Stockholm University

STOCKHOLM UNIVERSITY
Department of Statistics
Spring 2018, period C-D

Michael Carlson (examiner)
Ulf Högnäs

EXAM – BASIC STATISTICS FOR ECONOMISTS

2018-08-17

Time:	10.00 - 15.00 (10AM – 3PM)
Approved aid:	Hand-held calculator with no stored text, data or formulas
Provided aid:	*Formula Sheet and Probability Distribution Tables*, returned after the exam

- **Problems 1 – 5: MULTIPLE CHOICE QUESTIONS – max 60 points**
  - A total of 12 multiple choice questions with five alternative answers per question one of which is the correct answer. Mark your answers on the attached **answer form.**
  - Marking more than one alternative will result in zero points for that question.
  - Written solutions should not submitted; only your answers on the answer form will be considered in the assessment and final grading.

- **Problems 6 – 7: COMPLETE WRITTEN SOLUTIONS – max 40 points**
  - Use only the provided **answer sheets** when submitting your solutions and answers.
  - For full marks, clear, comprehensive and well-motivated solutions are required. Unclear and unexplained solutions may result in point deductions even if the final answer is correct.
  - Check your calculations and solutions before submitting. Careless mistakes may result in unnecessary point deductions.

- The maximum number of points is stated for each question. The maximum total number of points is 60 + 40 = 100. At least 50 points is required to pass (grades A-E). The grading scale is as follows:

  A: 90 – 100 points
  B: 80 – 89 points
  C: 70 – 79 points
  D: 60 – 69 points
  E: 50 – 59 points
  Fx: 40 – 49 points
  F: 0 – 40 points

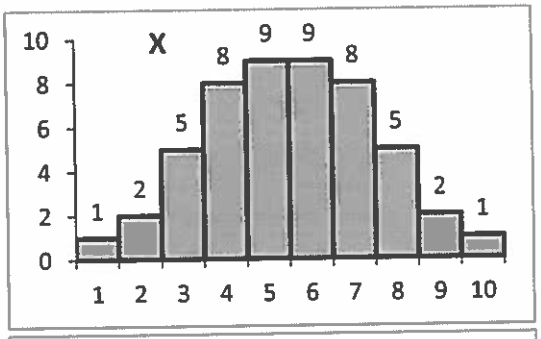
  NOTE! Fx and F are failing grades that require re-examination. Students who receive the grade Fx or F cannot supplement for a higher grade.

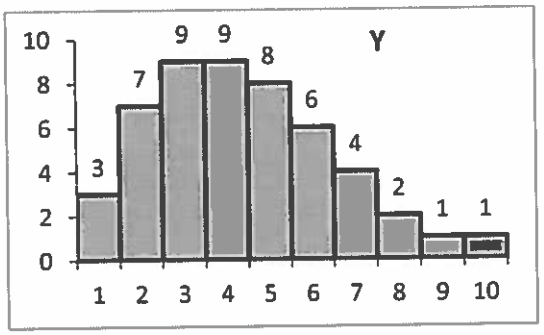
- Solutions will be posted on Mondo shortly after the exam.

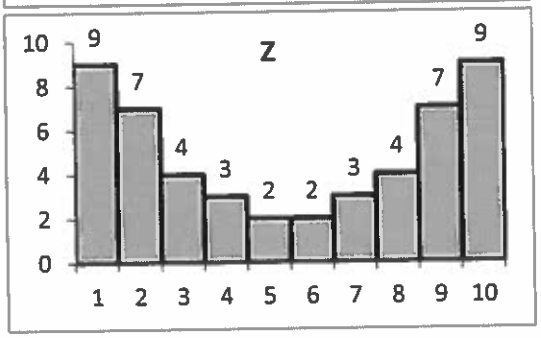
GOOD LUCK!

Problem 1

Three histograms describing the empirical distributions of three different variables X, Y and Z:







NOTE: The numbers above each of the bars are the corresponding absolute frequencies.

a) Which of the following statements is **true**? NOTE! Calculations are not needed! (5p)

A. The median of X and the median of Y are equal

B. The variance of Z is smaller than the variance of X

C. The interquartile range (IQR) of Z is larger than the IQR of X

D. It is possible to determine the correlation between X and Y using the graphs

E. All three distributions are symmetrical

The relationship between total household income (X) and saving (Y) was studied. A sample of size $n = 5$ households was sampled from a large population. The following total household incomes and savings per month was recorded (in 1000 SEK):

i	1	2	3	4	5
x_i	80	110	90	60	60
y_i	0,06	0,13	0,1	0,07	0,04

b) What are the values of the sample covariance and the sample correlation between X and Y? (5p)

A. $s_{xy} = 0{,}675$ $\quad r_{xy} = 1{,}2$

B. $s_{xy} = 0{,}675$ $\quad r_{xy} = 0{,}72$

C. $s_{xy} = 0{,}540$ $\quad r_{xy} = 0{,}72$

D. $s_{xy} = 0{,}540$ $\quad r_{xy} = 0{,}9$

E. $s_{xy} = 0{,}675$ $r_{xy} = 0{,}9$

Problem 2

Two Andorran football teams, Ordino and Santa Coloma, play a football match. Define the events

A = "Santa Coloma scores a goal in the first half"
B = "Santa Coloma wins the game"

From an on-line betting site you obtain the following probabilities for the corresponding events:

$P(A) = 0{,}40;$ $P(B) = 0{,}65;$ $P(A \cap B) = 0{,}35$

a) What is the probability that Santa Coloma scores a goal in the first half or Santa Coloma wins the game or both? (4p)

A. 1,00
B. 0,70
C. 0,75
D. 0,35
E. 0,26

b) What is the probability that Santa Coloma wins the game given that Santa Coloma scores a goal in the first half? (5p)

A. 0,5
B. 0,583
C. 0,6
D. 0,75
E. 0,875

TIP! Draw a Venn diagram or a table of the sample space and the events and their probabilities!

The sample space of a random variable X is $S_X = \{0,1,2,3\}$ and the probability function of X is defined by $P(x) = (x+1)/10$. Now define the random variable $Y = 10X + 5$.

c) What is the expected value $E(Y)$ and the variance $Var(Y)$? (6p)

A. $E(Y) = 105$ $Var(Y) = 10$
B. $E(Y) = 105$ $Var(Y) = 100$
C. $E(Y) = 25$ $Var(Y) = 10$
D. $E(Y) = 25$ $Var(Y) = 100$
E. $E(Y) = 2$ $Var(Y) = 1$

TIP! Y is a linear transformation of X.

Problem 3

A company is undergoing a tax audit. The auditor checks the company's outgoing invoices in order to detect unreported incomes. The company is in fact owned by a real crook and actually guilty of serious income tax evasion. The owner has deliberately failed to account for 55% of their invoices. The auditor samples 10 invoices.

a) What is the probability that no more than 2 invoices are unreported and unaccounted for? (5p)

A. 0,09956

B. 0,97261

C. 0,02739

D. 0,10199

E. 0,02289

Assume that X = the amount billed in each invoice is normal distributed (or at least approximately so) with mean value $\mu_X = 25\,000$ and standard deviation $\sigma_X = 10\,000$. A larger sample of size $n = 64$ is obtained.

b) What is the probability that the sample mean $\bar{X}$ is smaller than 24 000, i.e. $P(\bar{X} < 24\,000)$? (5p)

A. 0,21186

B. 0,50000

C. 0,78814

D. 0,20897

E. 0,57628

TIP! Draw sketches of the normal distribution and mark the area (event) of interest!

Problem 4

As a part of a human resource project, a large corporation collected sample data on X = the number of working hours per week for their employees. Only full time employees were included in the survey and weeks when they were either sick or on vacation were excluded. Based on $n = 400$ iid observations, it was found that the sample mean was 40,5 hours and the sample variance was 81.

a) Which of the following alternatives is a 95% confidence interval for μ_X, the mean of X? (5p)

A. (40,46; 40,54)

B. (40,10; 40,90)

C. (39,76; 41,24)

D. (39,62; 41,38)

E. (32,56; 48,44)

b) Suppose that your colleague calculated a confidence interval and the result was (39,34; 41,66). Which level of confidence level did your colleague use? (5p)

A. 90%

B. 95%

C. 98%

D. 99%

E. 99,9%

NOTE: The numbers in a) – b) have been rounded.

c) Which of these interpretations of a 95% confidence interval of the population mean μ_X is **incorrect**? (5p)

A. If we are able to repeat the experiment on the same population but with a new sample, a 95% confidence interval would capture the population mean with probability 0,95.

B. The proportion of the population captured by the 95% confidence interval goes to 0,95 as the sample size n goes to infinity.

C. The true population may lie entirely outside the confidence interval.

D. When we rely on the central limit theorem, the mid-point of the confidence interval is an unbiased estimate of the population mean μ_X.

E. If you set up the hypotheses $H_0: \mu_X = 45$ against $H_1: \mu_X \neq 45$ before the experiment, and it later turns out that the value 45 is outside the confidence interval, then you would reject the null hypothesis at the 5% level.

Problem 5

From previous studies it known that the average weight loss among newborn children one week after delivery is 0,2 kilos. A nurse measured the weight of eight randomly selected newborn children born to parents in a conflict area. The weight was measured first right after delivery (X), and then again, one week later (Y). The weight loss is defined as $D = X - Y$. The table below shows the two weight measurements in kilos for the eight infants:

Child no. i	1	2	3	4	5	6	7	8
Weight at delivery, x_i	2,9	3,4	3,3	2,9	3,9	3,3	3,9	4,3
Weight after 1 week, y_i	2,6	3,1	3,2	2,8	3,6	3,0	3,8	3,8

The nurse needs your help to test at a 5% significance level if the average weight loss μ_D for newborn in the conflict area is larger than 0,2 kilos.

a) Assuming normal distributions and independent observations, which of the following decision rules is appropriate for the test? (5p)

A. Reject H_0 if $t_{obs} > t_{7\,;\,0{,}05} = 1{,}895$

B. Reject H_0 if $|t_{obs}| > t_{7\,;\,0{,}025} = 2{,}365$

C. Reject H_0 if $t_{obs} < t_{7\,;\,0{,}05} = 1{,}895$

D. Reject H_0 if $|t_{obs}| > t_{14\,;\,0{,}025} = 2{,}145$

E. Reject H_0 if $t_{obs} > t_{14\,;\,0{,}05} = 1{,}761$

The nurse has already calculated some statistics for the data as seen below. However, note that you will not need all of these statistics to solve the problem, some are irrelevant to the test at hand!

$\bar{x} = 3{,}4875$ $\quad \bar{y} = 3{,}2375$ $\quad \bar{d} = 0{,}25$ $\quad \bar{x} - \bar{y} = 0{,}25$

$s_x^2 = 0{,}25268$ $\quad s_y^2 = 0{,}20554$ $\quad s_d^2 = 0{,}02$ $\quad s_p^2 = 0{,}22911$

$s_{xy} = 0{,}21911$ $\quad r_{xy} = 0{,}96145$

b) What is the correct outcome and conclusion of the test? (5p)

A. $t_{obs} = 2{,}8284$ and H_0 is rejected, μ_D is larger than 0,2 kilos

B. $t_{obs} = 1{,}0$ and H_0 is not rejected, μ_D is not larger than 0,2 kilos

C. $t_{obs} = 1{,}0$ and H_0 is rejected, μ_D is larger than 0,2 kilos

D. $t_{obs} = 0{,}20892$ and H_0 is not rejected, μ_D is larger than 0,2 kilos

E. $t_{obs} = 0{,}20892$ and H_0 is not rejected, μ_D is not larger than 0,2 kilos

Complete written solutions are required for Problems 6 and 7.
Use separate answer sheets for 6 and 7 respectively.

Problem 6

A candy company produces chocolate dipped peanuts. Each peanut has a colorful outer shell and there are $K = 4$ possible colors. The company claims that each color has the same probability of occurring, e.g. if a bag contains certain amount of peanuts, the expected frequency of each color is the same. You randomly select a bag containing $n = 200$ peanuts and count the colors: 47 are red, 37 are green, 49 are yellow, and 67 are brown. You decide to use a χ^2-test to determine whether each color is equally likely, or if the probabilities of the colors differ at a 5% significance level.

a) State your assumptions, hypotheses, test statistic, critical value and decision rule. (8p)

b) Finish your calculations, state your conclusions and give a verbal interpretation. (6p)

c) Explain briefly how the p-value can be used to determine the outcome of the test, no more than 2-4 sentences is required. Then use the χ^2-table to approximately determine the p-value of the observed value of the test statistic in b). (6p)

Problem 7

A company has a maintenance and service contract for a vital system component installed in various sites nationwide. Whenever the component fails they send a technician to repair it. They want you to identify factors (variables) that affect the time it takes to repair the component. You have obtained a small sample, $n = 10$, on four variables; data is provided on the following page.

Y = repair time in hours

X_1 = number of months since the latest maintenance service

X_2 = type of error (0 = mechanical error, 1 = electrical error)

X_3 = technician (0 = Jenny, 1 = John)

Three linear regression models are estimated and analyzed; Excel output for the estimated models are provided on the following pages.

Model 1: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

Model 2: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

Model 3: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

a) Illustrate the data for variables Y and X_1 in a suitable diagram. Briefly comment on your conclusions from this chart. Does there appear to be a linear relationship? (6p)

b) Use Model 1 and $\bar{x}_1 = 5{,}5$ and $s_{x_1}^2 = 6{,}72222$ to calculate a 90% prediction interval for the repair time given that it was four months since the latest maintenance. Interpret the result. (8p)

c) Model 3 and Model 2 differ in that variable X_3 is included in 3 but not in 2. Calculate a 95% confidence interval for β_3 and comment. Which of the two modes do you prefer? Explain.(6p)

DATA

Observation: i	1	2	3	4	5	6	7	8	9	10
Repair time: y_i	1,8	2,9	2,9	3,0	4,2	4,4	4,5	4,8	4,8	4,9
Maintenance: x_{1i}	3	2	2	6	9	4	6	8	8	7
Error type: x_{2i}	0	1	1	0	0	1	1	1	0	1
Technician: x_{3i}	1	1	1	1	0	0	1	0	0	0

ESTIMATED REGRESSION MODEL 1

$R^2 = 0{,}53418$ $R^2_{adj} = 0{,}47595$ $s_e = 0{,}78102$ $n = 10$

ANOVA	*df*	*SS*	*MS*	*F*	*P-value*
Regression	1	5,59603	5,59603	9,17389	0,01634
Residual	8	4,87997	0,61000		
Total	9	10,47600			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2,14727	0,60498	3,54934	0,00752
X1	0,30413	0,10041	3,02884	0,01634

ESTIMATED REGRESSION MODEL 2

$R^2 = 0{,}85919$ $R^2_{adj} = 0{,}81896$ $s_e = 0{,}45905$ $n = 10$

ANOVA	*df*	*SS*	*MS*	*F*	*P-value*
Regression	2	9,00092	4,50046	21,35700	0,00105
Residual	7	1,47508	0,21073		
Total	9	10,47600			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,93050	0,46697	1,99261	0,08656
X1	0,38762	0,06257	6,19540	0,00045
X2	1,26269	0,31413	4,01970	0,00506

ESTIMATED REGRESSION MODEL 3

$R^2 = 0{,}85919$ $R^2_{adj} = 0{,}81896$ $s_e = 0{,}45905$ $n = 10$

ANOVA	*df*	*SS*	*MS*	*F*	*P-value*
Regression	3	9,43049	3,14350	18,04002	0,00209
Residual	6	1,04551	0,17425		
Total	9	10,47600			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1,86016	0,72863	2,55294	0,04332
X1	0,29144	0,08360	3,48624	0,01304
X2	1,10241	0,30334	3,63418	0,01091
X3	-0,60909	0,38793	-1,57010	0,16744

CORRECTION! - RÄTTELSE!

For the Excel-output for Model 3 the correct values are as follows:

ESTIMATED REGRESSION MODEL 3

$R^2 = 0{,}90020$ $R^2_{adj} = 0{,}85030$ $s_e = 0{,}41743$ $n = 10$

The other values in the output are correct.

Michael Carlson

STOCKHOLM UNIVERSITY
Department of Statistics

Spring 2018 C-D

ANSWER FORM Exam – Basic statistics for economists
2018-08-17

Room: _____

Anonymous code: _____ *(write clearly!)*

Mark your answers with a clear cross (X) in the corresponding boxes below.

NOTE! Only one cross per question. If more than one alternative has been marked, zero points will be awarded for that question.

NOTE! If, after checking your calculations properly, you are convinced that the correct answer is not included among the given alternatives, write your answer in the margin to the right and explain you reasoning on the back.

		A	B	C	D	E
Problem 1	a)	☐	☐	☐	☐	☐
	b)	☐	☐	☐	☐	☐
Problem 2	a)	☐	☐	☐	☐	☐
	b)	☐	☐	☐	☐	☐
	c)	☐	☐	☐	☐	☐
Problem 3	a)	☐	☐	☐	☐	☐
	b)	☐	☐	☐	☐	☐
Problem 4	a)	☐	☐	☐	☐	☐
	b)	☐	☐	☐	☐	☐
	c)	☐	☐	☐	☐	☐
Problem 5	a)	☐	☐	☐	☐	☐
	b)	☐	☐	☐	☐	☐

Department of Statistics

Stockholms universitet

Correction sheet

Date: 180817

Room: Uggleviksalen

Exam: Statistics for Economists

Course: Basic Statistics for Economists

Anonymous code: 0093-KDK

☑ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X	X	X			3
Teacher's notes 5	15	10	15	10	17	18			

AN

Points	Grade	Teacher's sign.
90	A	ML

STOCKHOLM UNIVERSITY
Department of Statistics

Spring 2018 C-D

ANSWER FORM Exam – Basic statistics for economists
2018-08-17

Room: UG

Anonymous code: 0093-KDK *(write clearly!)*

Mark your answers with a clear cross (X) in the corresponding boxes below.

NOTE! Only one cross per question. If more than one alternative has been marked, zero points will be awarded for that question.

NOTE! If, after checking your calculations properly, you are convinced that the correct answer is not included among the given alternatives, write your answer in the margin to the right and explain you reasoning on the back.

		A	B	C	D	E
Problem 1	a)	☐	☐	☐	☒	☐
	b)	☐	☐	☐	☐	☒
Problem 2	a)	☐	☒	☐	☐	☐
	b)	☐	☐	☐	☐	☒
	c)	☐	☐	☐	☒	☐
Problem 3	a)	☐	☐	☒	☐	☐
	b)	☒	☐	☐	☐	☐
Problem 4	a)	☐	☐	☐	☒	☐
	b)	☐	☐	☐	☒	☐
	c)	☐	☒	☐	☐	☐
Problem 5	a)	☒	☐	☐	☐	☐
	b)	☐	☒	☐	☐	☐

✓

R

55

SU, DEPARTMENT OF STATISTICS

Room: UG Anonymous code: 0093-KDK Sheet number: 1

6

	obs.	exp. ($P_k = 0.25$)
K = 4		
red	47	50
green	39	50
yellow	49	50
brown	67	50

$n = 200$
$\alpha = 0.05$

a)

H_0: ~~The colors in each bag are evenly divided~~, and equally probable.

H_a: ~~The colors in each bag are unevenly divided~~, and unequally probable.

2

Assuming that...

- the bag is randomly selected out of a large and randomly selected sample of the ordinary stock.
- the counting and determination of colors are unbiased and confirmed by several individuals.

peanuts are idependent and eg. from the same color distribution (iid

1

Critical value

$$\chi^2_{crit.} = \chi^2_{n-2;\alpha/2} = \chi^2_{2;0.025} = 7.378$$

2

Decision rule: If $\chi^2_{obs} > \chi^2_{crit.}$, then H_0 lies outside of the confidence interval and must be rejected.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \text{where } E_i = n P_i$$

2

7

b) $$\chi^2_{obs} = \frac{(47-50)^2}{50} + \frac{(39-50)^2}{50} + \frac{(49-50)^2}{50} + \frac{(67-50)^2}{50} = 12.4$$

3

As $12.4 > 7.378$, H_0 must be rejected.

2

It can ~~be concluded~~ that if another sample were to be taken, ~~the probability that the sample average in question lies within the 95% confidence interval is known than 0.05~~ a confidence interval of less than 95%.

9

p-value The p-value determines the minimum α that rejects H_0. If the α is selected after the sample has been taken, the α can be manipulated. Therefore, it is best to allow an unbiased individual to select the appropriate α. The p-value can be assumed to be larger than 0.001, but smaller than 0.005.

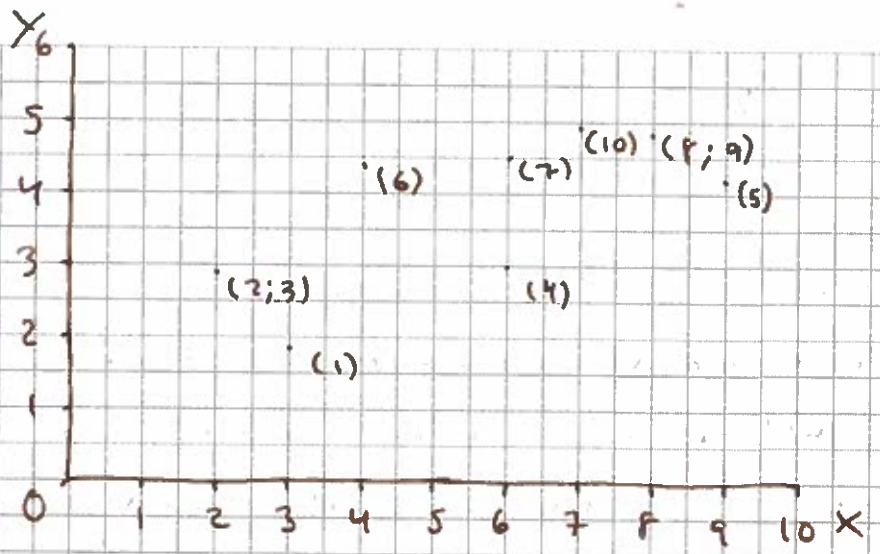
3

Consequence of wrong calculation (12.4) above but still not right, just almost

SU, DEPARTMENT OF STATISTICS

Room: UG Anonymous code: 0093-KDK Sheet number: 2

a)



There appears to be a positive ~~~~ linear relationship, meaning that the repair time increases linearly the longer it has been since a maintenance service. /6

b) Model I

$\bar{x}_i = 5.5$ $s_x^2 = 6.7222$ $s_e^2 = 0.78102^{\#2}$ $\alpha = 0.1$ $n = 10$

$t_{n-1;\alpha/2} \rightarrow t_{9;0.05} = 1.86$ $b_1 = 0.30413$ $b_0 = 2.14727$

$\hat{y} = 0.30413x + 2.14727$

Prediction interval:

$$(b_0 + b_1 x) \pm t_{n-2;\alpha/2} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)} \quad /3$$

~~~~ assuming / given $x = 4$

$$(2.14727 + 0.30413 \cdot 4) \pm 1.86 \sqrt{0.78102^2 \left(1 + \frac{1}{10} + \frac{(4-5.5)^2}{(10-1)\cdot 6.7222}\right)}$$

$= (1.81; 4.91)$ /3

It can be assumed that the probability of an $\bar{x}$ in another similar test lying inside the interval is 0.9. It is quite probable that if you hire the same maintenance personel, and have not had maintenance done in 4 months, that the issue will take between 1.81h and 4.91h to resolve.

ok /2

/8
~~~~

7 c) $\alpha = 0.05$ $b_3 = -0.60909$ $t_{n-2;\alpha/2} \rightarrow t_{F;0.025} = 2.306$ ✓

$df = n-K-1 = n-4 = 6$

These only apply to simple regress

$s_{x_3}^2 = \frac{5\cdot 1^2 + 5\cdot 0^2 + 10\cdot 1\cdot 5^2}{10-1} = 0.2777$ ok

$s_{b_3} = \sqrt{\frac{0,417443^2}{(10-1)\cdot 0.2777}} = 0.2640$

Interference for β_3

$b_3 \pm t_{10-2;\alpha/2} \cdot s_{b_3}$

$-0.60909 \pm 2.306 \cdot 0.2640 = (-1.22 ; -0.00029)$

Due to β_3's large relative confidence interval/interference, "Model II" would be recommended. Especially due to the fact that β_3 is ~~barely~~ not. (-0.00029) significant from zero. This is too large of a concern to only retain to "Model III"'s higher R^2 value.

→ Model II is recommended over Model III.

4

3