Chapter 7: Case 5 – Consumer Credit Counseling

1. Analyse the significance of the variables in Dorothy's regression model. Develop a regression model (be sure to include additive dummy variables for the seasonal component, if necessary), and use it to forecast the number of new clients for the first three months of 1993. Compare your forecasts with the actual observations.

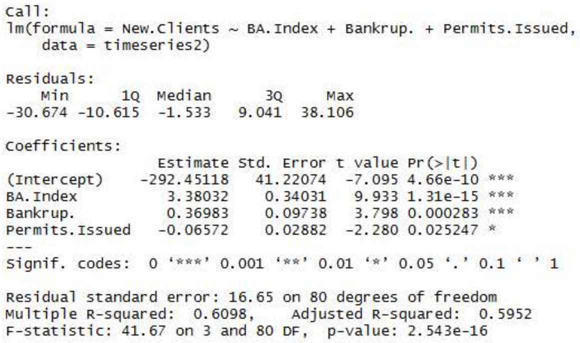
Firstly, we want to analyse Dorothy's **non-seasonal regression model** to determine its significance. The model has the following form:

$$Clients(\hat{Y}) = \beta_0 + \beta_1 Stamps + \beta_2 Index + \beta_3 Bankruptcies + \beta_4 Permits$$

Running the regression, we get to values for β_0, β_1, β_2, β_3 and β_4 of -291.02, -0.0016, 3.69, 0.47 and -0.0847 respectively. By looking at the analysis (see attachment) we find that the p-value for Stamps (0.538) is greater than the level of significance ($\alpha = 0.05$), which tells us that this independent variable is insignificant. This is because in this test, the p-value for each variable tells us the interval of confidence for which the variable's coefficient is equal to zero. If the p-value is more than 0.05 we can assume that we cannot reject the hypothesis that the coefficient is zero. Therefore, the variable is insignificant, because it does not contribute to the model. Thus, the former is omitted from the latter.

After the insignificant variable is removed from the model and the regression is run again, we obtain the multiple regression (also in Figure 1) equation of:

$$Clients(\hat{Y}) = \text{-}292.45 + 3.38 * Index + 0.369 * Bankruptcies\text{-}0.065 * Permits$$



```
Call:
lm(formula = New.Clients ~ BA.Index + Bankrup. + Permits.Issued,
    data = timeseries2)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.674 -10.615  -1.533   9.041  38.106 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -292.45118   41.22074  -7.095 4.66e-10 ***
BA.Index         3.38032    0.34031   9.933 1.31e-15 ***
Bankrup.         0.36983    0.09738   3.798 0.000283 ***
Permits.Issued  -0.06572    0.02882  -2.280 0.025247 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.65 on 80 degrees of freedom
Multiple R-squared:  0.6098,    Adjusted R-squared:  0.5952 
F-statistic: 41.67 on 3 and 80 DF,  p-value: 2.543e-16
```

Figure 1: Regression model without foodstamps in R.

It is implied by the p-value of the predictors, which is lesser than the level of significance ($\alpha = 0.05$), that the independent variables are significant and therefore that the non-seasonal model with three variables is the most appropriate one.

To check the Durbin-Watson test the following hypotheses are being tested:

H_0= There is no autocorrelation

H_1= There is positive autocorrelation present

We conduct the test by testing for:

If $D < D_L$, reject H_0

If $D > D_U$, do not reject H_0

If $D_L \leq D \leq D_U$, the test is inconclusive

From the attachment we obtain: $D = 1.605$. The critical values for α = 0.05, n=84 and k=3 are D_L=1.56 and D_U=1.72.

As $D_L \leq D \leq D_U$ we conclude that the test is inconclusive.

By looking at the autocorrelation of the residuals (Figure 2) we can clearly see that there is a significant autocorrelation with a spike at lag 4.

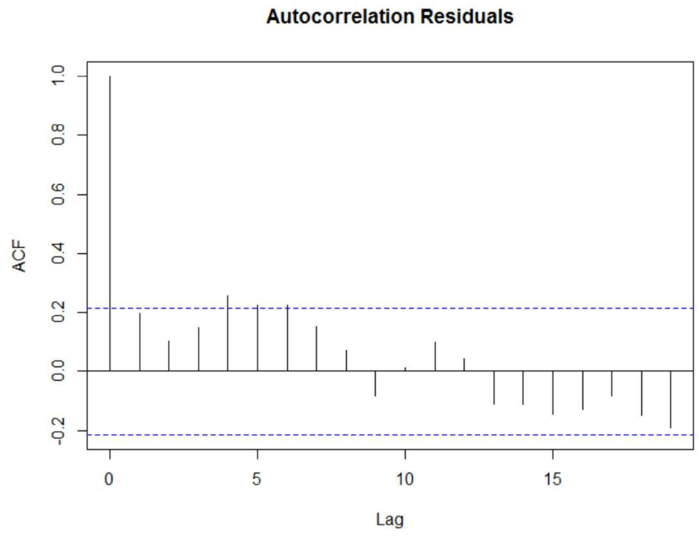


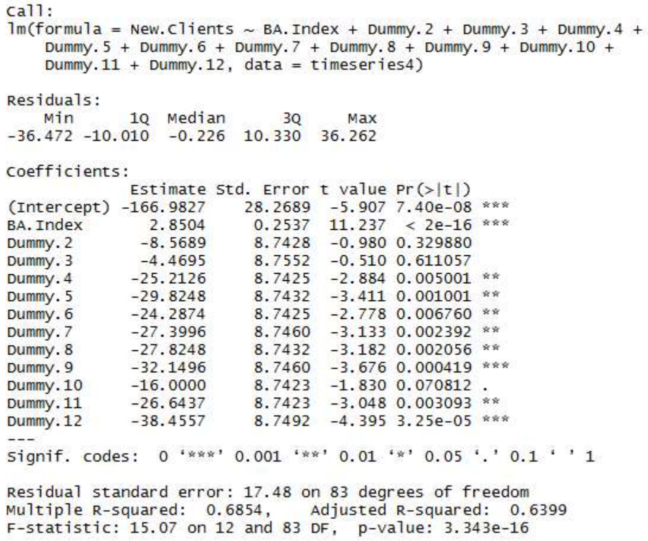
Figure 2: Autocorrelation of Residuals. Plotted with R

To develop the **seasonal regression model,** we included dummy variables for the monthly seasonal component and arrive to the following model:

$$Clients(\hat{Y}) = \beta_0 + \beta_1 Stamps + \beta_2 Index + \beta_3 Bankruptices + \beta_4 Permits + \beta_5 S2 + \beta_6 S3 + \beta_7 S4 + \beta_8 S5 + \beta_9 S6 + \beta_{10} S7 + \beta_{11} S8 + \beta_{12} S9 + \beta_{13} S10 + \beta_{14} S11 + \beta_{15} S12$$

By running and analysing the regression, we find that the p-values for Stamps (0.363), Bankruptcies (0.157) and Permits (0.091) are larger than the level of significance (α = 0.05). Thus these variables are omitted from the model.

By running the regression with the reduced amount of variables we arrive to the following seasonal regression model (Figure 3):



```
Call:
lm(formula = New.Clients ~ BA.Index + Dummy.2 + Dummy.3 + Dummy.4 +
    Dummy.5 + Dummy.6 + Dummy.7 + Dummy.8 + Dummy.9 + Dummy.10 +
    Dummy.11 + Dummy.12, data = timeseries4)

Residuals:
    Min      1Q  Median      3Q     Max
-36.472 -10.010  -0.226  10.330  36.262

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -166.9827    28.2689  -5.907 7.40e-08 ***
BA.Index       2.8504     0.2537  11.237  < 2e-16 ***
Dummy.2       -8.5689     8.7428  -0.980 0.329880
Dummy.3       -4.4695     8.7552  -0.510 0.611057
Dummy.4      -25.2126     8.7425  -2.884 0.005001 **
Dummy.5      -29.8248     8.7432  -3.411 0.001001 **
Dummy.6      -24.2874     8.7425  -2.778 0.006760 **
Dummy.7      -27.3996     8.7460  -3.133 0.002392 **
Dummy.8      -27.8248     8.7432  -3.182 0.002056 **
Dummy.9      -32.1496     8.7460  -3.676 0.000419 ***
Dummy.10     -16.0000     8.7423  -1.830 0.070812 .
Dummy.11     -26.6437     8.7423  -3.048 0.003093 **
Dummy.12     -38.4557     8.7492  -4.395 3.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.48 on 83 degrees of freedom
Multiple R-squared:  0.6854,    Adjusted R-squared:  0.6399
F-statistic: 15.07 on 12 and 83 DF,  p-value: 3.343e-16
```

Figure 3: Seasonal Regression Model with R.

The p-values imply that the independent variables are significant. Thus, the variables are appropriate to the model.

To check the Durbin-Watson test the following hypotheses are being tested:

H_0= There is no autocorrelation

H_1= There is positive autocorrelation present

We conduct the test by testing for:

If $D < D_L$, reject H_0

If $D > D_U$, do not reject H_0

If $D_L \leq D \leq D_U$, the test is inconclusive

From the attachment we obtain: $D = 1.337$. The critical values for $\alpha = 0.05$, n=95 and k=12 are D_L=1.39 and D_U=1.96.

As $D < D_L$ we reject H_0 and it can be concluded that there is positive autocorrelation amongst the residuals at a 5% level.

Figure 4 shows this autocorrelation of the residuals, where we can clearly see that there is significant autocorrelation with spikes from lag 1 through lag 6.

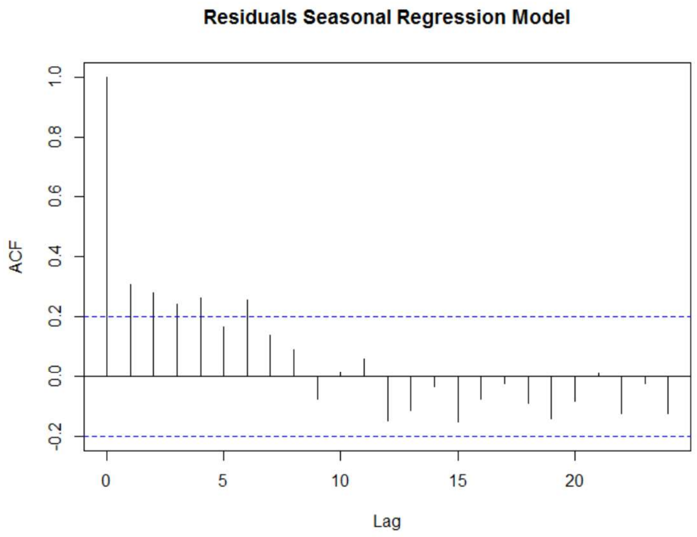


Figure 4: Autocorrelation of Residuals of seasonal model. Plotted with R.

The forecasted and actual values using this model for the first three months of 1993 can be found in Table 1:

Month	January 1993	February 1993	March 1993
Number of Clients forecasted	189.31	180.75	199.10
Actual number of clients	152	151	199

Table 1: Forecasted value for first three months of 1993

2. Develop an autoregressive model, and generate forecasts for the first three months of 1993.Which model (multiple regression or autoregression) do you feel is the better candidate for generating forecasts for the rest of 1993? Write Dorothy a memo that provides her with the information she has requested concerning the problem of serial correlation. Include an analysis of the results of your efforts to develop an appropriate model to forecast the number of new clients for the remainder of 1993.

We developed an autoregressive model with lag 1 (see attachment R Code or Excel) and arrive to the following regression equation:

$$Clients(\hat{Y}) = 61.4 + 0.487 * Lag\ Clients$$

As the p-value for the predictor variable is smaller than the level of significance (α = 0.05), we find the independent variable significant. Hence, the variable is appropriate to the model.

Figure 5 shows us the autocorrelation for the residuals, where we can see that there is significant autocorrelation with a spike at lag 12 present.

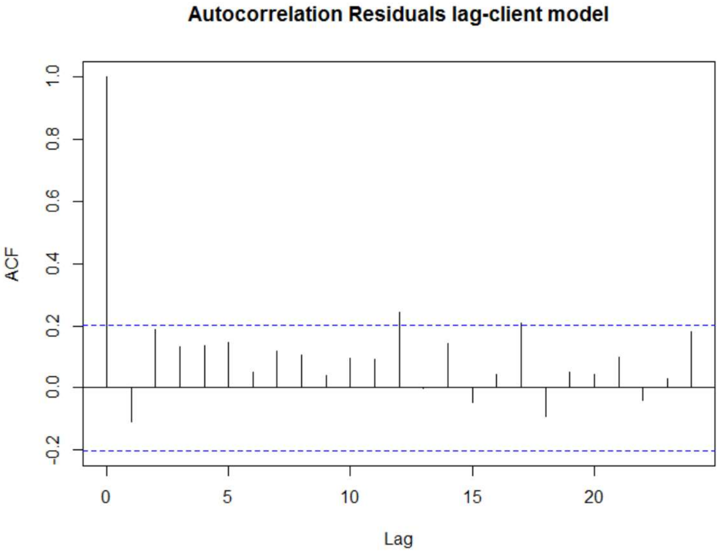


Figure 5: Autocorrelation of Residuals for the autoregressive model. Plotted with R.

By comparing the developed model, we can clearly see a difference in the values of R^2. In the autoregressive model 24% of the variability in *Clients* is explained by the *Lag Clients*, while the non-seasonal regression model and seasonal regression model have values for R^2 of 61% and 68.5%, respectively.

Therefore, the most appropriate model to use, among those we have analysed, is the seasonal regression model with the predictor variables Index and seasonal dummy variables.

Memo to Dorothy:

Dear Dorothy,

You were right to suspect serial correlation. First, we adapted your model to account for the monthly seasonal component present. Still the serious serial correlation poses a problem with the way we want to design our models. Further, we developed a regression model that considers the seasonality and can explain up to 68.5% of the variability in the number of new clients.

If we find the time, we suggest you take a look the Box-Jenkins (ARIMA) methodology, as this could be useful for our needs.