Computer Assignment 1

Uppsala University
Department of Statistics
Course: Econometrics, Fall 2019
Author: Claes Kock, Mayara Latrech, Yuchong Wu
Date: 9/10/2019

Contents

Introduction

The first purpose of this assignment is to get a better idea of the properties of various estimators both from a theoretical and practical perspective. The second is obtaining a deeper understanding of the OLS estimator by changing a model's parameters and seeing how the estimator would behave.

This assignment consists of five tasks that help us work with different data and models.

In task one, compare the theoretical assumptions to the practical values of the properties of various estimators.

Tasks two to four focus more on the properties of the OLS estimator.

In task five we evaluate the choice of a model to approximate certain data points.

1 Task 1

In this task we have four different estimators:

1) The OLS estimator:

$$\widehat{\beta}_{OLS} = \frac{\sum_{t=1}^{T} Y_t t}{\sum_{t=1}^{T} t^2}$$

2) The Mean estimator:

$$b_M = \frac{\bar{Y}}{\bar{t}}$$

3) The End estimator:

$$b_E = \frac{Y_T}{T}$$

4) The OLS estimator for the model with intercept:

$$\widehat{\beta}_{OLSI} = \frac{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right)(t - \bar{t})}{\sum_{t=1}^{T} (t - \bar{t})^2}$$

1.1 Task 1A

1.1.1 Unbiasedness

The following calculations will prove that the estimators are unbiased:

Calc 1):

$$\begin{aligned}
E\left(\hat{\beta}_{OLS}\right) &= E\left[\frac{\Sigma Y_t t}{\Sigma t^2}\right] \\
&= \frac{1}{\Sigma t^2} E\left[\Sigma Y_t t\right] \\
&= \frac{1}{\Sigma t^2} \sum E\left[\ Y_t t\ \right] \\
&= \frac{1}{\Sigma t^2} \Sigma\left[t^2 \beta + t E\left(u_t\right)\right] \qquad (\mathit{Assume\ that}\ E(u_t) = 0) \\
&= \beta + 0 \\
&= \beta
\end{aligned}$$

Calc 2):

$$
\begin{aligned}
E\left[b_M\right] &= E\left[\frac{\bar{Y}}{\bar{t}}\right] \\
&= \frac{1}{\bar{t}} E[\bar{Y}] \\
&= \frac{1}{\bar{t}} E\left[\frac{1}{T} \Sigma\left(\beta t + u_t\right)\right] \\
&= \frac{1}{\bar{t}} \frac{1}{T} \sum E\left(\beta t + u_t\right) \\
&= \frac{1}{\bar{t}T} \sum \beta_t \\
&= \frac{\beta}{\bar{t}T} \sum t \\
&= \beta
\end{aligned}
$$

Calc 3):

$$
\begin{aligned}
E\left[b_E\right] &= E\left[\frac{Y_T}{T}\right] \\
&= \frac{1}{T} E\left[\beta T + u_T\right] \\
&= \frac{1}{T} \beta T + \frac{1}{T} E\left[u_T\right] \\
&= \beta
\end{aligned}
$$

Calc 4):

$$
\begin{aligned}
E\left[\hat{\beta}_{OLSI}\right] &= E\left[\frac{\Sigma\left(Y_t - \bar{Y}\right)\left(t - \bar{t}\right)}{\sum (t - \bar{t})^2}\right] \\
&= \frac{1}{\sum (t - \bar{t})^2} E\left[\sum_{t=1}^{T}\left[\beta t + u_t - \frac{1}{T}\sum_{s=1}^{T}\left(\beta_s s + u_s\right)\right][t - \bar{t}]\right] \\
&= \frac{1}{\sum (t - \bar{t})^2} \sum_{t=1}^{T}\left[\beta t + E\left(u_t\right) - \frac{1}{T}\sum_{s=1}^{T}\left(\beta_s s + E\left[u_s\right]\right)\right][t - \bar{t}] \\
&= \frac{1}{\sum (t - \bar{t})^2} \sum (t\beta - \beta\bar{t})(t - \bar{t}) \\
&= \frac{\beta}{\sum (t - \bar{t})^2} \sum (t - \bar{t})(t - \bar{t}) \\
&= \beta
\end{aligned}
$$

1.1.2 Variance

The following calculations are the variances of the estimators:

Calc 1):

$$\begin{aligned}
Var(\hat{\beta_{OLS}}) &= Var(\frac{\sum_{t=1}^{T} Y_t t}{\sum_{t=1}^{T} t^2}) \\
&= \frac{1}{(\sum_{t=1}^{T} t^2)^2} Var(\sum_{t=1}^{T} \beta t^2 + u_t t) \quad \textit{we assume that } Cov(u_i, u_j) = 0\ (i \neq j),\ \textit{then}: \\
&= \frac{1}{(\sum_{t=1}^{T} t^2)^2} \sum_{t=1}^{T} [Var(\beta t^2 + u_t t) + 0] \\
&= \frac{1}{(\sum_{t=1}^{T} t^2)^2} \sum_{t=1}^{T} (t^2 Var(u_t)) \\
&= \frac{1}{(\sum_{t=1}^{T} t^2)^2} \sum_{t=1}^{T} (t^2 \sigma^2) \\
&= \sigma^2 \frac{1}{(\sum_{t=1}^{T} t^2)^2} \sum_{t=1}^{T} t^2 \\
&= \frac{\sigma^2}{\sum_{t=1}^{T} t^2}
\end{aligned}$$

Calc 2):

$$\begin{aligned}
Var(b_M) &= Var(\frac{\bar{Y}}{\bar{t}}) \\
&= \frac{1}{\bar{t}^2} Var(\frac{1}{T} \sum_{t=1}^{T} \beta t + u_t) \quad \textit{we assume that } Cov(u_i, u_j) = 0\ (i \neq j),\ \textit{then}: \\
&= \frac{1}{\bar{t}^2} \frac{1}{T^2} \sum_{t=1}^{T} [Var(\beta t + u_t) + 0] \\
&= \frac{1}{\bar{t}^2} \frac{1}{T^2} T \sigma^2 \\
&= \frac{\sigma^2}{T \bar{t}^2}
\end{aligned}$$

Calc 3):

$$\begin{aligned}
Var(b_E) &= Var(\frac{Y_T}{T}) \\
&= \frac{1}{T^2} Var(\beta T + u_T) \\
&= \frac{\sigma^2}{T^2}
\end{aligned}$$

Calc 4):

$$
\begin{aligned}
Var(\hat{\beta}_{OLSI}) &= Var(\frac{\sum_{t=1}^{T}(Y_t-\bar{T})(t-\bar{t})}{\sum_{t=1}^{T}(t-\bar{t})^2} \\
&= \frac{1}{(\sum_{t=1}^{T}(t-\bar{t})^2)^2)} Var(\sum_{t=1}^{T}(Y_t-\bar{Y})(t-\bar{t}) \\
&= \frac{1}{(\sum_{t=1}^{T}(t-\bar{t})^2)^2)} Var(\sum_{t=1}^{T}\beta t^2 + u_t t - \bar{Y}t - \beta\bar{t}t - \bar{t}u_t + \bar{t}\bar{Y}) \\
&= \frac{1}{(\sum_{t=1}^{T}(t-\bar{t})^2)^2)} \sum_{t=1}^{T} Var(\beta t^2 + u_t t - \bar{Y}t - \beta\bar{t}t - \bar{t}u_t + \bar{t}\bar{Y}) \\
& we\ assume\ that\ Cov(u_i, u_j) = 0\ (i \neq j),\ then: \\
&= \frac{1}{(\sum_{t=1}^{T}(t-\bar{t})^2)^2} \sum_{t=1}^{T} Var(tu_t - \bar{t}u_t) \\
&= \frac{1}{(\sum_{t=1}^{T}(t-\bar{t})^2)^2} \sum_{t=1}^{T} Var(u_t(t-\bar{t})) \\
&= \frac{1}{(\sum_{t=1}^{T}(t-\bar{t})^2)^2} \sigma^2 \sum_{t=1}^{T}(t-\bar{t})^2 \\
&= \frac{\sigma^2}{\sum_{t=1}^{T}(t-\bar{t})^2}
\end{aligned}
$$

1.2 Task 1B

This section contains a simulation study for T = 40, with 2000 replications. *We are using the r seed (931229) to generate our sample data.*

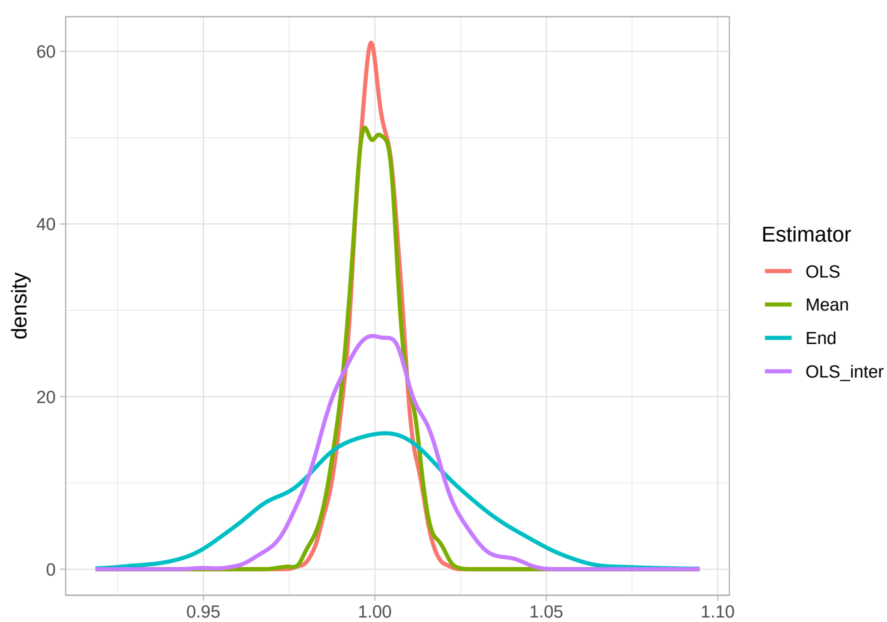

Figure 1: Sampling Distribution for Four Estimators (T = 40)

From a visual perspective, we can draw some conclusions regarding our four estimators. All four estimators seem to be unbiased, i.e. centered around the expected value $\beta = 1$, which is also supported by our calculations from the previous subtask.

The OLS estimator seems to have the lowest variance and the Mean estimator seems to have the second lowest variance. The End estimator has the largest variance, which again is supported by the theoretical calculations.

Our plots are not skewed, but some plots have “dents” at the peak of the graph. This can be explained by the random sample generated by our seed - when we changed our seed the dents also changed to different places.

Table 1: Descriptives of the sampling distribution for four estimators (T = 40)

Estimators	T	Mean	Median	Variance	Skewness	Kurtosis
OLS	40	1	1	0.00004533	-0.04	3.00
Mean	40	1	1	0.00005626	-0.03	3.06
End	40	1	1	0.00064169	0.03	2.99
OLS_inter	40	1	1	0.00020257	0.09	3.09

Looking at the table, we see that the mean is approximately equal to the expected value for the different estimators. The variances are also very close to 0 but there are some small differences between the different variances. We can conclude that the OLS estimator is the best one out of the four unbiased estimators in the linear regression, since it seems that its variance is the smallest. This is reasonable since graphically the density of the OLS estimator is the most narrow.

The “End” estimator is the **largest** variance out of all the estimators. This can be explained by the fact that this estimator only measures point T. While the other estimators take into consideration all the data points we have, the End estimator only measures one point.

None of the sampling distributions seem to be skewed since the values representing skewness are very low.

1.3 Task 1C

Task C is a repeat of Task 1B with a sample size of 400 instead of 40. The same seed and amount of replications apply.

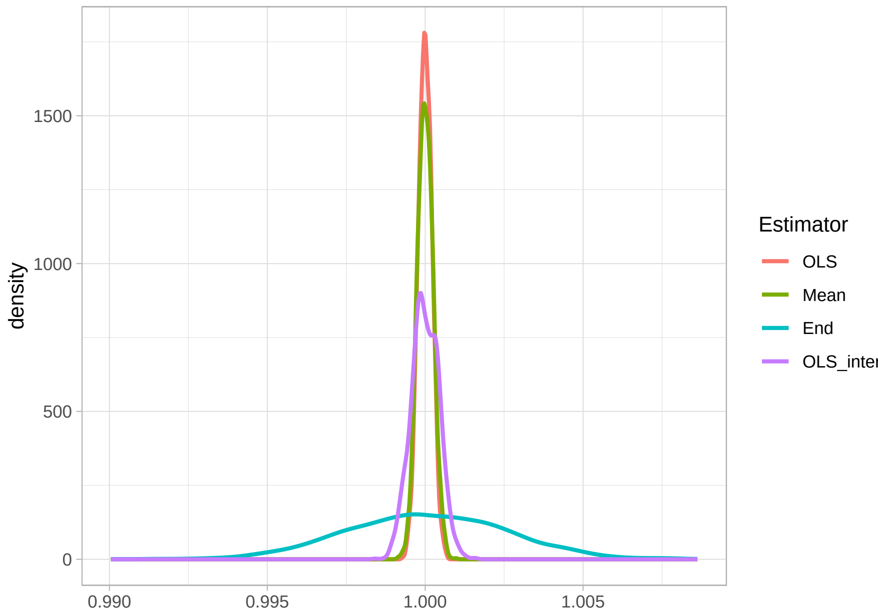

Figure 2: Sampling Distribution for Four Estimators (T = 400)

From a visual perspective we can tell that the estimators are still unbiased since they are all centered around the true value of $\beta = 1$. However the variances are now smaller, but the size of the variances compared to each other still have the same ordering as in 1B, i.e. OLS still has the smallest variance, followed by "Mean" estimator, etc. The difference of the variances are now much smaller compared to when the sample size was 40. This could mean that a further increase in sample size will lead to a smaller difference between the variances of the estimators.

Our plots are not skewed, but some plots have "dents" at the peak of the graph. This can be explained by the random sample generated by our seed - when we changed our seed the dents also changed to different places.

Table 2: Descriptives of the sampling distribution for four estimators (T = 400)

Estimators	T	Mean	Median	Variance	Skewness	Kurtosis
OLS	400	1	1	5.00e-08	0.01	2.87
Mean	400	1	1	6.00e-08	0.05	3.05
End	400	1	1	6.54e-06	-0.01	3.08
OLS_inter	400	1	1	1.90e-07	0.01	2.93

We can again see that the mean of all the estimators are **still** approximately equal to their expected values. The variances are also similar to the previous simulation, but far smaller. We can conclude that the OLS estimator is still the best one out of the four unbiased estimators in the linear regression, since it again seems like its variance is the smallest. This is reasonable since graphically the density of the OLS estimator is the most narrow, although the difference between all of the variances are far smaller.

The End estimator has the largest variance out of all the estimators. This can be explained by the same fact as in 1B.

Nothing has changed regarding the skewness of the estimators compared to the first simulation.

1.4 Task 1D

Table 3: Compare the results between the two simulation studies

Estimators	T	Mean	Median	Variance	Skewness	Kurtosis
OLS	40	1	1	0.00004533	-0.04	3.00
Mean	40	1	1	0.00005626	-0.03	3.06
End	40	1	1	0.00064169	0.03	2.99
OLS_inter	40	1	1	0.00020257	0.09	3.09
OLS	400	1	1	0.00000005	0.01	2.87
Mean	400	1	1	0.00000006	0.05	3.05
End	400	1	1	0.00000654	-0.01	3.08
OLS_inter	400	1	1	0.00000019	0.01	2.93

Since the variance of all of the estimators decreased when we increased the sample size from 40 to 400, we suspect that if we keep increasing the sample size to 4000, and then to 40000, the variances will get smaller and smaller. The shape of the distribution will become more narrow. We suspect that these changes are caused by the fact that the estimators are actually consistent; meaning that their variances will converge to 0 when we increase the sample size.

2 Task 2

2.1 Task 2A

We start with generating our two sets of 40 observations each, one for $\sigma = 1$ and the other $\sigma = 10$. The estimations of the model's parameters can be found in the table below.

Table 4: Estimations and variances for regression coefficients

	Estimate	Variance
β_1 ($\sigma = 1$)	10.09	0.09
β_2 ($\sigma = 1$)	1.00	0.00
β_1 ($\sigma = 10$)	10.92	9.06
β_2 ($\sigma = 10$)	0.97	0.02

We notice that the intercept and slope slightly changed while the variances drastically increased when the standard deviation increased.

1) When we increase the standard deviation from 1 to 10, we notice that the intercept, i.e. $\hat{\beta}_1$, has increased slightly. The change in the variable $\hat{\beta}_1$ is much smaller than the change in the estimator of the variance, $Var(\hat{\beta}_1)$, which has drastically changed. This is due to the change of the standard deviation. Thus, we can observe that a change in standard deviation affects the variance to a greater extent than the estimators of the variables. We assume that if we were to increase $\sigma = 20$, $Var(\hat{\beta}_1)$ would increase even further, and $\hat{\beta}_1$ will change.
2) When we increase the standard deviation from 1 to 10, we notice that the intercept, i.e. $\hat{\beta}_2$, has decreased slightly instead of increased. The change in the variable $\hat{\beta}_2$ is still much smaller than the change in the estimator of the variance. We can still observe that a change in standard deviation affects the variance to a greater extent than the estimators of the variables. We assume that if we were to increase $\sigma = 20$, $Var(\hat{\beta}_2)$ would increase even further, and $\hat{\beta}_1$ will change.

2.2 Task 2B

The detailed result of the regression can be seen in the appendix

Calculate (numerically) for the model with $\sigma = 10$ in Task 2A:

Table 5: Results For the Regression when $sd = 10$ (6 out of 40 heads)

y_hat	resid	y	t
11.886	5.927	17.814	1
12.853	-5.679	7.174	2
13.820	7.469	21.289	3
14.786	-12.007	2.779	4
15.753	14.978	30.731	5
16.719	0.561	17.281	6

The correlation between t and $\hat{u}_t$ is $1.1388496 \times 10^{-17}$.

The correlation between $\hat{Y}_t$ and $\hat{u}_t$ is $-3.6244593 \times 10^{-17}$.

The correlation between Y_t and $\hat{u}_t$ is 0.6322927.

The correlation between Y_t and t is 0.7747296.

Table 6: Comparison between Theoretical values and Simulated values

Correlation between	Theoretical values	Simulated values
t and $\hat{u}_t$	0	0.000
$\hat{Y}_t$ and $\hat{u}_t$	0	0.000
Y_t and $\hat{u}_t$	NA	0.632
Y_t and t	NA	0.775

Theoretically, there should **not** be a correlation between the residual ($\hat{u}$) and the independent variable (t), which has been proven in the OLS linear regression. $\hat{Y}_t$ and t should **not** be correlated because $\hat{Y}_t$ is a function of t. The first two correlations in this task are effectively zero, and thus match our expectations. The third and fourth correlation from the result of simulation are reasonable since the correlations do exist and we can never calculate them before we make a simulation.

3 Task 3

3.1 1

The plots for the two weights and the unweighted plot seem to look the same at a glance, however, they are scaled differently for the Y axis. This is natural because the weight only applies to the Y variable, so only the Y axis should be affected. If we scale all the plots to the same Y axis we should see much greater differences of the slope. The slope for w1 = 10 would be greater, and w1 = 0.1 would be flatter.

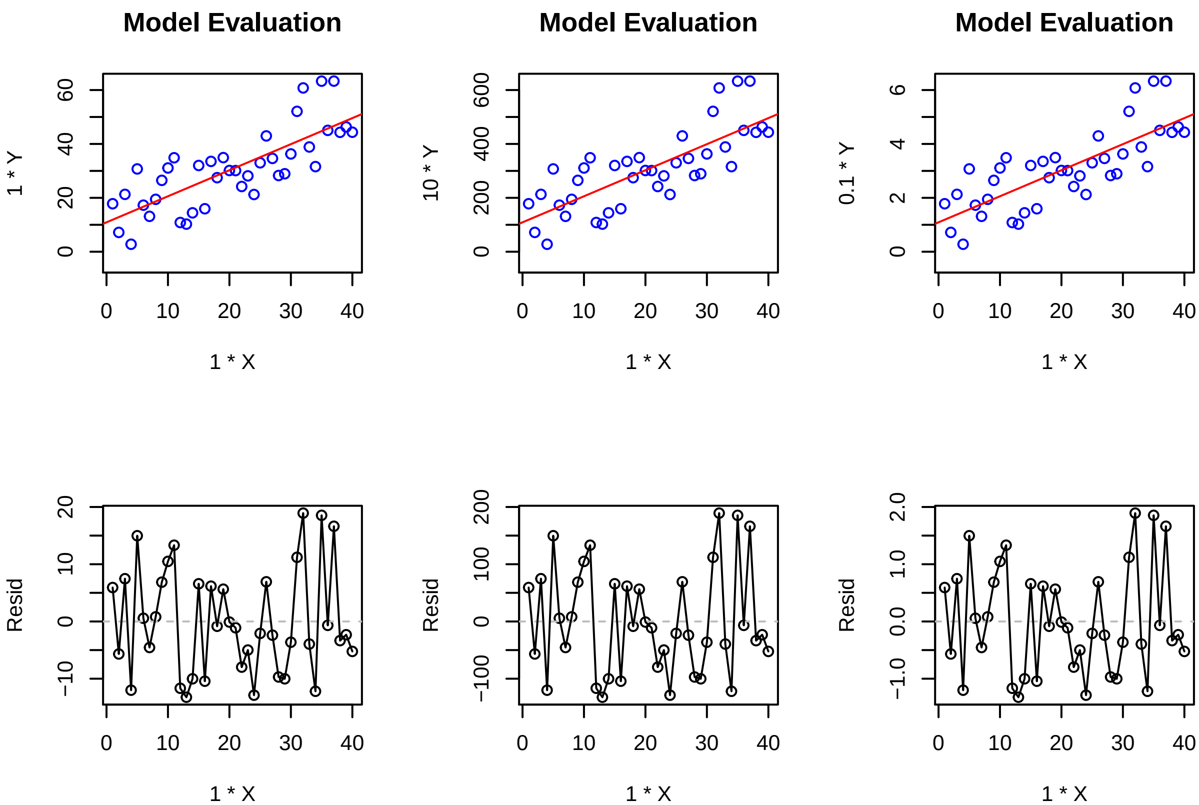

Figure 3: Plots of the actual and fitted for the rescaled and unscaled cases

3.2 2

Table 7: Estimations and variances for regression coefficients

	Estimate	Variance	T-test
β_1 $(w = 1)$	10.92	9.06	3.63
β_2 $(w = 1)$	0.97	0.02	7.55
β_1 $(w = 10)$	109.20	906.47	3.63
β_2 $(w = 10)$	9.67	1.64	7.55
β_1 $(w = 0.1)$	1.09	0.09	3.63
β_2 $(w = 0.1)$	0.10	0.00	7.55

3.3 3

When we apply the weight to the variable Y, β_1 and $Var(\beta_1)$ change accordingly i.e. the new β_1 becomes the old β_1 times w_1 and the new variance becomes the old variance multiplied by w_1^2.

$\hat{Y}_t = \hat{\beta}_1' + \hat{\beta}_2' t + \hat{\mu}_t$ where $\beta_1' = 10$ and $\beta_2' = 1$

$Y_t^{***} = \beta_1 + \beta_2 t + \mu_t$

$Y_t^{***} = \omega_1 Y_t = \omega_1 \beta_1' + \omega_1 \beta_2' t + \omega_1 \mu_t$

So, $\beta_1 = \omega_1 \beta_1'$ and $\beta_2 = \omega_1 \beta_2'$.

We also get, $Var(\beta_1) = Var(\omega_1 \beta_1') = \omega_1^2 Var(\beta_1')$ and $Var(\beta_2) = Var(\omega_1 \beta_2') = \omega_1^2 Var(\beta_2')$.

This can be seen in our table, as the estimate of $\hat{\beta}_1 w_1 = 10.92$, but the same beta times the weight of 10, $\hat{\beta}_1 w_{10} = 109.20$. This means that the only difference between the estimates are the weights we put on them. Same thing can be seen with the variances. The t-statistic does not change. Theoretically, when calculating the t-value we would multiply and divide by w_1, i.e

$t_{value}^{****} = \frac{Var(\hat{\beta}_1) - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{w_1 \hat{\beta}_1' - w_1 \beta_1'}{w_1 \sigma_{\hat{\beta}_1'}} = t_{value}$

Hence, we can simplify and thus not change the t-value. Intuitively, the t-value measures the size of the difference relative to the variation the data so it would not get affected by weighting the data. We can see the proof in our table by observing that all the t-tests for β_1 are the same number.

3.4 4

Similarly to the previous observation, when we apply our weight we multiply the estimates with our weight. $w_1 \hat{\beta}_2 = w_1 \hat{\beta}_2$. We can see that $w_1 \hat{\beta}_2 = 0.97$ and $w_1 = 10$ is equal to 9.67. The variance also follows this pattern but with larger differences due to the weights being squared. Theoretically, when calculating the t-value we would multiply and divide by w_1, i.e $t_{value} = w_1 t_{value} / w_1$ Hence, we can simplify and thus not change the t-value. Hence, it will be simplified and thus not change the t-value. Intuitively, the t-value measures the size of the difference relative to the variation the data so it would not get affected by weighting the data. We can see in our table that the t-value doesn't change.

4 Task 4

4.1 1

The plots for the two weights and the unweighted plot seem to look the same at a glance, however, they are scaled differently for the X axis. This is natural because the weight only applies to the X variable, so only the X axis should be affected. If we scale all the plots to the same X axis we should see much greater differences of the slope. The slope for w2 = 10 would be flatter, and w2 = 0.1 would be greater.

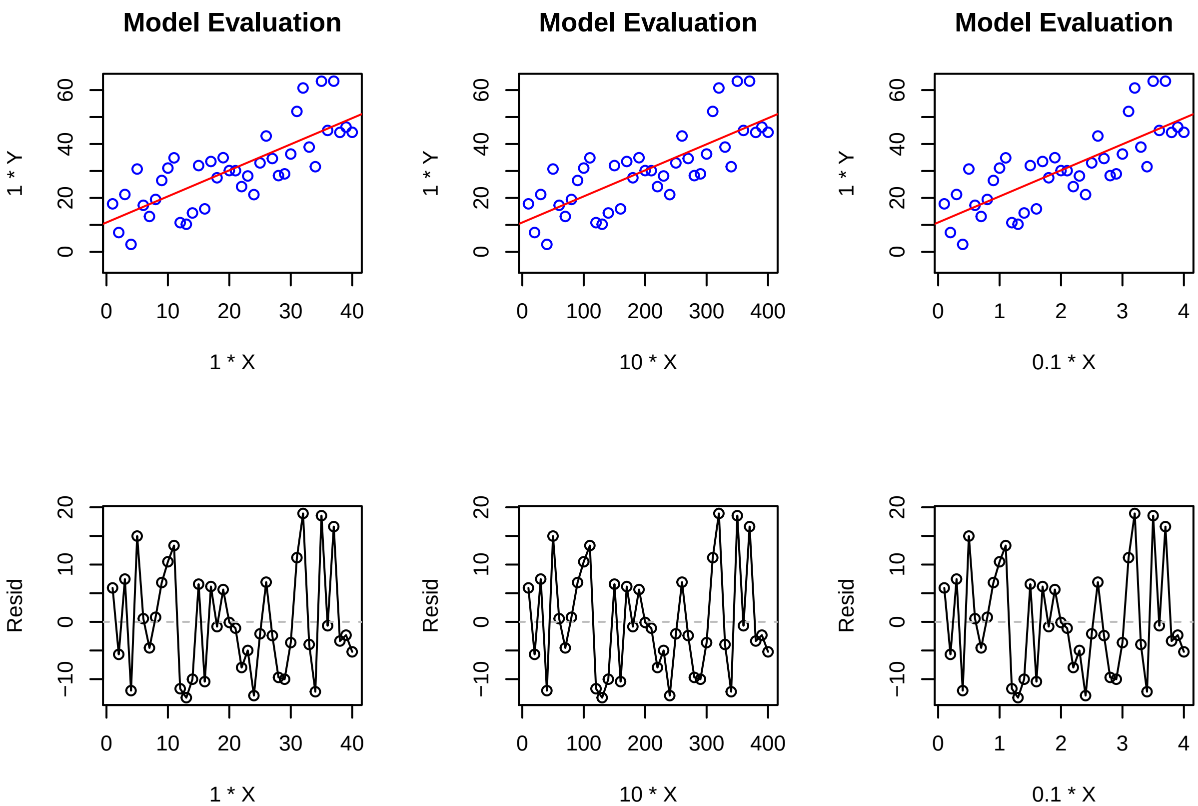

Figure 4: Plots of the actual and fitted for the rescaled and unscaled cases

4.2 2

Table 8: Estimations and variances for regression coefficients

	Estimate	Variance	T-test
β_1 $(w = 1)$	10.92	9.06	3.63
β_2 $(w = 1)$	0.97	0.02	7.55
β_1 $(w = 10)$	10.92	9.06	3.63
β_2 $(w = 10)$	0.10	0.00	7.55
β_1 $(w = 0.1)$	10.92	9.06	3.63
β_2 $(w = 0.1)$	9.67	1.64	7.55

4.3 3

When we apply the weight to the variable X, β_1 and $Var(\beta_1)$ do not really change, we are only applying the weight to β_2 in the model so β_1 and $Var(\beta_1)$ should not change.

$Y_t = \beta_1' + \beta_2' t + \mu_t$ where $\beta_1' = 10$ and $\beta_2' = 1$

$Y_t = \beta_1 + \beta_2 X_t + \mu_t = \beta_1 + \beta_2 \omega_2 t + \mu_t$

So, $\beta_1 = \beta_1'$ and $\beta_2 \omega_2 = \beta_2'$.

Hence, $Var(\beta_1) = Var(\beta_1')$ and $Var(\beta_2') = \omega_2^2 Var(\beta_2)$.

The t-statistic does not change as well.

$t_{value}^{****} = \frac{Var(\hat{\beta}_2) - \beta_2}{\sigma_{\hat{\beta}_2}} = \frac{w_2 \hat{\beta}_2' - w_2 \beta_2'}{w_2 \sigma_{\hat{\beta}_2'}} = t_{value}$

Theoretically, when calculating the t-value we would multiply and divide by w_2. Hence, it will be simplified and thus not change the t-value. Intuitively, the t-value measures the size of the difference relative to the variation the data so it would not get affected by weighting the data.

4.4 4

Similarly to the previous observation, when we apply our weight we multiply the estimates with our weight. $w_2 \hat{\beta_2} = w_2 \hat{\beta_2}$. We can see that $w_2 = 1$ which is in the table 0.97, and $w_2 = 10$ which is in the table 9.67. It is approximately $10(w_2 = 1)$. The variance also follows this pattern but with larger differences due to the weights being squared. Theoretically, when calculating the t-value we would multiply and divide by w_2, i.e $t_{value} = w_2 t_{value} / w_2$ Hence, we can simplify and thus not change the t-value. Hence, it will be simplified and thus not change the t-value. Intuitively, the t-value measures the size of the difference relative to the variation the data so it would not get affected by weighting the data. We can see in our table that the t-value doesn't change.

5 Task 5

For each of the four data sets, we estimate a simple regression model with intercept using the standard OLS-estimate for intercept and slope:

$$Y_i = \beta_0 + \beta_1 X_i + u \quad (i = 1, 2, 3, 4)$$

Then we get the estimated models:

Table 9: Regression Results for 4 Models

Model	β_0	β_1
1	3.0001	0.5001
2	3.0009	0.5000
3	3.0025	0.4997
4	3.0000	0.5000

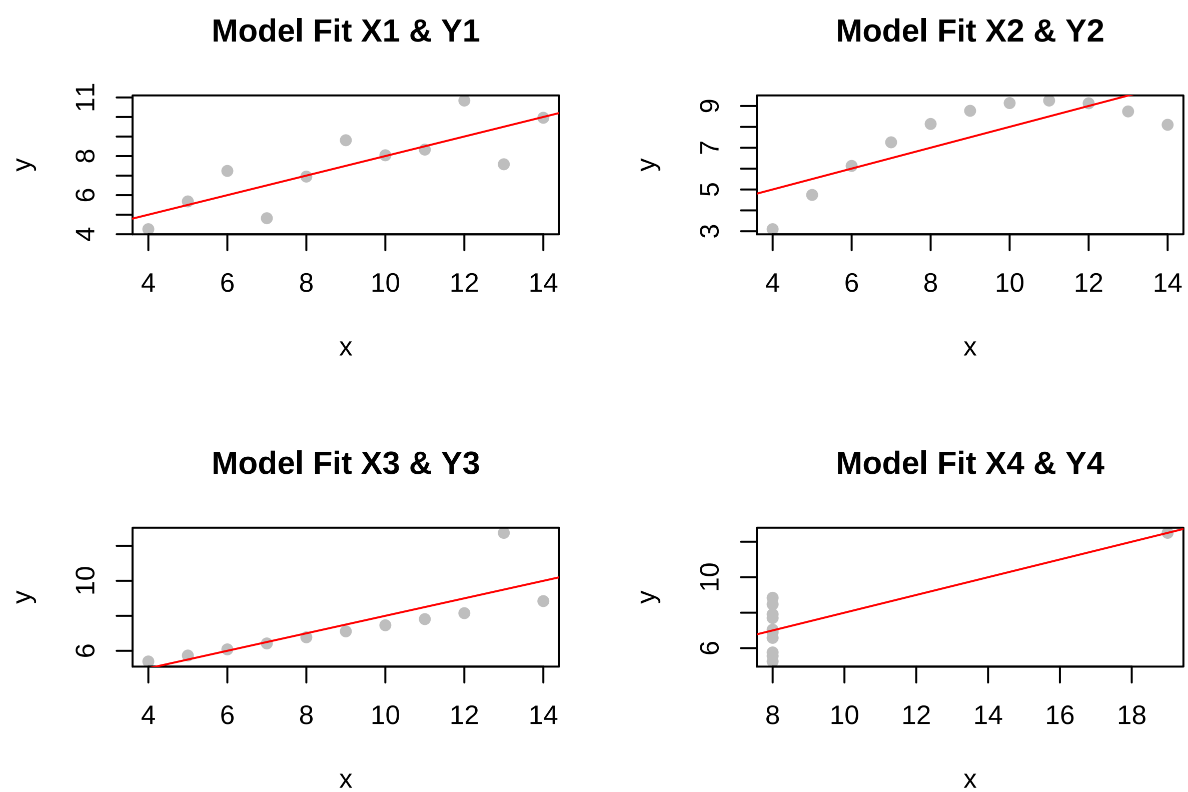

Figure 5: Simple regression models with intercept using the standard OLS-estimate

For X1 & Y1, the spread of the data points seem to give the impression of a pattern; thus it seems to be homoscedastic. The linear plot seems to fit the data points to an extent; but there is still some spread, although this spread seems to be loosely focused around the linear function. This one seems to be the one with the most consistent spread if we were to keep the outlier in the third model. It is a much better tool of estimation for the data compared to model two and model four.

The second model, which is X2 & Y2, seems to be a much poorer fit compared to the first model, as the data points seem to follow a completely different pattern, which we suspect to be a multiple non linear regression.

Table 10: Comparison among 4 models

	Dependent variable:			
	Y1	Y2	Y3	Y4
	(1)	(2)	(3)	(4)
X1	0.500^{***}			
	(0.118)			
X2		0.500^{***}		
		(0.118)		
X3			0.500^{***}	
			(0.118)	
X4				0.500^{***}
				(0.118)
Constant	3.000^{**}	3.001^{**}	3.002^{**}	3.000^{**}
	(1.125)	(1.125)	(1.124)	(1.123)
Observations	11	11	11	11
R^2	0.667	0.666	0.666	0.667
Adjusted R^2	0.629	0.629	0.629	0.630
Residual Std. Error (df = 9)	1.237	1.237	1.236	1.235
F Statistic (df = 1; 9)	17.990^{***}	17.966^{***}	17.972^{***}	18.028^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01		

In the second model the data is not centered around the line, instead following a curve. This makes the data a poor fit for a linear estimation. Compared to the the first and third models it is a worse estimatior, but probably a better one than the fourth model, as that model lacks almost any sort of slope.

For the third model, X3 & Y3, we can see that a linear model is a rather good fit, although one of the data points is an outlier. Due to this outlier the slope of the linear regression doesn't quite match the slope of the data points. If we were to remove the outlier then we would get a better fit for the model. Despite this the data is approximately linear distributed. If we remove the outlier this might become an even better fit than the first model, due to the fact that the data points seem to be very closely grouped together and seem to follow the same path, with the exception of the outlier who pulls the line upwards towards itself, thus causing the line to have a higher slope. Still a much better model than the second or the fourth, as it is still a very good fit.

For the forth model, The linear approximation seems like a poor fit. The data points seem to be concentrated around $X = 8$ with only one extreme value of $X = 19$ thus do not follow a linear model. We suspect that $X = 8$ is an influential point for the data points because the majority of them are concentrated around it. Due to this data, almost all the observations are centered around a single point, which makes this the worst estimator by far, as any sort of slope the model has is solely generated by the outlier $X = 19$. The fourth model is thus the worst model by far, as the data is far to clustered around the same value for linear approximation to work well.

6 Conclusion

In conclusion, we worked with various estimators in this assignment and compared their theoretical properties to their practical properties. We took a deeper dive into the properties of the OLS estimators when we change the parameters of the model.

Finally, we evaluated the choice of a model itself when considering different kinds of data points. This assignment helped us have a deeper understanding of how the properties of an estimator behave to different changes in the model. It also helped us to critically assess the choice of a model.

7 Appendix

7.1 Task 2A

A simulate for 2000 times in order to find the variance.

```
beta = c()
for(i in 1:1000)
{
  x <- 1:n # 1, ..., 40

  set_sd = 1
  u_t = rnorm(n, mean = 0, sd = set_sd)
  y = 10 + x + u_t

  # solution
  estimate <- lm(y ~ x) # ignore data argument of lm as we're calling two seperate vectors
  estimate_beta = estimate$coefficients
  beta = rbind(beta, estimate_beta)
}
kable(head(beta), caption = "Regression Coefficients for Simulations when $sd=1$ (Head 6 times)")
```

Table 11: Regression Coefficients for Simulations when $sd = 1$ (Head 6 times)

	(Intercept)	x
estimate_beta	10.419816	0.9913172
estimate_beta	10.386344	0.9899547
estimate_beta	10.303615	0.9850730
estimate_beta	9.470161	1.0161210
estimate_beta	9.902117	0.9985484
estimate_beta	10.041251	1.0011409

```
variance_sd1 = apply(beta, 2, var)

beta = c()
for(i in 1:1000)
{
  x <- 1:n # 1, ..., 40

  set_sd = 10
  u_t = rnorm(n, mean = 0, sd = set_sd)
  y = 10 + x + u_t

  # solution
  estimate <- lm(y ~ x) # ignore data argument of lm as we're calling two seperate vectors
  estimate_beta = estimate$coefficients
  beta = rbind(beta, estimate_beta)
}
kable(head(beta), caption = "Regression Coefficients for Simulations when $sd=10$ (Head 6 times)")
```

Table 12: Regression Coefficients for Simulations when $sd = 10$ (Head 6 times)

	(Intercept)	x
estimate_beta	13.195828	0.9451503
estimate_beta	7.977999	1.0474246
estimate_beta	6.196924	1.1099261
estimate_beta	9.424137	0.9665133
estimate_beta	11.555186	0.9551074
estimate_beta	8.079400	1.2126743

```
variance_sd10 = apply(beta, 2, var)

variance = rbind(variance_sd1, variance_sd10)
kable(variance, caption = "variances for the two regression coefficients", digits = 4)
```

Table 13: variances for the two regression coefficients

	(Intercept)	x
variance_sd1	0.1112	0.0002
variance_sd10	10.5773	0.0196

7.2 Task 2B

Detailed result of the regression for task 2B

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3236 -0.6253 -0.1596  0.6654  1.8928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0920     0.3011   33.52   <2e-16 ***
## x             0.9967     0.0128   77.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9343 on 38 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9936
## F-statistic:  6065 on 1 and 38 DF,  p-value: < 2.2e-16
```

Table 14: Results For the Regression ($sd = 1$)

y_hat	resid	y	t
11.089	0.593	11.681	1
12.085	-0.568	11.517	2
13.082	0.747	13.829	3
14.079	-1.201	12.878	4
15.075	1.498	16.573	5
16.072	0.056	16.128	6
17.069	-0.455	16.614	7
18.065	0.080	18.145	8
19.062	0.687	19.748	9
20.059	1.049	21.108	10
21.055	1.332	22.387	11
22.052	-1.167	20.885	12
23.049	-1.324	21.725	13
24.045	-1.002	23.044	14
25.042	0.658	25.700	15
26.039	-1.044	24.995	16
27.035	0.617	27.652	17
28.032	-0.084	27.948	18
29.028	0.564	29.593	19
30.025	-0.010	30.015	20
31.022	-0.110	30.912	21
32.018	-0.798	31.221	22
33.015	-0.499	32.516	23
34.012	-1.288	32.724	24
35.008	-0.209	34.800	25
36.005	0.693	36.698	26
37.002	-0.239	36.762	27
37.998	-0.970	37.028	28

y_hat	resid	y	t
38.995	-1.002	37.993	29
39.992	-0.363	39.629	30
40.988	1.121	42.109	31
41.985	1.893	43.878	32
42.982	-0.395	42.587	33
43.978	-1.220	42.759	34
44.975	1.854	46.829	35
45.972	-0.068	45.904	36
46.968	1.663	48.631	37
47.965	-0.335	47.630	38
48.962	-0.232	48.730	39
49.958	-0.524	49.435	40

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.236  -6.253  -1.596   6.654  18.928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.9199     3.0108   3.627 0.000839 ***
## x             0.9666     0.1280   7.553 4.45e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.343 on 38 degrees of freedom
## Multiple R-squared:  0.6002, Adjusted R-squared:  0.5897
## F-statistic: 57.05 on 1 and 38 DF,  p-value: 4.446e-09
```

Table 15: Results For the Regression ($sd = 10$)

y_hat	resid	y	t
11.886	5.927	17.814	1
12.853	-5.679	7.174	2
13.820	7.469	21.289	3
14.786	-12.007	2.779	4
15.753	14.978	30.731	5
16.719	0.561	17.281	6
17.686	-4.548	13.138	7
18.653	0.798	19.451	8
19.619	6.865	26.485	9
20.586	10.493	31.078	10
21.552	13.322	34.875	11
22.519	-11.673	10.846	12
23.485	-13.236	10.249	13
24.452	-10.016	14.436	14
25.419	6.584	32.002	15
26.385	-10.437	15.949	16
27.352	6.166	33.518	17

y_hat	resid	y	t
28.318	-0.837	27.481	18
29.285	5.641	34.926	19
30.252	-0.103	30.148	20
31.218	-1.103	30.115	21
32.185	-7.977	24.207	22
33.151	-4.994	28.157	23
34.118	-12.879	21.239	24
35.085	-2.088	32.996	25
36.051	6.929	42.980	26
37.018	-2.393	34.624	27
37.984	-9.700	28.284	28
38.951	-10.019	28.932	29
39.917	-3.626	36.291	30
40.884	11.209	52.093	31
41.851	18.928	60.779	32
42.817	-3.949	38.868	33
43.784	-12.197	31.587	34
44.750	18.540	63.290	35
45.717	-0.679	45.038	36
46.684	16.630	63.314	37
47.650	-3.346	44.304	38
48.617	-2.318	46.299	39
49.583	-5.235	44.348	40