

Understanding econometric modeling with simple Consumption Function for Pakistan

Zahid Asghar

Table of contents

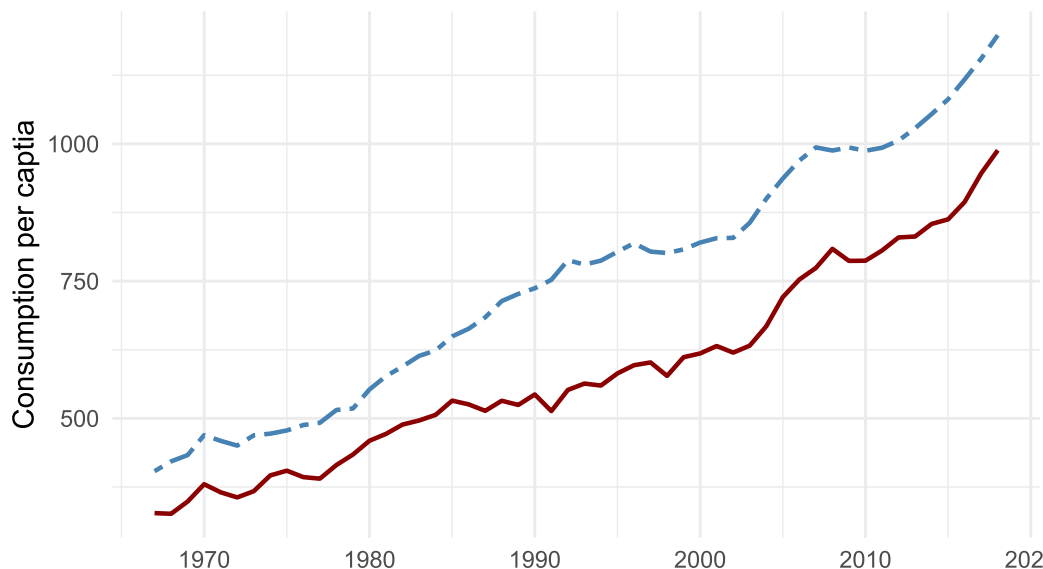
0.1 Keynesian Consumption Function for Pakistan	1
0.1.a Load data	1
0.2 2. What to do when errors are not independent?	3
0.3 STATIC VERSUS DYNAMIC MODELS	4
0.3.a General-to-Simple vs. Simple-to-General Modelling:	4
0.4 Dynamic Model	4
0.5 Multicollinearity and economic rationale	7
0.6 Summary and conclusion	8

0.1 Keynesian Consumption Function for Pakistan

Time series models are those where the data occurs in sequential time periods, so that there is a concept of progression of time. In contrast to this we have cross section models where the data is from the same period but taken from different places. In time series models, the simplest type is the static models, all the action takes place in one time period, and there is no change in activity in one time period on the next. The simplest consumption function is the static consumption function, which assumes that current consumption is determined only on current income, so that the function is $C_t = \alpha + \beta Y_t + \epsilon_t$. Assuming that the function is static, there can be a random error, ϵ_t , in the simplest Keynesian consumption function.

initiation issues of consumption expenditure over time (main source :WDI) has been used for this lecture. Figure 1 indicates that both consumption and income are trending variables.

Consumption Income pattern over time for Pakistan



0.1.a Load data

Data on annual income and consumption for Pakistan with some approximations due to def-

Figure 1: Consumption and Income per capita over time

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): $GDP \Rightarrow Con$. The simplest regression model which embodies this relationship is: $C = \alpha + \beta Y$. Running this regression on the data Table 1 gives the first five observations of the regression results.

Table 1

```
# A tibble: 5 × 5
  ID      C      Y C_
<int> <dbl> <dbl> <dbl>
1     1  327.  404.  3
2     2  326.  422.  3
3     3  348.  433.  3
4     4  380.  469.  3
5     5  365.  459.  3
```

Now we have regression results

Table 2: Regression results of Income

(Intercept)	(-4.15)
Y	(0.785)
Num.Obs.	5
R2	0.970
R2 Adj.	0.969
AIC	509.1
BIC	515.0
Log.Lik.	-251.556
F	1589.518
RMSE	30.53

The regression has R^2 of 0.97, which is interpreted to mean that 97% of the variation in Pakistani Consumption can be explained by the Pakistani GDP. The $t - stat = \frac{0.785}{0.020} = 39$

shows that the coefficient income is highly significant. The $p - value$ of 0.000 means we can reject the null hypothesis that the true coefficient is zero, corresponding to the idea that Y has no influence on C . Validity of regression results depends on a large number of assumptions, which are discussed in econometrics textbooks.

Here is the plot of residuals of regressing Per Capita consumption in Pakistan on Per Capita GDP from 1967 to 2018.

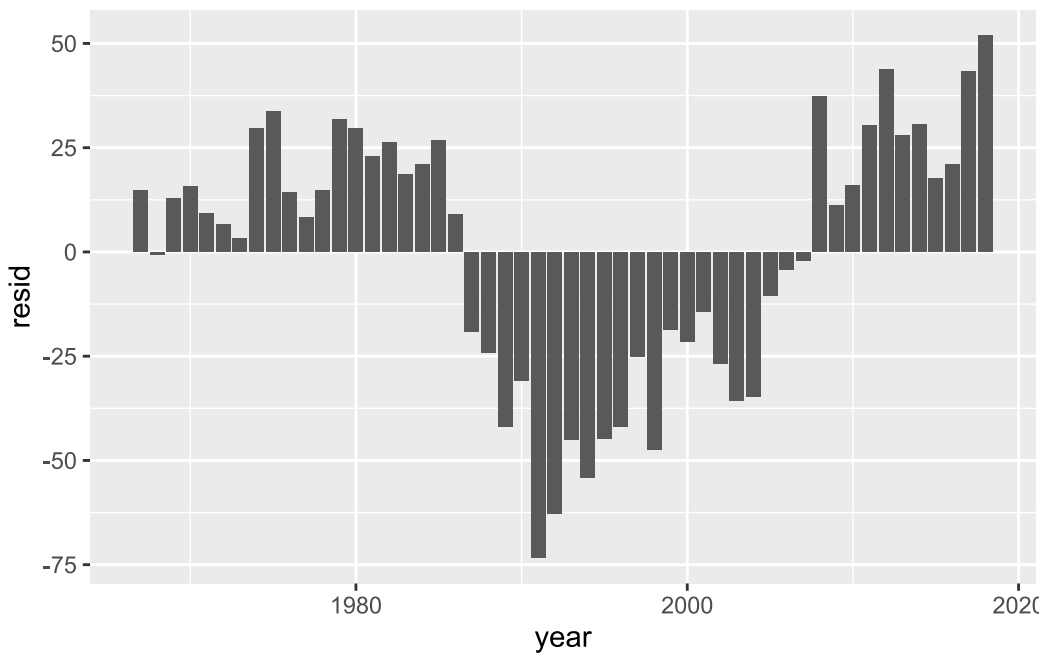


Figure 2

One of the central assumptions is that the regression residuals should be random, and should come from a common distribution. Figure 2 indicates high correlation among the residuals and these don't follow random pattern. 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, estimated regression model state Keynesian consumption function most widely accepted and estimated models. To verify further, we analysis of these residuals to heteroscedasticity and independence Figure 3.

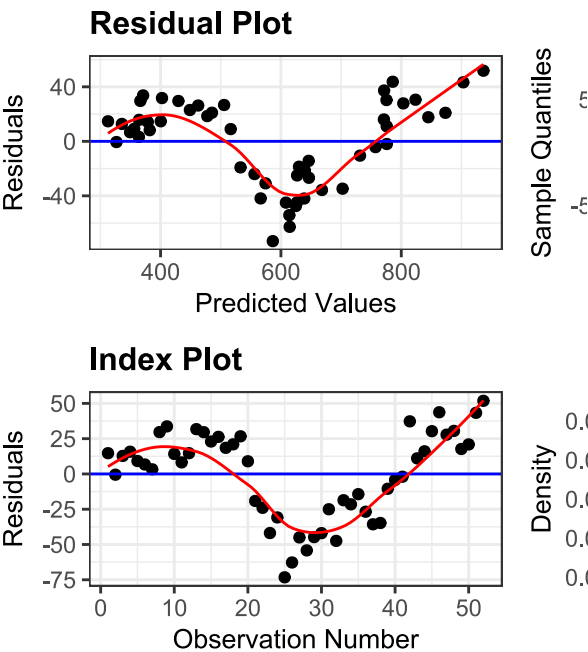


Figure 3

The causal hypothesis is that Income (GDP) determines Consumption (Con): $GDP \Rightarrow Con$. There may be a deterministic trend and now I regress consumption on trend and show graphical results of trend model and residuals are given in Figure 4:

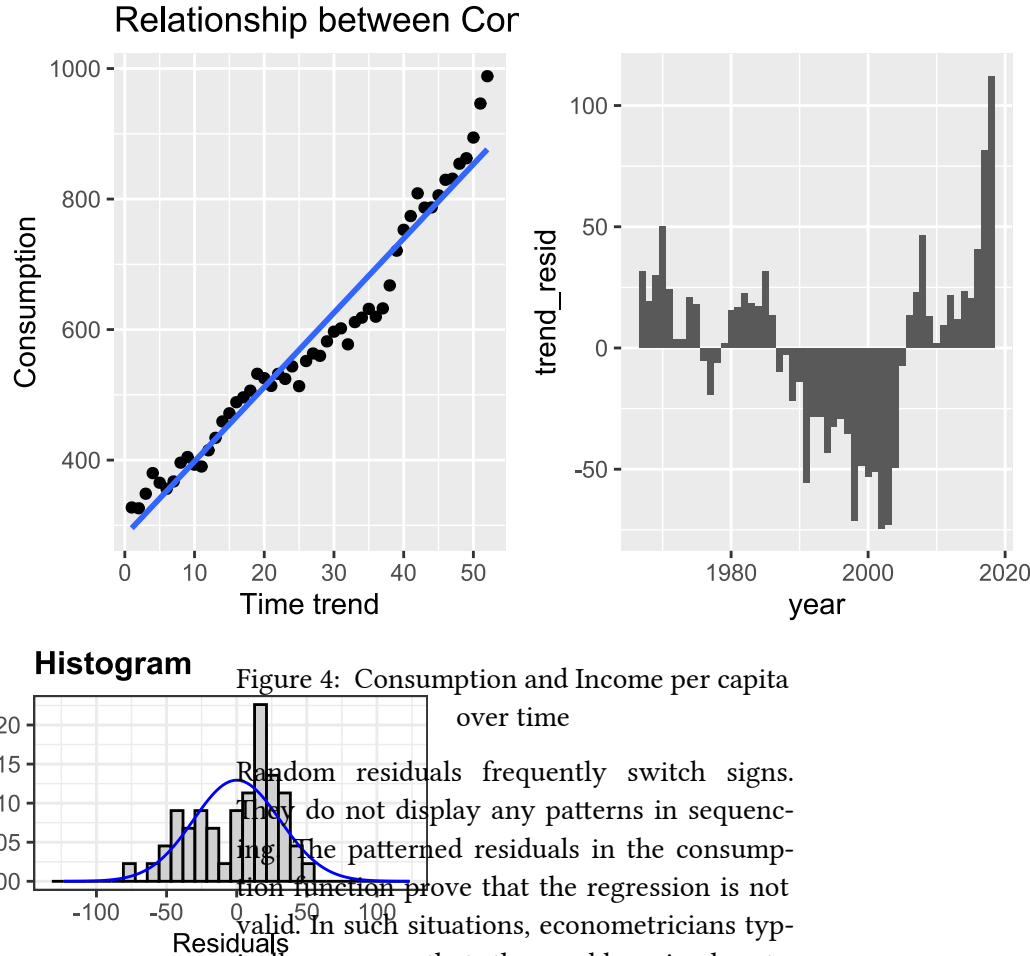


Figure 4: Consumption and Income per capita over time

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.

0.2 2. What to do when errors are not independent?

Note that $e_t = y_t - \hat{\beta}x_t$ and $e_{t-1} = y_{t-1} - \hat{\beta}x_{t-1}$. A relationship between the two errors shows that what happens in period $t - 1$ has an effect on what happens in period t . In other words, we have a dynamic model instead of a static model.

0.3 STATIC VERSUS DYNAMIC MODELS

A model of the type $C_t = \alpha + \beta Y_t + e_t$ is called a static model. Events in time period t are isolated from those of period $t - 1$ and $t + 1$. As opposed to this, if data from period $t - 1$ enters into the equation for period t , then we have dynamic model.

Suppose that we conduct tests and learn that e_t is correlated with e_{t-1} . Note that $e_{t-1} = C_{t-1} - \alpha - \beta Y_{t-1}$. It follows that C_{t-1} and Y_{t-1} have an effect on current consumption.

The main message of correlated residuals is that variables from one period ago have an effect on current period. **?@sec-static-dynamic** shows that the static model is not valid.

Thus in order to fix up the problem of correlated residuals, we need to build a dynamic model. The simplest dynamic extension of the Keynesian model is the following:

$$C_t = \alpha + \beta Y_t + \gamma C_{t-1} + \delta Y_{t-1} + e_t$$

We put in ALL the information we have about ALL the variables in the model at period $t - 1$. Similarly we could go back to period $t - 2$. As in Section 0.2, we could have e_t correlated with e_{t-1} and e_{t-2} . This would suggest a second order dynamic model. We could keep increasing the order of the model until we get to a point where the errors are not correlated.

0.3.a General-to-Simple vs. Simple-to-General Modelling:

In fixing problems detected by tests, there are two strategies which are commonly used. The older and more traditional strategy is called the bottom-up or simple-to-general strategy. In this strategy, we would start with the simplest model, which is a static Keynesian model. Then we note that the errors are correlated. Then we make the simplest dynamic extension. Next we

test the errors. If the problem of autocorrelation has been eliminated then we stop here. If there is still autocorrelation, it follows that a higher order dynamic model is needed. So we go to a second order model. We keep increasing the order (and complexity) of the model until we get to a point where the errors are not autocorrelated. This is one aspect of simple-to-general modelling.

The alternative to this which has recently been proposed, and shown to have superior properties, is general to simple modelling. In this method, we note that errors seem to be correlated upto the fourth order. So we start with a fourth order dynamic model. In general the goal is to start with the biggest model you might possibly need and simplify it down to a simpler model. This is called the top-down or general to simple strategy. We have illustrated this on the data set as well. The final model which emerges from our analysis is $C_t = \alpha + \beta Y_t + \gamma C_{t-1} + e_t$. However, this model suffers from both structural change and heteroskedasticity. This suggests that we should try a log transformation to get rid of the heteroskedasticity. Perhaps this will eliminate or reduce the structural change problem as well.

the simple Keynesian consumption function $C_t = \alpha + \beta Y_t + e_t$ failed several tests on Pakistani data – it has autocorrelated errors, and maybe structural stability.

How should we proceed? Experience with consumption functions shows that they very often have dynamic properties. This is indicated by the autocorrelation of the errors as well. Thus we try the simplest dynamic extension:

0.4 Dynamic Model

We start with **General to Specific** approach and include upto three lags of both consumption and income. There is no hard and fast rule

but it is suggested that number of regressors should not exceed number of observations. We have in total 52 observations we can select 3 or 4 lags of each (had it been quarterly data we would have selected 4 lags of each). Some studies suggest to have regressor should not exceed one third of the total number of observations.

Table 3: Dynamic Consumption Function

(1)	
(Intercept)	-8.881 (8.590)
	(8.590)
Y	0.618 (0.169)***
	(0.169)
lag(Y, 1)	-0.118 (0.296)
	(0.296)
lag(C, 1)	0.798 (0.156)***
	(0.156)
lag(Y, 2)	-0.544 (0.297)+
	(0.297)
lag(C, 2)	0.114 (0.189)
	(0.189)
lag(Y, 3)	0.105 (0.214)
	(0.214)
lag(C, 3)	0.025 (0.159)
	(0.159)
Num.Obs.	49
R2	0.993
R2 Adj.	0.992
AIC	417.0
BIC	434.0
Log.Lik.	-199.494
F	817.725
RMSE	14.19

Results in Table 3 indicate that Y and first lag of C are highly significant and second lag of Y is significant at 10%. So by standard statistical

procedure, one should drop all variables and re-run the regression with significant variables as follows:

```

Rows: 50
Columns: 11
$ year      <dbl> 1969, 1970, 1971,
1972, 1973, 1974, 1975, 1976, 1977,
1978...
$ C         <dbl> 348.4950, 380.0568,
365.3227, 356.0221, 367.2798,
396.0909...
$ Y         <dbl> 433.0744, 469.4377,
459.0634, 450.3759, 469.0675,
472.1552...
$ lnc       <dbl> 5.853624, 5.940320,
5.900781, 5.874993, 5.906124,
5.981644...
$ lny       <dbl> 6.070910, 6.151536,
6.129189, 6.110083, 6.150747,
6.157308...
$ c_y       <dbl> 0.9642087,
0.9656647, 0.9627345, 0.9615243,
0.9602288, 0.9...
$ cdividy   <dbl> 0.8047000,
0.8096000, 0.7958000, 0.7905000,
0.7830001, 0.8...
$ Trend     <int> 3, 4, 5, 6, 7,
8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 1...
$ resid     <dbl> 12.763683, 15.764683, 9.178845,
6.701742, 3.278557, 29.664...
$ trend_resid <dbl> 30.063569,
50.242900, 24.126271, 3.443237,
3.318460, 20.74...
$ g2slres   <dbl> 12.222517,
12.470682, -20.045428, 4.372311,
9.566801, 22.8...

```

Table 4: Dynamic Consum

(Intercept)	-7.094
Y	0.527
lag(C, 1)	0.896
lag(Y, 2)	-0.438
Num.Obs.	5
R2	0.9
R2 Adj.	0.9
AIC	41
BIC	42
Log.Lik.	-204.854
F	2082.114
RMSE	14.56

Oh great now we have all the variables are highly significant in Table 4 and if residuals satisfy all standard assumptions, it will be perfect fit. So lets have a look at the residuals behavior.

Lets have first look at residuals plot over time.

Residuals of Dynamic Consumption Function

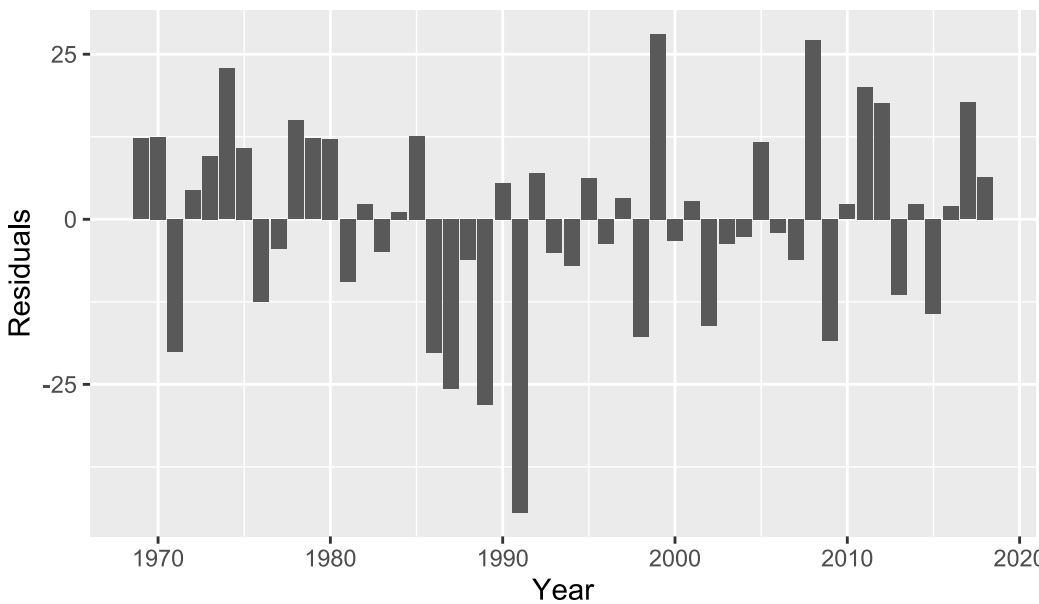
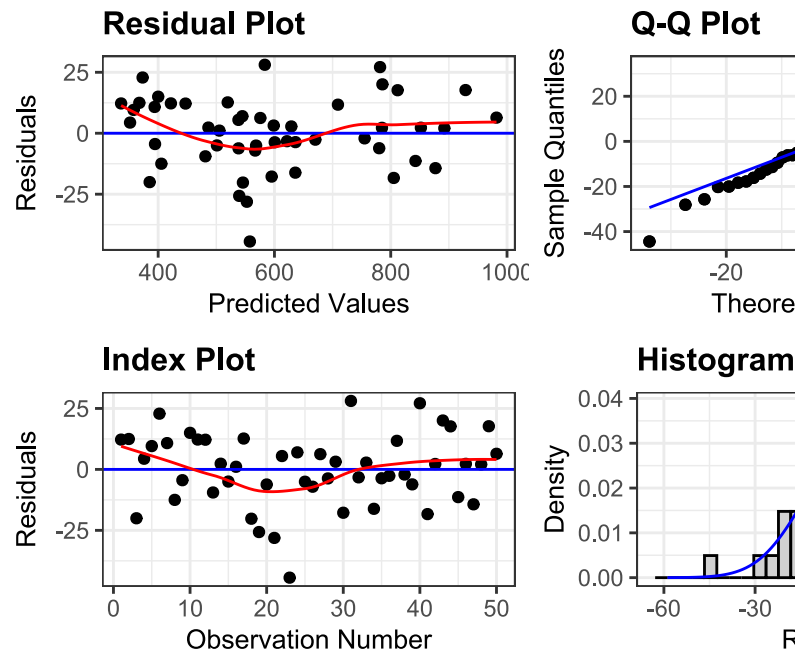


Figure 5: Residuals of Dynamic Consumption Function

From Figure 5, it seems residuals are behaving randomly. Lets apply some more tests before I finalize the model.



This seems quite a good job. But what about **theoretical interpretation**.

0.5 Multicollinearity and economic rationale

Now if we interpret and explain our model, it will be fine to say that consumption is function of current income, but little difficult to justify **2 years** lagged income effect on consumption while no effect of **One year lagged income**. As first two coefficient makes sense and we shall also learn about these later on as well. But instead of first lag of income, what does explain second lag to be in the model either needs a very solid reason or needs econometric model to be explored little more.

As in time series data there is very high correlation of variables with their own lagged values. So lets have a look at the correlation matrix and observe how highly they are correlated with each other.

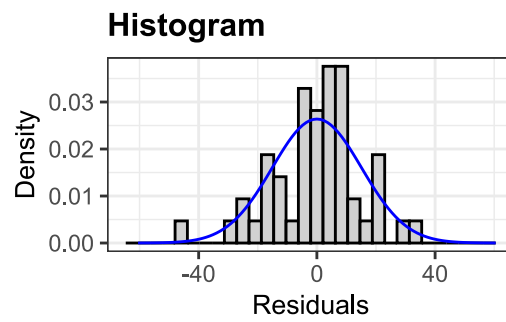
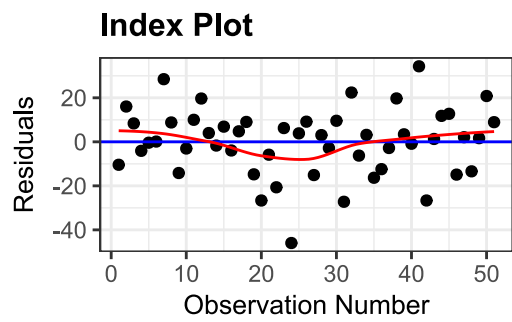
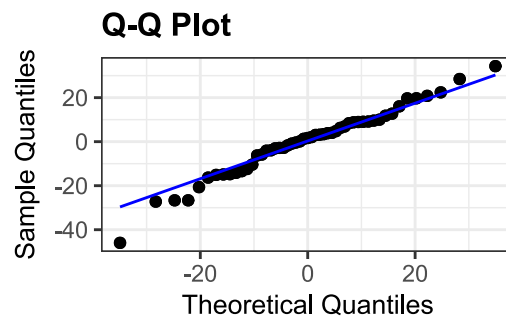
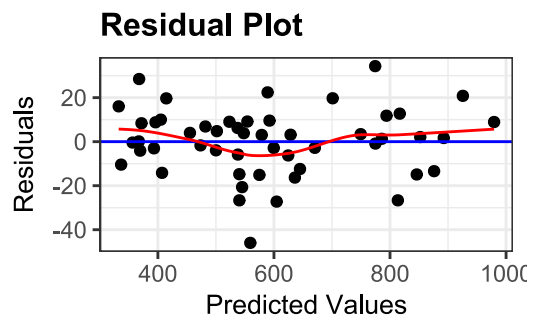
	C	Y	lagY1	lagY2	lagY3
lagC1	lagC2	lagC3			
C	1.000	0.983	0.979	0.973	0.968
	0.994	0.987	0.983		
Y	0.983	1.000	0.998	0.994	0.989
	0.982	0.980	0.979		
lagY1	0.979	0.998	1.000	0.998	0.994
	0.982	0.981	0.979		
lagY2	0.973	0.994	0.998	1.000	0.998
	0.979	0.982	0.981		
lagY3	0.968	0.989	0.994	0.998	1.000
	0.974	0.980	0.982		
lagC1	0.994	0.982	0.982	0.979	0.974
	1.000	0.994	0.988		
lagC2	0.987	0.980	0.981	0.982	0.980
	0.994	1.000	0.994		
lagC3	0.983	0.979	0.979	0.981	0.982
	0.988	0.994	1.000		

This correlation plot indicated very correlation between variables and their lags which implies these variables have very high multicollinearity and indicating they are almost per-

fect substitutes. Economic rationale suggests that first lag of income should be more relevant in explaining current consumption than that of second lag of income. Correlation between lag(Y,1) and lag(Y,2) is 0.998, therefore, lets use first lag of Y in the model.

	(1)
(Intercept)	-7.375 (7.985)
	(7.985)
Y	0.727 (0.156)***
	(0.156)
lag(C, 1)	0.903 (0.074)***
	(0.074)
lag(Y, 1)	-0.639 (0.161)***
	(0.161)
Num.Obs.	51
R2	0.992
R2 Adj.	0.992
AIC	430.8
BIC	440.4
Log.Lik.	-210.392
F	2067.854
RMSE	14.98

Oh that makes sense and current consumption is explained by current income, past year consumption and past year income and all other variables are insignificant. This model pass all diagnostic tests as well.



0.6 Summary and conclusion