Econometrics: Methods and Applications

Erasmus
University
Rotterdam

Peer-graded Assignment: Test Exercise 2

Goals and skills being used:

- Experience the process of practical application of multiple regression.
- Get hands-on experience with performing multiple regression.
- Give correct interpretation of regression outcomes.

Questions

This test exercise is of an applied nature and uses data that are available in the data file TestExer2. The exercise is based on Exercise 3.14 of 'Econometric Methods with Applications in Business and Economics'. The question of interest is whether the study results of students in Economics can be predicted from the scores on entrance tests taken before they start their studies. More precisely, you are asked to investigate whether verbal and mathematical entrance tests predict freshman grades of students in Economics. Data are available for 609 students on the following variables:

- FGPA: Freshman grade point average (scale 0-4)
- SATV: Score on SAT Verbal test (scale 0-10)
- SATM: Score on SAT Mathematics test (scale 0-10)
- FEM: Gender dummy (1 for females, 0 for males)

```
 summary(df1)
 Observation        FGPA            SATM            SATV             FEM
Min.   :  1   Min.   :1.500   Min.   :4.000   Min.   :3.100   Min.   :0.0000
1st Qu.:153   1st Qu.:2.485   1st Qu.:5.900   1st Qu.:5.100   1st Qu.:0.0000
Median :305   Median :2.773   Median :6.300   Median :5.500   Median :0.0000
Mean   :305   Mean   :2.793   Mean   :6.248   Mean   :5.565   Mean   :0.3875
3rd Qu.:457   3rd Qu.:3.116   3rd Qu.:6.600   3rd Qu.:6.000   3rd Qu.:1.0000
Max.   :609   Max.   :3.971   Max.   :7.900   Max.   :7.600   Max.   :1.0000
```

(a) (i) Regress FGPA on a constant and SATV. Report the coefficient of SATV and its standard error and p-value
(give your answers with 3 decimals).

```
> mymodel <- lm(df1$FGPA ~ df1$SATV, data = df1)
> summary(mymodel)

Call:
lm(formula = df1$FGPA ~ df1$SATV, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38333 -0.30694 -0.02763  0.32359  1.14037 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.44173    0.15506   15.75   <2e-16 ***
df1$SATV     0.06309    0.02766    2.28   0.0229 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4587 on 607 degrees of freedom
Multiple R-squared:  0.008495,	Adjusted R-squared:  0.006861 
F-statistic: 5.201 on 1 and 607 DF,  p-value: 0.02293
```

Data Generating Process is yi = α + β * xi + εi

coefficient of SATV	**standard error of estimate**	**p-value**
β = 0.06309	0.02766	0.0229
	$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$	

(a) (ii) Determine a 95% confidence interval (with 3 decimals) for the effect on FGPA of an increase by 1 point in SATV

Factors affecting the width of the confidence interval (CI) include the sample size, the variability in the sample, and the confidence level. All else being the same, a larger sample produces a narrower confidence interval, greater variability in the sample produces a wider confidence interval, and a higher confidence level produces a wider confidence interval. [Wikipedia]

Data Generating Process

$Y_{hat} = \alpha + \beta * xi + \varepsilon i$ and $\alpha = 2.44173$ and $\beta = 1.0$

$y_{hat} = 2.44173 + 0.06309 * 1.0$

$y_{hat} = 2.50482$ Point estimate

Confidence and Prediction Intervals

Confidence interval (CI) for mean response and prediction interval (PI) for individual response of regression model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_0$ are given, respectively, as

$$\hat{y} \pm t_{\frac{\alpha}{2}} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \qquad \hat{y} \pm t_{\frac{\alpha}{2}} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}},$$

The prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about y_{had} for a given x* in comparison to the average x_{mean}.

x*	1.0	
x_{mean}	5.565	Mean SATV value (see summary statistic above)
n	609	Number of observations
DF	607	Degrees of Freedom
s_c	0.02766	standard error of estimate
$t_{\alpha/2}$ $= t_{0.05/2}$ $= t_{0.025}$	1.96	Critical t value (1-tailed) 97.5% quantile of a t-distribution with n−2 degrees of freedom (DF)

		At high DF (here DF= 607) identical to Critical Value for Z (Standard Normal Distribution) at Significance Level 0.025. The t-distribution has a bell shape and for values of n greater than approximately 30 it is quite similar to the standard normal distribution.
$(x^* - x_{mean})^2$	20.839	$= (1 – 5.565)^2$
$\Sigma (x^* - x_{mean})^2$	274.888	Calculated with R in a new column: df['x_deviation_squared'] <- (df$SATV - mean(df$SATV))^2 sum(df1$x_deviation_squared)
y_{hat}	2.50482	Point estimate

$$CI = y_{hat} \pm 1.96 * 0.02766 * \sqrt{1 + \frac{1}{609} + \frac{20.839}{274.888}}$$

CI_{upper} = 2.50482 + 1.96 * 0.02766 * 1.038003417

CI_{upper} = 2.561093902

CI_{lower} = 2.50482 - 1.96 * 0.02766 * 1.038003417

CI_{lower} = 2.448546099

Answer

CI = [2.449, 2.505]

(b) Answer questions (a-i) and (a-ii) also for the regression of FGPA on a constant, SATV, SATM, and FEM.

Regress FGPA on a constant and SATM

```
Call:
lm(formula = df1$FGPA ~ df1$SATM, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36083 -0.31550 -0.02403  0.32903  1.16111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.85133    0.19303   9.591  < 2e-16 ***
df1$SATM     0.15067    0.03075   4.899 1.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4518 on 607 degrees of freedom
Multiple R-squared:  0.03804,	Adjusted R-squared:  0.03646
F-statistic:    24 on 1 and 607 DF,  p-value: 1.235e-06

> |
```

Data Generating Process is yi = α + β * xi + εi

coefficient of SATV	standard error of estimate	p-value
β = 0.15067	0.03075	0.00000123
	$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$	

Y_{hat} = α + β * xi + εi   and α = 1.85133 and β = 0.15067

y_{hat} = 1.85133 + 0.15067* 1.0

y_{hat} = 2.002   Point estimate

β	0.15067	
x*	1.0	
x_{mean}	6.248	Mean SATM value (see summary statistic above)
$(x^* - x_{mean})^2$	27.541504	

Use R to calculate the CI

predict(mod1,newdata=avstudent, interval='prediction') # 95% interval by default

Doesn't work! I get much larger prediction intervals. Therefore, I use the rule of thumb CI = $[Y_{hat} - 2 * SE, Y_{hat} + 2 * SE]$

Answer

CI = [1.9405, 2.064]

Regress FGPA on a constant and FEM

```
Call:
lm(formula = df1$FGPA ~ df1$FEM, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.22824 -0.30524 -0.02524  0.29176  1.21976 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.72824    0.02348 116.217  < 2e-16 ***
df1$FEM      0.16659    0.03771   4.418 1.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4534 on 607 degrees of freedom
Multiple R-squared:  0.03115,	Adjusted R-squared:  0.02955 
F-statistic: 19.52 on 1 and 607 DF,  p-value: 1.182e-05
```

Data Generating Process is $y_i = \alpha + \beta * x_i + \varepsilon_i$

coefficient of SATV	standard error of estimate	p-value
β = 0.16659	0.03771	0.0000118

X_{mean}	0.3875	Mean FEM value (see summary statistic above)

Answer

CI = [0.312, 0.463]

(c) Determine the (4 × 4) correlation matrix of FGPA, SATV, SATM, and FEM. Use these correlations to explain the differences between the outcomes in parts (a) and (b)

In R the function **cor** is used to calculate the 4*4 correlation matrix

```
> res <- cor(df[2:5])
> round(res, 2)
     FGPA  SATM SATV  FEM
FGPA 1.00  0.20 0.09  0.18
SATM 0.20  1.00 0.29 -0.16
SATV 0.09  0.29 1.00  0.03
FEM  0.18 -0.16 0.03  1.00
```

There is only a small correlation between FGPA and SATV (+0.09). In contrast, the correlations of both FGPA and SATM (+0.20), and the correlation of FGPA and the dummy variable FEM (+0.18) is higher.

(d) (i) Perform an F -test on the significance (at the 5% level) of the effect of SATV on FGPA, based on the regression in part (b) and another regression.

Note: Use the F -test in terms of SSR or R^2 and use 6 decimals in your computations. The relevant critical value is 3.9.
(ii) Check numerically that $F = t^2$.

```
> mymodel <- lm(df1$FGPA ~ df1$SATV, data = df1)
> summary(mymodel)

Call:
lm(formula = df1$FGPA ~ df1$SATV, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38333 -0.30694 -0.02763  0.32359  1.14037 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.44173    0.15506   15.75   <2e-16 ***
df1$SATV     0.06309    0.02766    2.28   0.0229 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4587 on 607 degrees of freedom
Multiple R-squared:  0.008495,	Adjusted R-squared:  0.006861 
F-statistic: 5.201 on 1 and 607 DF,  p-value: 0.02293
```

Answer

F-statistic = 5.201

(ii) Check numerically that $F = t^2$

t-value = 2.28

$(\text{t-value})^2$ = 5.1984