Free University Brussels

Class: Econometrics

Assignment Econometrics

Professor:
Jeroen Kerkhof
Faculty of Economic Sciences

Group:
Marcio Reverbel
Max Temmerman

December 30, 2022

Contents

List of Figures

List of Tables

1 Cheating in the Sumo world

1.1 Context

A sumo tournament involves 66 sumo wrestlers (rikishi) fighting one match per day for 15 days. An overall win (8 wins) by a wrestler guarantees his advancement in the sumo ranking, while having 7 or less wins means that the sumo wrestler will fall in the ranking. Additionally, the difference in quality of life between the high and low ranked wrestlers is significantly large. Therefore, one can argue that there is an incentive to cheat if, after 14 days, a wrestler who has 7 wins and 7 losses (7-7) faces another wrestler who already has 8 wins: Since a wrestler with 8 wins will still advance in ranking, the expected gain for the 7-7 wrestler for a victory is higher than the expected loss for a wrestler with 8 victories already.

1.2 Evidence for cheating

One would expect a fair match between a 7-7 wrestler and another who achieved 8 wins and 6 losses (8-6) to be pretty even. In other words, if we were to take a random sample of 7-7 wrestler vs 8-6 wrestler fights (on the 15th day of sumo tournaments), we would expect the 7-7 wrestler to win approximately 50% of the time. Thus, our null hypothesis of would be that the mean proportion of wins by the 7-7 wrestler is be $p_0 = 50\%$ in the case that there is no cheating. Instead, the data has shown that 7-7 wrestlers win close to 75% of the matches against 8-6 wrestlers. That is a statistically significant deviation, indicating that it is very unlikely that this happens purely by chance. Therefore, it can be seen as strong evidence that matches are being rigged, and that the 7-7 wrestlers are buying wins from 8-6 wrestlers.

1.3 Selection bias in police investigation

The sumo wrestling world has also witnessed some strange deaths of sumo wrestlers, two of whom were going to be witnesses against the cheating scandal that had arrised and died both in the same day in the same hospital. The other was a young wrestler who had visible signs of mutilation, with a later autospy identifying that the wrestler had been burnt with cigarretes and beaten with baseball bats and glass bottles.

The strange thing is that for none of the three were the police willing to open a murder investigation, despite the very strange circumstances in which deaths had occurred. The autopsy performed on the young wrestler was only done after the victim's father insisted. It was later explained that the police boast a surprisingly

high and consistent success rate (above 96%) for murder investigations, which can be explained in the frame of selection bias. We can separate murders in two broad categories: murders with easily identifiable suspects (x_{1i}) and murders without easily identifiable suspects (x_{2i}). What the police in Japan is accused of is selecting cases of victims with easily identifiable suspects (x_{1i}) to be classified as murder victims, while the other victims, (x_{2i}), which would be naturally more difficult to solve, are closed as "abandoned body" cases. This selection process allows the police to boast an unusually high success rate. Let $p_{m,1}$ be the probability of a murder case being solved successfully in Japan, and $p_{m,2}$ the probability of a murder case being solved successfully in another country, such as Belgium. Then we would be interested in is:

$$E(p_{m,1}) - E(p_{m,2}), \tag{1}$$

while we can observe:

$$E(p_{m,1}|x = x_{1i}) - E(p_{m,2}|x = x_{1i} \cup x_{2i}). \tag{2}$$

If we subtract and add $E(p_{m,2}|x = x_{1i})$ to this equation, we get:

$$E(p_{m,1}|x = x_{1i}) - E(p_{m,2}|x = x_{1i}) + E(p_{m,2}|x = x_{1i}) - E(p_{m,2}|x = x_{1i} \cup x_{2i}). \tag{3}$$

The first half of this equation gives us the difference in success rate of murder investigations between two countries, given that the cases were easier to solve, while the second half of the equation makes the selection bias evident: It is the difference in a same country between the success rate given that the cases are easy to solve and the success rate considering all the cases.

2 Simulation study

2.1 Question 1

We start by creating a matrix $X_{n\times 3}$, with $n = 100$ observations. For the remainder of this section, when $n = 10$, the first 10 rows of X were used.

2.2 Question 2

Next, error terms ε were created following 4 different scenarios.

$n = 10$ and $\varepsilon \sim N(0, \sigma^2)$	$n = 10$ and $\varepsilon \sim Laplace(0, \sigma^2)$
$n = 100$ and $\varepsilon \sim N(0, \sigma^2)$	$n = 100$ and $\varepsilon \sim Laplace(0, \sigma^2)$

Table 1: The simulation scenarios.

Each scenario was simulated 2^{16} times, with $\sigma = 0.25$. The variance, σ^2 is considered to be known only to the "creator", and thus unknown to us when running the simulation.

2.3 Question 3

The dependent variable y was created for each simulation, such that $y = X\beta + \varepsilon$, where the column vector $\beta = (10, 0.5, 2.2)^t$ consists of the true values for the parameters (known only to the "creator").

2.4 Question 4

In Figure 1 we plot histograms for each simulation scenario of ε across all 2^{16} simulations. This allows us to see the difference between the Laplace and Normally distributed error terms. The top two graphs match the normal distribution extremely well, while the bottom two graphs match the Laplace distribution.

2.5 Question 5

Here we calculated the OLS estimates, the t-values for each $\hat{\beta}$ associated to the following null hypothesis $H_0 : \beta_i = [10, 0.5, 2.2]_{i+1}$, and the f-values associated with the model test, where $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$. The OLS estimates were calculated both manually, via the equation equation $\hat{\beta} = (X^tX)^{-1}X^ty$ and via the stats.models.OLS. The t-values were calculated manually using the residuals ($e = y - \hat{y}$) the estimate the variance of the error terms, such that $\hat{\sigma} = s = \frac{e^te}{\sqrt{n-3}}$. We can then calculate the

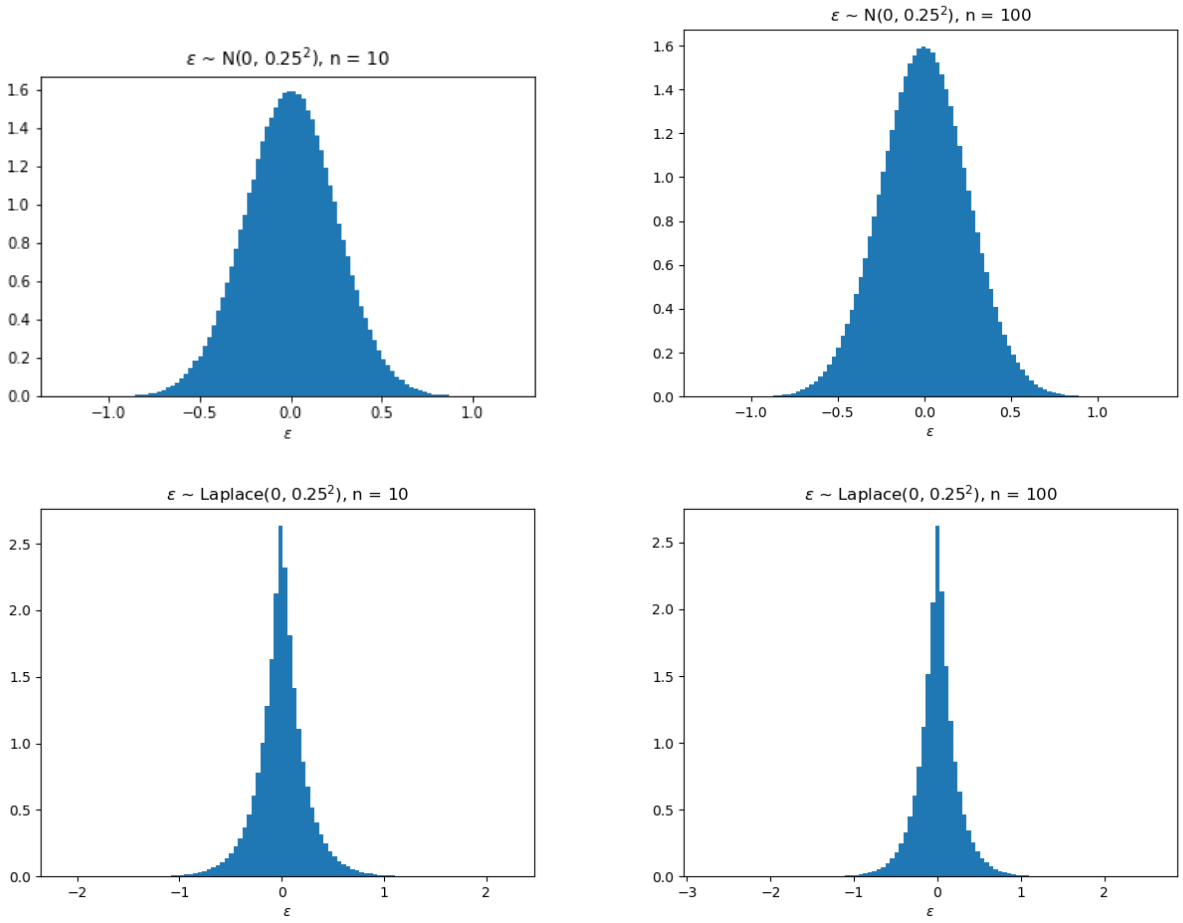

Figure 1: The histograms for the error terms for each simulation scenario.

t-values, which are t-distributed with $n - 3$ degrees of freedom (since we have two independent variables and are estimating the variance):

$$t_{\beta_i} = \frac{(\hat{\beta}_i - \beta_{i,0})}{s\sqrt{(X^t X)^{-1}_{i,i}}} \sim t_{n-3}. \tag{4}$$

Finally, the f-values were calculated using the OLS function from the stats.models.api package in Python.

2.6 Question 6

Based on the estimated quantities, we constructed histograms for the OLS estimates, t-values and the f-value under all simulation scenarios.

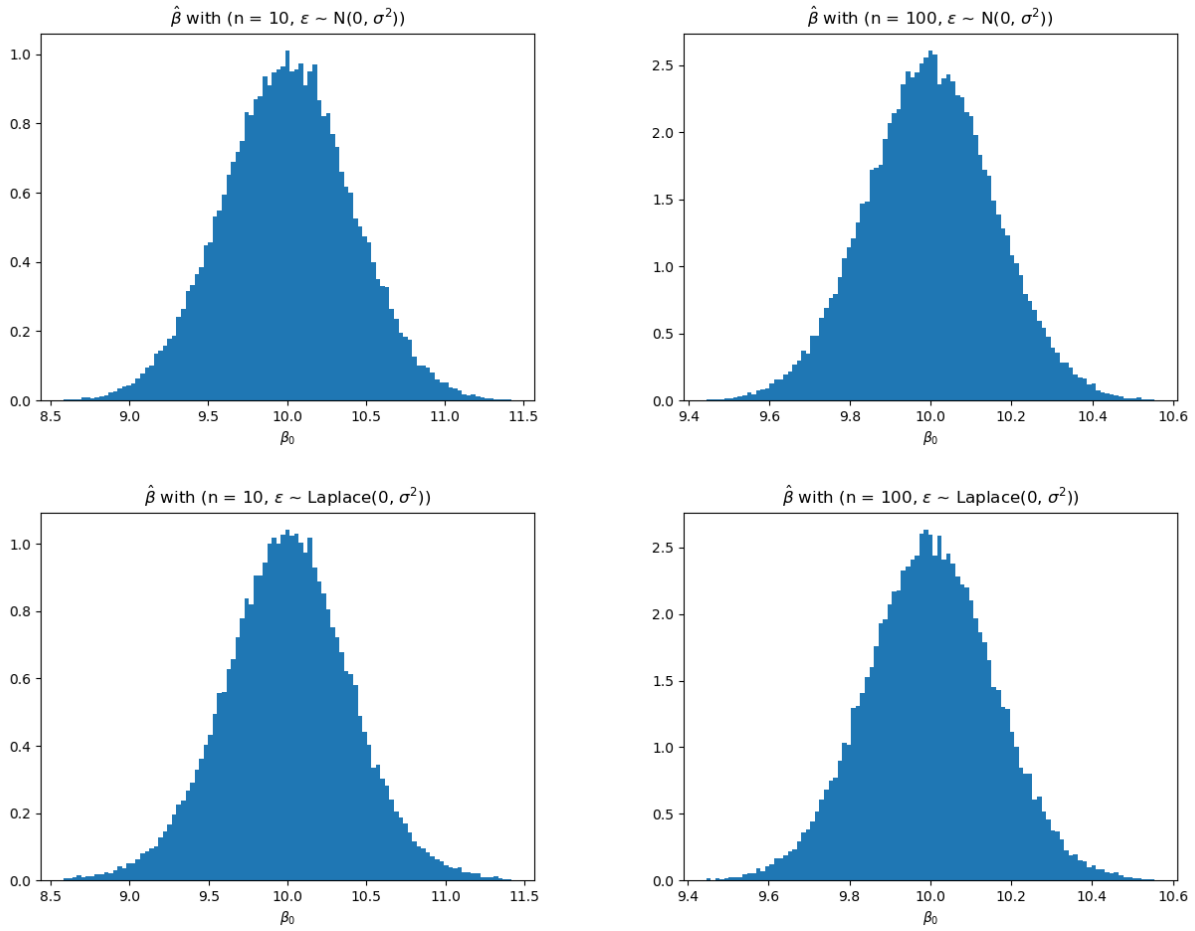

Figure 2: The histograms for $\hat{\beta}_0$.

As can be seen on Figures 2, 3 and 4, all OLS estimates are (approximately) normally distributed.

2.7 Questions 7 and 8

Similarly, the t-values associated with the estimated parameters and their respective null hypotheses can also be shown to follow approximately normal distributions (Figures 5, 6 and 7). When adding the probability density function (pdf) of the appropriate t-distribution in the graph, two things become immediately clear: First, how the data really matches the appropriate t-distribution extremely well, and the effect of the Central Limit Theorem in bringing the t-distribution increasingly closer to the normal distribution as the number of observations increases, visible when com-

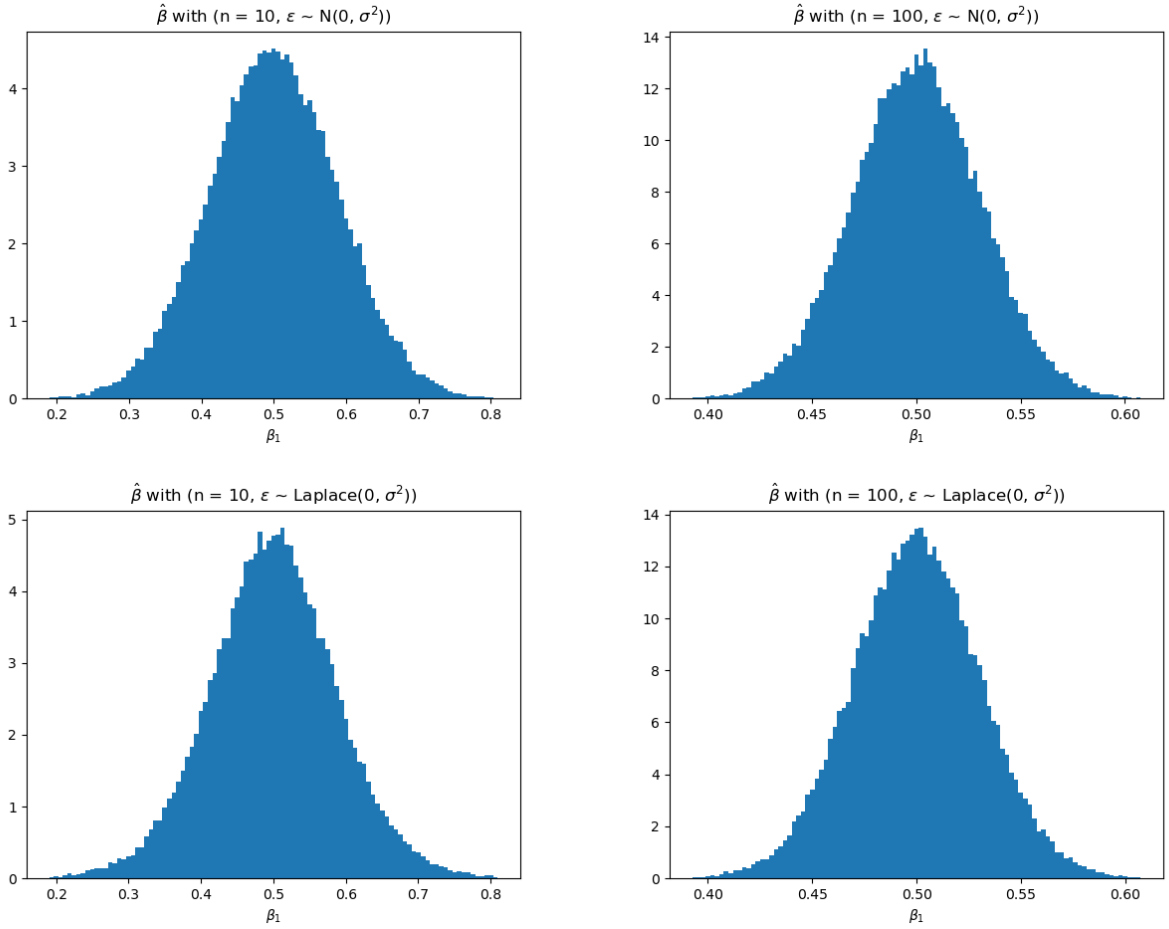

Figure 3: The histograms for $\hat{\beta}_1$.

paring the simulations of 10 observations with the simulations of 100 observations. Another remark is how the difference between Laplace and normally distributed error terms seems rather small in how they impact these distributions. Naturally, in all of these cases, we would fail to reject the null hypothesis most of the time since our data falls neatly within the area delineated by the respective pdf. Therefore, the proportion with which we would (wrongly) reject the null hypothesis aproximates α, marking the area under the graph associated with the relevant critical value, which also contains a minor amount of the data results.

However, when looking at the f-values (Figure 8), one can see that the data falls really far away from the appropriate F and X^2 distributions. The distributions are barely visible touching the horizontal line. This of course makes sense, since the true parameter values are all non-zero and the model test's null hypothesis is that

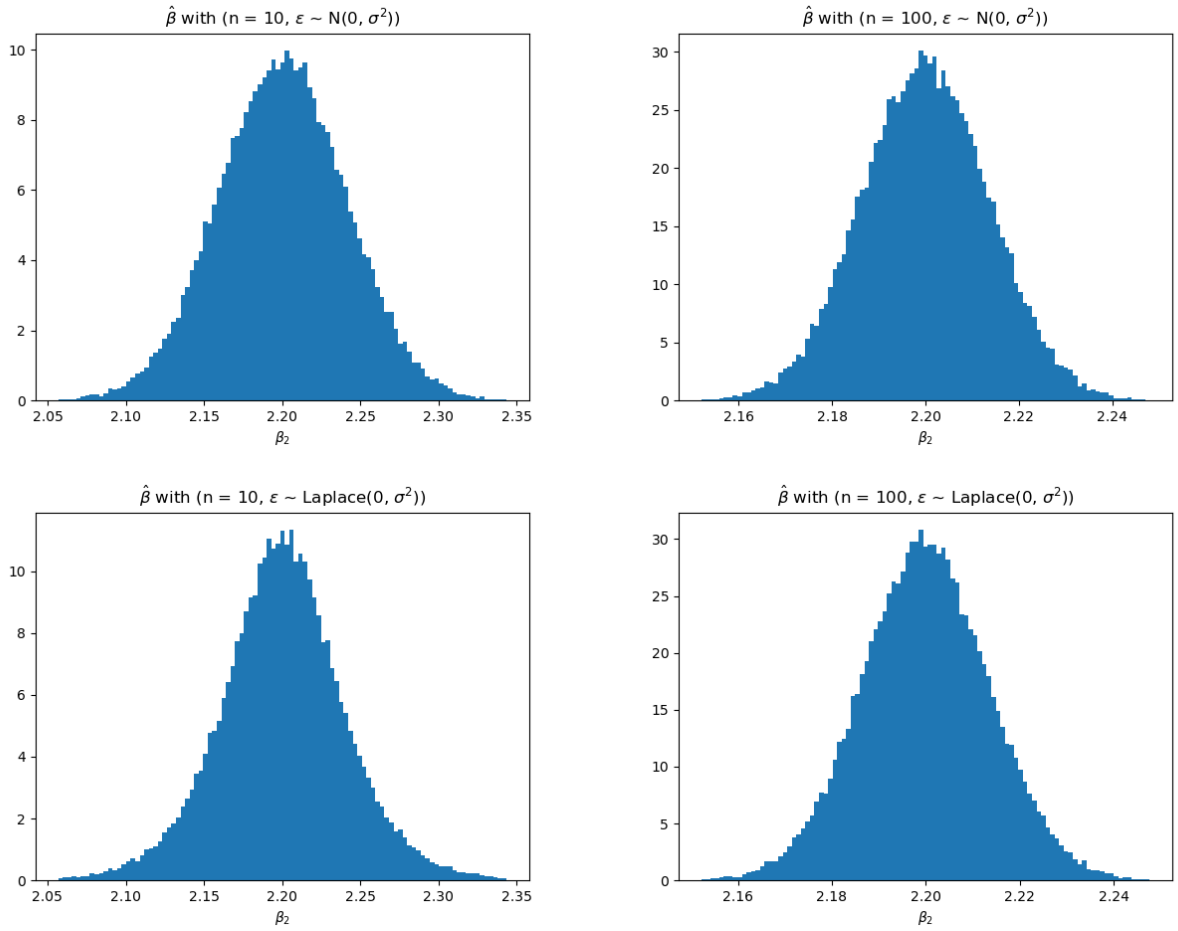

Figure 4: The histograms for $\hat{\beta}_2$.

they are all zero. (OBS: for the cases in which $n = 10$, a non-zero part of the pdfs should be visible. However, since there is a large difference in proportion between the histogram and the pdfs, they occupy such little space that the size of the bins is to large to make them appear. See the python code for more details.)

2.8 Question 9

These results show that parameter estimation and hypothesis testing are reliable under the conditions of normally and Laplace distributed errors, given that we were able to reject the null hypothesis when needed (model test), and we failed to reject the null hypothesis in all instances in which we would've liked to fail to reject the null hypothesis (when we tested against the true parameter values). Under these

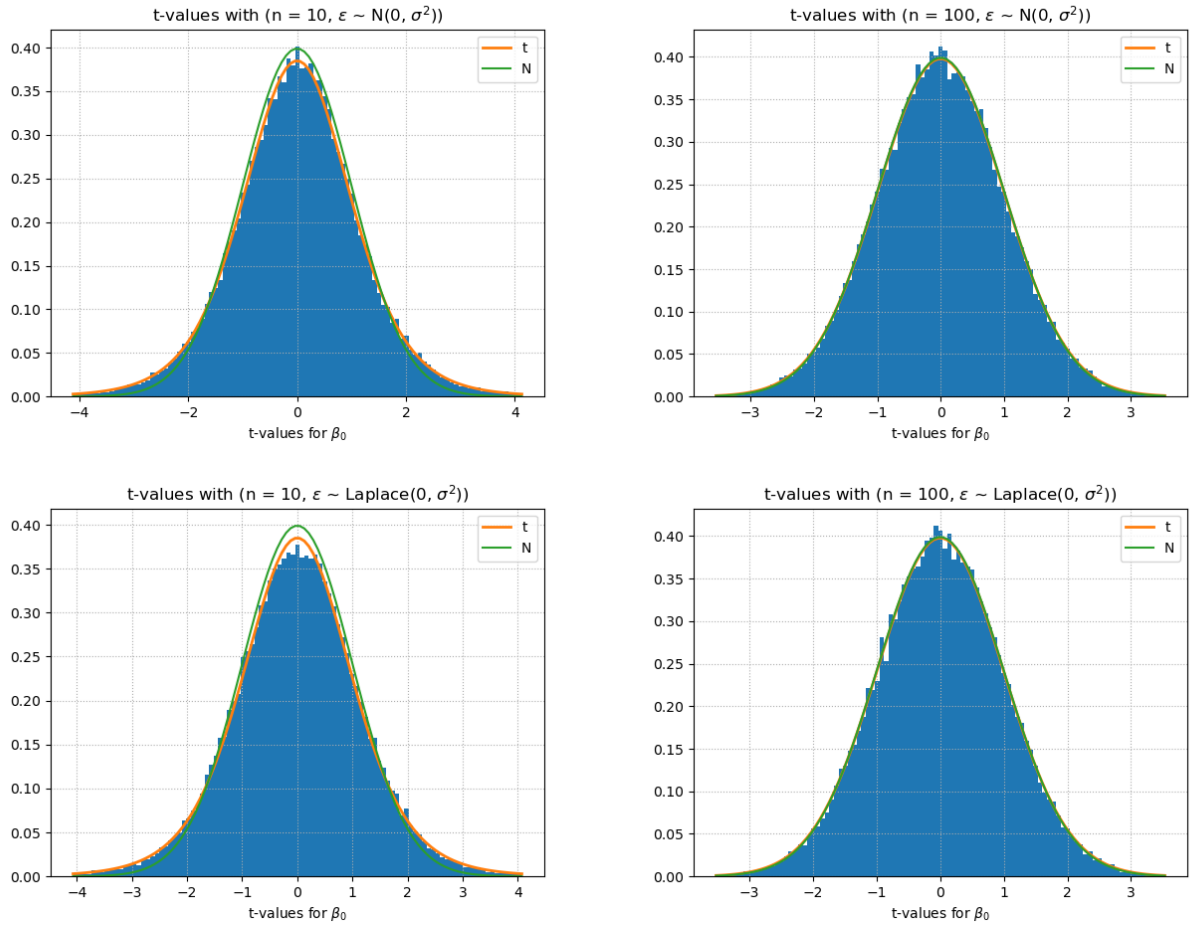

Figure 5: The histograms for $\hat{\beta}_0$'s t-values.

circumstances, we can't really say that any of these simulation combinations was problematic. A few remarks, though: Firstly, upon inspecting Figure 7, we note how the t-values under the Laplace distributed errors seem to peak higher than the appropriate t-distribution does, bringing it closer to the asymptotic normal distribution. Could this mean that under conditions in which errors are Laplace distributed, it becomes easier to rely on asymptotics? If that's the case, Laplace distributed errors become less interesting for larger sample sizes (due to overestimating the peak) and more interesting for smaller sample sizes.

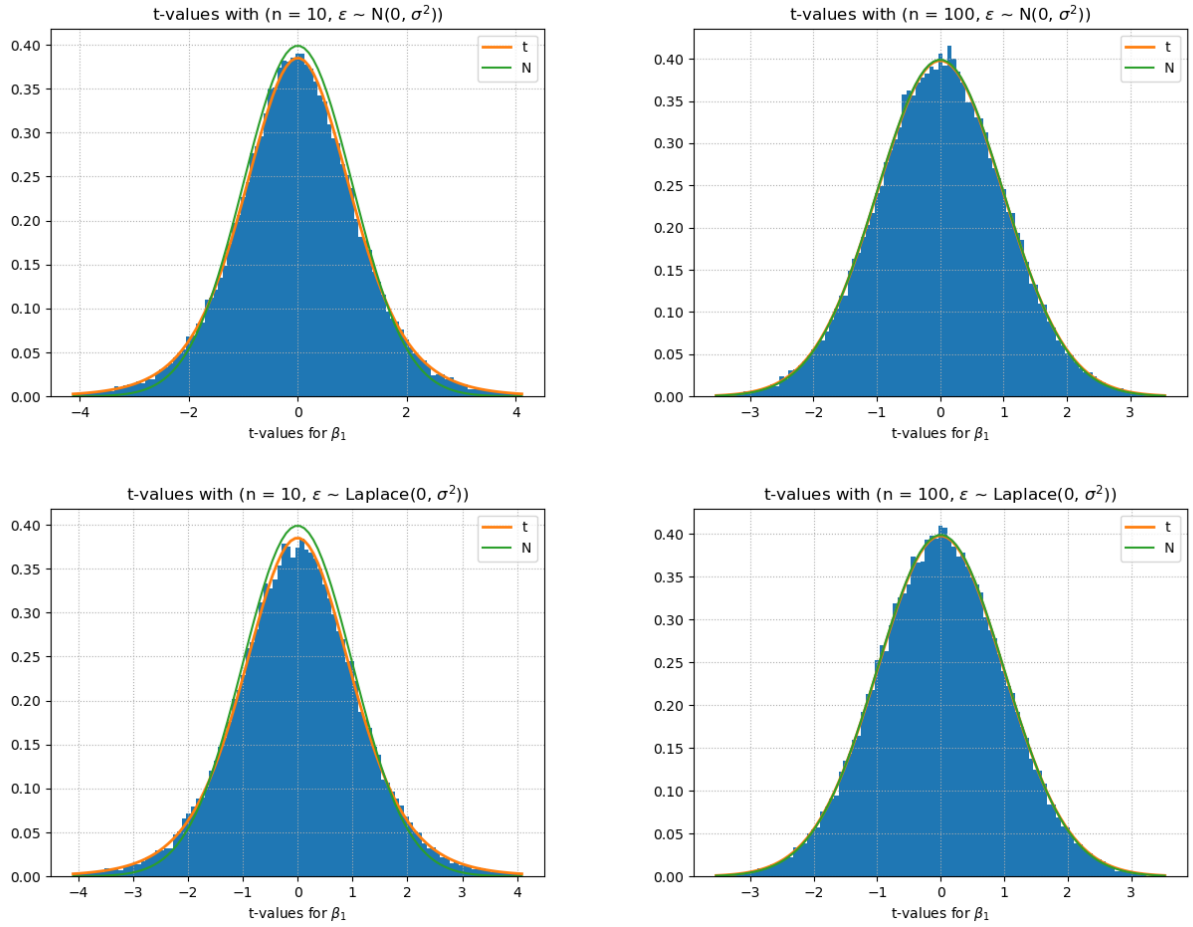

Figure 6: The histograms for $\hat{\beta}_1$'s t-values.

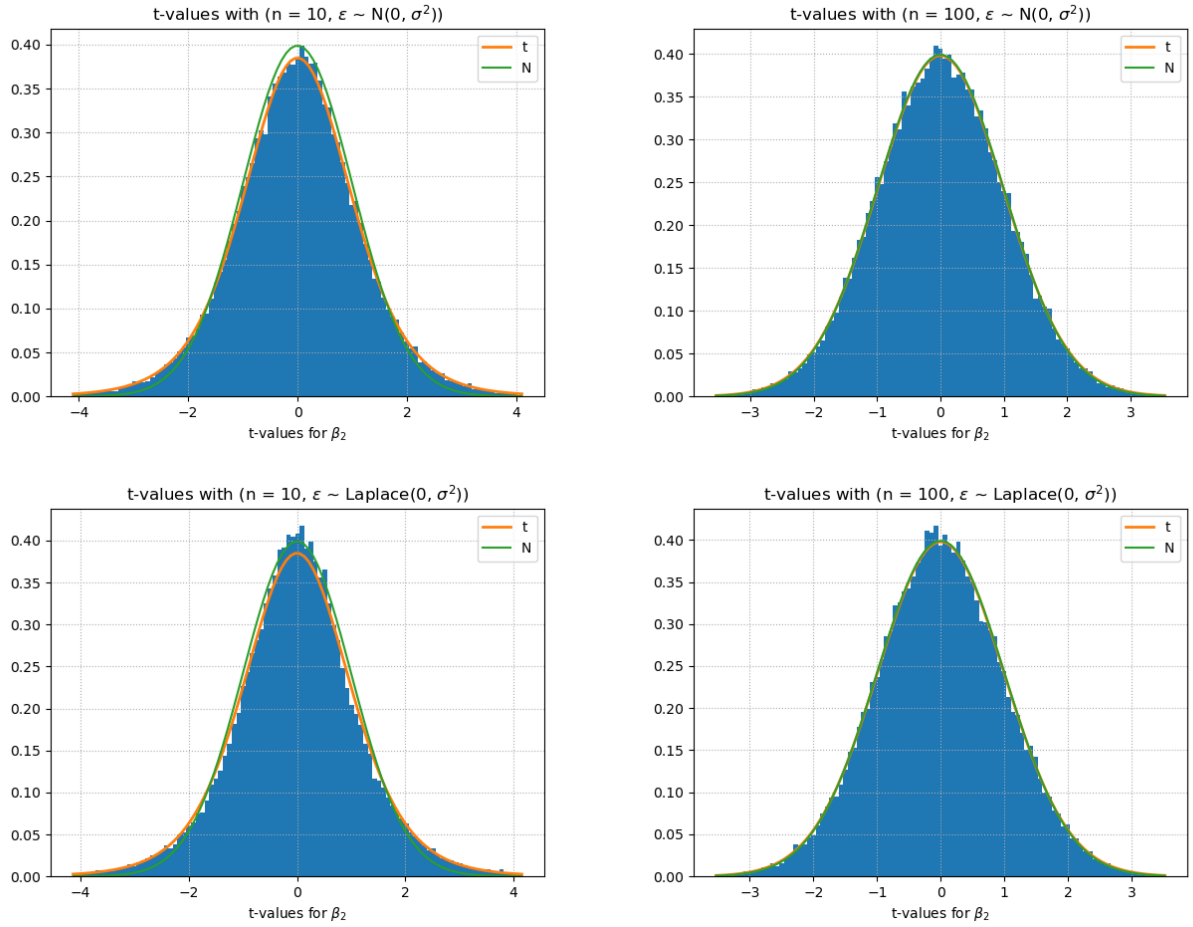

Figure 7: The histograms for $\hat{\beta}_2$'s t-values.

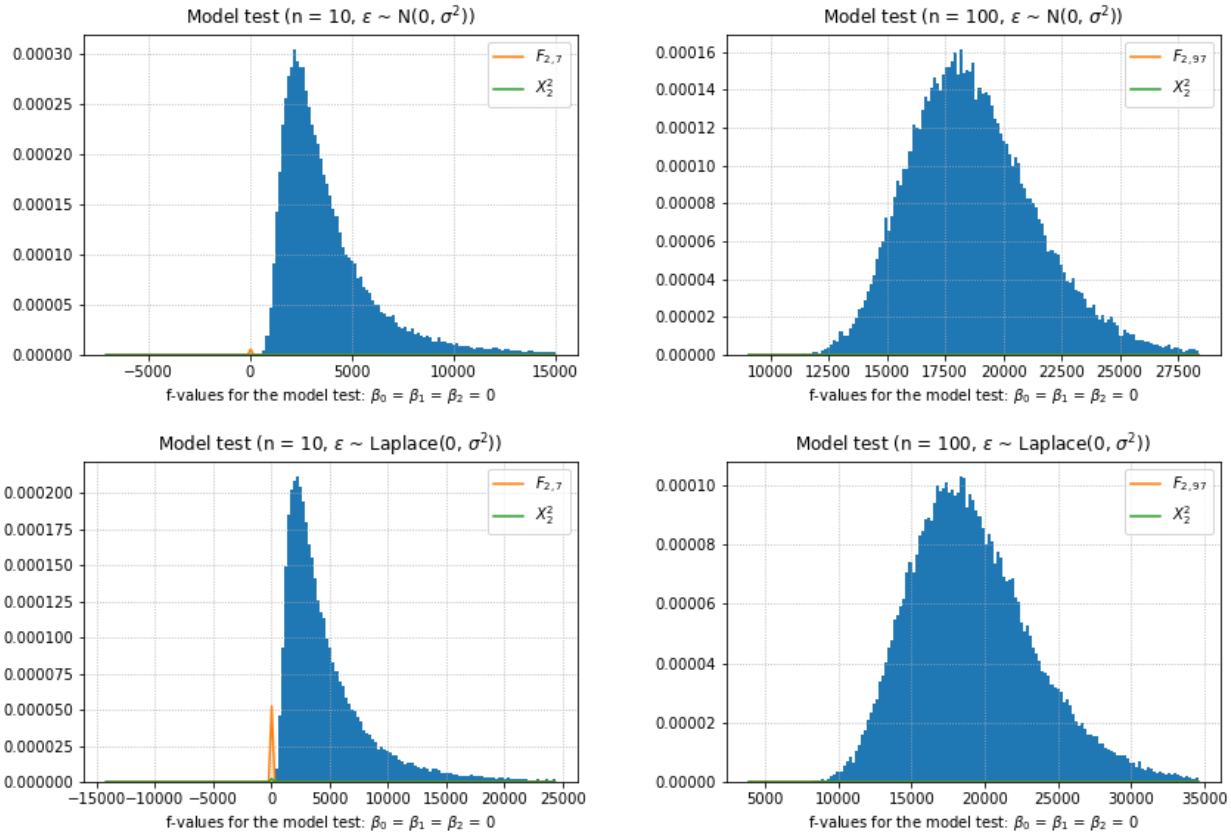

Figure 8: The histograms for the f-values for the Model test.

3 Empirical investigation

3.1 Question 1

The descriptive statistics are given in Table 2.

	count	mean	std	min	25%	50%	75%	max
female	6000.0	0.52	0.50	0.00	0.00	1.00	1.00	1.00
phsrank	6000.0	56.28	24.26	0.00	45.00	50.00	76.00	99.00
BA	6000.0	0.30	0.46	0.00	0.00	0.00	1.00	1.00
AA	6000.0	0.04	0.21	0.00	0.00	0.00	0.00	1.00
black	6000.0	0.10	0.29	0.00	0.00	0.00	0.00	1.00
hispanic	6000.0	0.05	0.21	0.00	0.00	0.00	0.00	1.00
id	6000.0	40640.08	24983.66	19.00	19429.00	39290.50	58838.50	89958.00
exper	6000.0	122.60	33.44	3.00	104.00	129.00	149.00	166.00
jc	6000.0	0.34	0.78	0.00	0.00	0.00	0.00	3.83
univ	6000.0	1.93	2.30	0.00	0.00	0.20	4.20	7.50
lwage	6000.0	2.25	0.49	0.56	1.93	2.28	2.60	3.91
stotal	6000.0	0.06	0.85	-3.32	-0.30	0.00	0.61	2.24
smcity	6000.0	0.29	0.45	0.00	0.00	0.00	1.00	1.00
medcity	6000.0	0.12	0.32	0.00	0.00	0.00	0.00	1.00
submed	6000.0	0.07	0.25	0.00	0.00	0.00	0.00	1.00
lgcity	6000.0	0.10	0.29	0.00	0.00	0.00	0.00	1.00
sublg	6000.0	0.09	0.28	0.00	0.00	0.00	0.00	1.00
vlgcity	6000.0	0.06	0.24	0.00	0.00	0.00	0.00	1.00
subvlg	6000.0	0.06	0.24	0.00	0.00	0.00	0.00	1.00
ne	6000.0	0.21	0.41	0.00	0.00	0.00	0.00	1.00
nc	6000.0	0.30	0.46	0.00	0.00	0.00	1.00	1.00
south	6000.0	0.33	0.47	0.00	0.00	0.00	1.00	1.00
totcoll	6000.0	2.27	2.33	0.00	0.00	1.51	4.37	10.07

Table 2: Descriptive Statistics for the data.

The first thing we notice upon looking at this table is that the 'id' variable is useless and carries no significant information. Secondly, we notice that most of the variables are binary, many of them composing groups of disjoint variables.

3.2 Question 2

This regression model is a log-lin model, meaning that the logarithm of the dependent variable can be written as a linear combination of the explanatory variables.

3.3 Question 3

$$\mathbf{log}(wage_i) = \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + \varepsilon_i$$

	Coef.	Std.Err.	z	P¿—z—	[0.025	0.975]
const	1.4579	0.0225	64.8366	0.0	1.4139	1.5020
jc	0.0674	0.0072	9.3228	0.0	0.0532	0.0816
univ	0.0765	0.0025	31.1583	0.0	0.0717	0.0813
exper	0.0051	0.0002	30.2261	0.0	0.0047	0.0054

Table 3: Results coefficient estimates model 3

All three variables are statistically significant since their p-values are far below the threshold of 0.05, this means that for each variable we reject the null hypothesis that it's coefficient is equal to zero. The three variables are also economically significant because the size of their coefficients is big enough for a realistic change in the variable to lead to an economically significant change in wage. For instance the variable experience has a very small coefficient but being continuous with a mean value of 122.60 means that a change in its size has an economically significant impact on the *wage*, the dependent variable.

3.4 Question 4

To test whether β_1 is equal to β_2 we use a t-test under the null hypothesis $H_0 : \beta_1 = \beta_2$:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

Since this statistic is t-distributed, we performed a t-test in Python using the *stats.modelsOLSResults.t_test*. The results are in Table 4.

	0
t test	-1.240
p-value	0.215

Table 4: t-test results

The result of our t-test has a p-value far above our α of 0.05 which leads us to fail to reject the null hypothesis.

3.5 Question 5

Now we want to test for $\beta_2 > \beta_1$. This means that our null hypothesis is $H_0 : \beta_2 \leq \beta_1$ but for this one-sided t-test we can't use the *stats.modelsOLS.results.t_test*, thus performed the one-sided t-test manually in Python.

To get the standard error of (β_2 - β_1) we use a small trick. We write $\theta = \beta_2$ - β_1 (this is the same as $\beta_1 = \beta_2 + \theta$). Plugging this into our regression equation, we get:

$$\mathbf{log}(wage_i) = \beta_0 + (\beta_2 + \theta)jc_i + \beta_2 univ_i + \beta_3 exper_i + \varepsilon_i \tag{5}$$

$$\mathbf{log}(wage_i) = \beta_0 + \theta jc_i + \beta_2(jc_i + univ_i) + \beta_3 exper_i + \varepsilon_i \tag{6}$$

We can now use the existing variable $totcoll = jc + univ$ and re-estimate our model with OLS:

$$\mathbf{log}(wage_i) = \beta_0 + \theta jc_i + \beta_2 totcoll_i + \beta_3 exper_i + \varepsilon_i \tag{7}$$

	Coef.	Std.Err.	z	P¿—z—	[0.025	0.975]
const	1.4579	0.0225	64.8366	0.0000	1.4139	1.5020
jc	-0.0091	0.0073	-1.2403	0.2149	-0.0235	0.0053
totcoll	0.0765	0.0025	31.1583	0.0000	0.0717	0.0813
exper	0.0051	0.0002	30.2261	0.0000	0.0047	0.0054

Table 5: Results coefficient estimates model 5

When we compare the results of the estimation of this with our prior model and the results of the t-test we performed above we see that the values for β_0, β_2 and β_3 and their standard errors haven't changed. We also observe that θ has the same z-score and p-value as the results of the t-test above, this makes sense because $\theta = \beta_2$ - β_1 and the z-score is calculated for H_0:$\theta = 0$

Now we can calculate the one-sided t-test for θ:

$$t = \frac{\hat{\theta}}{SE(\hat{\theta})}$$

This is equivalent to calculating a one-sided t-test between the difference of β_2 and β_1:

	0
t test	-1.240
p-value	0.107

Table 6: Results one-sided t-test

The resulting t-value is not very big and thus the p-value is not small enough for us to be able to reject the null hypothesis (α=0.05). In other words, we cannot ascertain that the difference between β_2 and β_1 is positive.

3.6 Question 6

Here we add the variable *phsrank* to our model and re-estimate the coefficients.

$$\mathbf{log}(wage_i) = \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + \beta_4 phsrank_i + \varepsilon_i \tag{8}$$

	Coef.	Std.Err.	z	P¿—z—	[0.025	0.975]
const	1.4456	0.0252	57.4582	0.0000	1.3963	1.4950
jc	0.0669	0.0072	9.2368	0.0000	0.0527	0.0811
univ	0.0752	0.0027	27.6405	0.0000	0.0699	0.0806
exper	0.0051	0.0002	30.1801	0.0000	0.0047	0.0054
phsrank	0.0003	0.0003	1.0898	0.2758	-0.0002	0.0008

Table 7: Results coefficient estimates model 6

In the results we see that *phsrank* has a relatively small t-value and a p-value far above 0.05. This means that we cannot reject the null hypothesis that it's coefficient β_4 is equal to zero. We conclude that *phsrank* is not statistically significant.

For a log-lin model, the effect of an increase in a variable which is measured as a percentage can be calculated algebraically. The absolute increase in the dependent variable x_i by 10 will result in a relative increase in the dependent variable, y_i, by β_i. We can calculate this more specifically, for an absolute increase by 10 in 'phsrank', y_i will increase by 0.028%.

From Equation 8, we can derive that:

$$wage_i = e^{(\beta_0+\beta_1 x_{1_i}+\beta_2 x_{2_i}+\beta_3 x_{3_i}+\beta_4 x_{4_i})}$$

If $phsrank(x_4)$ increases by 10%, remembering that it is already expressed as a percentage:

$$\Rightarrow wage_i = e^{(\beta_0+\beta_1 x_{1_i}+\beta_2 x_{2_i}+\beta_3 x_{3_i}+\beta_4 (x_{4_i}+10))} = wage_i \times e^{10\beta_4}$$

Therefore, an increase of 10% results in an increase of $e^{10\beta_4}$, or approximately 0.28%.

3.7 Question 7

Adding *phsrank* to our model doesn't substantively change our conclusions about the returns to two- and four-year colleges because if we compare the results from Table 3to Table 7, the size of the coefficients for *jc* and *univ* changes very little and their p-values remain far below 0.05.

3.8 Question 8

To test whether the overall model makes sense, we performed a model test:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$
$$H_A : \neg H_0$$

The reason why we chose the model test is because a model only retains some sense if the variables in the model are relevant. This essentially means that, if we fail to reject the null hypothesis in the model test, the proposed model doesn't add value to the analysis.

The test was calculated in two ways. We used the $OLSResults.wald\_test$ as well as the $OLSResults.f\_test$ in Python. The statistic is F-distributed with $4, n-5$ degrees of freedom, which is already very close to a X_4^2 distribution, given the large sample size. The value encountered for the F-test was F-val ≈ 433.30, which is much larger than the critical value, and thus we safely rejected the null hypothesis. The wald test (which is also F-distributed) arrives at the same conclusion as expected, with a statistic $W \approx 1733.20$, and corresponding $p-value \approx 0.0$. (In essence, these two tests are the same, but for some reason the calculation and statistics associated to each one is different in $stats.models.api$.)

3.9 Question 9

Next we computed the confidence and prediction intervals around $\hat{y}$ for all observations (see Figure 9).

3.10 Question 10

To account for possible heteroskedasticity, we performed the White test: H_0 : $Var[\varepsilon_1] = ... = Var[\varepsilon_n] = \sigma^2$.

	0
F test	3.24
p-value	0.00

Table 8: White test results.

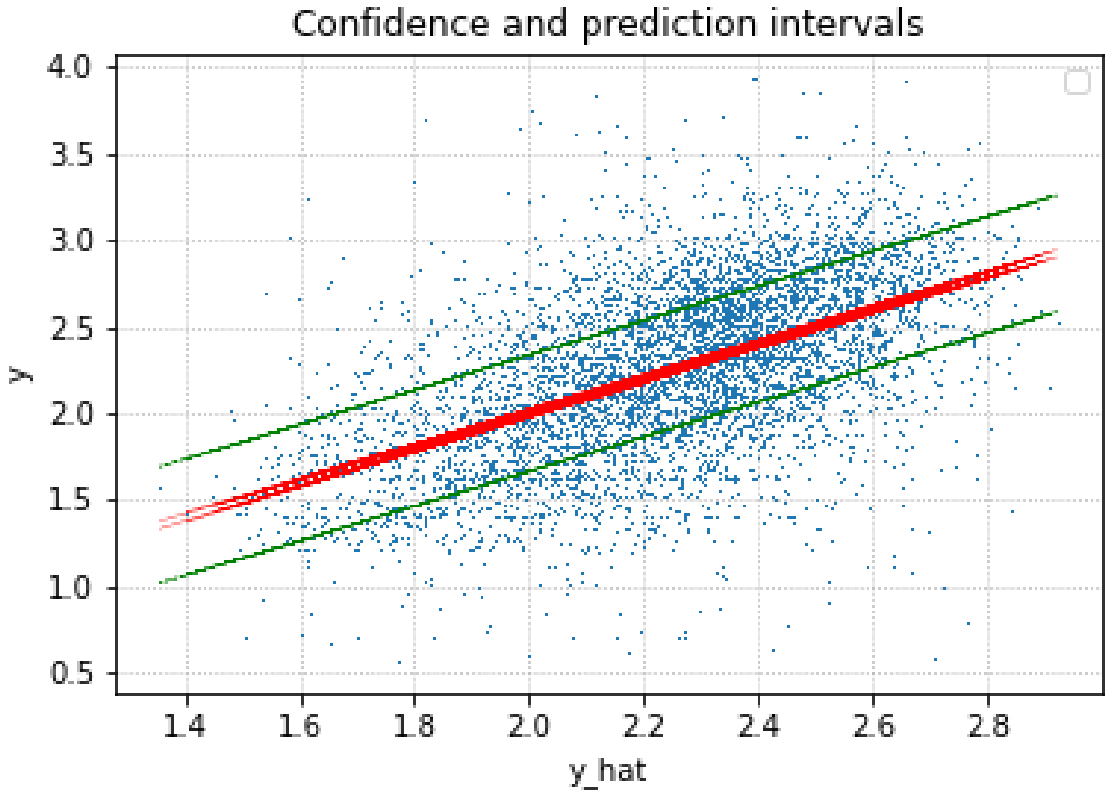

Figure 9: Describing the figure

As can be seen on in Table 8, the F statistic was found to be high enough to reject the null hypothesis ($p-value < 0.05$), which means we now need to adjust our standard errors to heteroskadasticity consistent (HC) standard errors. For this we used the *stats.models.api* $HC0\_se$ and cov_HC0 to calculate the adjusted standard errors, and re-perform the model test, taking the adjusted standard errors into account. After re-performing the model test (using the same methods described in Question 8), we found no significant differences in the results.

3.11 Question 11

When considering possibilities on how to improve the model, we first took note of how many of the independent variables were disjoint, such as all the city variables, or the ethnicity variables. Thus, we separated all the remaining variables in groups of disjoint variables, and performed regression analysis on each of these groups together with the original 'jc', 'univ' and 'exper'. BA and AA variables are particular cases because they very likely measure the very similar things as 'jc' and 'univ'. The difference is that one pair is binary, while the other is continuous. Otherwise, we might stipulate that they would be collinear (or measure an added effect when graduating).

From each of these analysis, we separated the variables which were found to be both statistically and economically significant, and created a model including all of these. We performed a White test, and since we rejected the null hypothesis, we then proceeded to adjust the standard errors for heteroskedasticity. Finally, we performed a model test to verify whether the model adds any value. We achieved a test statistic result of 244.01 (F test) and $p - value \approx 0.0$. What is curious is that this time the same result was achieved via both calculations $Results.wald_test$ and $Results.f - test$, and we suppose this might be related to the fact that the standard errors were adjusted for heteroskedasticity.

	0
F test	2.24
p-value	0.00

Table 9: White test for the new model

The final model presented is thus:

$$\begin{aligned}\mathbf{log}(wage_i) = \beta_0 + \beta_1 jc_i + \beta_2 univ_i + \beta_3 exper_i + \beta_4 female_i + \beta_5 black_i \\ + \beta_6 vlgcity_i + \beta_7 subvlg_i + \beta_8 nc_i + \beta_9 south_i + \varepsilon_i\end{aligned} \tag{9}$$

	Coef.	Std.Err.	t	P¿—t—	[0.025	0.975]
const	1.7649	0.0259	68.1492	0.0000	1.7141	1.8156
jc	0.0565	0.0070	8.1092	0.0000	0.0428	0.0701
univ	0.0610	0.0027	22.9066	0.0000	0.0558	0.0662
exper	0.0041	0.0002	24.1824	0.0000	0.0037	0.0044
female	-0.2206	0.0113	-19.5089	0.0000	-0.2428	-0.1984
black	-0.0460	0.0197	-2.3398	0.0193	-0.0846	-0.0075
stotal	0.0553	0.0073	7.5476	0.0000	0.0410	0.0697
vlgcity	0.1134	0.0228	4.9687	0.0000	0.0687	0.1582
subvlg	0.0745	0.0221	3.3707	0.0008	0.0312	0.1178
nc	-0.0697	0.0131	-5.3050	0.0000	-0.0955	-0.0439
south	-0.0789	0.0132	-5.9767	0.0000	-0.1048	-0.0530

Table 10: New model results

We also analysed the difference between using 'AA' and BA or using 'jc' and 'univ', and came to the conclusion the each pair of variables is addressing some of the main things, due to how strongly their coefficients change when in the presence one of another. In the end, we opted to use only 'jc' and 'univ' because, since they are continuous variables, we thought that they offer more nuance to the model. Addidionally, the existance of the totcoll variable, as previously demonstrated, allows us to make one-sided tests when comparing these 'jc' and 'univ'. That is not possible with the AA and BA variables. Additionally, the (adjusted) R-squared value is higher when using 'jc' and 'univ'. An argument could be made for the inclusion of all four variables, in such a way that 'jc' and 'univ' measures the effects of studies, while the BA and AA variables measure the added effect of getting a diploma when graduating. Here we opted for a more simple approach. We conclude the assignment

by plotting the confidence and prediction intervals around $\hat{y}$ for all observations for the new model (see Figure 10).

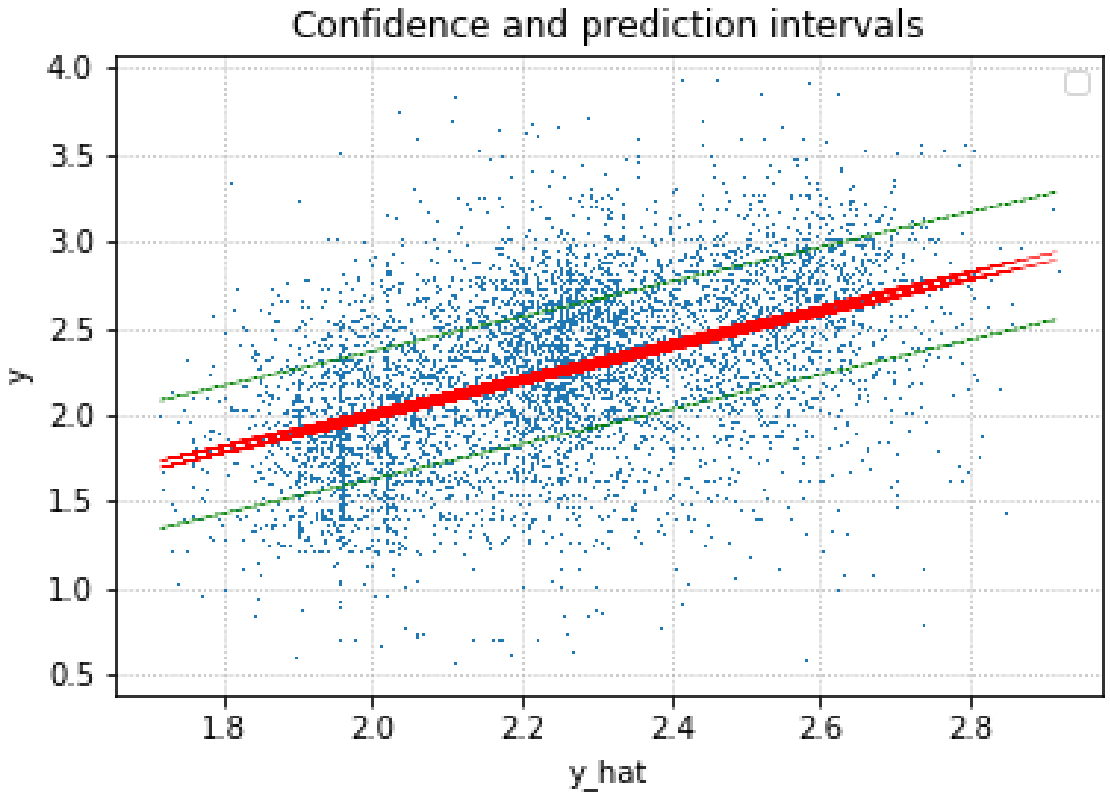

Figure 10: y vs $\hat{y}$ plot of the final model.