Free University Brussels

Class: Econometrics

Assignment Econometrics 2023

Professor:
Jeroen Kerkhof
Faculty of Economic Sciences

Group:
Aleksandr Medvedev
Mahina Lynn Lokwa

December 20, 2023

Contents

List of Figures

List of Tables

1 Selection bias

1.1 Real life examples

Selection bias is a critical issue in econometrics, occurring when vital information about a sample is neglected, leading to wrong conclusions. This bias can significantly distort the validity of econometric studies.

Example 1: Impact of Job Training Programs on Employment

Consider a study assessing the effectiveness of job training programs. If the sample only includes individuals who voluntarily enrolled in these programs, there is a high likelihood that these participants are more motivated or skilled than the average unemployed person.

Impact: This selection bias leads to overestimating the effectiveness of the training programs, as the results need to account for the inherent differences between those who choose to enroll and those who do not.

Example 2: Health Insurance and Overall Health Levels

In a study assessing the health levels of individuals with health insurance, there is a tendency to overlook those without insurance. People with a healthy lifestyle will probably be less interested in health insurance. Because their health quality is already high, the improvement in health due to treatment (getting insurance) might not be as profound as for the other group. Consequently, the study might falsely conclude that health insurance leads to better health outcomes.

Impact: Such bias could lead to misguided health policies that overestimate the effectiveness of health insurance on health outcomes.

Example 3: Analysis of Subsidized Housing Benefits on Employment Rates

In this example, consider a study assessing the impact of subsidized housing on employment rates. The study focuses on individuals who have successfully applied for and received subsidized housing. This group may inherently differ from the general population in ways that could affect employment rates, such as having lower incomes, different family compositions, or varying levels of education.

Impact: By only examining individuals in subsidized housing, the study might conclude that such housing leads to higher or lower employment rates. However, this conclusion could be skewed because it fails to account for the pre-existing differences between those who receive housing subsidies and those who do not.

Handling selection bias

Across econometric studies, randomized trials are a fundamental strategy to combat selection bias, the goal being to compare ”apples to apples”. We want two groups that are the same on average, but only differ in the treatment provided.

1.2 Elon Musk on Rogan

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, emerged in late 2019 and rapidly evolved into a global health crisis. Characterized by symptoms ranging from mild flu-like signs to severe respiratory distress, the virus's profound impact on the lungs. This severe lung condition, seen in critical COVID-19 cases, impairs the body's ability to supply oxygen to vital organs. Consequently, many patients required mechanical ventilation support. Ventilators became a critical lifeline, providing necessary oxygen by taking over the body's breathing process when the lungs failed, allowing patients to survive the most severe phase of the illness while their immune systems fought the virus.

Let p_{1i} be the probability of death on the ventilator, p_{0i} be the probability of death without the ventilator, and $x_i = \{1, 0\}$ is the dummy variable where 1 indicates the patient being put on the ventilator and 0 indicates the patient not put on the ventilator, for all severe cases.

We would be interested in the difference between the mortality rates of the group of people who were put on the ventilators and those who were not:

$$E[p_{1i}|x_i = 1] - E[p_{0i}|x_i = 0]. \tag{1}$$

These are the values that we can observe. Subtracting and adding an unobserved outcome $E[p_{0i}|x_i = 1]$ gives:

$$E[p_{1i}|x_i = 1] - E[p_{0i}|x_i = 1] + E[p_{0i}|x_i = 1] - E[p_{0i}|x_i = 0]. \tag{2}$$

The first part of the equation is the average causal effect between the two groups, and the second part of the equation is the selection bias, namely the difference between the mortality rates of people who were not on the ventilation system, have they been put on the ventilation and people who were not on the ventilation system and were not put on the ventilation.

$$\underbrace{E[p_{1i}|x_i = 1] - E[p_{0i}|x_i = 1]}_{\text{Average causal effect}} + \underbrace{E[p_{0i}|x_i = 1] - E[p_{0i}|x_i = 0]}_{\text{Selection bias}}. \tag{3}$$

Elon Musk does not consider the selection bias mentioned above thus arriving at a wrong conclusion.

2 Simlulation Study

2.1 Question 1

We start by creating the x values for $n = 100$ and $n = 1000$ normally distributed with an expected value of 20 and a variance of 2, $x_1 \sim N(20, 2)$. Then we proceeded with adding a constant, creating a n x 2 matrix X.

The figures below provide a visual description of the X matrices.

$$X_{100} = \begin{bmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{1,100} \end{bmatrix} \quad X_{1000} = \begin{bmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{1,1000} \end{bmatrix}$$

2.2 Question 2

Here we created the error terms for each simulation for the heteroskedastic and homoskedastic case each with $n = 100$ and $n = 1000$, set the true values for β, namely $\beta_1 = 1$, $\beta_2 = 2$. Each scenario was simulated 2^{16} times with $\sigma = 0.3/\sqrt{20}$.

num obs	heteroskedastic	homoskedastic
$n = 1000,$	$\epsilon_i \sim N(0, \sigma^2 x_i)$	$\epsilon_i \sim N(0, \sigma^2 \bar{x}_i)$
$n = 100,$	$\epsilon_i \sim N(0, \sigma^2 x_i)$	$\epsilon_i \sim N(0, \sigma^2 \bar{x}_i)$

Table 1: Simulation combinations

2.3 Question 3

Given the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4}$$

We have generated dependent variable y for each of the simulations combinations, the resulting generalized matrix is the following:

$$Y = \begin{bmatrix} y_1^{(1)} & \cdots & y_1^{(S)} \\ \vdots & \ddots & \vdots \\ y_n^{(1)} & \cdots & y_n^{(S)} \end{bmatrix}$$

2.4 Question 4

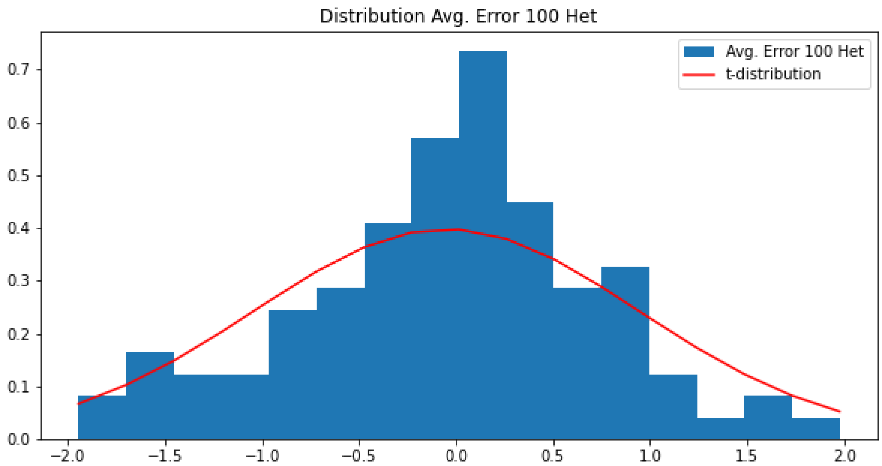

Figure 1

Figure 1 and Figure 2 plots the average error term for the heteroskedastic case. As we can see both distributions are centred around zero as expected by the definition of the error term above. The smaller sample size, namely $n = 100$ seem to follow the t-distribution by having a bit fatter tails than the distribution in the Figure 2. The latter gets closer to normal distribution as the sample size grows, which is also expected by the statistical theory.

Results for the homoskedastic average errors are reported in the Figure 3 and Figure 4, the description of the behaviour of these graphs are similar to those in the heteroskedastic case.

We can observe the heteroskedasticity with the scatter plot, such as in Figure 5 and Figure 6. The higher the x-axis of the fitted value, the higher the peak values will be for the squared error term. This represents the definition of heteroskedasticity.

In contrast to previous figures Figure 7 and Figure 8 represent the homoskedastic squared error terms, where peaks are on average the same and do not depend on the fitted values.

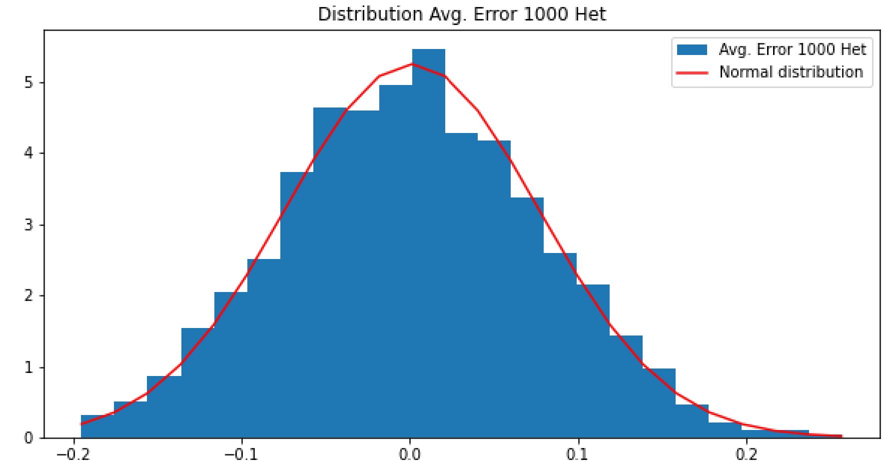

Figure 2

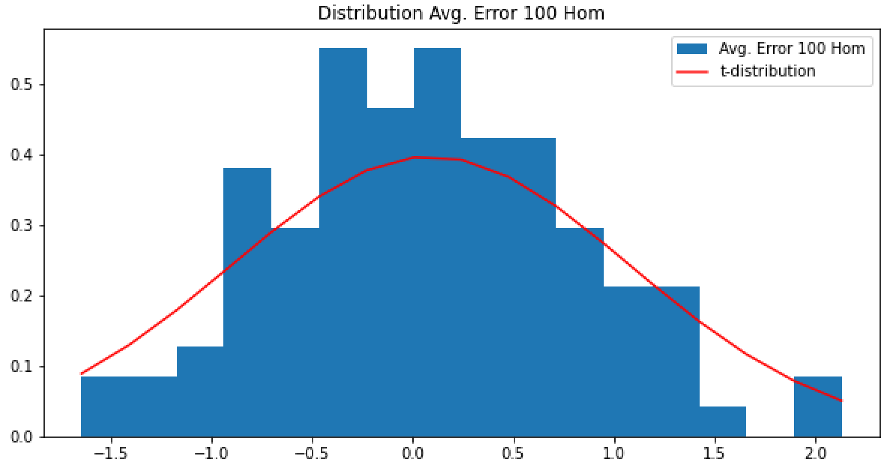

Figure 3

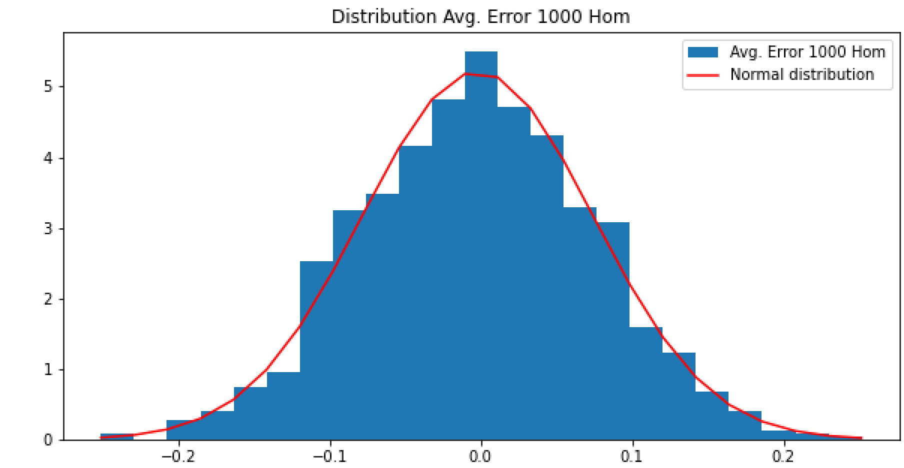

Figure 4

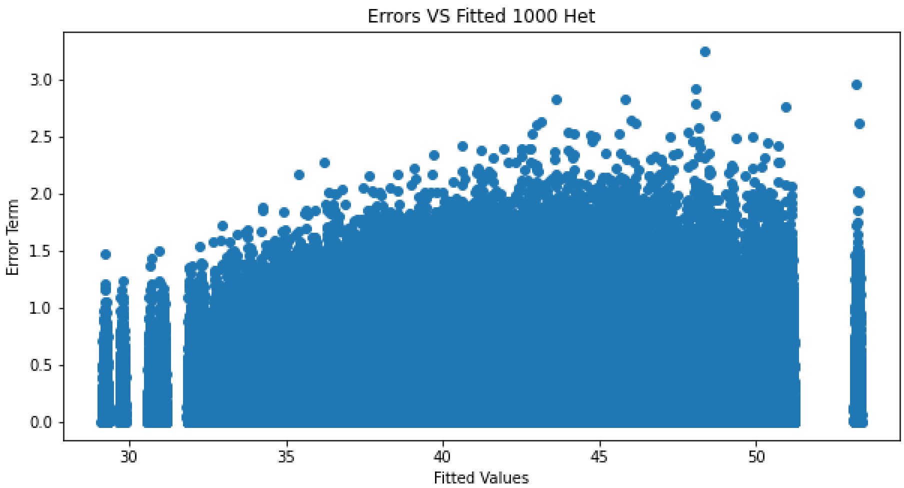

Figure 5

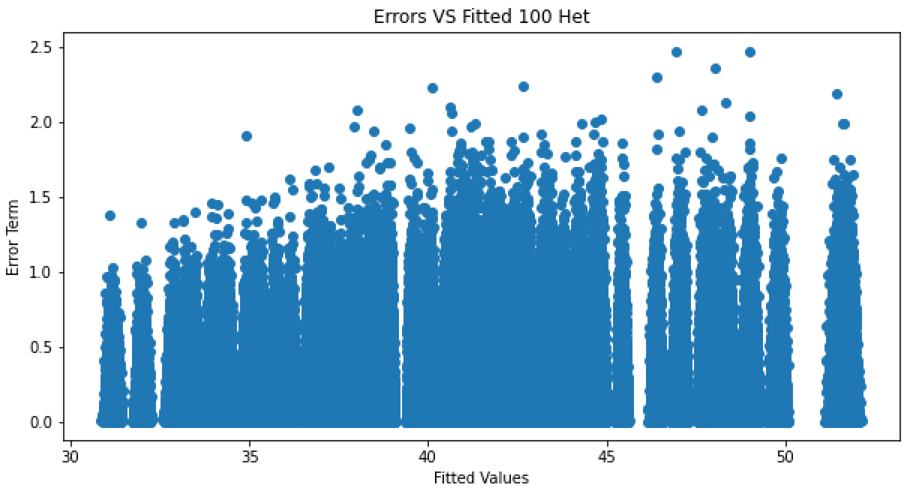

Figure 6

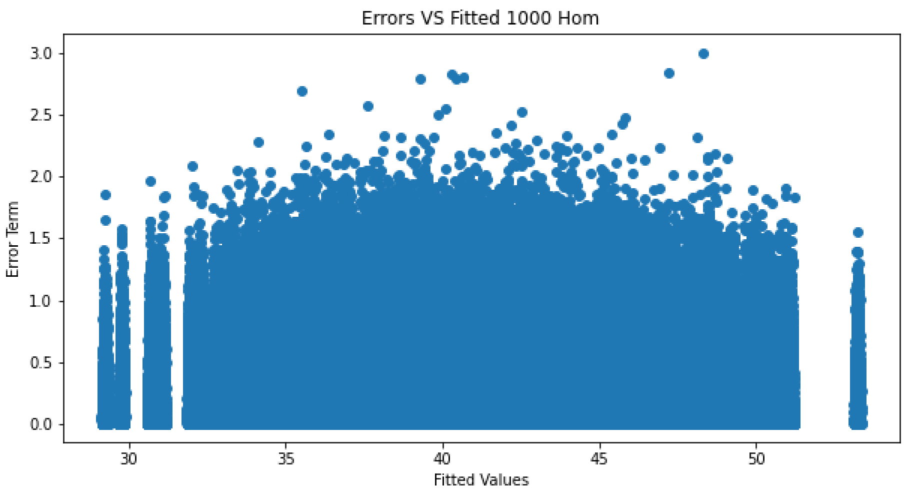

Figure 7

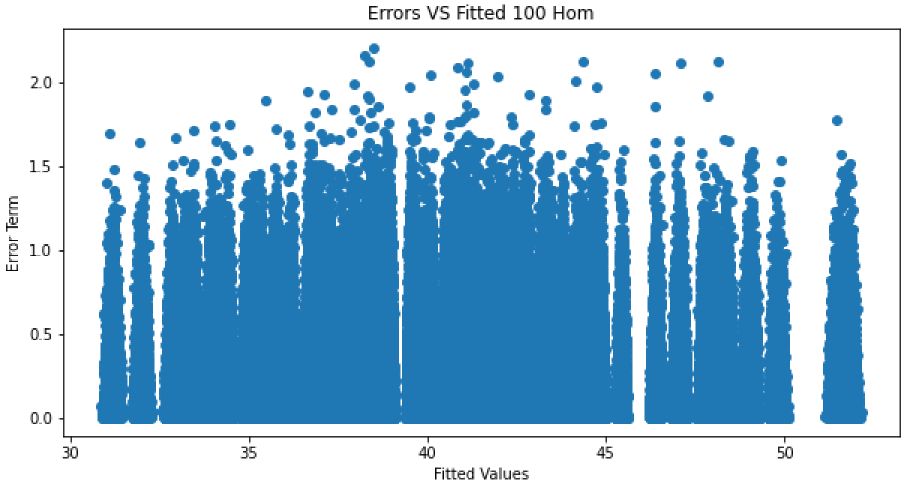

Figure 8

2.5 Question 5

In case of the heteroskedastic case the Ω matrix would be the product of the σ^2 and the corresponding x values x_i.

$$\Omega_{Het} = \begin{bmatrix} \sigma^2 x_1 & 0 & \cdots & 0^{(n)} \\ 0 & \sigma^2 x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0^{(n)} & 0 & \cdots & \sigma^2 x_n \end{bmatrix}$$

For the homoskedastic case we will have:

$$\Omega_{Hom} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0^{(n)} \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0^{(n)} & 0 & \cdots & 1 \end{bmatrix}$$

The difference being in the values of the diagonal, since the error term is homoskedastic the values on the diagonal would always be (approximately) 1, whereas in the heteroskedastic case the variance would be different for each observation.

2.6 Question 6&7

For question 6 and 7 we have calculated 2 x S $\widehat{\beta_{OLS}}$ and 2 x S $\widehat{\beta_{GLS}}$ estimates, where S was the number of simulations using homoskedastic and heteroscedastic case. Next we have generated 2 x S t - tests also for GLS and OLS and homoscedastic and heteroscedastic case. The formula for the t values was:

$$t = \frac{\hat{\beta} - c}{SE(\hat{\beta})} \tag{5}$$

Where c is a constant that is used for a t-test (true value) and 0 for a model test. This is additionally explained with an example.

For β_0 the t - test would look like this:

$$t = \frac{\hat{\beta}_0 - 1}{SE(\hat{\beta}_0)} \tag{6}$$

And the model test:

$$t = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \tag{7}$$

The standard error of $\hat{\beta}$ is the square root of the diagonal of the covariance matrix for β.

When calculating the GLS standard errors for all simulations we had to resort to some rewritings of the original formula to get the desired 2 x S shape instead of 2 x 1.

Given the covariance matrix for GLS:

$$Cov(\beta_{GLS}) = \sigma^2(X^t\Omega_x^{-1}X)^{-1} \tag{8}$$

We looked at $X^t\Omega_x^{-1}$ and created the new variable XO_1. The result of that matrix is:

$$\text{XO\_1} = \begin{bmatrix} \frac{1}{x_1} & \cdots & \frac{1}{x_n} \\ 1 & \cdots & 1 \end{bmatrix}$$

Next we proceeded with the rest of the formula to get the standard errors.

$$GLS(SE) = \sqrt{(\sigma^2(\mathrm{XO\_1} \cdot X)^{-1})} \tag{9}$$

Where σ^2 is now 1 x S row vector and $(\mathrm{XO\_1} \cdot X)^{-1}$ is 2 x 1 resulting in the final result of 2 x S GLS standard errors.

The graphics below show us the distributions of the t-statistics for the regular t-test and the model test. Homoskedastic results for $\widehat{\beta_{OLS}}$ are not present as they are identical to the GLS results.

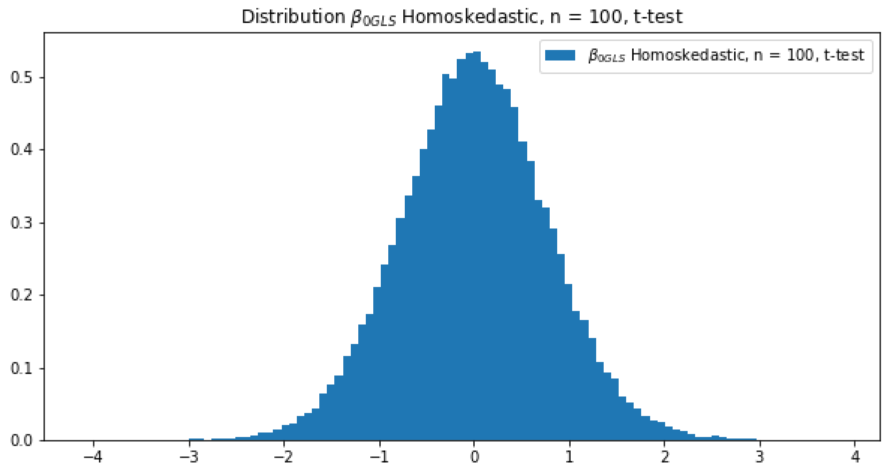

Figure 9

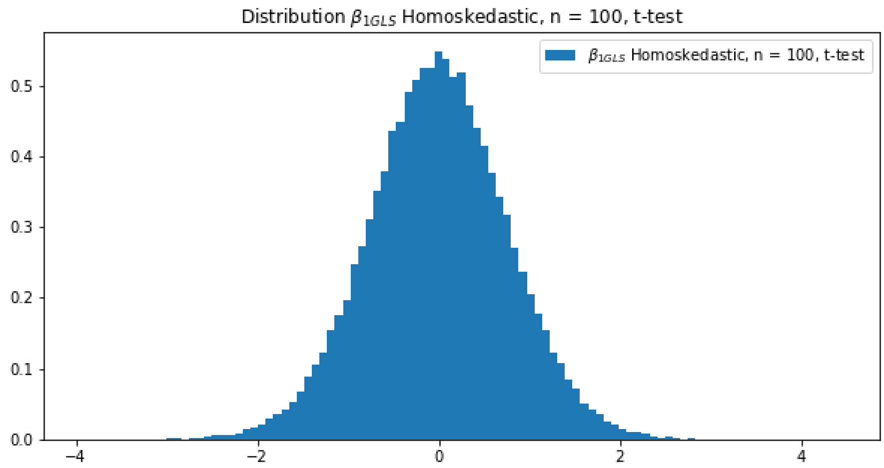

Figure 10

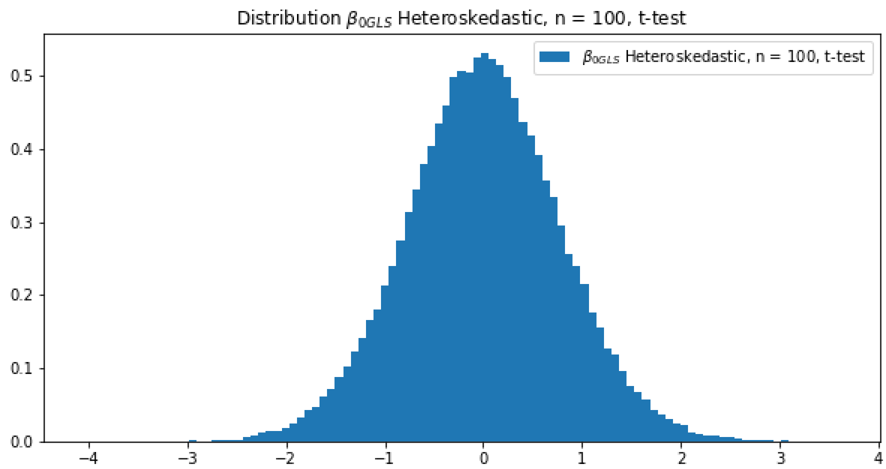

Figure 11

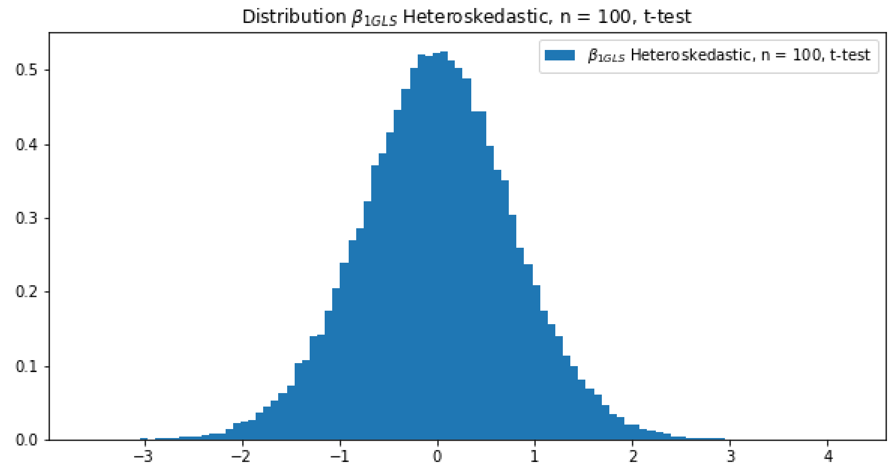

Figure 12

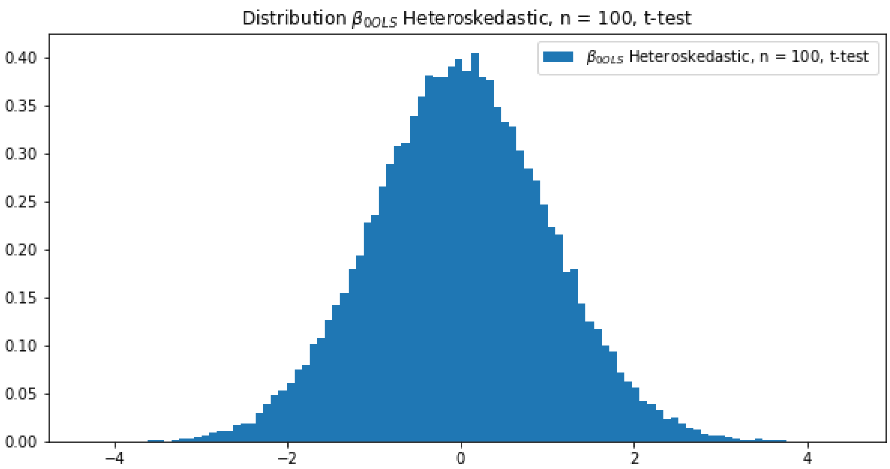

Figure 13

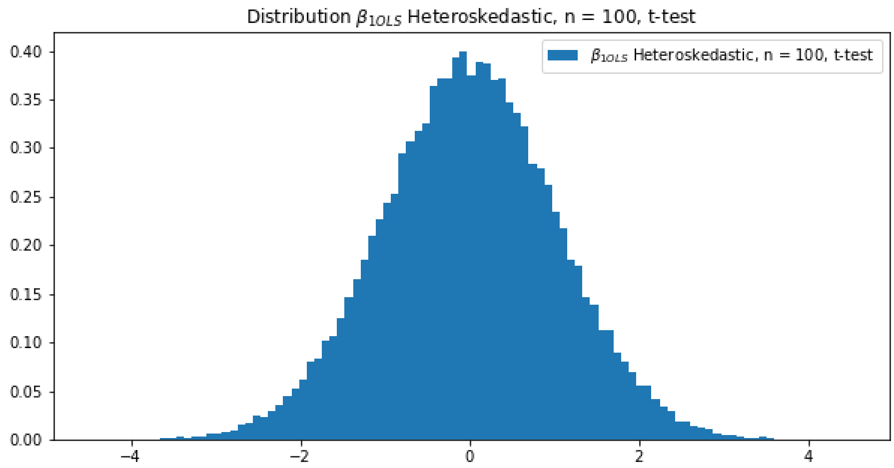

Figure 14

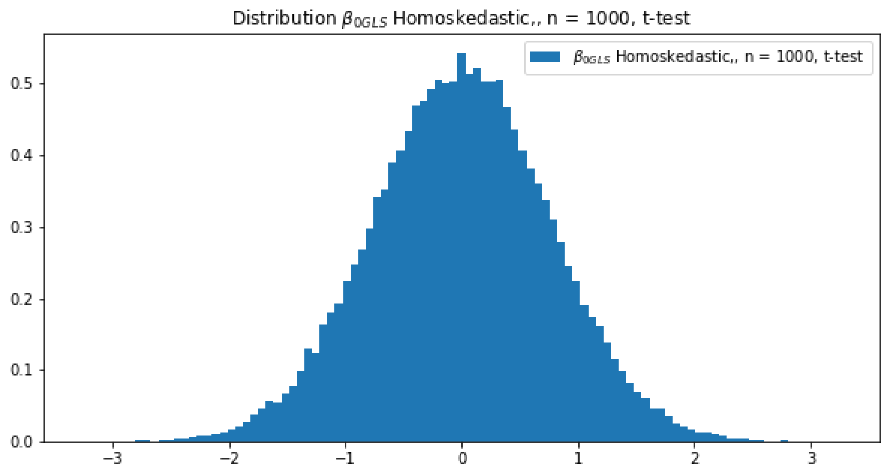

Figure 15

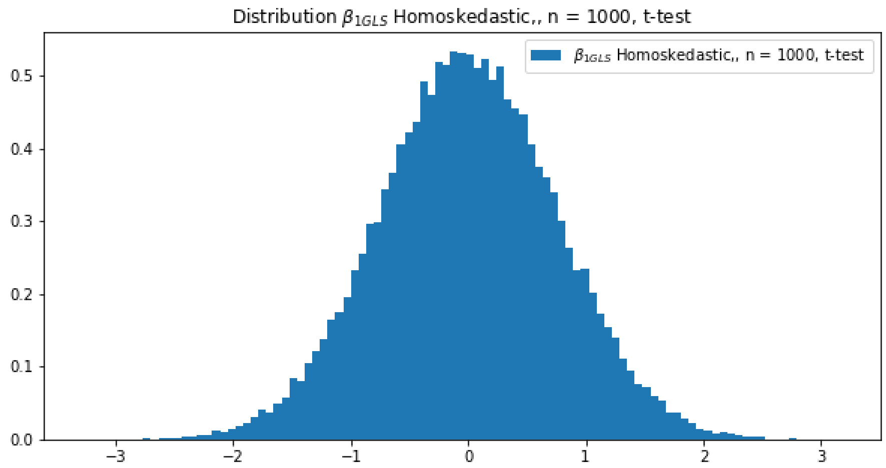

Figure 16

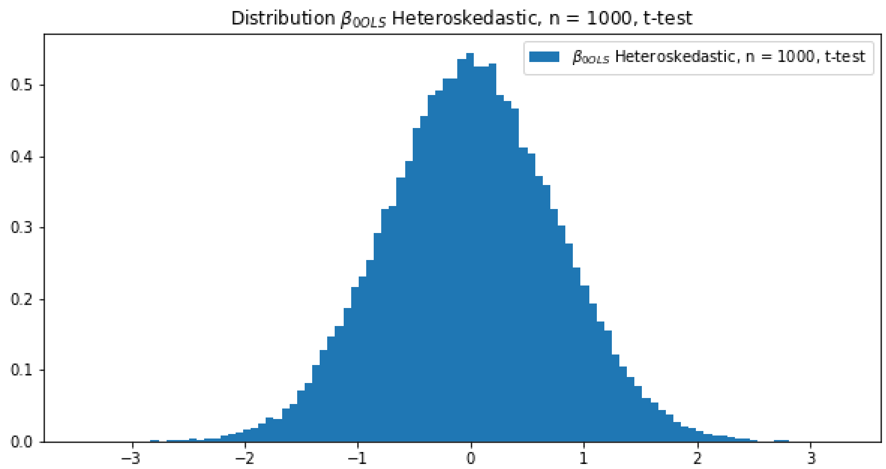

Figure 17

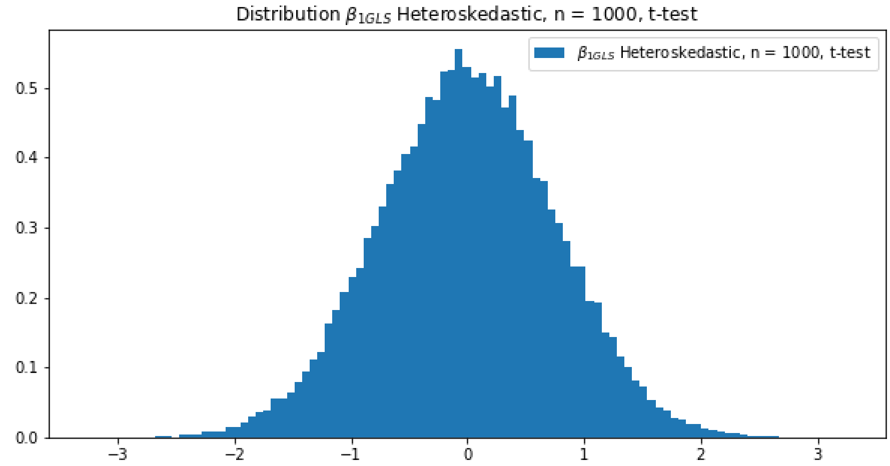

Figure 18

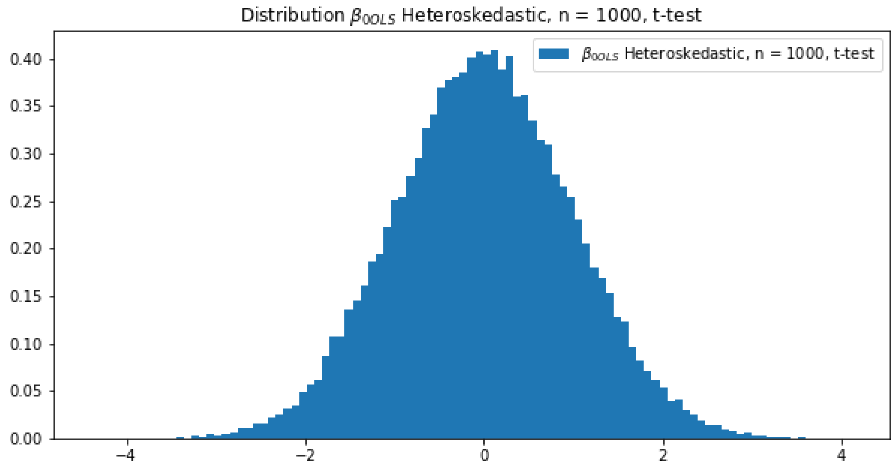

Figure 19

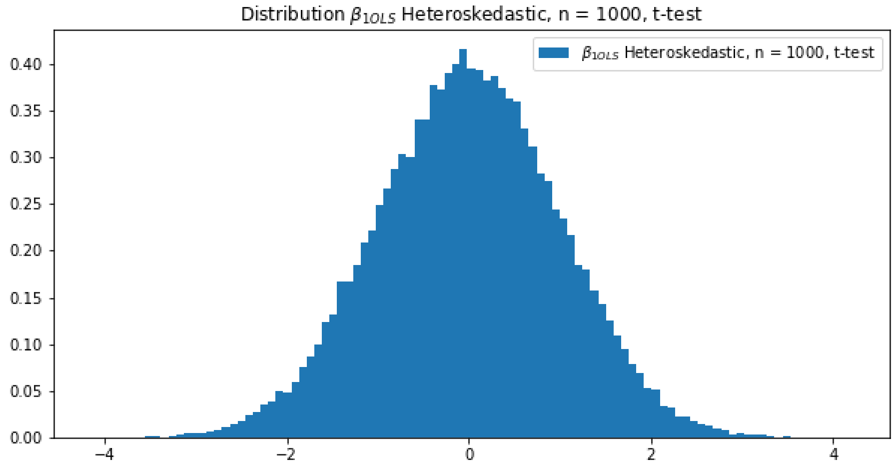

Figure 20

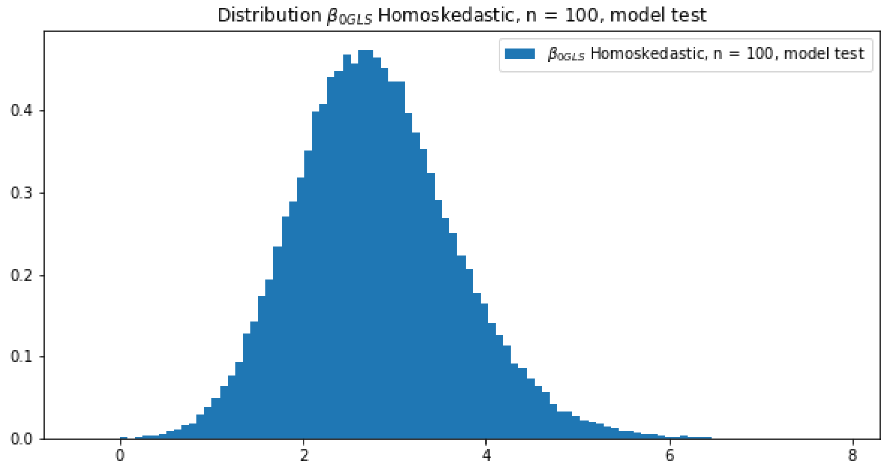

Figure 21

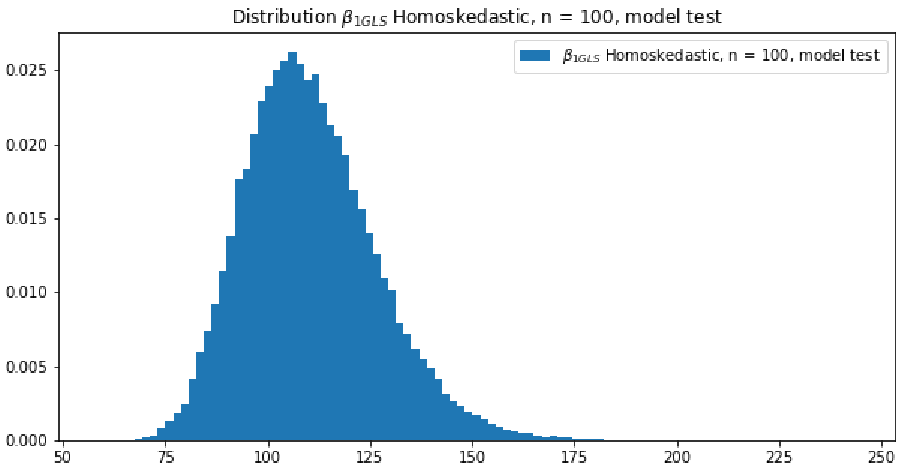

Figure 22

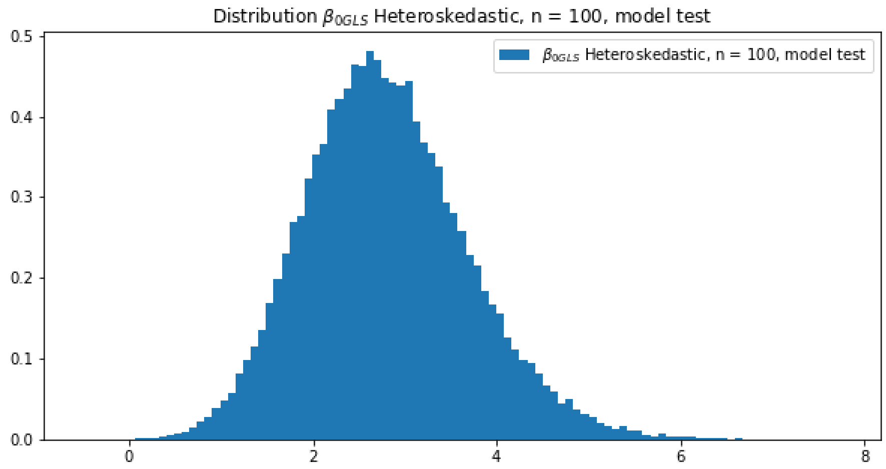

Figure 23

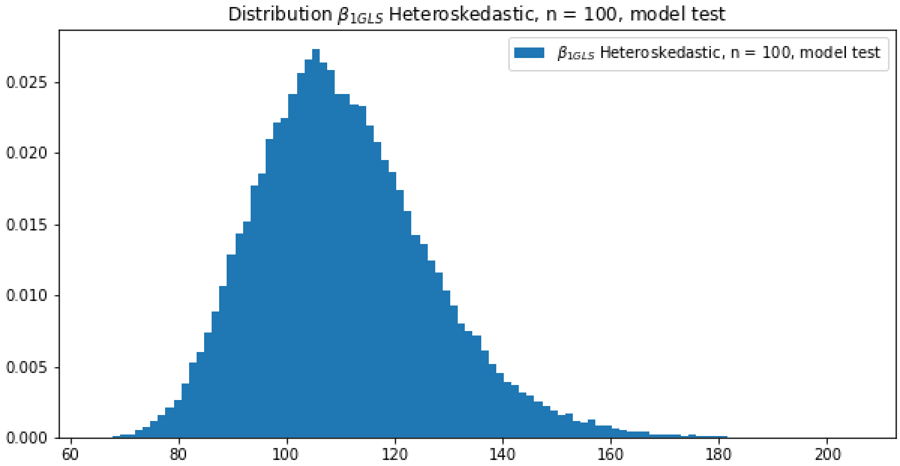

Figure 24

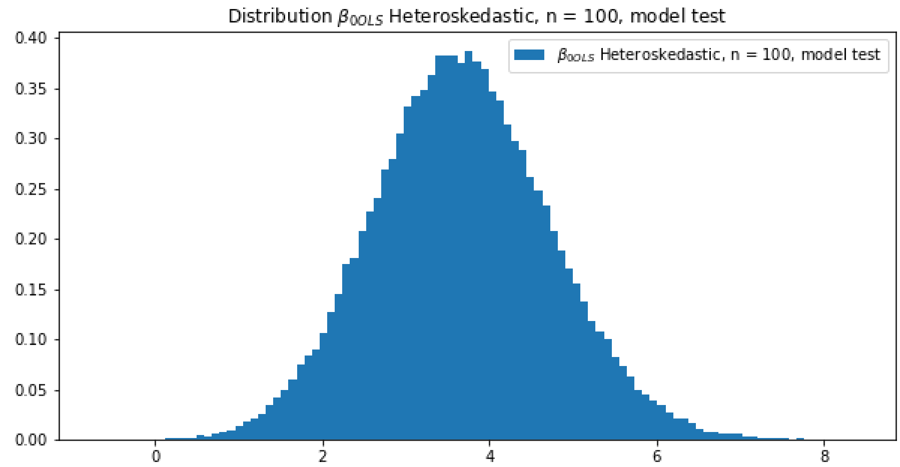

Figure 25

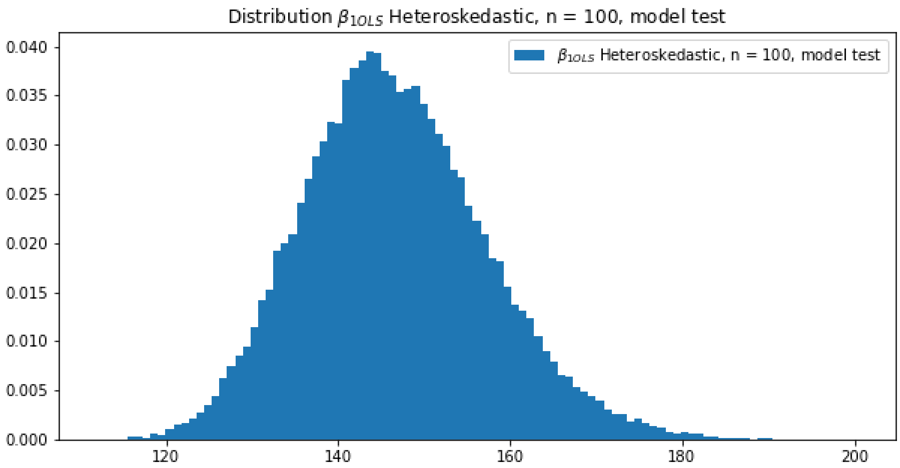

Figure 26

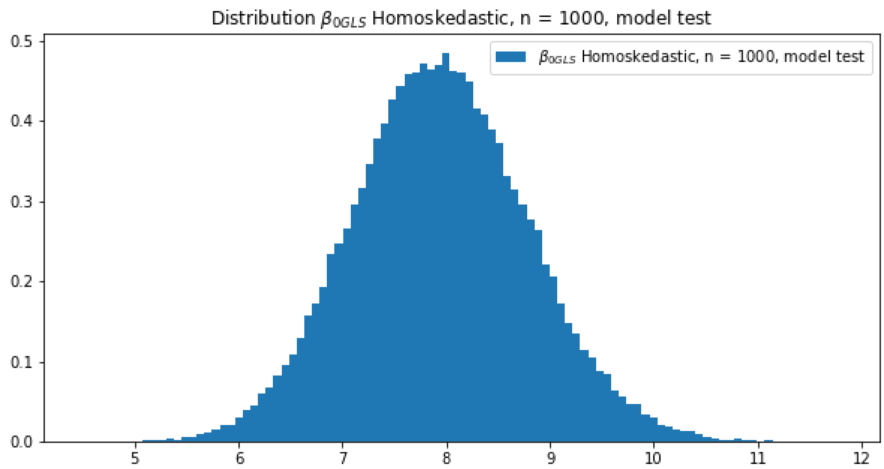

Figure 27

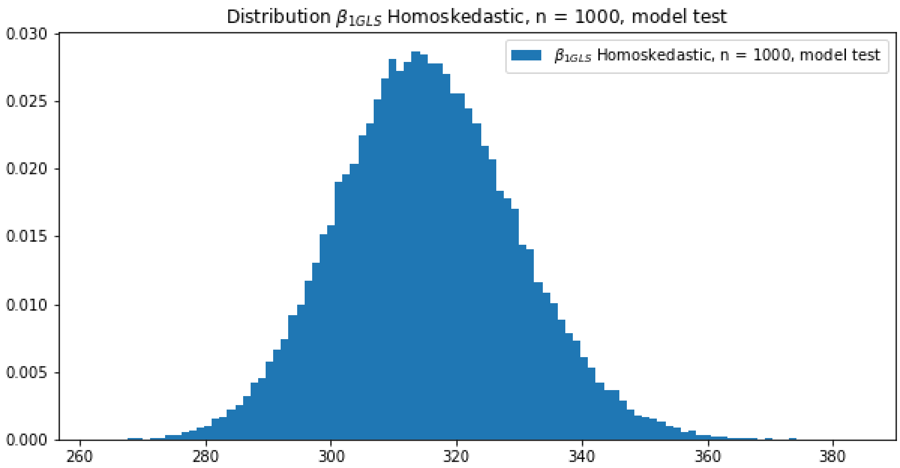

Figure 28

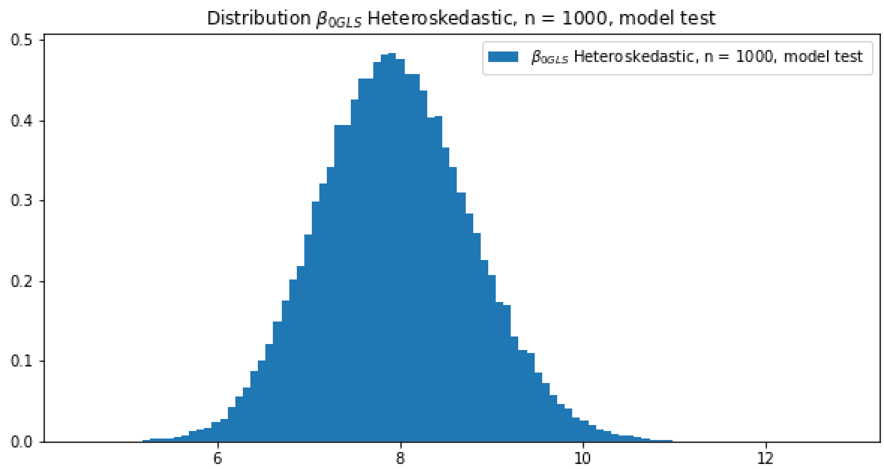

Figure 29

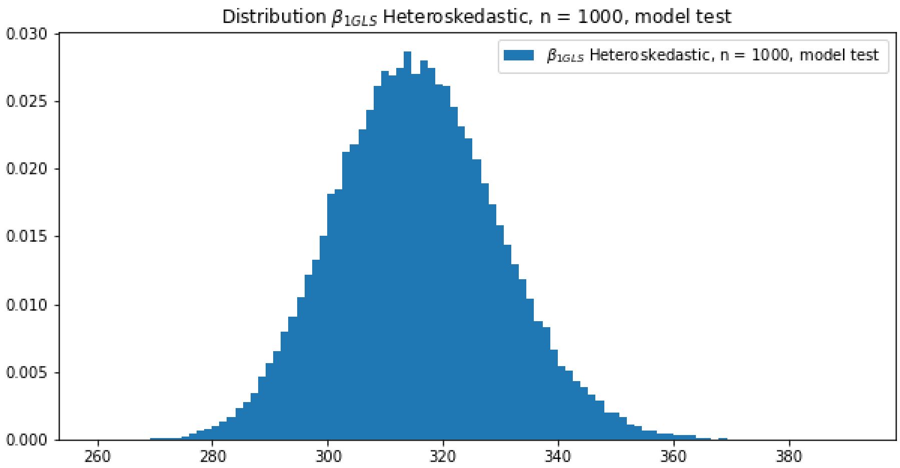

Figure 30

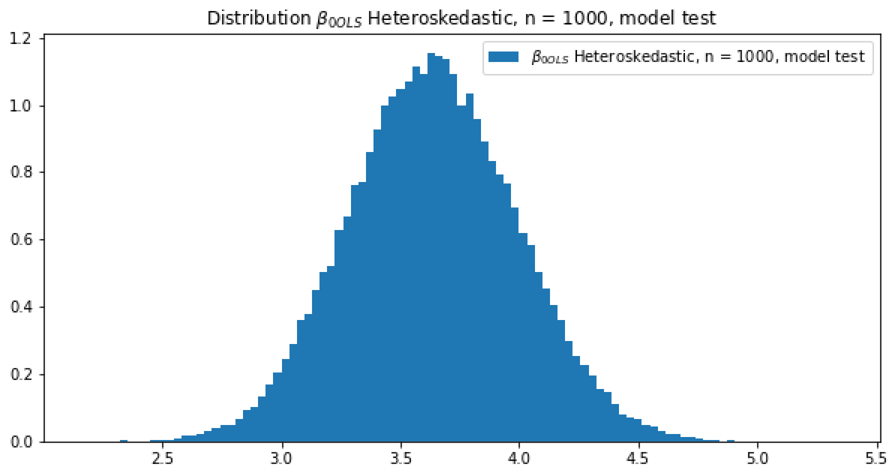

Figure 31

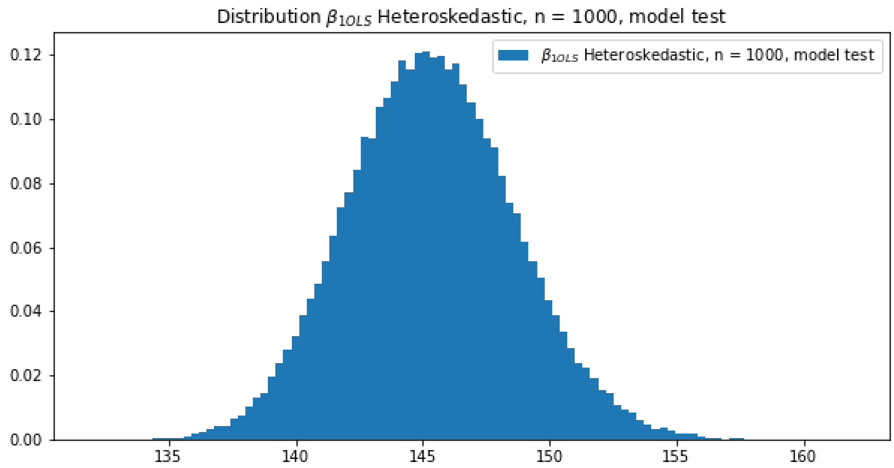

Figure 32

As we can see the distributions for the t-test seem to follow the normal distribution with the range from -3 to 3 which is expected by the standard normal distribution. Sometimes the values are above or below the expected region but this is due to randomness in the simulation. We can also see that the model test distributions look normal but skewed. The econometric theory says that the f-values should be distributed χ^2_p with p degrees of freedom asymptotically, however this will happen in the case if the true values of β vector are indeed 0, otherwise we get the results that are present in the study, namely the distributions for the model test are not perfectly χ^2_p but rather skewed t / normal depending on the sample size.

2.7 Question 8

β_0	β_1	Simulation combination
0.05235291	0.05204773	hom_100_GLS
0.0151062	0.01507568	het_100_GLS
0.05654907	0.05938721	het_100_OLS
0.05235291	0.05204773	hom_100_OLS
0.04962158	0.05021667	hom_1000_GLS
0.00985718	0.00952148	het_1000_GLS
0.04676819	0.04893494	het_1000_OLS
0.04962158	0.05021667	hom_1000_OLS

Table 2: Rejection rates at $\alpha = 0.05$

As we can see in the Table 2 when naively using OLS in heteroscedastic case the rejection rate is close to the desired 5% but doesn't truly represent the reality. It is important to note that the β_{OLS} covariance matrix is estimated using the assumption of homoscedasticity, which is obviously does not hold in heteroscedastic case. For the GLS in homoscedastic case the rejection rates are identical, since both OLS and GLS have the same covariance matrix.

$$Cov(\beta_{OLS}) = \sigma^2(X^tX)^{-1} \tag{10}$$

GLS rejection rates show more realistic results for the heteroscedastic simulations, as the estimation for the covariance of β_{GLS} includes the Ω matrix which is not equal to identity matrix in the heteroscedastic case.

$$Cov(\beta_{GLS}) = \sigma^2(X^t\Omega_x^{-1}X)^{-1} \tag{11}$$

2.8 Question 9

When adjusting the standard errors to White standard errors it makes the t-tests in the heteroscedastic case more precise, in other words the desired rejection rate goes to α asymptotically.

We have adjusted the standard errors using this formula:

$$White(SE) = (X^tX)^{-1}X^tdiag(e_1^2, e_2^2, ..., e_n^2)X(X^tX)^{-1} \tag{12}$$

Where $diag(e_1^2, e_2^2, ..., e_n^2)$ is the diagonal of the squared residuals. This gave us the White standard errors for 1 simulation, however we had to have the standard errors for all simulations. We used the following manipulations to get to the desired format:

$$A = (X^tX)^{-1}X^t \tag{13}$$

$$A^2 = A \odot A \tag{14}$$

$$\text{White}(SE) = \sqrt{(A^2 \cdot e^2)} \tag{15}$$

Where e^2 is 1 x S vector of squared residuals. Using the obtained White standard errors yields the following rejection rate table.

β_0	β_1	Simulation combination
0.06466675	0.06607056	het_100_GLS_white
0.06466675	0.06607056	het_100_OLS_white
0.04930115	0.04942322	het_1000_GLS_white
0.04930115	0.04942322	het_1000_OLS_white

Table 3: Rejection rates at $\alpha = 0.05$, White SE

The findings clearly show the convenience of White standard errors as they remove the gap between the OLS and GLS rejection rates and show rejection rates closer to 0.05 and grow towards 0.05 as sample size increases, which is what we want to see given α rejection rate.

3 Empirical Investigation

3.1 Question 1

	count	mean	std	min	25%	50%	75%	max
age	600.000000	38.820000	11.370000	23.000000	29.000000	36.000000	48.000000	65.000000
black	600.000000	0.050000	0.220000	0.000000	0.000000	0.000000	0.000000	1.000000
case	600.000000	356.260000	205.200000	1.000000	177.750000	356.500000	533.750000	706.000000
clerical	600.000000	0.180000	0.330000	0.000000	0.000000	0.000000	0.180000	1.000000
construc	600.000000	0.030000	0.150000	0.000000	0.000000	0.000000	0.030000	1.000000
educ	600.000000	12.850000	2.820000	1.000000	12.000000	12.000000	16.000000	17.000000
earns74	600.000000	9899.580000	9528.590000	0.000000	2500.000000	8250.000000	13750.000000	42500.000000
gdhlth	600.000000	0.880000	0.320000	0.000000	1.000000	1.000000	1.000000	1.000000
inlf	600.000000	0.750000	0.430000	0.000000	0.000000	1.000000	1.000000	1.000000
leis1	600.000000	4718.190000	916.290000	2090.000000	4108.250000	4649.500000	5244.000000	7417.000000
leis2	600.000000	4601.040000	919.060000	1677.000000	3999.250000	4546.000000	5116.000000	7297.000000
leis3	600.000000	4542.540000	915.030000	1677.000000	3938.500000	4489.000000	5053.250000	7282.000000
smsa	600.000000	0.400000	0.490000	0.000000	0.000000	0.000000	1.000000	1.000000
lhrwage	449.000000	1.440000	0.650000	-1.050000	1.060000	1.510000	1.840000	3.570000
lothinc	600.000000	6.240000	4.220000	0.000000	0.000000	8.610000	9.330000	10.660000
male	600.000000	0.560000	0.500000	0.000000	0.000000	1.000000	1.000000	1.000000
marr	600.000000	0.820000	0.380000	0.000000	1.000000	1.000000	1.000000	1.000000
prot	600.000000	0.660000	0.470000	0.000000	0.000000	1.000000	1.000000	1.000000
rlxall	600.000000	3434.700000	531.050000	1380.000000	3137.250000	3415.500000	3712.750000	6110.000000
selfe	600.000000	0.130000	0.330000	0.000000	0.000000	0.000000	0.000000	1.000000
sleep	600.000000	3259.040000	448.370000	755.000000	3008.000000	3255.000000	3525.000000	4695.000000
slpnaps	600.000000	3376.200000	507.180000	1335.000000	3095.000000	3355.000000	3653.500000	6110.000000
south	600.000000	0.190000	0.390000	0.000000	0.000000	0.000000	0.000000	1.000000
spsepay	600.000000	5233.720000	8361.310000	0.000000	0.000000	0.000000	9000.000000	75000.000000
spwrk75	600.000000	0.480000	0.500000	0.000000	0.000000	0.000000	1.000000	1.000000
totwrk	600.000000	2102.760000	958.910000	0.000000	1538.000000	2272.500000	2686.500000	6415.000000
union	600.000000	0.200000	0.400000	0.000000	0.000000	0.000000	0.000000	1.000000
worknrm	600.000000	2079.160000	961.430000	0.000000	1506.750000	2263.000000	2638.000000	6415.000000
workscnd	600.000000	23.610000	126.340000	0.000000	0.000000	0.000000	0.000000	1205.000000
exper	600.000000	19.970000	12.450000	0.000000	10.000000	17.000000	30.000000	55.000000
yngkid	600.000000	0.140000	0.340000	0.000000	0.000000	0.000000	0.000000	1.000000
yrsmarr	600.000000	11.850000	11.640000	0.000000	0.000000	9.000000	19.250000	43.000000
hrwage	449.000000	5.180000	3.840000	0.350000	2.890000	4.510000	6.290000	35.510000
agesq	600.000000	1636.370000	956.360000	529.000000	841.000000	1296.000000	2304.000000	4225.000000

Table 4: Descriptive statistics for the sleep data

We proceeded with the loading of the data set and looking at the descriptive statistics which are summarized in Table 4.

Upon the investigation of the descriptive statistics we have found a few nuances that should be mentioned. First, the 'hrwage' and 'lhrwage' rows contained only 449 values out of a sample of 600 and had different sample statistics due to log transformation, second we noticed that the data set contained a lot of dummy variables.

3.2 Question 2

Given the model:

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \varepsilon_i \tag{16}$$

The sign of the β_1 will be negative, as people trade off work and sleep. In other words if the amount of total work will increase it would negatively affect the sleep time amount.

3.3 Question 3

		R-squared:		0.107		
		No. Observations:		600		
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3580.2901	41.759	85.738	0.000	3498.279	3662.301
totwrk	-0.1528	0.018	-8.454	0.000	-0.188	-0.117

Table 5: Regression results

Running the first regression model yields the results summarized in the Table 5.

The constant means the amount of sleep in minutes per week that people get no matter the amount of time they work. An average person sleeps a certain amount of hours a day regardless of the work hours and this fact is reported in the constant β_0.

3.4 Question 4

If the *totwrk* will increase by 2 hours (120 minutes) then the effect on *sleep* would be $120\beta_1$.

$$120(-0.1528) = -18.336 \tag{17}$$

That means if we increase the work hours by 2 then it would reduce the amount of sleep in minutes by 18.336 per week, or approximately (18.336/7) 2.62 minutes per day. We do not find this to be a large effect.

3.5 Question 5

Adding the new variables to our model, namely *educ* and *age* will result in the following equation:

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 educ_i + \beta_3 age_i + \varepsilon_i \tag{18}$$

We expect the signs for β_2 and β_3 to be negative. The level of education might negatively affect sleep because during schooling, the amount of time available for sleep is limited. After graduation, people with higher education will usually get more demanding jobs, hence indirectly reducing the amount of sleep. When people are born, they usually sleep for 10 - 12 hours a day, but this reduces with time as the children grow older. As an adult, you have more responsibilities than a child, affecting sleep time. It is important to remember that the signs of the β coefficients are subject to change as we add new variables to the model and be aware that existing model might not perfectly explain *sleep* variable.

3.6 Question 6

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3621.2172	123.116	29.413	0.000	3379.423	3863.012
totwrk	-0.1494	0.018	-8.301	0.000	-0.185	-0.114
educ	-12.2426	6.358	-1.926	0.055	-24.729	0.244
age	2.8177	1.579	1.784	0.075	-0.284	5.919

Table 6: Regression results with *totwrk*, *age* and *educ*.

The *educ* has a negative effect on the sleep amount as expected. However, the *age* seems to have an inverse relationship. Both variables are statistically insignificant at $\alpha = 0.05$, suggesting that they should not be used for predicting the sleep time amount.

3.7 Question 7

If someone were to work 5 more hours per week (300 minutes) then the change in sleep would be as follows.

Holding other parameters constant apart from the increase in total work amount:

$$sleep_i = \beta_0 + 300\beta_1 + \beta_2 educ_i + \beta_3 age_i + \varepsilon_i \tag{19}$$

Substituting β_1 into equation gives:

$$-44.82 = 300(-0.1494) \tag{20}$$

An additional 5 hours per week will reduce weekly sleep time by approximately 45 minutes. Since we are talking about weekly data, this means on a daily scale, the sleep will be reduced by 45/7, which is approximately 6 minutes. We do not find this value a large trade-off since potential benefits from additional work will outweigh 6 minutes lost in sleep.

3.8 Question 8

As previously expected, the education (*educ*) variable has a negative sign, meaning that the more years of schooling you have, the less sleep time you get. The magnitude suggests that you will lose approximately 12 minutes of weekly sleep time for each year of schooling. This also makes sense as the more sophisticated your education becomes, the more time people might need to invest to graduate, get the desired results, etc. As discussed above, there is a higher chance that people with better education might get more demanding jobs, thus indirectly reducing sleep time.

3.9 Question 9

Given the output result below the current model explains approximately 12% of the variation in the data. This is because other potential factors might affect the sleeping behaviour such as the place where you live or your health.

R-squared:	0.121
Adj. R-squared:	0.116

Omnibus:	67.297	**Durbin-Watson:**	2.047
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	192.388
Skew:	-0.545	**Prob(JB):**	1.67e-42
Kurtosis:	5.551	**Cond. No.**	1.66e+04

Notes:
[2] The condition number is large, 1.66e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Table 7: Regression results with *totwrk*, *age* and *educ*.

The *totwrk* variable can potentially correlate with age and especially with the level of education. Python output also suggests that there is a presence of strong multicollinearity meaning that the variables in the model can potentially be highly correlated.

3.10 Question 10

Adding a new dummy variable *yngkid* and *agesq* will give us:

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 educ_i + \beta_3 age_i + \beta_4 yngkid_i + \beta_5 agesq_i + \varepsilon_i \tag{21}$$

With the following estimation results:

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3828.7290	262.316	14.596	0.000	3313.550	4343.908
totwrk	-0.1466	0.018	-8.028	0.000	-0.182	-0.111
educ	-12.0644	6.369	-1.894	0.059	-24.572	0.443
age	-8.6049	12.499	-0.688	0.491	-33.152	15.942
agesq	0.1383	0.148	0.933	0.351	-0.153	0.429
yngkid	9.5107	53.274	0.179	0.858	-95.118	114.140

Table 8: Regression results with *totwrk*, *age*, *educ*, *yngkid* and *agesq*.

We can now see that the added variables do not help us to improve the model, as they are both statistically insignificant at the level $\alpha = 0.05$.

3.11 Question 11

Now we have to estimate the same model but for men and women separately.

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 educ_i + \beta_3 age_i + \beta_4 yngkid_i + \beta_5 agesq_i + \beta_6 male = 1 + \varepsilon_i \quad (22)$$

We have chosen a subset of existing data where the *male* dummy variable equals to 1, meaning we haven chose all the males from the sample and estimated the model again.

Dep. Variable:	sleep	**R-squared:**	0.167
Model:	OLS	**Adj. R-squared:**	0.155
Method:	Least Squares	**F-statistic:**	13.35
Date:	Sat, 16 Dec 2023	**Prob (F-statistic):**	7.30e-12
Time:	18:51:11	**Log-Likelihood:**	-2504.6
No. Observations:	338	**AIC:**	5021.
Df Residuals:	332	**BIC:**	5044.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3640.2288	337.088	10.799	0.000	2977.131	4303.326
totwrk	-0.1865	0.027	-6.956	0.000	-0.239	-0.134
educ	-11.6004	8.149	-1.424	0.156	-27.630	4.430
age	5.7909	15.845	0.365	0.715	-25.378	36.960
agesq	-0.0213	0.186	-0.114	0.909	-0.388	0.345
yngkid	82.5022	62.612	1.318	0.189	-40.664	205.669

Omnibus:	22.490	**Durbin-Watson:**	2.022
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	47.818
Skew:	-0.339	**Prob(JB):**	4.13e-11
Kurtosis:	4.713	**Cond. No.**	4.65e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.65e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Table 9: Regression results for the sleep model for males

Table 9 summarized the regression results for males. Only total work minutes per week seem to have a statistically significant effect on the *sleep* variable. The adjusted

R-squared has increased, meaning that now the model explains approximately 16% of the variation in sleep for males, which is an improvement from previous results. On the other hand, the existing model still suffers from multicollinearity, meaning that some of the independent variables correlate.

For the females all we have to do is to select data, where the *male* dummy variable is equals to 0. Namely:

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 educ_i + \beta_3 age_i + \beta_4 yngkid_i + \beta_5 agesq_i + \beta_6 male = 0 + \varepsilon_i \tag{23}$$

Dep. Variable:	sleep	**R-squared:**	0.112
Model:	OLS	**Adj. R-squared:**	0.094
Method:	Least Squares	**F-statistic:**	6.446
Date:	Sat, 16 Dec 2023	**Prob (F-statistic):**	1.16e-05
Time:	18:51:11	**Log-Likelihood:**	-1962.4
No. Observations:	262	**AIC:**	3937.
Df Residuals:	256	**BIC:**	3958.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	4277.0584	418.076	10.230	0.000	3453.753	5100.364
totwrk	-0.1400	0.029	-4.791	0.000	-0.198	-0.082
educ	-14.2456	10.166	-1.401	0.162	-34.266	5.775
age	-29.8172	20.060	-1.486	0.138	-69.321	9.687
agesq	0.3628	0.241	1.508	0.133	-0.111	0.837
yngkid	-201.3997	99.856	-2.017	0.045	-398.044	-4.755

Omnibus:	57.031	**Durbin-Watson:**	1.747
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	215.068
Skew:	-0.849	**Prob(JB):**	1.99e-47
Kurtosis:	7.101	**Cond. No.**	3.95e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.95e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Table 10: Regression results for the sleep model for females

Estimating the model for females results in Table 10. We can now observe the difference between the male model. First, the *yngkid* dummy variable is now statistically significant at $\alpha = 0.05$, which means that it helps predict females' sleep time. Although we now have an additional helpful variable in explaining sleep, the adjusted R-squared dropped to 9.4%; thus, the current model explains less sleep variance than males.

What is also interesting is that the coefficient size and sign vary between males and females. The variables *age* and *yngkid* are positive for males and negative for females. This means that as men grow older, they actually sleep more, and having young children does not negatively impact their sleep, in fact, it is positive, so men sleep more when they have young children (*yngkid* is one if children < 3 years old are present). In contrast, females sleep less with age and sleep ≈ 30 minutes less per day if a young child is in the family. One of the possible explanations might be that females care more about their children and spend more time with them.

3.12 Question 12 & 13

Now we would like to estimate the model accounting for both male and female at the same time. The male is now a binary variable, being 1 when the observation is for a male and 0 for female. $male = 1, 0$.

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 educ_i + \beta_3 age_i + \beta_4 yngkid_i + \beta_5 agesq_i + \beta_6 male_i + \varepsilon_i \quad (24)$$

Dep. Variable:	sleep	**R-squared:**	0.130
Model:	OLS	**Adj. R-squared:**	0.121
Method:	Least Squares	**F-statistic:**	14.79
Date:	Sat, 16 Dec 2023	**Prob (F-statistic):**	8.59e-16
Time:	18:51:11	**Log-Likelihood:**	-4472.4
No. Observations:	600	**AIC:**	8959.
Df Residuals:	593	**BIC:**	8990.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3842.8329	261.403	14.701	0.000	3329.444	4356.222
totwrk	-0.1637	0.020	-8.350	0.000	-0.202	-0.125
educ	-13.0459	6.359	-2.052	0.041	-25.534	-0.558
age	-8.8835	12.452	-0.713	0.476	-33.340	15.573
agesq	0.1371	0.148	0.928	0.354	-0.153	0.427
yngkid	-8.2672	53.616	-0.154	0.878	-113.568	97.034
male	88.2285	37.717	2.339	0.020	14.153	162.304

Omnibus:	66.472	**Durbin-Watson:**	2.071
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	193.463
Skew:	-0.532	**Prob(JB):**	9.78e-43
Kurtosis:	5.570	**Cond. No.**	4.30e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.3e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Table 11: Regression results for the sleep model for males and females

This model is an improvement to the previous two because it accounts for both males and females. It is easier to make conclusions for the whole population with this model. For example, we have found a few significant differences between the males and females in the previous section, which is also consistent with what we find in Table 11, namely, being a male is statistically significant in predicting the sleeping time. However, what has changed now is that *yngkid* and *age* do not play much of a role in predicting sleep, but education *educ* does. The *agesq* variable has not helped in predicting sleep, not for males or females, and did not help with the existing model, suggesting this variable might not be needed in the model.

Total minutes worked (*totwrk*), as expected, still plays a significant role in predicting sleep, as it has been a critical variable in every model, suggesting that work is significant at the population level and does not vary in importance between genders.

The *const* variables remain significant for biological reasons. Namely, people need the same basic level of sleep in order to function correctly, so it makes sense that this variable has been statistically significant at all times during the study.

Adjusted R-squared has improved for females but worsened for males. This is because the model improves the prediction at the population level at the cost of the individual level.

3.13 Question 14

F-statistic:	14.79
Prob (F-statistic):	8.59e-16

Table 12: F-statistics

Taking the result reported in Table 12 from the regression summary we can see that the probability of F-statistic is basically 0 that means that the model in fact adds value and helps in predicting sleep.

3.14 Question 15

Before we start with the last question it is important to say that we would have started improving the model by trying to remove multicollinearity by removing some variables, but since in the assignment document we were asked to only add variables to the existing model we decided to not change the original set up.

After several experiments with different variables, we have pursued the model described in Table 13. First of all, with our new model including the new variable *south*, all key indicators, such as adjusted R-squared and AIC, have improved slightly, second, the new variable is not only statistically significant but also has the biggest coefficient out of all other variables besides the constant, which may signify the importance of living in the south.

The reasoning behind this strong effect needs to be clarified. One possible explanation might be that the sunset is usually earlier near the equator; hence, people might get to sleep earlier; also, if talking on a state level, the difference between governments, for example in the US, might play a role. Since the original document does not explain the exact definition of the *south* variable, we would need more information on the *south* variable to explain why the effect is so strong.

Dep. Variable:	sleep	**R-squared:**	0.143
Model:	OLS	**Adj. R-squared:**	0.132
Method:	Least Squares	**F-statistic:**	14.06
Date:	Sat, 16 Dec 2023	**Prob (F-statistic):**	6.31e-17
Time:	18:51:11	**Log-Likelihood:**	-4468.1
No. Observations:	600	**AIC:**	8952.
Df Residuals:	592	**BIC:**	8987.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3759.7722	261.313	14.388	0.000	3246.558	4272.986
totwrk	-0.1681	0.020	-8.605	0.000	-0.207	-0.130
educ	-12.1322	6.326	-1.918	0.056	-24.557	0.293
age	-6.2186	12.408	-0.501	0.616	-30.587	18.150
agesq	0.1065	0.147	0.724	0.470	-0.182	0.395
yngkid	-17.6216	53.376	-0.330	0.741	-122.450	87.207
male	95.6125	37.565	2.545	0.011	21.835	169.390
south	127.7070	43.738	2.920	0.004	41.806	213.608

Omnibus:	64.903	**Durbin-Watson:**	2.047
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	189.982
Skew:	-0.516	**Prob(JB):**	5.57e-42
Kurtosis:	5.556	**Cond. No.**	4.33e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.33e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Table 13: Regression results for Question 15