Økonometri I

Efterår 2021

Obligatorisk Opgave 2

Deadline: 11. november, 23:59

Lavet af:

Emil Møller Rasmussen **mgn531**

1 Problem 1

We are given a dataset containing nine different variables. As my student number ends with one (1), I will be using the dataset **groupdata1.dta**, and the results will therefore reflect this. As we throughout this assignment only will work with four of the variables, we will only include these in our descriptive statistics. Relevant statistics can be found in Table 1.

Variables	*mean*	*sd*	*min*	*max*	*count*	*datatype*	$datatype_{compress}$
OMS	112.12	128.65	4.73	539.59	250	double	double
KONK	1.3	.79	-2	2	250	double	byte
NYPR	.6	.49	0	1	250	double	byte
PRMRES	6.44	12.41	-23.94	76.46	250	double	double

Table 1: Variables, descriptive statistics and data types of **groupdata1.dta**

A total number of 250 observations are available for each variable. The included variables desribe factors that may influence the operating profits of a given company, i. Both $KONK$ and $NYPR$ have been coded as floating point variables (a double), but contain categorical and boolean data types respectively - both of which would have been better suited for integers. The compress command is used to ensure better suited data types.

The OMS and $PRMRES$ variables have high standard deviations, and their means are significantly closer to their *min*-values than their *max*-values, indicating they are positively skewed. This is further illustrated by plotting histograms of the variables, which can be found in Figure 1.

The value of $KONK_i$ is very left skewed, this makes sense either due a sampling bias, but also because few companies probably indicate that their area of business is lacking competetion - it is rare to hear a business say how easy they have it in the market. Though, it could be that our sample only consists of companies in very competetive markets, but that would require a different analysis on a larger population.

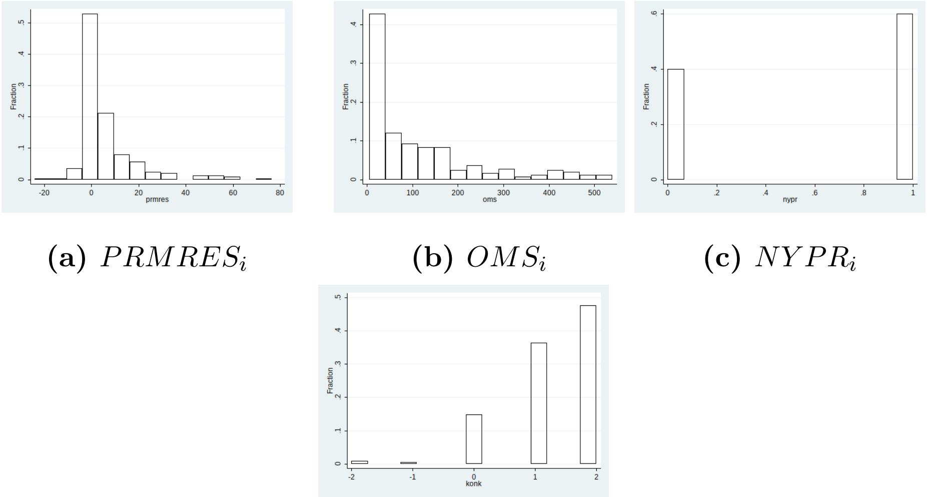
We have a 60%/40% split between companies that sell new products and companies that do not. Interestingly, a majority of companies in our dataset have an operating profit of zero, and a revenue around zero as well. Obviously revenue cannot be negative which makes it hard to have a uniform distribution, given the data congreation around zero.

2 Problem 2

We are given the following model:

$$PRMRES = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_i + \beta_3 NYPR_i + \beta_4 (NYPR_i \times OMS_i) + u_i \tag{2.1}$$

MLR.1-4 are assumed to hold.



(a) $PRMRES_i$ **(b)** OMS_i **(c)** $NYPR_i$

(d) $KONK_i$

Figure 1: Histograms of variables

1. The model given in (2.1) describes how the operating profits of company i are determined, based on a series of variables and their coefficients. The variables $KONK_i$ and $NYPR_i$ are categorical variables, whilst β_4 is an interaction term between a continous variable OMS_i and the binary variable $NYPR_i$.

   The ceteris paribus interpretation of β_3 is the expected average difference on $PRMRES$ between companies that innovate and companies that do not. We expect the sign of this variable to be negative, to reflect the initial cost associated with this.

   The ceteris paribus analysis of β_1 is how much the operating profits change when the revenue OMS_i changes by one unit. Ceteris paribus of β_4 indicates how much the operating profits change when the revenue (OMS_i) changes *if* the company also innovates - i.e. when $NYPR_i = 1$. In the case where it doesn't innovate $\beta_4(0 \times OMS_i) = 0$. Another way of looking at this would be to rewrite equation (2.1) to the following form:

$$PRMRES_i = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_i + (\beta_3 + \beta_4 OMS_i)NYPR_i + u_i \quad (2.2)$$

   In equation (2.2) it becomes apparent that when $NYPR_i$ changes 1 unit, the mean of $PRMRES_i$ changes by $(\beta_3 + \beta_4 OMS_i)$. A term that in turn is dependent on the value of variable OMS_i. Rewriting the equation also makes it apparent that the regressionmodel has a nested model.

   The coefficient β_2 indicates how much $PRMRES_i$ will change as the competetiveness of the environment the company is in, increases, all else fixed. Intuitively we would expect *more* competition to decrease the operating profits of the company, and we therefore expect the sign of β_2 to be negative - as a positive change (i.e. a more competetive environment) will negatively impact the operating profit of a company.

   It *could* be conceived that the value of β_2 is positive, if more competetive environments are likely to cause higher revenue - or if the causality is reversed, and higher revenues cause higher competition over time.

2. The estimated parameters of the regression on $PRMRES_i$ can be found in Table 2.

Variables	*coeff*	*se*	*t*	$P > t$
β_1*	.0566653	.0090631	6.25	0.000
β_2	.111483	.7438074	0.15	0.881
β_3	-.8530814	1.55664	-0.55	0.584
β_4	.0121786	.0105253	1.16	0.248
β_0	-.5131678	1.447223	-0.35	0.723
$R^2 = 0.4716$				

Table 2: Estimated parameters of equation (2.1) with standard error and R^2. * indicates statistical significance.

Replacing the coefficients in equation (2.1) with our estimates from Table 2 the regression model looks like this:

$$\begin{aligned} PRMRES_i = -0.513 + 0.0567OMS_i + 0.111KONK_i - 0.853NYPR_i \\ +0.012(NYPR_i \times OMS_i) + u_i \end{aligned} \tag{2.3}$$

It seems that competition has a positive effect on the operating profits of a given company and that β_4 is positive whilst β_3 is negative gives credence to the explanation that there is an initial investment tied to developing new products - a cost that diminishes as revenue increases. This could be explained by the fact that smaller percentages of the R&D costs eat away at larger companies profit margins.

The R^2 value associated with the regression indicates an average explanatory power of the variables - the current model explains $\approx 50\%$ of the SST. The standard errors indicate some issues with the estimators, as even though β_1 is statistically significant, all of the other coefficients are statistically insignificant, down to at least a 25% level.

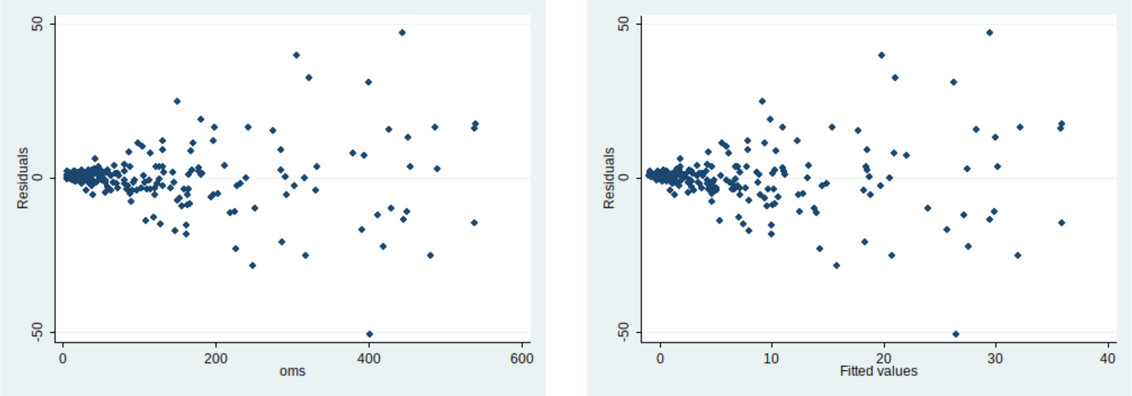
We are told that MLR.1 - 4 hold for the model, and we can therefore assume that the model is unbiased and consistent. We cannot say if the model is efficient until we analyse our residuals.

3. We have plotted $\hat{u}$ against OMS_i and $\hat{u}$ against $PRE\hat{M}RES_i$ in Figure 2.

The plots indicate that the model suffers from heteroskedasticity - i.e. that the residuals have non-constant variance. The interpretation of this result comes from the trumpet-like shape of the scatter plots. To verify this analysis we perform a **Breusch Pagen** test.

(a) Using the **Breusch Pagen** test, we use the residuals from equation 2.1. We then use the same predictor values from that equation, and set it equal to our residuals squared, giving us the following formula:

$$\hat{u}^2 = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_i + \beta_3 NYPR_i + \beta_4(NYPR_i \times OMS_i) \tag{2.4}$$



(a) $\hat{u}$ plotted against OMS_i

(b) $\hat{u}$ on $\hat{y}$

Figure 2: Testing for hommoskedasticity

Then we test

$$H_0 : \text{homoskedasticity is present}; H_A : \text{heteroskedasticity is present}$$

By running (2.4) in STATA we are given $R^2 = 0.28$. The χ^2 test statistic can be found by nR^2 which in this case is $250 \cdot 0.2818 = 70.45$. This gives $\chi^2 = 0.133529$. As $\chi^2 \not< 0.05$, we fail to reject the null hypothesis, thus heteroskedasticity is present.

We run a similar analysis in STATA using *hettest* and get $\chi^2 = 70.45$ with a p-value of 0.0000, further cementing our belief in the heteroskedasticity of the data. If the terms of the model had not been linear we could use the **Whites**-test - but as we are given that MLR.1 holds, we can safely use the **Breusch Pagen** test.

4. As the model is not homoskedastic we calculate the robust standard errors. To facilitate comparison between the robust and regular standard errors we include them in Table 3. As could be expected, our robust standard errors are lower than their the regular standard errors - which of course is expected as we are trying to select a method that better minimizes the standard errors of our regression. Still, it has not changed the statistical significance enough to create new insights about

Variable	*Std. Err.*	*Robust std. err*	*t*	*P >t*
β_1*	.0090631	.0065666	8.63	0.000
β_2	.7438074	.634616	0.18	0.861
β_3	1.55664	.8782156	-0.97	0.332
β_4	.0105253	.0126306	0.96	0.336
β_0	1.447223	.8721705	-0.59	0.557

Table 3: Regular and robust standard errors of (2.1)

the model. Du to heteroskedasticity our model might suffer from a misspecification error.

5. We do not know the form of the heteroscedastic error, so we will try to "fix" these issues by running a Feasible GLS procedure. We follow the steps on page 279 of Wooldridge and get the following equation (weighting our WLS by $\frac{1}{h}$):

$$\tilde{PRMRES_i} = .013 + 0.05OMS_i - 0.14KONK_i - 0.11NYPR_i + 0.0027(NYPR_i \times OMS_i) \tag{2.5}$$

This result is more consistent and more efficient than the OLS result from equation (2.3). Relevant statistics alongside this regression are included in Table 4

Variables	*coeff*	*se*	*t*	$P > t$
$\beta_1 *$	.0514788	.0058315	8.83	0.000
β_2	-.1440191	.2077447	-0.69	0.489
β_3	-.1079648	.4779372	-0.23	0.821
β_4	.0027416	.0092623	0.30	0.767
β_0	.125691	.3692735	0.34	0.734
$R^2 = 0.3612$				

Table 4: Estimated parameters of equation (2.5) with standard error and R^2. * indicates statistical significance.

We will now test two hypotheses:

i $\beta_2 = 0$: This hypthoses indicates that once revenue and the innovative factor have been accounted for, the competetive environment has no effect on the operating profits of a company.

As we only are testing *one* hypothesis - i.e. if a particular variable has no partial effect on the dependent variable - we choose to perform a t-test with $H_0 : \beta_2 = 0, H_1 \neq 0$.

The t-statistic for a double-sided null hypothesis has been calculated by STATA and is included in Table 2. The $t - value$ and $P > t$ are given as -0.69 and 0.489 respectively. Interpreting these values puts our t-value far from a critical value at any significance level - which also is indicated by $P > t$. This means that the variable is statistically insignificant. We therefore fail to reject the null hypothesis at every level, meaning the competitive environment, once all other included variables have been accounted for, has no effect on the operating profits of a company.

ii $\beta_3 = 0, \beta_4 = 0$: We now want to test the hypothesis that once revenue and competition have been acounted for, innovation, and the additional revenue caused by innovation, does not have an effect on the operating profits of a company.

As we are testing multiple restrictions on our regressionmodel we will perform an $F - test$ with $H_0 : \beta_3 = \beta_4 = 0$. $H_1 : H_0$ is not true. This means that if *any* of our restrictions are $\neq 0$ we reject H_0 in favor of H_1.

In this case equation (2.1) is the unrestricted model, and the restricted model can be seen in (2.6).

$$\begin{gathered} PRMRES_i = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_i + 0 \cdot NYPR_i + 0 \cdot (NYPR_i \times OMS_i) \\ PRMRES_i = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_i + u_i \end{gathered} \tag{2.6}$$

We calculate the F statistic by the following formula (using regression models fitted with WLS as above), $F = \frac{\frac{SSR_r - SSR_{ur}}{df_r - df_{ur}}}{\frac{SSR_{ur}}{df_{ur}}}$. To calculate this we estimate the relevant values from equation (2.1) and (2.6) giving us $SSR_r = 2031, SSR_{ur} = 1434$. We plug the values into the F equation: $F = \frac{\frac{2031-1434}{247-245}}{\frac{1434}{245}} = 51$. With $q = df = df_r - df_{ur} = 2$ in the numerator and $df = 245$ in the denominator. The critical value in a 10% F distribution is 2.35 and we reject H_0 if our F-value is $>$ the critical value. As $51 > 2.35$ we reject our H_0 and the variables are thus jointly significant. All the variables have a joint effect on the operating profits of companies. Also even though most of the estimators individually have no statistical significance.

iii We therefore see that we can test the parameters of the regression individually or in groups. The results might also show that even if some parameters individually are insignificant, together they might contain explanatory power.

3 Problem 3

1. To test if there is one unified model describing companies that innovate and companies that do not, we wish to test the null hypothesis if the groups follow the samme regression function, against the alternative that one or more of the slopes differ across the groups. To test this we must allow a model where the intercept and all slopes can be different for the groups involved. We therefore define the following model:

$$\begin{gathered} PRMRES_i = \beta_0 + \delta_0 NYPR_i + \beta_1 OMS_i + \\ \delta_1 NYPR_i OMS_i + \beta_2 KONK_i + \delta_2 NYPR_i KONK_i \end{gathered} \tag{3.1}$$

To that $PRMRES$ follows this model our hypothesis is $Ho : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0$ The unrestricted model is in (3.1), whilst the restricted model will be:

$$PRMRES_i = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_i \tag{3.2}$$

The $R^2_{ur} =$ and $R^2_r =$, $SSR_{ur} = 20178.45$,$SSR_r = 20389.48$, $df_{ur} = 244$,$df_r = 247$,$q = 247 - 244 = 3$. $F = \frac{20389.48 - 20178.45}{20178.45}\frac{244}{3} = 0.85$. With this F-value we fail to reject our null hypothesis, and the two groups therefore do not follow different models.

The coefficients and relevant statistics for can be found in Table 5. Looking at the $NYPR_i$ variable, we would assume that, holding all other factors fixed, operating profits are about 1.12 units higher.

Variables	*coeff*	*se*	*t*	$P > t$
δ_0	1.115434	2.40033	0.46	0.643
β_1	0.0568639	0.0090621	6.27	0.000
δ_1	0.0126376	0.0105305	1.20	0.231
β_2	0.8865055	1.034706	0.86	0.392
δ_2	-1.602676	1.487931	-1.08	0.282
β_0	-1.404965	1.666908	-0.84	0.400
$R^2 = 0.4741$	$\overline{R}^2 = 0.4633$			

Table 5: Estimated parameters of equation (3.1) with standard error and R^2. * indicates statistical significance.

4 Problem 4

1. It is most likely not appropriate to measure competitive environment as an index variable. The interpretation and encoding of the variable is completely subjective - we are not aware of any criteria the surveyee was provided to base their evaluation on, there are also no objective ways of differentiating what the difference between tough competetition and very tough competetion is.

2. An alternative specification of the given model, would be to provide a **dummy variable** for each category of the scale, this could be $KONK_m$, $KONK_n$,$KONK_t$, $KONK_{vt}$. To not fall into the dummy variable trap, we define the base of the model as $KONK_n = KONK_m = KONK_t = KONK_{vt} = 0 = KONK_{verymild}$. giving us the following formula:

$$PRMRES_i = \beta_0 + \beta_1 OMS_i + \delta_2 KONK_{i,m} + \delta_3 KONK_{i,n} + \delta_4 KONK_{i,t} + \delta_5 KONK_{i,vt} + u_i$$

This means that δ_0 is the expected response when we have the reference value, i.e. when we are at a very mild competetive environment. With estimators included we get:

$$PRMRES_i = -10.33 + 0.66 OMS_i + 9.66 KONK_{i,m} + 9.84 KONK_{i,n} + 9.44 KONK_{i,t} + 9.3 KONK_{i,vt} + u_i \tag{4.1}$$

For reference the old version is included here:

$$PRMRES_i = -1.20 + 0.066 OMS_i + 0.20 KONK_i \tag{4.2}$$

Immediatly we see that one value change in (4.3) in $KONK_i$ results in an increase in $PRMRES_i$ of 0.20, whilst every individual $KONK_{value}$ in equation (4.2) has it's own coefficient value. The R^2 value of the unrestricted regression is slightly higher than for the restricted model, but $\overline{R}^2$ values show the opposite picture. I.e. we are in a situation where, all else fixed, we can choose between two models that have circa the same explanatory power, but where the ease of interpretation or parsimoniousness differ.

3. Another way of writing three restrictions that imply a constant partial effect of competition is $\delta_3 = 2\delta_2, \delta_4 = 3\delta_2, \delta_5 = 4\delta_2$. We can then insert these into (4.1) and rearrange, which gives us:

$$\begin{aligned} PRMRES_i = \beta_0 + \beta_1 OMS_i + \delta_1(KONK_{i,m} + 2KONK_{i,n} \\ +3KONK_{i,t} + 4KONK_{i,vt}) + u_i \end{aligned} \tag{4.3}$$

This makes is to that the term multiplied by δ_1 is the original value of $KONK_i$. We will test $H_0 : \delta_2 = 0, \delta_3 = 0, \delta_4 = 0, \delta_5 = 0$, $H_1 : H_0 isnottrue$ and Our F-test in this case is then holding (4.1) as the unrestricted model, and (4.3) as the restricted model. $R^2_{ur} = 0.4733$, $R^2_r = 0.0136$, $df_{ur} = 244$,$df_r = 248$,$q = 4$. The F-value is thus

$$\frac{0.4733 - 0.0136}{1 - 0.4733}\frac{248}{4} = 54.11 \tag{4.4}$$

As this value is higher than the critical value at all significance levels we reject our null hypothesis and the individual parameters therefore have an effect on $PRMRES_i$ and should be in the model.

5 Problem 5

1. Throughout this assignment we have been introduced to the concepts of heteroscedasticity, nested models, interaction terms and dummy variables. The model provided, suffered from hereoscedasticity implying that it might have suffered from a misspecification error. First we tried to gain relevant insights by using robust standard errors, estimating the model with WLS and interpretating the outputs using R^2 and $\overline{R}^2$. As these tools did not provide significant extra explanatory power, different forms of respecification were attempted.

   First, evaluation of the used parameters was done - to see if some coefficients could be excluded either singularly or jointly. Then, more advanced respecifications were tried by exploring if innovative companies were adequately described by the current model, and finally the resegmentation of a categorical variable into a series of binary variables. The final model in equation (5.1) and results in table 6 try to summarize the knowledge gained throughout this assignment in a succint way.

$$\begin{aligned} PRMRES_i = \beta_0 + \beta_1 OMS_i + \beta_2 KONK_{i,m} + \beta_3 KONK_{i,n} + \beta_4 KONK_{i,t} + \\ \beta_5 KONK_{i,vt} + \beta_6(NYPR_i \times OMS_i) + u_i \end{aligned} \tag{5.1}$$

   The relationship between the operating profits, the competetive environment and innovation is hard to catch in one simple model. We have tried to create a more intuitive model with better explanatory power, but have still failed to describe the complexity adequately. When working with a small sample of data, and no access to the population, makes it hard to properly correct, or check, for issues in the underlying data that might seep into our validatory work here.

Variables	*coeff*	*se*	*robust std. err*	*t*	$P > t$
$\beta_1 *$	.0570656	.0090899	.0066226	8.62	0.000
$\beta_2 *$	9.594142	11.24483	3.91649	2.45	0.015
$\beta_3 *$	9.563612	6.641665	3.933068	2.43	0.016
$\beta_4 *$	9.155883	6.556622	3.925859	2.33	0.021
$\beta_5 *$	8.896665	6.551803	3.912299	2.27	0.024
β_6	-.9439366	1.569173	.8837346	-1.07	0.287
β_7	.0121008	.0105576	.0126799	0.95	0.341
$\beta_0 *$	-9.374185	6.51461	3.872485	-2.42	0.016
$R^2 = 0.4762$	$\overline{R}^2 = 0.4610$	$Prob > F = 0.0000$	$R^2_{robust} = 0.4762$		

Table 6: Estimated parameters of equation (5.1) with standard error and R^2. * indicates statistical significance.