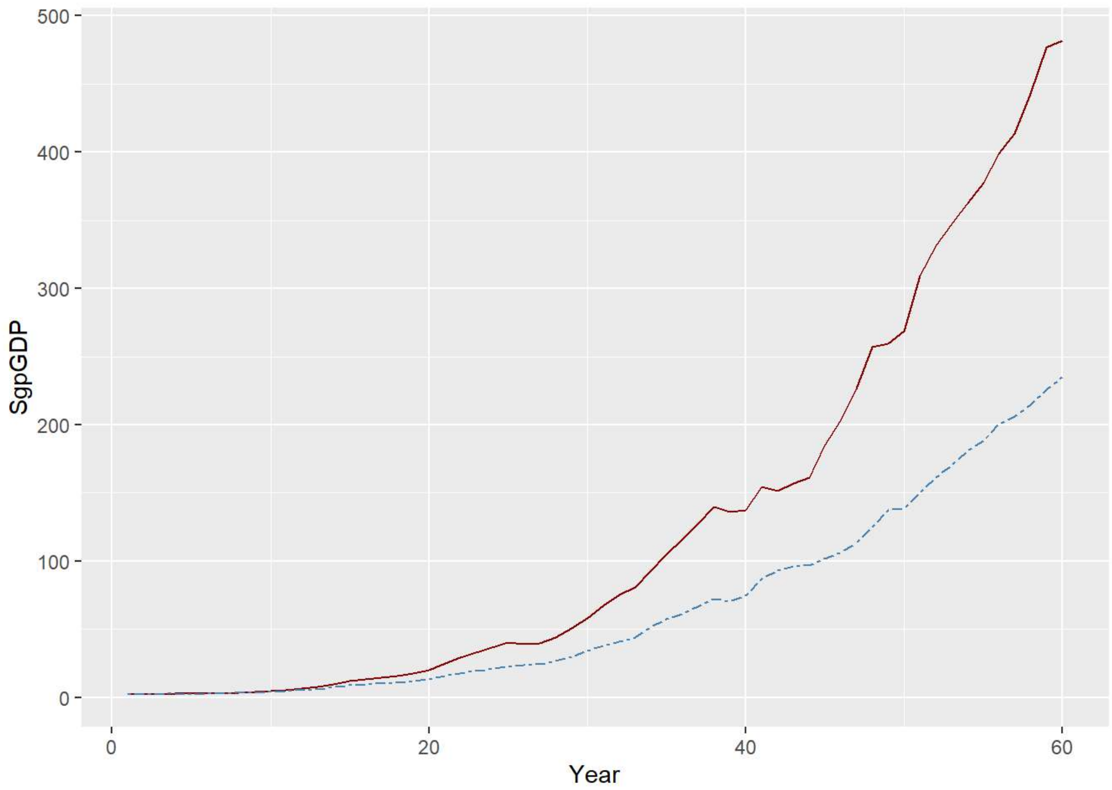
Introduction

In previous lectures Causal Explanation of Autonomy & Invariance of Regression Relationships, we have seen the essential importance of analyzing the causal structure of the variables under study via path diagrams. In this lecture, we study one aspect of this which creates serious problems for understanding and interpreting regression results. This is the problem of a common cause.

is a common cause of and if and . When this happens, and will be correlated, but neither variable causes the other. In such situations, a regression of on will give results which show, according to standard methods for analysis, that is a strong determinant of . This is called a *spurious* regression, or a *nonsense* regression. In this lecture, we will study some examples of this phenomena in real world data.

A Causal Relationship: A Consumption Function

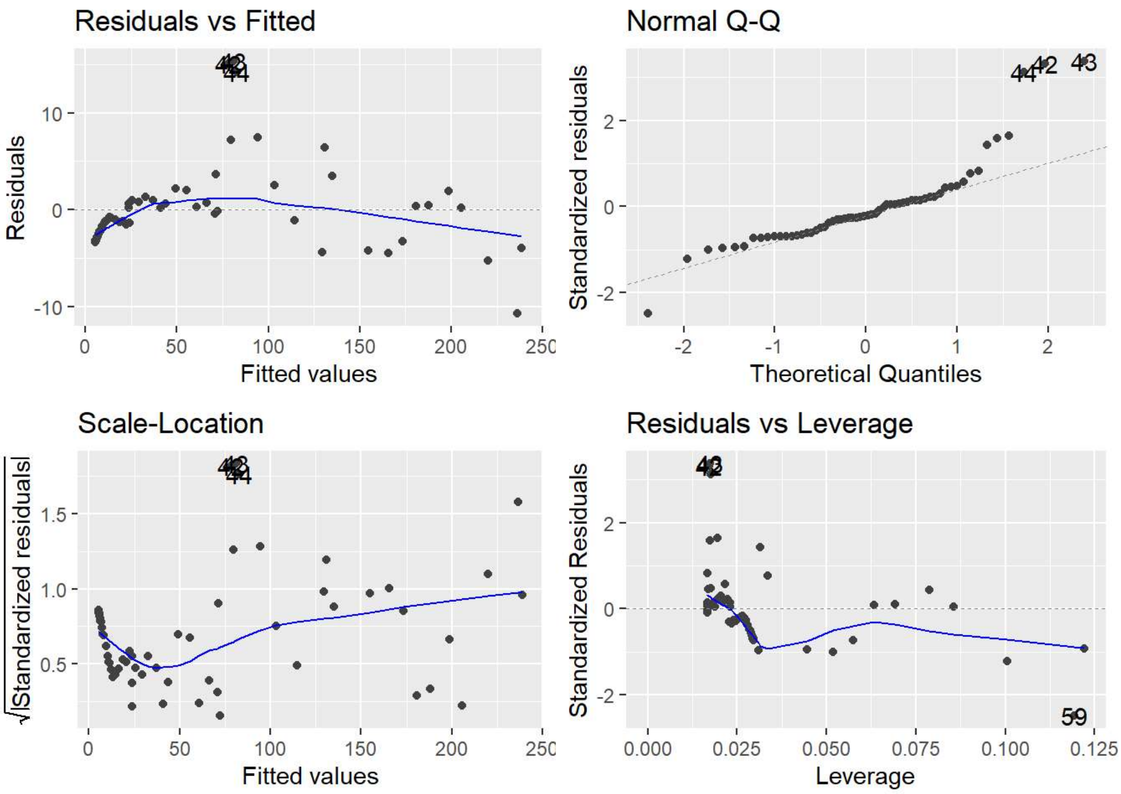
Data on annual GDP, Consumption, and Investment, for Singapore, taken from the WDI data set of the World Bank, is plotted below:

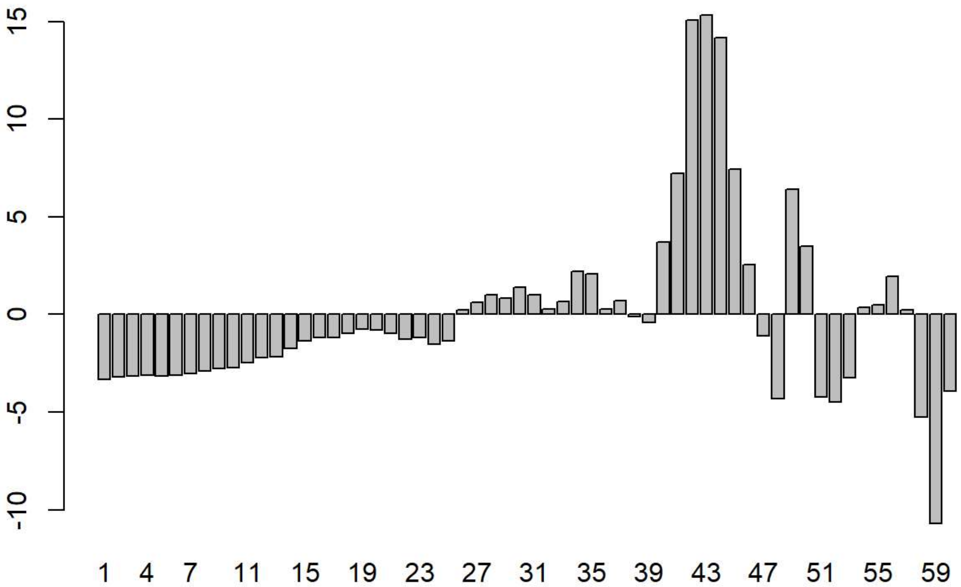


The Keynesian consumption function is one of the most widely accepted and estimated regression models. The causal hypothesis is that Income (GDP) determines Consumption (Con): GDP Con. The simplest regression model which embodies this relationship is: Con = a + b GDP. Running this regression on the data leads to the following results:

```
# A tibble: 60 × 5
      ID SgpCon SgpGDP SgpCon_hat residual
   <int>  <dbl>  <dbl>      <dbl>    <dbl>
 1     1   2.09   2.04       5.42    -3.33
 2     2   2.33   2.22       5.51    -3.18
 3     3   2.44   2.40       5.60    -3.15
 4     4   2.61   2.68       5.73    -3.12
 5     5   2.53   2.60       5.69    -3.16
 6     6   2.66   2.82       5.80    -3.13
 7     7   2.92   3.17       5.97    -3.04
 8     8   3.25   3.60       6.18    -2.92
 9     9   3.65   4.15       6.44    -2.79
```

```
10     10   4.02   4.82        6.77    -2.74
# ℹ 50 more rows
```



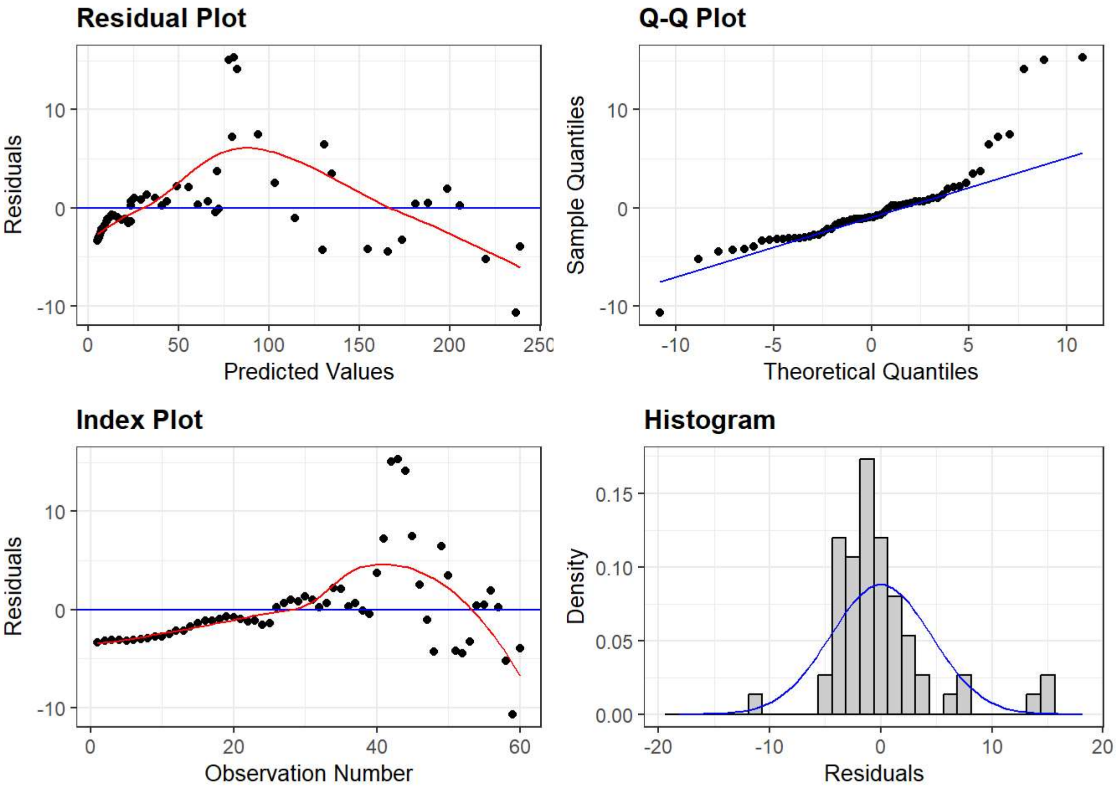


	(1)
(Intercept)	4.425 ***
	(0.794)
SgpGDP	0.486 ***

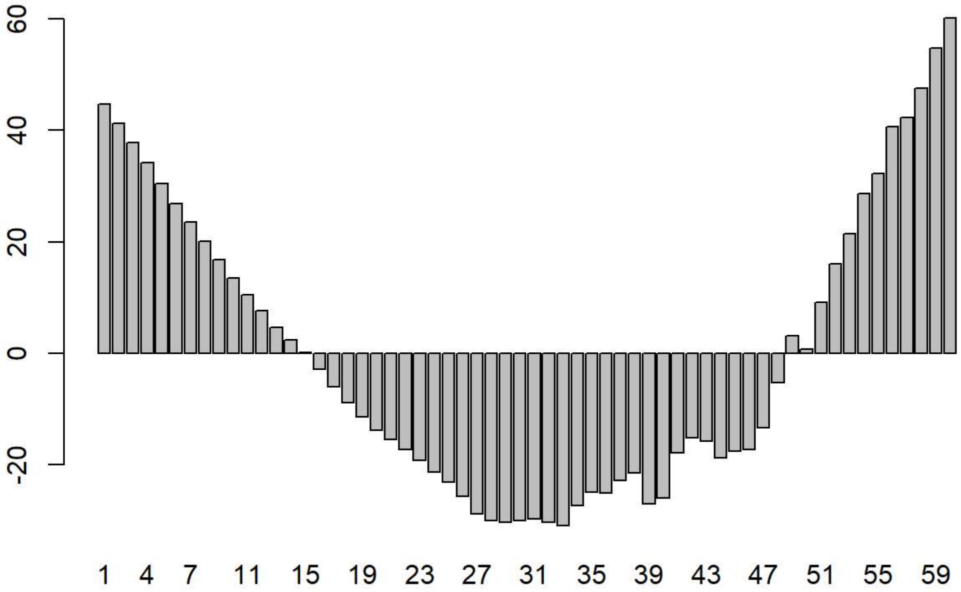
	(0.004)
N	60
R2	0.996
logLik	-175.410
AIC	356.821

*** p < 0.001; ** p < 0.01; * p < 0.05.

The regression has of 99.6%, which is interpreted to mean that 99.7% of the variation in Singapore Consumption can be explained by the Singapore GDP. The t-stat of 116 shows that the coefficient 0.486 of SgpGDP is highly significant. The p-value of 0.000 means we can reject the null hypothesis that the true coefficient is 0.0, corresponding to the idea that SgpGDP has no influence on SgpCon. Validity of regression results depends on a large number of assumptions, which are discussed in econometrics textbooks. One of the central assumptions is that the regression residuals should be random, and should come from a common distribution. To check whether or not this holds, we graph the regression residuals, the differences between the actual value and the regression fit:



This plot shows serious problems, since these residuals display systematic behavior. They are all negative and small early. To see how these patterns differ from independent random variables, we provide a graph of independent random variables with mean 0 and standard error 4.579, matching the estimated regression model standard error. The Keynesian consumption function is one of the most widely accepted and estimated regression models. The causal hypothesis is that Income (GDP) determines Consumption (Con): GDP Con. The simplest regression model which embodies this relationship is: Con = a + b GDP. Running this regression on the data leads to the following results:



Random residuals frequently switch signs. They do not display any patterns in sequencing. The patterned residuals in the consumption function prove that the regression is not valid. In such situations, econometricians typically assume that the problem is due to missing regressors or wrong functional form. By adding suitable additional regressors, and modifying the functional form, one can generally ensure that the residuals appear to satisfy the assumptions made about them. But, solving the problem of random residuals by this search over regression models creates a serious problem. This can be explained as follows. Some of the fits in the data reflect a genuine real-world relationship. Other patterns are only accidental, and do not reflect any genuine relationship. The more we search, the more likely we are to end up with an accidental pattern. The fact that most regression relationships breakdown very frequently is due to the fact that most of them are accidental patterns in the data without any counterpart in reality.

The coefficient of the regressor is supposed to be a measure of the causal effect. According to the regression above, if SgpGDP goes up by , then of it will be spent on consumption. This means that the enormously high proportion of 51% of the income will be saved. Since savings translate to investments, this could account for the dramatic growth of Singapore over the period in question. However, the patterns shown in the errors show that we cannot rely on the validity of this estimate. Vast amounts of experience with estimating this kind of regression function leads to two major conclusions.

> 1.The estimate 0.49 of the causal effect is highly unstable – by adding many different plausible variables, we can make it change over a great range of values. Thus, these regressions do not provide accurate estimates of the key parameter we want to estimate. 2.The best fitting regressions very often make huge forecasting errors, and therefore are not very useful for policy.

```
Even though the regression is not very useful in estimating the size of the causal effect, at least the

## A Spurious Regression: Correlation without Causation

First, let us just demonstrate the problem by running a nonsense regression of SgpCon on SAfGDP -- what

A regression of SgpCon on SAfGDP yields: $SgpCon = 16.99 + 0.517 SAfGDP$: $R^2$ is 97%, and the $t-stat

## The Distinction Between Nominal and Real Econometrics

The central problem with conventional econometrics textbooks is the failure to clarify the difference b
```

````
It is useful to clarify the correct interpretation of the coefficient 0.517 of SAfGDP in the second reg

## The Nominalist Solutions

The deeper point is that assessing whether or not a relationship is causal ALWAYS requires going beyond

The real solution to the problem of nonsense regressions can only be found by deeper study of causal st

## Missing Variables

In the second regression, which regresses SgpCon on SAfGDP, the regression has a very high R-squared, a

## Variable Transformation to Stationarity:

The reason for the failure of trend fit may be due to the reasonable assumption that the growth rate is

::: {.cell}
::: {.cell-output-display}
![](spurious_regression_files/figure-html/unnamed-chunk-5-1.png){width=672}
:::

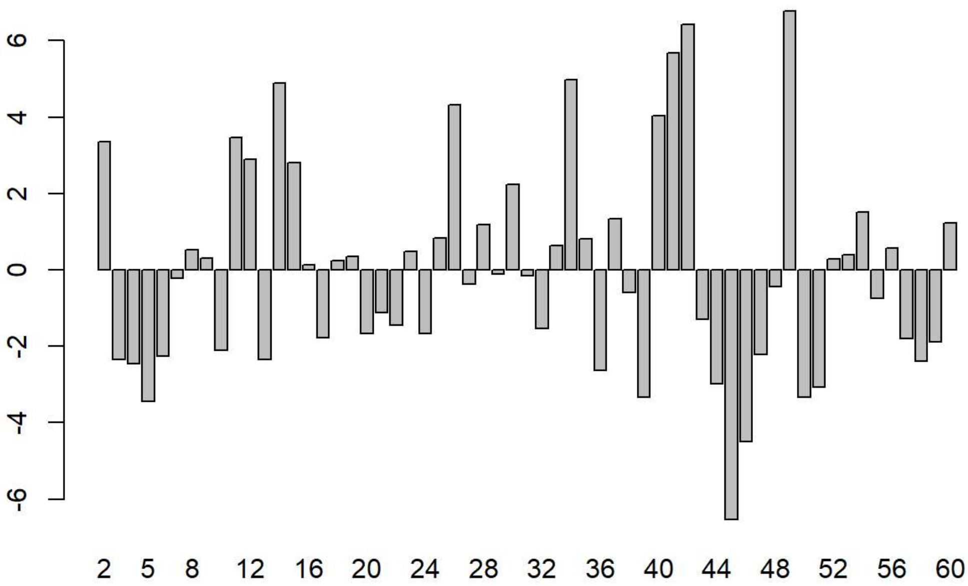
::: {.cell-output-display}
![](spurious_regression_files/figure-html/unnamed-chunk-5-2.png){width=672}
:::

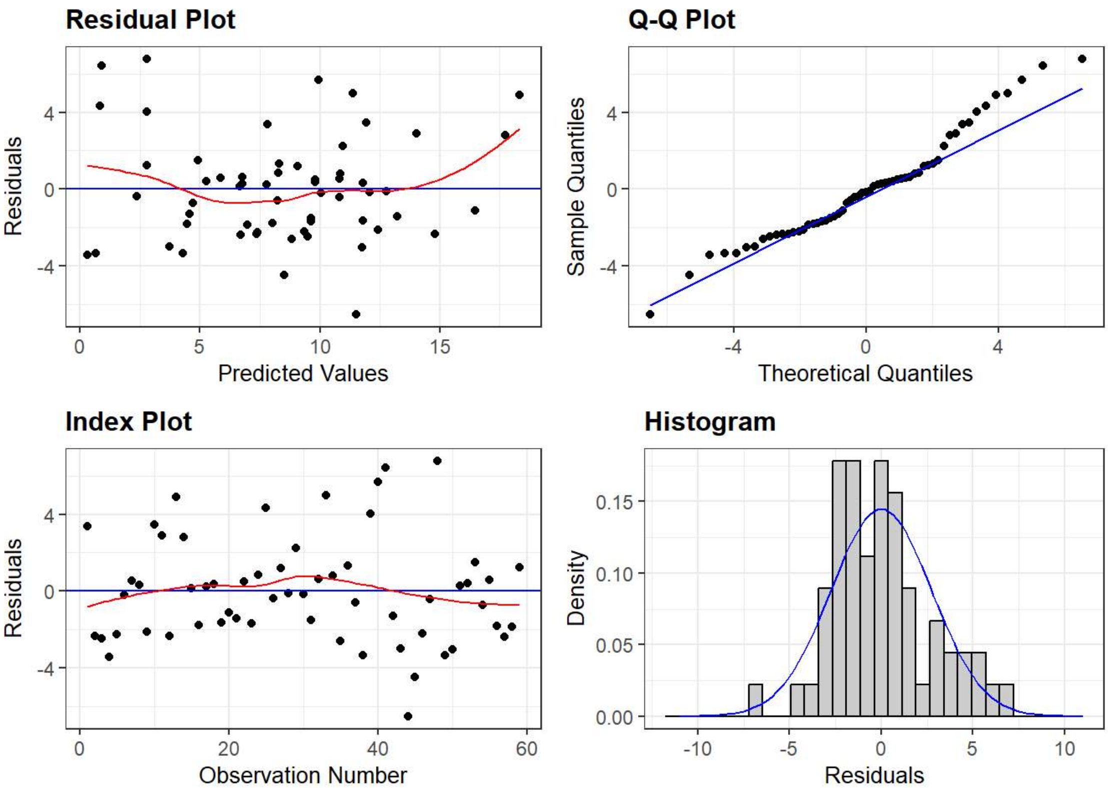
::: {.cell-output-display}

```{=html}
<table class="huxtable" data-quarto-disable-processing="true" style="border-collapse: collapse; border:
<col><col><tr>
<th style="vertical-align: top; text-align: center; white-space: normal; border-style: solid solid soli
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th style="vertical-align: top; text-align: left; white-space: normal; border-style: solid solid solid
<tr>
<th colspan="2" style="vertical-align: top; text-align: left; white-space: normal; border-style: solid
</table>
````

::: :::

This variable is also uncorrelated with time, at least visually. The variable transformation of taking growth rates has removed the common time trend from both variables. Now it should be possible to run a regression of Consumption on Income which does not suffer from the spurious correlation between the two variables created by the common trend. Unfortunately, this strategy also fails to differentiate between genuine and spurious regressions as we will see. First, let us look at the genuine causal regression of Gr_SgpC on Gr_SgfG: , is now 69%. It has gone down because we have removed the correlation due to the common factor of time-trend, but it is still very high. The regression provides a good fit. Also, for the first time, the residuals from this regression look random:
````





This transformation from levels to growth rates has substantially improved the regression results. There is now some hope that the coefficient 0.634 might be a correct measure of a causal effect, if other aspects of the specification are also correct.

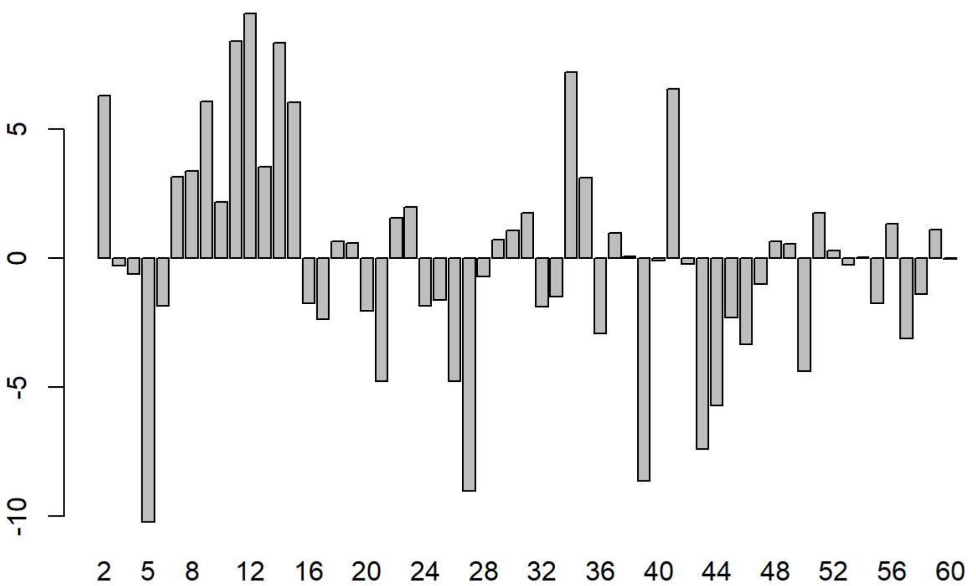
However, our main concern is to see if this transformation can distinguish between genuine causal relationships and spurious ones. For this purpose, we run the regression of the growth rate of SgpCon on the growth rate of SAfGDP. This yields the following results:

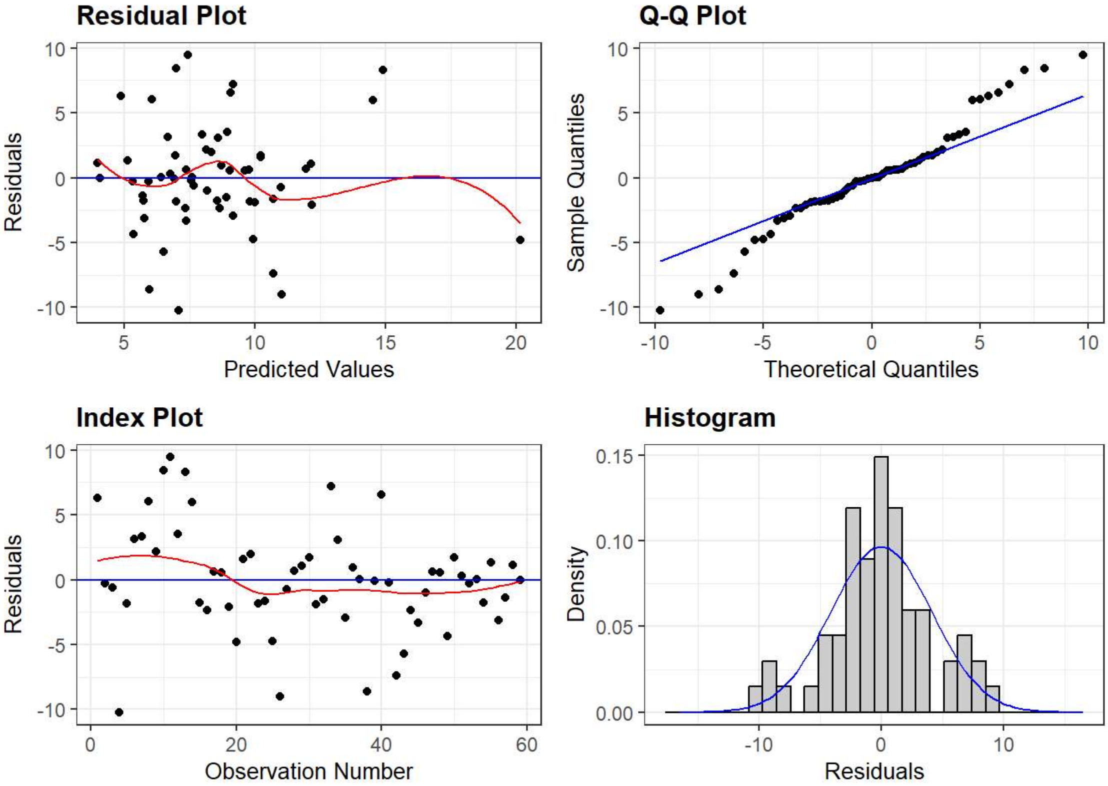
	(1)

(Intercept)	1.822
	(1.407)
gdp_gr_Saf	0.539 ***
	(0.106)
N	59
R2	0.313
logLik	-166.857
AIC	339.714

*** p < 0.001; ** p < 0.01; * p < 0.05.

This regression also shows that growth rates of South African GDP are highly significant in explaining growth rates of Singaporean Consumption. The regression residuals look random, and support the validity of this regression:



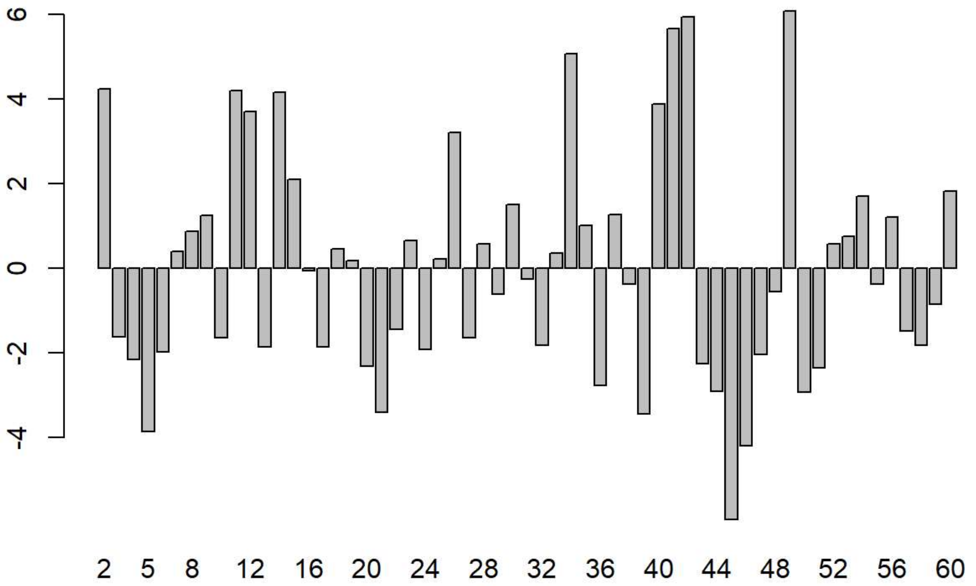


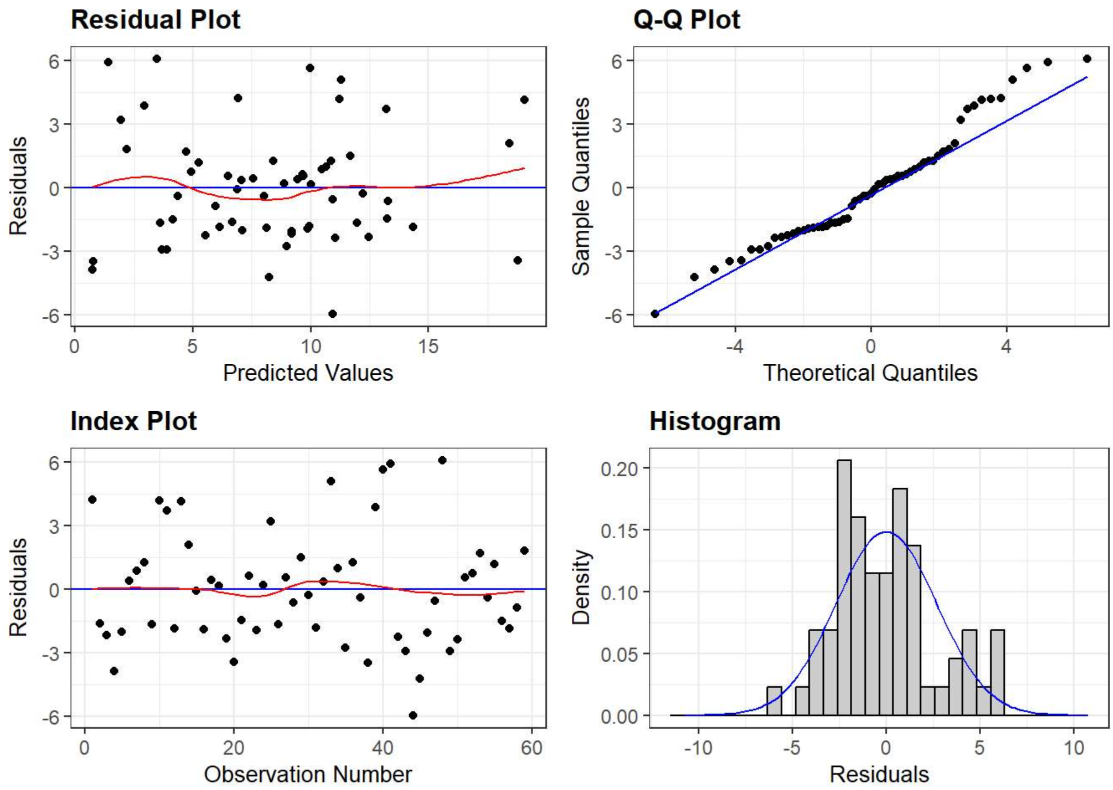
According to conventional econometrics, the causal regression of growth rates of SgpGDP on growth rates of SgpCon allows us to conclude that 63% of changes in Gr_SgpGDP are transmitted to Gr_SgpCon. But, by exactly the same reasoning, we can use the spurious regression of Gr_SgpCon on Gr_SAfGDP to conclude that 56.6% of changes in the growth rate of the South African economy are transmitted to growth rates of Singaporean consumption. The first measurement has some chance being correct, while the second statement is ridiculous.

In this particular case, the missing variables strategy can solve the problem. If we think that the source of the problem in this last equation is that the primary determinant of SgpCon is missing from the equation, we can fix the problem by adding this variable. When we do so, we get the following equation: Gr_SgpC = 0.004 + 0.163 Gr_SAfG + 0.569 Gr_SgpG

	(1)
(Intercept)	1.032
	(0.929)
gdp_gr_Saf	0.141
	(0.083)
gdp_gr	0.573 ***
	(0.066)
N	59
R2	0.709
logLik	-141.561
AIC	291.121

*** p < 0.001; ** p < 0.01; * p < 0.05.





Putting in growth rates of both Singapore and South Africa leads to the right result: Singapore GDP is a highly significant determinant, while South African GDP is not significant at the 95% level. Nominal econometrics consists of this kind of exercise, where we try one equation after another in order to get a match to our a priori ideas about the size and strength of the causal effects. This is quite the opposite of what students are taught about this methodology. It is not that we allow the data to tell us about the causal structures in the world. Rather, we know these structures in advance, and try many different formulations until we get one which matches our preconceptions.

Concluding Remarks

Econometricians have been working on finding methods to discriminate between nonsense and sensible regressions for decades, without success. We have discussed three strategies for doing so in this lecture. All three lead to failure, although the third one can be salvaged. Currently, it is the third strategy which dominates the scene. However, it also suffers from failures in many different cases. The reason for this failure is that solutions to causal problems cannot be found in nominal econometrics. Without having a good grasp of the causal structures which relate the regressors to the dependent variables, it is not possible to estimate causal effects.

POSTSCRIPT: This lecture is mainly taken from Prof.Dr. Asad Zaman post on spurious regression and reproduced using R after his permission. The WDI Data Set for Singapore and South Africa, Consumption and GDP, is available from Singapore Data