NTNU

KANDIDAT

10107

PRØVE

SØK1005 1 Økonometri

Emnekode	SØK1005
Vurderingsform	Skriftlig eksamen
Starttid	14.05.2025 07:00
Sluttid	14.05.2025 11:00
Sensurfrist	05.06.2025 21:59
PDF opprettet	11.09.2025 12:25

Seksjon 1

Oppgave	Tittel	Oppgavetype
ℹ	Front page	Informasjon eller ressurser
1	1	Langsvar
2	2	Langsvar
3	3	Langsvar

1 1

Spørsmål 1

Diskuter kort (3-5 setninger på hver) følgende begrep eller fenomen (12 poeng):

1. Utelatt variabel skjevhet
2. Gauss Markov Theorem
3. Multikollinearitet
4. Dummy variabel fellen (Dummy variable trap)

Skriv ditt svar her

1. **Omitted variable bias**

Omitted variable bias is the bias that occurs when a variable that affects both one or more of the explanatory variables x *and* the explained variable y is omitted from the model. By omitting the variable the effect of the x that is correlated with the omitted variable is biased, and does not show its true effect on the explained variable. This is a common violation of the MLR/SLR 4 about zero conditional mean.

2. **Gauss Markov Theorem**

Gauss Markov Theorem is a theorem that refers to the Ordinary Least Squares (OLS) method as the Best Linear Unbiased Estimator (BLUE) when all assumptions for the model holds. MLR/SLR 1-4 are crucial, while MLR 5 is an additional factor to making the estimator the "best", with little variation in the error term as well.

3. **Multicollinearity**

Multicollinearity appears when two or more of the explanatory variables are closely related to each other. This means when one explanatory variable can be written as a linear function of another. When explanatory variables are closely related, it is hard to distinguish the effect of each variable separately, on the explained variable. This is why assumption MLR/SLR 3 about no perfect collinearity is so important to be able to estimate a model.

4. **Dummy variable trap**

Dummy variable trap occurs when using categorical data, included as dummies, in a model. Dummies take the value 1 when the category is included in the model, and the value zero if not. For example female: female = 1, and therefore male: female = 0. When doing so one must always have a reference category, so the coefficient can be interpreted in relation to this. If all categories are included in the model, the dummy variable trap occurs, and there will be perfect collinearity and the model cannot be estimated.

Ord: 305

Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

1 8 1 1 6 3 4

2 2

Skriv ditt svar her

a) To use the OLS method to estimate the relationship between height and income in this sample there are some assumptions that must hold.

SLR 1: Linearity in the parameters, meaning that the β's must be included in the model linearly.

SLR 2: Random sampling. The sample must be independently and identically distributed so that the results can be generalized to the population. From the task it is clear that this assumption holds.

SLR 3: No perfect collinearity. As previously mentioned in question 1, for the model to be estimated there must be no perfect collinearity between the explanatory variables to be able to get estimates of the betas.

SLR 4: Zero conditional mean $(E(u|X) = 0$. This assumption is about how the error term cannot be correlated with any of the x's and the explained variable y. The expected value of u given x must be zero. This is the most important assumption to have an unbiased model. This is very difficult to not violate, especially using *simple* linear regression models, and in this model this is definitely violated, as there are factors, not included in the model (hence included in u) that affects both height and income, for example age or gender.

The purpose of the OLS estimator is to minimize the residuals. This means to minimize the errors of the model. In the drawing on sheet number 1, the spaces between the regression line and the observations, are notated with u, as this is the residuals. The OLS method sums the residuals, and squaring them to not let negative and positive residuals null each other out, and then minimizing this to get a model that explains a relationship to the best of its abilities, with as little residuals as possible.

b) The β_1 is the slope parameter. This tells us how much the explained variable, income, changes when the explanatory variable, height, changes by one unit, here: one centimeter.

c) This model is now a *multiple* linear regression model, meaning there is more than one explanatory variable. When interpreting β_1 , one must remember to do this holding the other variables equal (ceteris paribus). This is essential to understand the causal relationship between only height and income. In addition the β_1 will most likely change. This is because in the previous model, one can assume the coefficient was biased by omitted variables in the error term. By including one of those previously omitted variables, like gender, it makes the coefficient take on a value closer to its true value. This is because some of the effect of gender on income could have been absorbed in the effect of the variable height.

d) Assuming the effect of height is non linear, one can extend the model with the squared term of height. Like this:

$$Income_i = \beta_0 + \beta_1 height_i + \beta_2 height_i^2 + u_i$$

If β_2 is significant one can assume there is a non-linear effect of height on income. Depending on the sign of β_2 one can say wether the effect is diminishing or increasing. Most likely the effect is diminishing, and one can further estimate where the turning point is, and when the effect of height changes sign.

Ord: 524

Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

2 7 4 1 1 9 5

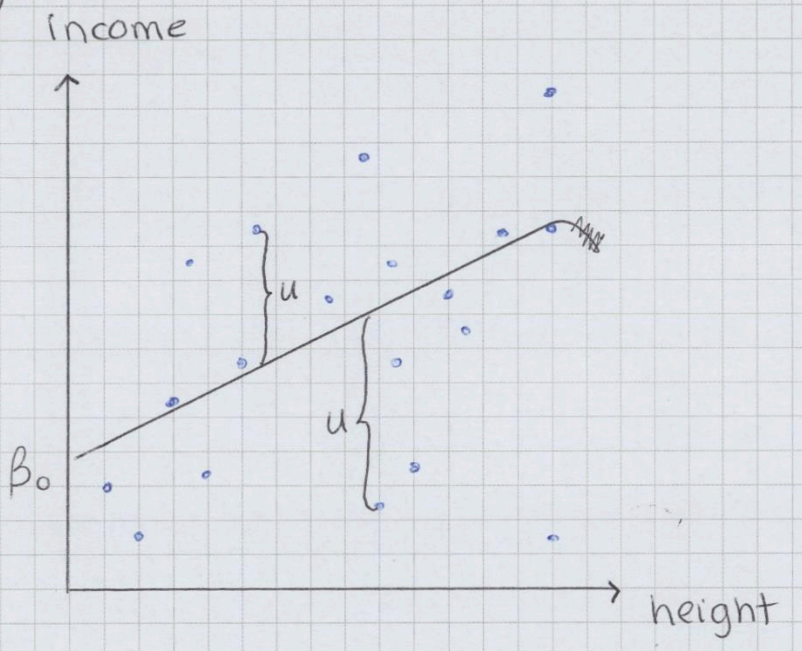
Håndtegning 1 av 1

Fill out Question Code and Test Information on every sheet. Fyll inn oppgavekode og emneinformasjon på alle skisseark.

Question Code Oppgavekode	Date Dato	Subject code Emnekode	Candidate ID KandidatID	Question no Oppgavenr	Page number Sidetall	Number of pages Antall ark
274 1 1 95	14.05.25	SØK1005	10107	2	1	1

Writing area Skriveområde

2a)



Form Number:INSPERA-02-160919-1 Provider:CADA Sistemas Informaticos www.cada.es

3 3

Skriv ditt svar her

For this task, because of no *, **, or *** in the coefficients in the tables, which is usually included to tell whether the estimates are significant or not, I will interpret all coefficients as if they are significant, except for when I calculate significance myself, that may prove otherwise.

a) An increase in education of one year is expected to increase monthly earnings by $ 60.21 on average, holding all else equal

b) $H_0: \ educ = 0, \ H_1: \ educ \ is \ not \ 0$
$t = \frac{60.21-0}{5.695} \approx 10.572$
The critical value is found in t table:
$t_{\frac{\alpha}{2}} = t_{0.0025} = c = 1.96$
As $c < t - stat$ we can reject the null hypothesis, and say that the estimated effect of education on wage is significantly different from zero at a 5% significance level and 935 degrees of freedom.

c) 95% confidence interval using the formula:
$\beta_1 \pm t_{\frac{\alpha}{2}} \cdot se(\beta_1)$
$60.21 \pm 1.96 \cdot 5.695$
$CI: \ [49.0478, \ 71.3722]$
This means we can say with 95 % confidence that the true population value is between 49.0478 and 71.3722.
Because 0 is not within the confidence interval, we know the estimates are statistically significant.

d) $R^2 = \frac{SSE}{SST} \implies \frac{16340644.5}{152716168} \approx 0.107$
The R^2 is a measure of the models goodness of fit. This means how much of the variation in y (here wages) is being explained in the model. The measure ranges from 0 to 1, so a R^2 of 0.107 implies that the model explains approximately 10.7 % of the variation in wages. Researchers do not normally focus a lot on the R^2 and a high R^2 does not necessarily make a model better than others. When using an estimator like OLS, the most important factor is unbiasedness so the results can be generalized to the population.

e) Column (1) IQ and experience is included in the model
All else equal:
An increase of 1 in IQ score is expected to increase monthly earnings by $5.609 on average
A one year increase in experience is expected to increase monthly earnings by $17.42 on average

f) By adding further variables that are not relevant to the model the R-squared will increase automatically. Therefore when running regressions, it is common to include an adjusted R-squared that takes the amount of coefficients into account. Including other variables could be helpful to increase the unbiasedness of the model. At the same time it is not helpful to add many variables that are very similar, as this will cause multicollinearity and "confuse" the effect of other variables.

g) Three new variables: married, black and sibs

Married and black are dummy variables and must be interpreted in relation to their reference categories.

Married - the monthly earnings are expected to be $175.1 higher for married people than non-married people, holding all else equal.

Black - monthly earnings are expected to be $112.9 lower for black people than for non-black people, holding other variables constant

Sibs - an increase in siblings by one is expected to reduce monthly earnings by $4.318 on average, holding all else equal

h) To figure out if the effect of education on monthly earnings is higher for individuals with high IQ, one should add an interaction term to the model.

From column 1:

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 (educ \cdot IQ) + u$$

By calculating the effect and significance of β_3 one is able to calculate how different levels of IQ affect the effect of education. on monthly earnings.

The effect of education without interaction term:

$$\frac{dwage}{deduc} = \beta_1$$

The effect of education with interaction term:

$$\frac{dwage}{deduc} = \beta_1 + \beta_3 \cdot IQ$$

i) Individual statistical significance:

Mother's education's significance:

$$H_0 : meduc = 0, \quad H_1 : meduc \; is \; not \; 0$$

$$t - stat = \frac{7.007 - 0}{6.156} \approx 1.138$$

Critical value from t-table at 5% significance level

$$t_{\frac{\alpha}{2}} = t_{0.0025} = c = 1.96$$

Because $c > t - stat$, we do not have sufficient evidence in the data to reject H_0 and cannot say that the effect of mothers education on monthly earnings is statistically different from 0.

Father's education's significance:

$$H_0 : feduc = 0, \quad H_1 : feduc \; is \; not \; 0$$

$$t - stat = \frac{11.20}{5.342} \approx 2.096$$

Critical value from t-table at 5% significance level

$$t_{\frac{\alpha}{2}} = t_{0.0025} = c = 1.96$$

Because $c < t - stat$, we reject the H_0 and we can say that the effect of fathers education on monthly earnings is statistically different from 0.

j) Joint significance between mothers and fathers education

$H_0: \ meduc = feduc = 0$
$H_1: \ not \ H_0$

Must conduct an F-test.

$$F = \frac{\frac{\left(R_{ur}^2 - R_r^2\right)}{q}}{\frac{\left(1 - R_{ur}^2\right)}{n-k-1}}$$

Unrestricted model is column (3), while restricted model is column (2).
$R_{ur}^2 = 0.194, \quad R_r^2 = 0.190$
q is the number of coefficient removed from the unrestricted model, here $meduc, \ feduc$
n is the number of observations
k is the number of coefficients in the unrestricted model excluding the intercept
$q = 2, \ n = 935, \ k = 8$

$$F = \frac{\frac{(0.194 - 0.190)}{2}}{\frac{(1 - 0.194)}{935-8-1}} = \frac{0.002}{0.00087} \approx 2.2988$$

Finding critical value in F-table at 5% significance, $df = 2$ in numerator and $df = 926$ in denominator
$c = 2.9957$

Because $c > F - stat$, we do not have sufficient evidence in the data to reject the null hypothesis that the joint effect of mothers' and father's education is significantly different from zero.

k) Column 4
All else equal:
A one year increase in education is expected to increase monthly earnings by 5.13% on average
One extra point in IQ is expected to increase monthly earnings by 4.27% on average.

l) If you were to include a dummy with non-married people in the model, one would incur in the dummy variable trap as explained in task 1.
This would take away the reference category and the coefficient would not have anything to be compared to and therefore make you unable to interpret or even estimate the coefficient.
There would be perfect multicollinearity as married and non-married directly correlate and could be written as linear functions of each other.

Ord: 903

Knytte håndtegninger til denne oppgaven?
Bruk følgende kode:

8 7 8 9 6 0 1