian; zero otherwise;

MALE: Dummy variable: 1 if male; zero if female;

MAT: Entry points obtained for matric mathematics;

MCQTOT: Mark obtained for all multiple-choice questions in tests and exam (weighted appropriately);

POINTS: Number of UCT entry points (double weight for English and maths);

POINTS_MAT_ENG: Number of UCT entry points for the four matric subjects other than English and maths;

POINTSLESSMAT: POINTS less UCT entry points for maths (where maths is given double weighting);

PRIVSCHOOL: Dummy variable: 1 if attended private school; zero otherwise;

TOTALMARK: Final mark achieved for ECO1010F;

WHITE: Dummy variable: 1 if white; zero otherwise;

YEARMARK: Year mark (in percentage), based on three tests (weight of 45 per cent of TOTALMARK)

(a) Regress MCQTOT on three racial dummy variables, and the intercept. Does MCQTOT differ significantly by race? What is the average value for MCQTOT for blacks, coloureds, Indians, and whites, respectively?

(b) Repeat part (a), but include all four racial categories and exclude the intercept. What is the interpretation of the coefficients? Explain why the R^2 is different to the one obtained in the regression from part (a). Now use the `tsscons` option to force Stata to calculate the centred R^2 (i.e. execute the command: `reg mcqtot black coloured indian white, nocons tsscons`) and compare this to the R^2 for the regression from part (a).

(c) What happens if you run the following command: `reg mcqtot black coloured indian white`?

(d) Regress MCQTOT on the three racial dummy variables, and an intercept, as well as ENGLISH (i.e. English home language). Based on this output complete the following table (at this point, ignore the significance of the coefficients):

Demographics	*Average mark*	*Demographics*	*Average mark*
Black & English		Indian & English	
Black & non-English		Indian & non-English	
Coloured & English		White & English	
Coloured & non-English		White & non-English	

(e) You suspect that there may possibly be significant interaction effects between race and English home language. Estimate the model

$$\begin{aligned} \text{MCQTOT} = {} & \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{COLOURED} + \beta_3 \text{INDIAN} + \beta_4 \text{ENGLISH} \\ & + \beta_5 \text{BLACK} * \text{ENGLISH} + \beta_6 \text{COLOURED} * \text{ENGLISH} \\ & + \beta_7 \text{INDIAN} * \text{ENGLISH} + u \end{aligned}$$

What happens?

(f) Apparently there is a multicollinearity problem, which we have inadvertently created. What has happened is that the INDIAN and INDIAN∗ENGLISH variables are perfectly correlated because all Indians at UCT proclaim that they are English speakers. To verify this, one could change the sample as follows:

`preserve` (This is important!)

`drop if indian != 1` (or, equivalently: `keep if indian = 1`)

and then consider the INDIAN and ENGLISH series. You will notice that they all have a value of one. This means that INDIAN and INDIAN∗ENGLISH are perfectly correlated. Use the `restore` command to restore the data set to the full sample of observations (you can only do this if you used the `preserve` command earlier).

(g) Stata automatically solves the problem by dropping INDIAN∗ENGLISH from the regression equation in (e). Based on the output obtained in (e), complete the following table (at this point, ignore the significance of the coefficients):

Demographics	*Average mark*	*Demographics*	*Average mark*
Black & English		Indian & English	
Black & non-English		Indian & non-English	
Coloured & English		White & English	
Coloured & non-English		White & non-English	

(h) How does the table obtained in (g) compare to the table obtained in (d)? Which table better indicates the impact of race and home language characteristics on the average performance in ECO1010F?

(i) Regress MCQTOT against the racial dummy variables, a gender dummy, MAT, ENGMARK, POINTS_MAT_ENG and AGE. On the basis of the regression results, test whether the impact of the school subjects (maths, English, or any of the others captured in POINTS_MAT_ENG) are the same. You could do this by using the command `test mat=engmark=points_mat_eng`. Looking at the coefficients, which school subject has the biggest impact on MCQTOT? Given your experience of ECO1010F, is this what you would have expected?

(j) In all the previous regressions you would have found a significant positive coefficient on POINT_MAT_ENG. Is it possible that the impact of this variable is different for the various racial groups and/or whether the student's home language is English or not? You can test for this with interaction variables, where you include POINT_MAT_ENG∗ENGLISH or POINT_MAT_ENG∗Race group, where Race group = {BLACK, COLOURED, INDIAN} in the regression equation. The coefficients on these variables are often called differential slope coefficients. You can then use the t-values on the coefficients on these interaction variables to determine whether the impact of POINT_MAT_ENG on MCQTOT differs for various race and language groups. What do you find?

(k) Using the data available, try to build a good model that explains the variation in MCQTOT. Are the results as you expect them to be? Are there any variables that are likely to affect MCQ performance but are not available in the data set? How does this affect the results you have obtained?

(l) Up to this point we have only considered the determinants of performance of multiple-choice questions. You may want to consider the determinants of performance in essay questions (EXAMLQ). You will probably find that some important determinants of performance for multiple-choice questions suddenly seem less important determinants of performance in essay questions. Have fun.

TUTORIAL 10 SOLUTIONS

11 October 2010

ECO3021S

Part A: Problems

1. Consider the following regression output:

$$\begin{array}{ccccc} Q & = & 300 & - & 5 \quad P \\ (se) & & (60) & & (0.5) \\ (t) & & (5) & & (-10) \end{array}$$

where Q is the quantity of cheese demanded in the Waterfront Pick 'n Pay (measured in kg per day) and P is the price of cheese (measured in R/kg)

Explain how the regression output will change if:

(a) Q is measured as the number of tons (i.e. 1000 kg) of cheese demanded per week (a shopping week has seven days)

(b) P is measured in cents per kg

(c) Q is measured as the number of tons (i.e. 1000 kg) of cheese demanded per week (a shopping week has seven days) and P is measured in cents per kg

SOLUTION:

(a)

$$\begin{array}{ccccc} Q & = & 2.1 & - & 0.035 \quad P \\ (se) & & (0.42) & & (0.0035) \\ (t) & & (5) & & (-10) \end{array}$$

We multiply the standard errors and beta coefficients by 7/1000 for both the intercept and slope coefficient, as we have re-scaled the dependent variable by dividing it by 1000 to reflect tons instead of kg, while weekly changes in quantity demanded should be greater than daily changes. The t-statistics thus do not change.

(b)

$$\begin{array}{ccccc} Q & = & 300 & - & 0.05 \quad P \\ (se) & & (60) & & (0.005) \\ (t) & & (5) & & (-10) \end{array}$$

We have re-scaled the independent variable price by converting from Rands to cents. We divide the coefficient and standard error for price by 100, as the partial change in quantity demand from a unit change in cents per kg should clearly be smaller than a unit change in Rands per kg. Again the t-statistic is unaffected as both the sample coefficient and standard error is adjusted.

(c) This simply combines the two re-scalings from (a) and (b). The appropriate calculations are described below:

$$\begin{array}{ccccc} Q & = & 300 \times [7/1000] & - & 5 \times [7/(1000 \times 100)] \quad P \\ (se) & & (60) \times [7/1000] & & (0.5) \times [7/(1000 \times 100)] \\ (t) & & (5) & & (-10) \end{array}$$

Thus,

$$\begin{array}{ccccc} Q & = & 2.1 & - & 0.00035 \quad P \\ (se) & & (0.42) & & (0.000035) \\ (t) & & (5) & & (-10) \end{array}$$

Even though we re-scaled both the dependent and independent variables the t-statistics have not changed. However we must be careful when interpreting the coefficients. For example, the coefficient of price gives the partial change in quantity demand in tons per week from a unit change in cents per tons.

2. Assume the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + u_i$$

where Y is personal consumption expenditure, X_1 is personal income, and X_2 is personal wealth.

(a) The term $(X_{1i}X_{2i})$ is known as the *interaction term*. What is meant by this expression?

(b) Show that the marginal propensity to consume, holding wealth constant, in this model is $\beta_1 + \beta_3 X_{2i}$.

(c) Explain how you would test whether the marginal propensity to consume is significantly different from zero.

(d) How would you test the hypothesis that the marginal propensity to consume is independent of the wealth of the consumer?

SOLUTION:

(a) The term $(X_{1i}X_{2i})$ is called the interaction term because it reflects an interaction between the two variables X_1 and X_2 (it is the product of the two variables). The interaction term means that the change in Y for a unit change in X_1 (X_2), holding X_2 (X_1) constant, depends on the magnitude of X_2 (X_1) .

It seems likely that the marginal propensity to consume differs across different levels of personal wealth, and the interaction term allows to capture this possibility.

(b) Recall that the marginal propensity to consume is simply the partial derivative of personal consumption expenditure with respect to personal income.

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 + \beta_3 X_{2i}$$

(c) The marginal propensity to consume is $\beta_1 + \beta_3 X_{2i}$ and will not be significantly different from zero if β_1 AND β_3 are not significantly different from zero (or, $\beta_1 + \beta_3 X_{2i}$ will be significantly different from zero if at least one of β_1 or β_3 is significantly different from zero).. This amounts to a test for the joint significance of β_1 and β_3. (Note that β_1 gives the partial change in consumption for a unit change in personal income when personal wealth is zero.)

Thus, the relevant test is the F-test:

$H_0 : \beta_1 = 0, \beta_3 = 0$

H_1 : At least one of β_1 or β_3 is different from zero.

We would estimate the unrestricted model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + u_i$$

and the restricted model:

$$Y_i = \beta_0 + \beta_2 X_{2i} + u_i$$

We can then calculate the F-statistic by using the formula

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where $q = 2$, and $k = 3$.

We would then compare our calculated F-statistic to the critical value (c) from the F-tables. If $F > c$, we can reject the null hypothesis (at the chosen significance level) and conclude that the marginal propensity to consume is significantly different from zero. If $F < c$, we cannot reject the null hypothesis (at the chosen significance level) and conclude that the marginal propensity to consume is not significantly different from zero.

(d) The marginal propensity to consume is $\beta_1 + \beta_3 X_{2i}$ and will be independent of the wealth of the consumer if β_3 is not significantly different from zero. This amounts to a test of the significance of β_3.

Thus, the relevant test is the t-test:

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

We would estimate the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + u_i$$

and calculate the t-statistic

$$t = \frac{\widehat{\beta}_3}{se\left(\widehat{\beta}_3\right)}$$

We would then compare our calculated t-statistic to the critical value (c) from the t-tables. If $|t| > c$, we can reject the null hypothesis (at the chosen significance level) and conclude that the marginal propensity to consume is not independent of the wealth of the consumer. If $|t| < c$, we cannot reject the null hypothesis (at the chosen significance level) and conclude that the marginal propensity to consume is independent of the wealth of the consumer.

(**Note** how (d) is different from question (c))

3. For a sample of firms in the chemical industry, the following equation was obtained by OLS (standard errors in parentheses):

$$\begin{aligned} \widehat{rdintens} &= \underset{(0.429)}{2.613} + \underset{(0.00014)}{0.00030}\ sales - \underset{(0.0000000037)}{0.0000000070}\ sales^2 \\ n &= 32, \quad R^2 = 0.1484 \end{aligned}$$

where *rdintens* denotes research and development (R&D) expenditure as a percentage of sales and *sales* denotes annual sales in millions of Rands.

(a) At what point does the marginal effect of *sales* on *rdintens* become negative?

(b) Would you keep the quadratic term in the model? Explain.

(c) Define *salesbil* as sales measured in billions of Rands: $salesbil = sales/1,000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the R-squared. (*Hint*: Note that $salesbil^2 = sales^2/\left(1,000\right)^2$.)

(d) For the purpose of reporting the results, which equation do you prefer?

SOLUTION:

(a) The turnaround point is given by $\widehat{\beta}_1 / \left(2\left|\widehat{\beta}_2\right|\right)$, or $.0003/(.000000014) \approx 21,428.57$; remember, this is sales in millions of dollars.

(b) Probably. Its t statistic is about -1.89, which is significant against the one-sided alternative $H_0 : \beta_2 < 0$ at the 5% level (cv ≈ -1.70 with $df = 29$). In fact, the p-value is about .036.

(c) Because *sales* gets divided by 1,000 to obtain *salesbil*, the corresponding coefficient gets multiplied by $1,000 : (1,000)(.00030) = .30$. The standard error gets multiplied by the same factor. As stated in the hint, $salesbil^2 = sales/1,000,000$, and so the coefficient on the quadratic gets multiplied by one million: $(1,000,000)(.0000000070) = .0070$; its standard error also gets multiplied by one million. Nothing happens to the intercept (because *rdintens* has not been re-scaled) or to the R^2:

$$\begin{aligned} \widehat{rdintens} &= \underset{(0.429)}{2.613} + \underset{(0.14)}{0.30}\ salesbil - \underset{(0.0037)}{0.0070}\ salesbil^2 \\ n &= 32, \quad R^2 = 0.1484 \end{aligned}$$

(d) The equation in part (c) is easier to read because it contains fewer zeros to the right of the decimal. Of course the interpretation of the two equations is identical once the different scales are accounted for.

Part B: Computer Exercises

1. The data set NBASAL.DTA contains salary information and career statistics for 269 players in the National Basketball Association (NBA).

   (a) Estimate a model relating points-per-game (*points*) to years in the league (*exper*), *age* and years played in college (*coll*). Include a quadratic in *exper*; the other variables should appear in level form. Interpret your results.

   (b) Holding college years and age fixed, at what value of experience does the next year of experience actually reduce points-per-game? Does this make sense?

   (c) Why do you think *coll* has a negative and statistically significant coefficient? (*Hint*: NBA players can be drafted before finishing their college careers and even directly out of high school.)

   (d) Add a quadratic in *age* to the equation. Is it needed? What does this appear to imply about the effects of age, once experience and college years are controlled for?

(e) Now regress log(*wage*) on *points*, *exper*, $exper^2$, *age*, and *coll*. Interpret your results in full. (Use the command: `gen expersq = exper^2` to create the $exper^2$ variable.)

(f) Find the predicted value of log(*wage*), when *points* = 10, *exper* = 5, *age* = 27 and *coll* = 4. Using the methods in Section 6.4 of Wooldridge, find the predicted value of *wage* at the same values of the explanatory variables.

(g) Test whether *age* and *coll* are jointly significant in the regression from part (e). What does this imply about whether age and college years have separate effects on wage, once productivity and seniority are accounted for?

SOLUTION:

(a) The estimated equation is

$$\begin{aligned} \widehat{points} &= \underset{(6.99)}{35.22} + \underset{(.405)}{2.364}\ exper - \underset{(.0235)}{.0770}\ exper^2 - \underset{(.295)}{1.074}\ age - \underset{(.451)}{1.286}\ coll \\ n &= 269, \quad R^2 = .141, \quad \overline{R}^2 = .128 \end{aligned}$$

(b) The turnaround point is $2.364/[2(.0770)] \approx 15.35$. So, the increase from 15 to 16 years of experience would actually reduce salary. This is a very high level of experience, and we can essentially ignore this prediction: only two players in the sample of 269 have more than 15 years of experience.

(c) Many of the most promising players leave college early, or, in some cases, forego college altogether, to play in the NBA. These top players command the highest salaries. It is not more college that hurts salary, but less college is indicative of super-star potential.

(d) When age^2 is added to the regression from part (a), its coefficient is $.0536(se = .0492)$. Its t statistic is barely above one, so we are justified in dropping it. The coefficient on *age* in the same regression is $-3.984(se = 2.689)$. Together, these estimates imply a negative, increasing, return to *age*. The turning point is roughly at 74 years old. In any case, the linear function of *age* seems sufficient.

(e) The OLS results are

$$\begin{aligned} \widehat{log(wage)} &= \underset{(.85)}{6.78} + \underset{(.007)}{.078}\ points + \underset{(.050)}{.218}\ exper - \underset{(.0028)}{.0071}\ exper^2 - \underset{(.035)}{.048}\ age - \underset{(.053)}{.040}\ coll \\ n &= 269, \quad R^2 = .488, \quad \overline{R}^2 = .478 \end{aligned}$$

(f)

$$\begin{aligned} \widehat{log(wage)} &= 6.78 + .078\,(10) + .218\,(5) - .0071\,(5)^2 - .048\,(27) - .040\,(4) \\ &= 7.0165 \end{aligned}$$

We cannot just exponentiate the predicted value for $\widehat{log(wage)}$ in order to find $\widehat{wage}$, as this will systematically underestimate the expected value of *wage*. Instead, we must use the following equation, where $\widehat{\sigma}$ is the standard error of the regression (also called the Root MSE in Stata) :

$$\begin{aligned}\widehat{wage} &= \exp\left(\widehat{\sigma}^2/2\right)\exp\left(\widehat{log(wage)}\right) \\ &= \exp\left(\left(.63673\right)^2/2\right)\exp\left(7.0165\right) \\ &= 1365.4\end{aligned}$$

(g) The joint F statistic produced by Stata is about 1.19. With 2 and 263 df, this gives a p-value of roughly .31. Therefore, once scoring and years played are controlled for, there is no evidence for wage differentials depending on age or years played in college.

2. Consider the data provided in MARKSANALYSIS.DTA. The file consists of variables that were used to investigate the determinants of performance in UCT's standard first-year course in microeconomics.

The matric subjects are weighted as follows:

Points	*Requirements*
10	A at Cambridge System A levels
9	B at Cambridge system A levels
8	A at HG or C at Cambridge System A levels
7	B at HG or D at Cambridge System A levels
6	A at SG or C at HG or E at Cambridge System A levels
5	B at SG or D at HG or F at Cambridge System A levels

To calculate the total number of entry points, the points received for maths and English are doubled. For UCT entry purposes only English and maths and the best four other subjects are considered. Thus a South African student who gets six As at HG level will get 64 entry points. Students who follow the Cambridge system can in principle obtain more entry points.

The TOTALMARK consists of 72.5 per cent multiple-choice questions (MCQTOT) and 27.5 per cent essays (EXAMLQ).

The interpretation of the data series is as follows:

AFRIKAANSMARK: Entry points received for matric Afrikaans, where Afrikaans first language is fully weighted and Afrikaans second language is the entry point less 25 per cent;

AFRSCHOOL: African school (e.g. student from Malawi, Mauritius, Zimbabwe, etc.);

AGE: Age of student at start of academic year (28 Feb. 2002, continuous scale);

ATTENDANCE: Number of monitored lectures attended by student (out of 6);

BLACK: Dummy variable: 1 if black, zero otherwise;

COLOURED: Dummy variable: 1 if coloured, zero otherwise;

DACT: Dummy variable: 1 if passed Accounting at matric level; zero otherwise;

DADM: Dummy variable: 1 if passed Additional Maths at matric level; zero otherwise;

DAFL: Dummy variable: 1 if passed an African language (e.g. Xhosa, Zulu, etc.) at matric level (1st, 2nd or 3rd language); zero otherwise;

DART: Dummy variable: 1 if passed Art at matric level; zero otherwise;

DBEC: Dummy variable: 1 if passed Business Economics at matric level; zero otherwise;

DBIO: Dummy variable: 1 if passed Biology at matric level; zero otherwise;

DCST: Dummy variable: 1 if passed Computer Studies at matric level; zero otherwise;

DECS: Dummy variable: 1 if passed Economics at matric level; zero otherwise;

DEGM: Dummy variable: 1 if passed English first language at matric level; zero otherwise;

DEGS: Dummy variable: 1 if passed English second language at matric level; zero otherwise;

DGEO: Dummy variable: 1 if passed Geography at matric level; zero otherwise;

DHIS: Dummy variable: 1 if passed History at matric level; zero otherwise;

DLNA: Dummy variable: 1 if passed a non-African language other than English, e.g. German, Chinese, or French at matric level; zero otherwise;

DMUS: Dummy variable: 1 if passed Music at matric level; zero otherwise;

DPSC: Dummy variable: 1 if passed Physical Science at matric level; zero otherwise;

ENGLISH: Dummy variable: 1 if English home language; zero otherwise;

ENGMARK: Entry points received for matric English, where English first language is fully weighted and English second language is the entry point less 25 per cent;

EXAMLQ: Mark obtained for the essays in the exam;

FEMALE: Dummy variable: 1 if female; zero if male;

INDIAN: Dummy variable: 1 if Indian; zero otherwise;

MALE: Dummy variable: 1 if male; zero if female;

MAT: Entry points obtained for matric mathematics;

MCQTOT: Mark obtained for all multiple-choice questions in tests and exam (weighted appropriately);

POINTS: Number of UCT entry points (double weight for English and maths);

POINTS_MAT_ENG: Number of UCT entry points for the four matric subjects other than English and maths;

POINTSLESSMAT: POINTS less UCT entry points for maths (where maths is given double weighting);

PRIVSCHOOL: Dummy variable: 1 if attended private school; zero otherwise;

TOTALMARK: Final mark achieved for ECO1010F;

WHITE: Dummy variable: 1 if white; zero otherwise;

YEARMARK: Year mark (in percentage), based on three tests (weight of 45 per cent of TOTALMARK)

(a) Regress MCQTOT on three racial dummy variables, and the intercept. Does MCQTOT differ significantly by race? What is the average value for MCQTOT for blacks, coloureds, Indians, and whites, respectively?

(b) Repeat part (a), but include all four racial categories and exclude the intercept. What is the interpretation of the coefficients? Explain why the R^2 is different to the one obtained in the regression from part (a). Now use the `tsscons` option to force Stata to calculate the centred R^2 (i.e. execute the command: `reg mcqtot black coloured indian white, nocons tsscons`) and compare this to the R^2 for the regression from part (a).

(c) What happens if you run the following command: `reg mcqtot black coloured indian white`?

(d) Regress MCQTOT on the three racial dummy variables, and an intercept, as well as ENGLISH (i.e. English home language). Based on this output complete the following table (at this point, ignore the significance of the coefficients):

Demographics	*Average mark*	*Demographics*	*Average mark*
Black & English		Indian & English	
Black & non-English		Indian & non-English	
Coloured & English		White & English	
Coloured & non-English		White & non-English	

(e) You suspect that there may possibly be significant interaction effects between race and English home language. Estimate the model

$$\begin{aligned} \text{MCQTOT} &= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{COLOURED} + \beta_3 \text{INDIAN} + \beta_4 \text{ENGLISH} \\ &\quad + \beta_5 \text{BLACK} * \text{ENGLISH} + \beta_6 \text{COLOURED} * \text{ENGLISH} \\ &\quad + \beta_7 \text{INDIAN} * \text{ENGLISH} + u \end{aligned}$$

What happens?

(f) Apparently there is a multicollinearity problem, which we have inadvertently created. What has happened is that the INDIAN and INDIAN∗ENGLISH variables are perfectly correlated because all Indians at UCT proclaim that they are English speakers. To verify this, one could change the sample as follows:

`preserve` (This is important!)

`drop if indian != 1` (or, equivalently: `keep if indian = 1`)

and then consider the INDIAN and ENGLISH series. You will notice that they all have a value of one. This means that INDIAN and INDIAN∗ENGLISH are perfectly correlated. Use the `restore` command to restore the data set to the full sample of observations (you can only do this if you used the `preserve` command earlier).

(g) Stata automatically solves the problem by dropping INDIAN∗ENGLISH from the regression equation in (e). Based on the output obtained in (e), complete the following table (at this point, ignore the significance of the coefficients):

Demographics	*Average mark*	*Demographics*	*Average mark*
Black & English		Indian & English	
Black & non-English		Indian & non-English	
Coloured & English		White & English	
Coloured & non-English		White & non-English	

(h) How does the table obtained in (g) compare to the table obtained in (d)? Which table better indicates the impact of race and home language characteristics on the average performance in ECO1010F?

(i) Regress MCQTOT against the racial dummy variables, a gender dummy, MAT, ENGMARK, POINTS_MAT_ENG and AGE. On the basis of the regression results, test whether the impact of the school subjects (maths, English, or any of the others captured in POINTS_MAT_ENG) are the same. You could do this by using the command `test mat=engmark=points_mat_eng`. Looking at the coefficients, which school subject has the biggest impact on MCQTOT? Given your experience of ECO1010F, is this what you would have expected?

(j) In all the previous regressions you would have found a significant positive coefficient on POINT_MAT_ENG. Is it possible that the impact of this variable is different for the various racial groups and/or whether the student's home language is English or not? You can test for this with interaction variables, where you include POINT_MAT_ENG∗ENGLISH or POINT_MAT_ENG∗Race group, where Race group = {BLACK, COLOURED, INDIAN} in the regression equation. The coefficients on these variables are often called differential slope coefficients. You can then use the t-values on the coefficients on these interaction variables to determine whether the impact of POINT_MAT_ENG on MCQTOT differs for various race and language groups. What do you find?

(k) Using the data available, try to build a good model that explains the variation in MCQTOT. Are the results as you expect them to be? Are there any variables that are likely to affect MCQ performance but are not available in the data set? How does this affect the results you have obtained?

(l) Up to this point we have only considered the determinants of performance of multiple-choice questions. You may want to consider the determinants of performance in essay questions (EXAMLQ). You will probably find that some important determinants of performance for multiple-choice questions suddenly seem less important determinants of performance in essay questions. Have fun.

SOLUTION:

(a)

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  3,  1339) =   23.95
       Model |  17517.8118     3  5839.27059           Prob > F      =  0.0000
    Residual |  326406.561  1339  243.768903           R-squared     =  0.0509
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  343924.373  1342  256.277476           Root MSE      =  15.613

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |   7.671421   1.639221     4.68   0.000     4.455699    10.88714
    COLOURED |  -.0140622   1.464697    -0.01   0.992    -2.887413    2.859289
       WHITE |    7.43208   1.044933     7.11   0.000     5.382197    9.481963
       _cons |   52.95548   .8673945    61.05   0.000     51.25388    54.65708
------------------------------------------------------------------------------
```

The p-value of the F-statistice for the overall significance of the regression is 0.0000. Therefore we can reject the null hypothesis that the beta coefficients associated with the explanatory variables is jointly equal to zero.

MCQTOT for Indians and whites are significantly higher compared to the African group. This may be due to historical disadvantages in the schooling system from which matriculants graduate.

The average for African students is 52.9, that of Indian students is $52.9 + 7.67 = 60.62690\ldots$, for coloured students it is $52.9 - 0.014 = 52.94142\ldots$ and for white students it is $52.9 + 7.43 = 60.38756\ldots$

(b)

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  4,  1339) = 4595.03
       Model |  4480504.97     4  1120126.24           Prob > F      =  0.0000
    Residual |  326406.561  1339  243.768903           R-squared     =  0.9321
-------------+------------------------------           Adj R-squared =  0.9319
       Total |  4806911.53  1343   3579.2342           Root MSE      =  15.613

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |    60.6269   1.390926    43.59   0.000     57.89827    63.35553
    COLOURED |   52.94142   1.180239    44.86   0.000      50.6261    55.25674
       WHITE |   60.38756   .5826757   103.64   0.000      59.2445    61.53062
       BLACK |   52.95548   .8673945    61.05   0.000     51.25388    54.65708
------------------------------------------------------------------------------
```

The coefficient associate dwith each race dummy now gives the average MCQ score of that race category.

When an intercept is not included, Stata calculates the uncentred R-squared, R_0^2. This R_0^2 is not generally a suitable measure of goodness of fit. Here, R_0^2 is much larger than the correct R-squared. Using the `tsscons` option forces Stata to calculate the centred R-squared when the constant is omitted.

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  3,  1339) =   23.95
       Model |  17517.8118     3  5839.27059           Prob > F      =  0.0000
    Residual |  326406.561  1339  243.768903           R-squared     =  0.0509
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  343924.373  1342  256.277476           Root MSE      =  15.613

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |    60.6269   1.390926    43.59   0.000     57.89827    63.35553
    COLOURED |   52.94142   1.180239    44.86   0.000      50.6261    55.25674
       WHITE |   60.38756   .5826757   103.64   0.000      59.2445    61.53062
       BLACK |   52.95548   .8673945    61.05   0.000     51.25388    54.65708
------------------------------------------------------------------------------
```

This R-squared is identical to the one in the regression from (a).

(c)

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  3,  1339) =   23.95
       Model |  17517.8118     3  5839.27059           Prob > F      =  0.0000
    Residual |  326406.561  1339  243.768903           R-squared     =  0.0509
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  343924.373  1342  256.277476           Root MSE      =  15.613

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |   7.685483   1.824182     4.21   0.000     4.106918    11.26405
    COLOURED |  (dropped)
       WHITE |   7.446142   1.316236     5.66   0.000     4.864034    10.02825
       BLACK |   .0140622   1.464697     0.01   0.992    -2.859289    2.887413
       _cons |   52.94142   1.180239    44.86   0.000      50.6261    55.25674
------------------------------------------------------------------------------
```

Stata automatically drops one of the dummy variables to prevent you from falling into the dummy variable trap.

(d)

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  4,  1338) =   19.16
       Model |  18634.8679     4  4658.71697           Prob > F      =  0.0000
    Residual |  325289.505  1338  243.116222           R-squared     =  0.0542
-------------+------------------------------           Adj R-squared =  0.0514
       Total |  343924.373  1342  256.277476           Root MSE      =  15.592

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |   5.274076   1.982596     2.66   0.008     1.384741    9.163412
    COLOURED |  -2.259663   1.799192    -1.26   0.209    -5.789208    1.269882
       WHITE |   5.173429   1.482988     3.49   0.001     2.264195    8.082663
     ENGLISH |     3.3194   1.548564     2.14   0.032      .281523    6.357278
       _cons |   52.03343   .9671574    53.80   0.000     50.13612    53.93074
------------------------------------------------------------------------------
```

Demographics	*Average mark*	*Demographics*	*Average mark*
Black & English	$52 + 3.319$	Indian & English	$52 + 5.27 + 3.319$
Black & non-English	52	Indian & non-English	$52 + 3.319$
Coloured & English	$52 - 2.26 + 3.319$	White & English	$52 + 5.17 + 3.319$
Coloured & non-English	$52 - 2.26$	White & non-English	$52 + 5.17$

(e)

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  6,  1336) =   15.36
       Model |  22195.6456     6  3699.27427           Prob > F      =  0.0000
    Residual |  321728.727  1336  240.814915           R-squared     =  0.0645
-------------+------------------------------           Adj R-squared =  0.0603
       Total |  343924.373  1342  256.277476           Root MSE      =  15.518

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       BLACK |  -13.53519   3.009367    -4.50   0.000    -19.43879   -7.631596
    COLOURED |  -5.436981   6.174871    -0.88   0.379    -17.55048    6.676517
      INDIAN |   .4117608   1.503746     0.27   0.784    -2.538199    3.361721
     ENGLISH |  -4.126581   2.894337    -1.43   0.154    -9.804521    1.551358
BLACK_ENGL~H |   11.86281   3.475922     3.41   0.001     5.043952    18.68167
COLOURED_E~H |   -2.12241   6.318311    -0.34   0.737     -14.5173    10.27248
INDIAN_ENG~H |  (dropped)
       _cons |   64.34172   2.833225    22.71   0.000     58.78367    69.89978
------------------------------------------------------------------------------
```

Stata automatically drops the INDIAN$*$ENGLISH interaction term.

(f) Stata commands:

```
preserve
drop if INDIAN != 1
browse INDIAN ENGLISH
restore
```

(g)

```
      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  6,  1336) =   15.36
       Model |  22195.6456     6  3699.27427           Prob > F      =  0.0000
    Residual |  321728.727  1336  240.814915           R-squared     =  0.0645
-------------+------------------------------           Adj R-squared =  0.0603
       Total |  343924.373  1342  256.277476           Root MSE      =  15.518

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |   2.084144   2.141717     0.97   0.331     -2.11735    6.285638
    COLOURED |   8.098213   5.579515     1.45   0.147    -2.847351    19.04378
       WHITE |   13.53519   3.009367     4.50   0.000     7.631596    19.43879
     ENGLISH |   7.736229   1.924797     4.02   0.000     3.960275    11.51218
COLOURED_E~H |  -13.98522   5.937063    -2.36   0.019     -25.6322    -2.33824
WHITE_ENGL~H |  -11.86281   3.475922    -3.41   0.001    -18.68167   -5.043952
       _cons |   50.80653   1.014457    50.08   0.000     48.81643    52.79663
------------------------------------------------------------------------------
```

Demographics	*Average mark*
Black & English	$50.8 + 7.73$
Black & non-English	50.8
Coloured & English	$50.8 + 8.09 + 7.73 - 13.9$
Coloured & non-English	$50.8 + 8.09$
Indian & English	$50.8 + 7.73 + 2.084$
Indian & non-English	N/A
White & English	$50.8 + 7.73 + 13.5 - 11.86$
White & non-English	$50.8 + 13.5$

(h) Table in (g) is better at explaining variation in average performance: Consider the adjusted R-squared and F-stats.

(i)

```
      Source |       SS       df       MS              Number of obs =    1298
-------------+------------------------------           F(  8,  1289) =  151.02
       Model |  161962.796     8  20245.3495           Prob > F      =  0.0000
    Residual |  172800.788  1289   134.05802           R-squared     =  0.4838
-------------+------------------------------           Adj R-squared =  0.4806
       Total |  334763.584  1297  258.106079           Root MSE      =  11.578

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      INDIAN |   1.444247   1.279078     1.13   0.259    -1.065056     3.95355
    COLOURED |   1.622591   1.133192     1.43   0.152    -.6005114    3.845694
       WHITE |   1.646031   .8627371     1.91   0.057    -.0464923    3.338553
      FEMALE |  -7.717172   .6692306   -11.53   0.000    -9.030073   -6.404271
         MAT |   3.555371   .2700451    13.17   0.000     3.025595    4.085147
     ENGMARK |   1.927489   .3845365     5.01   0.000     1.173103    2.681875
POINTS_MAT~G |   1.690067   .1181483    14.30   0.000     1.458283    1.921851
         AGE |   1.065098   .2649588     4.02   0.000     .5452998    1.584895
       _cons |  -41.54227   6.472732    -6.42   0.000    -54.24052   -28.84403
------------------------------------------------------------------------------
```

```
 ( 1)  MAT - ENGMARK = 0
 ( 2)  MAT - POINTS_MAT_ENG = 0

       F(  2,  1289) =   15.53
            Prob > F =    0.0000
```

We can reject the null hypothesis and conclude that the impact of the school subjects is not the same.

Maths has the biggest impact (in terms of magnitude).

(j)

```
      Source |       SS       df       MS              Number of obs =    1298
-------------+------------------------------           F(  9,  1288) =   78.63
       Model |  118707.248     9  13189.6942           Prob > F      =  0.0000
    Residual |  216056.336  1288  167.745603           R-squared     =  0.3546
-------------+------------------------------           Adj R-squared =  0.3501
       Total |  334763.584  1297  258.106079           Root MSE      =  12.952

------------------------------------------------------------------------------
      MCQTOT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       BLACK |   28.89729   6.562292     4.40   0.000     16.02334    41.77124
    COLOURED |   10.09422    8.46152     1.19   0.233    -6.505651     26.6941
      INDIAN |  -5.648567   10.48092    -0.54   0.590    -26.21012    14.91298
     ENGMARK |   .4778676    .416448     1.15   0.251    -.3391231    1.294858
POINTS_MAT~G |   2.751405   .1646868    16.71   0.000     2.428321    3.074488
POINTS_MAT~H |    .064989   .0511024     1.27   0.204    -.0352641    .1652421
POINTS_MAT~K |  -1.193568   .2463176    -4.85   0.000    -1.676796   -.7103403
POINTS_MAT~D |  -.5141414   .3173214    -1.62   0.105    -1.136665    .1083821
POINTS_MAT~N |   .1280032   .3692735     0.35   0.729    -.5964403    .8524466
       _cons |  -20.27969   4.375461    -4.63   0.000     -28.8635   -11.69587
------------------------------------------------------------------------------
```