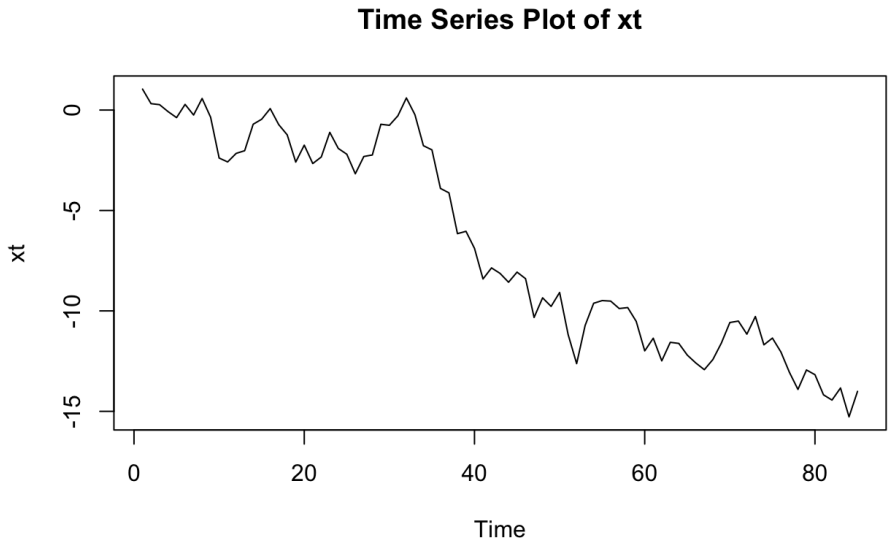
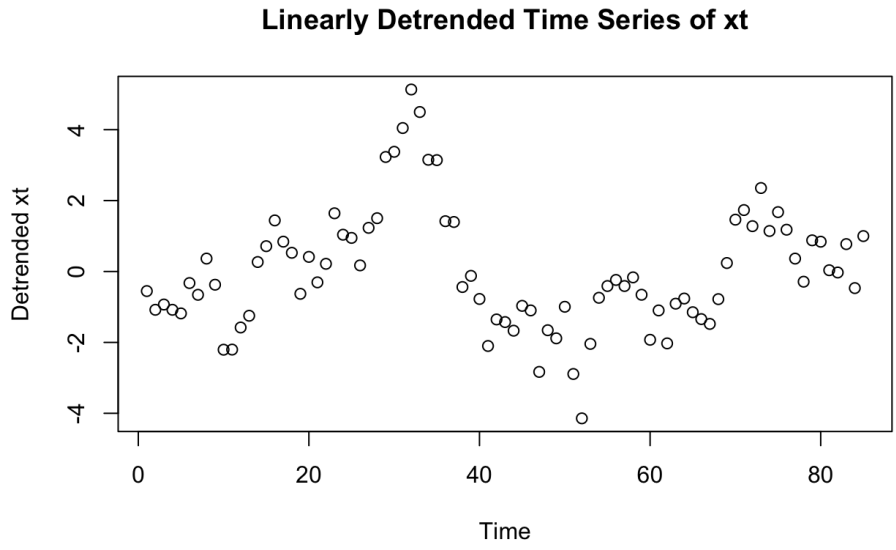
STAT 5350 – Assignment #2
Name: Mohammed Raza Syed
Dept: SEAS (Data Science)
PennID: 37486255

Q1)

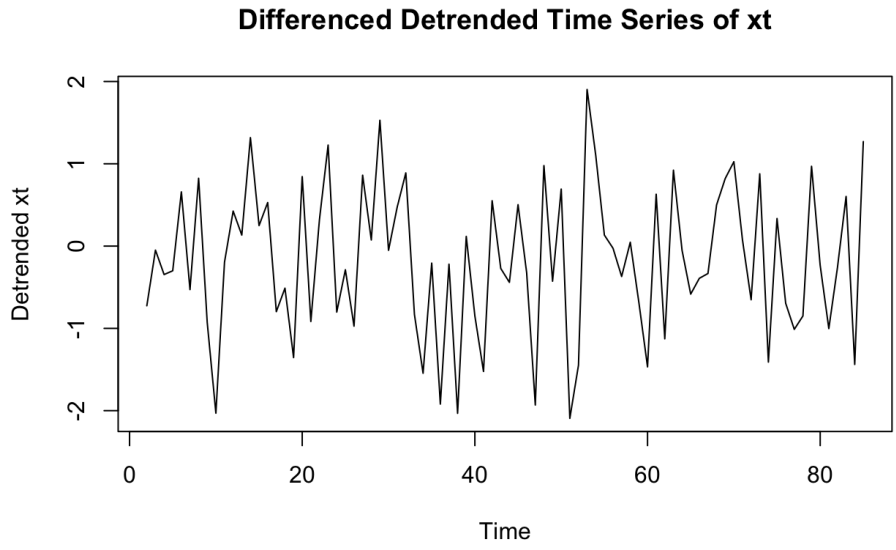
Initial X_t Plot:



Detrend using Linear method:



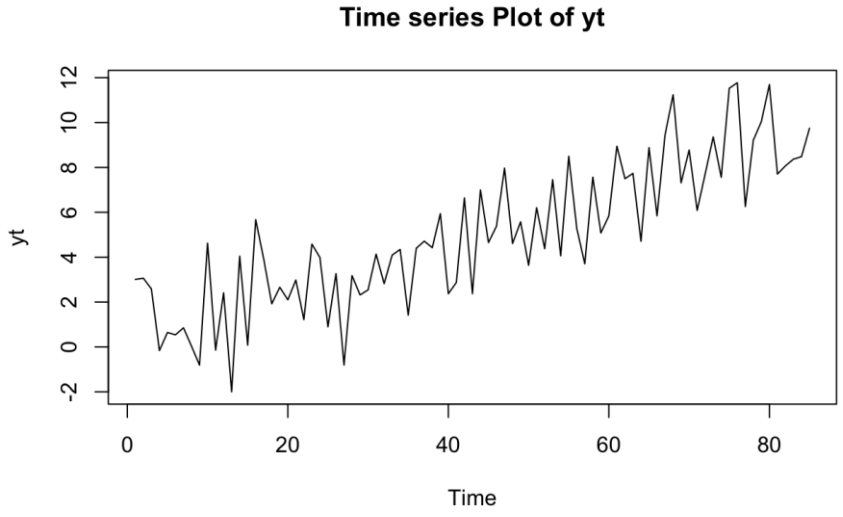
Detrend using Difference method:



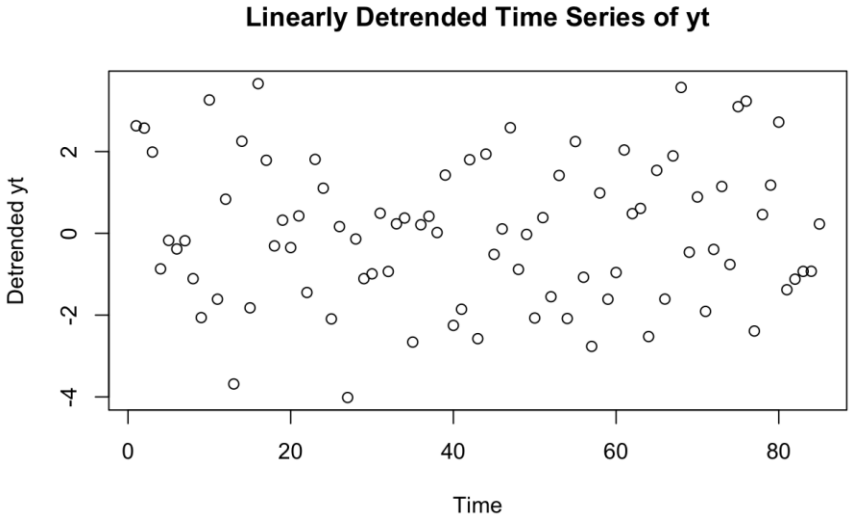
The time series X_t exhibited a clear downward trend, indicating nonstationarity. On trying both linear detrending and differencing to remove the trend I found that linear detrending worked to some extent, while differencing was more effective in removing all nonstationary patterns. Therefore, I chose differencing as the more suitable method for detrending Xt.

Q2

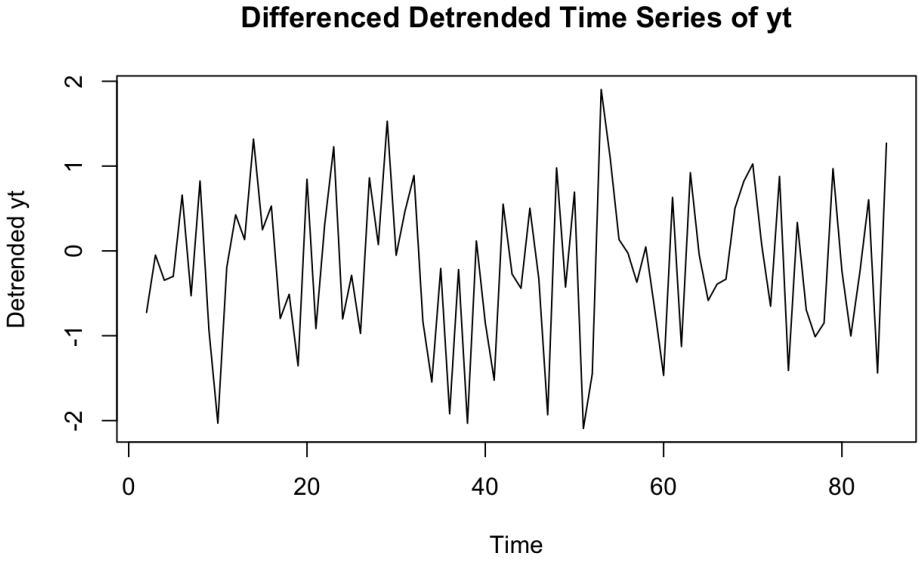
Initial Y_t Plot:



Detrend using Linear method:



Detrend using Difference method:



For Y_t, I explored both linear detrending and differencing. Unlike X_t, Y_t showed no clear trend, and both methods seemed equally effective in removing any potential nonstationarity. Since no significant trend was present, either method could be justified, but I opted for differencing due to its general applicability.

Q3

```
> raw_correlation <- cor(d$xt, d$yt)
> print(raw_correlation)
[1] -0.7966989
>
> detrended_correlation <- cor(diff_xt, diff(d$yt))
> print(detrended_correlation)
[1] 0.00519516
```

The initial correlation between X_t and Y_t was strong at -0.79, suggesting a potential relationship. However, after detrending both series, the correlation dropped to near zero (0.005), indicating that the original correlation was likely spurious. The initial correlation appears to have been driven by shared trends in the nonstationary series, rather than a true underlying relationship. Once the trends were removed, no meaningful correlation remained, suggesting the two processes are independent.

Q4

```
dynlm(formula = occRate ~ L(occRate, 1) + L(occRate, 4) + quarter +
    L(unRate, 1) + L(unRate, 2))

Residuals:
    Min      1Q  Median      3Q     Max
-11.7023  -1.4658   0.2759   1.4534   5.7736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.83020    6.18905   3.366 0.000986 ***
L(occRate, 1)  0.44251    0.07326   6.040 1.32e-08 ***
L(occRate, 4)  0.32723    0.06801   4.811 3.83e-06 ***
quarter2      -7.64273    1.08495  -7.044 7.72e-11 ***
quarter3      -1.46903    0.68852  -2.134 0.034618 *
quarter4      -6.94353    1.00353  -6.919 1.50e-10 ***
L(unRate, 1)  -4.09161    0.98427  -4.157 5.59e-05 ***
L(unRate, 2)   4.18144    0.95345   4.386 2.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 2.859 on 140 degrees of freedom
Multiple R-squared: 0.7349, Adjusted R-squared: 0.7217
F-statistic: 55.45 on 7 and 140 DF, p-value: < 2.2e-16

The coefficient for quarter2 is **-7.64**, meaning that, all else equal, the occupancy rate in the second quarter is about 7.64 units lower than in the first quarter. This likely reflects seasonal trends, where factors like reduced tourism, changes in business travel, or general financial concerns may cause a dip in hotel occupancy during the second quarter.

Q5

Summary of Regr2:

```
Call:
dynlm(formula = occRate ~ L(occRate, 1) + L(occRate, 4) + quarter +
    L(unRate, 1) + L(unRate, 2) + L(pctChgGNP, 1) + L(pctChgGNP,
    2))

Residuals:
    Min      1Q  Median      3Q     Max
-11.4713  -1.3764   0.3079   1.4286   5.4829

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     19.64140    6.24396   3.146  0.00203 **
L(occRate, 1)    0.44325    0.07340   6.039 1.36e-08 ***
L(occRate, 4)    0.33285    0.06816   4.883 2.84e-06 ***
quarter2        -7.52045    1.08886  -6.907 1.66e-10 ***
quarter3        -1.55259    0.69436  -2.236  0.02696 *
quarter4        -6.90026    1.00272  -6.882 1.90e-10 ***
L(unRate, 1)    -3.23461    1.21528  -2.662  0.00870 **
L(unRate, 2)     3.32051    1.18593   2.800  0.00584 **
L(pctChgGNP, 1) 66.26557   43.68360   1.517  0.13157
L(pctChgGNP, 2) -7.05974   43.29093  -0.163  0.87070
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 2.855 on 138 degrees of freedom
Multiple R-squared: 0.7395, Adjusted R-squared: 0.7225
F-statistic: 43.52 on 9 and 138 DF, p-value: < 2.2e-16

First, the small difference between the Multiple R-Squared and Adjusted R-Squared values indicates that the GNP predictors did not meaningfully enhance the model's predictive power.

Second, the ANOVA test results show an F-statistic of 1.1981 and a corresponding p-value of 0.3049. Since the p-value is greater than 0.05, the difference between the models is not statistically significant.

Analysis of Variance Table

Model 1: occRate ~ L(occRate, 1) + L(occRate, 4) + quarter + L(unRate,1) + L(unRate, 2)
Model 2: occRate ~ L(occRate, 1) + L(occRate, 4) + quarter + L(unRate,1) + L(unRate, 2) + L(pctChgGNP, 1) + L(pctChgGNP, 2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	140	1144.5				
2	138	1124.9	2	19.532	1.1981	0.3049

Hence after interpreting these two points it suggests that adding the GNP predictors does not significantly improve the model's ability to predict occupancy rates.

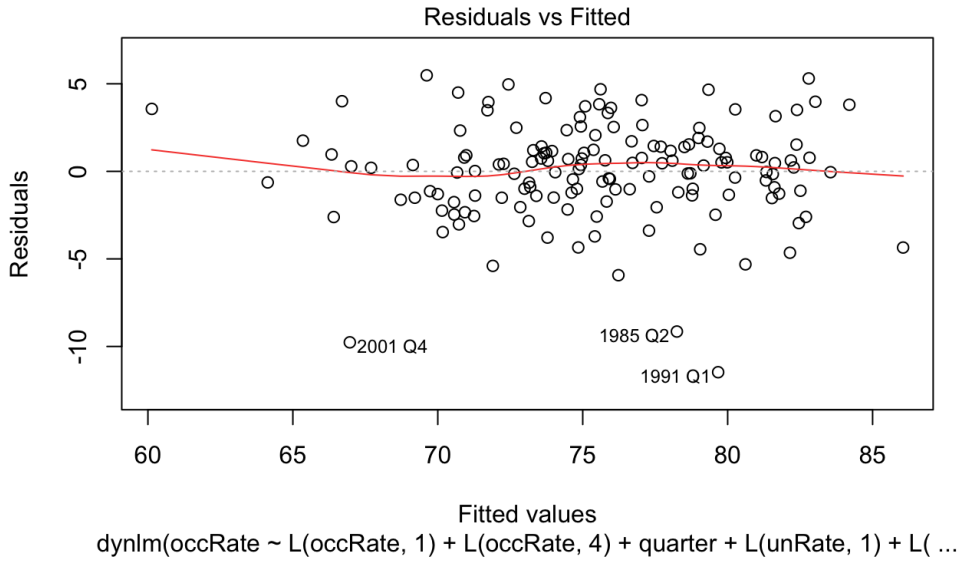
Q6)

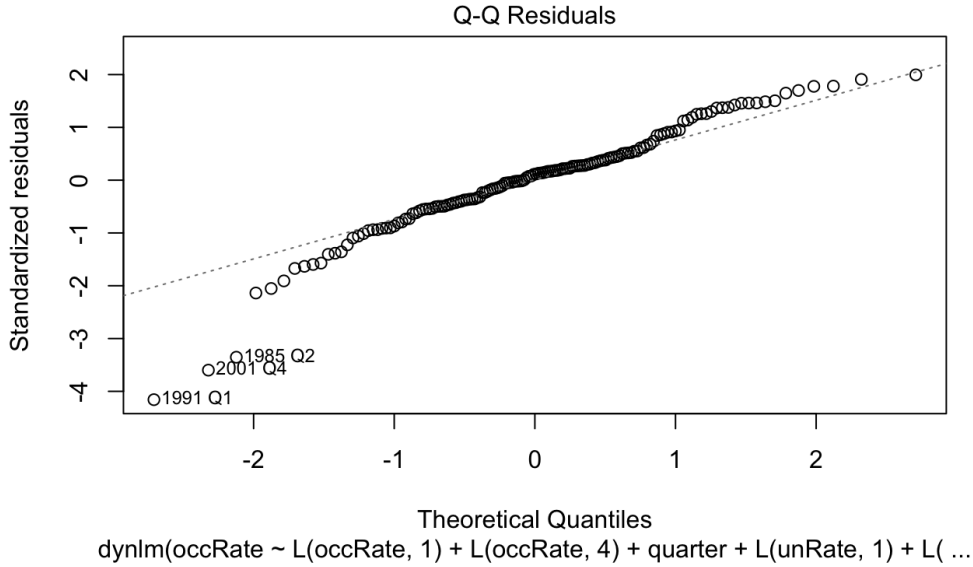
After calculating the Variance Inflation Factor values of the GNP predictors to understand the collinearity, it was found the values are really small which is less than 5 or 10 which is required to confirm the significance of collinearity. But since it was found to be in the small which is less than 5. Hence we cannot attribute the lack of significance to collinearity.

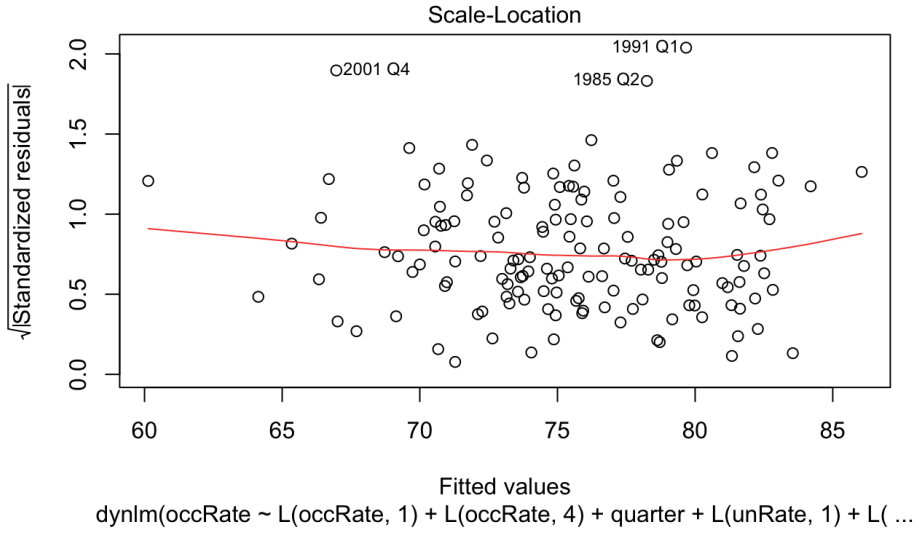
	GVIF	Df	GVIF^(1/(2*Df))
L(occRate, 1)	2.894693	1	1.701380
L(occRate, 4)	2.508258	1	1.583748

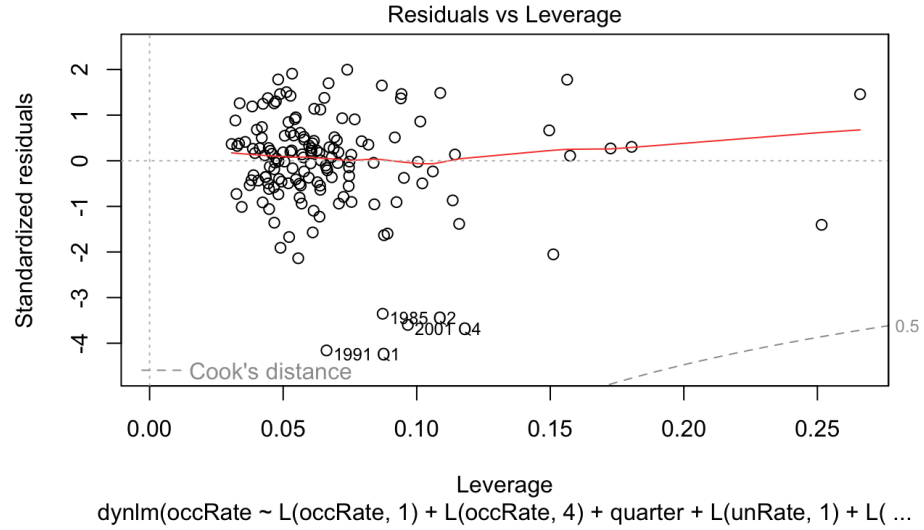
quarter	3.680063	3	1.242537
L(unRate, 1)	72.135199	1	8.493244
L(unRate, 2)	70.086787	1	8.371785
L(pctChgGNP, 1)	1.723635	1	1.312873
L(pctChgGNP, 2)	1.696629	1	1.302547

Q7









After running the diagnostic plots, 2001 Q4, 1985 Q2, and 1991 Q1 stood out across several of them. In the Residuals vs Fitted plot, these points have large residuals, indicating that the model made significant prediction errors.

In the Normal Q-Q plot, they deviate from the expected normal distribution, confirming they are outliers.

The Scale-Location plot shows relatively consistent variance, and these points don't cause notable issues in terms of residual spread.

Finally, in the Residuals vs Leverage plot, while these points have moderate leverage, they do not exceed the Cook's distance threshold, meaning they are not highly influential.

Although 2001 Q4, 1985 Q2, and 1991 Q1 are clear outliers with large negative residuals, they are not highly influential and do not have undue influence on the model's coefficients.

Q8)

The residuals from regr_2 generally meet the conditions needed for making conclusions about the regression coefficients. The model follows a linear relationship, the residuals are normally distributed, they have consistent variance, and they are independent with no signs of autocorrelation, as confirmed by the Durbin-Watson test giving D-W Statistic around 1.97 which is close to 2 for confirming no signs of autocorrelation.

lag Autocorrelation D-W Statistic p-value
1 0.00435664 1.979168 0.826
Alternative hypothesis: rho != 0

Q10)

When we simplified the model by using the difference in unemployment (diff(unRate, 1)) instead of the lagged unemployment values, it didn't reduce the model's accuracy. The ANOVA test showed no significant difference between the original and simplified models, and the residuals actually improved slightly in the simplified version.

This means we can use the simplified model without losing predictive accuracy, making it a good, simpler alternative to the original.

```
Call:
dynlm(formula = occRate ~ L(occRate, 1) + L(occRate, 4) + quarter +
    diff(unRate, 1))

Residuals:
     Min       1Q   Median       3Q      Max
-11.0227  -1.6505  -0.0126   1.6783   5.7160

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.06143    4.37944   4.581 1.01e-05 ***
L(occRate, 1)    0.46420    0.06253   7.424 9.82e-12 ***
L(occRate, 4)    0.32281    0.06014   5.368 3.19e-07 ***
quarter2        -7.70935    1.00171  -7.696 2.22e-12 ***
quarter3        -1.49462    0.65001  -2.299    0.023 *
quarter4        -7.08732    0.92526  -7.660 2.71e-12 ***
diff(unRate, 1) -5.17057    0.85037  -6.080 1.06e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.712 on 141 degrees of freedom
Multiple R-squared:  0.7598,     Adjusted R-squared:  0.7495
F-statistic: 74.32 on 6 and 141 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Model 1: occRate ~ L(occRate, 1) + L(occRate, 4) + quarter + L(unRate,
    1) + L(unRate, 2) + L(pctChgGNP, 1) + L(pctChgGNP, 2)
Model 2: occRate ~ L(occRate, 1) + L(occRate, 4) + quarter + diff(unRate,
    1)
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1    138 1124.9
2    141 1037.3 -3    87.658
```