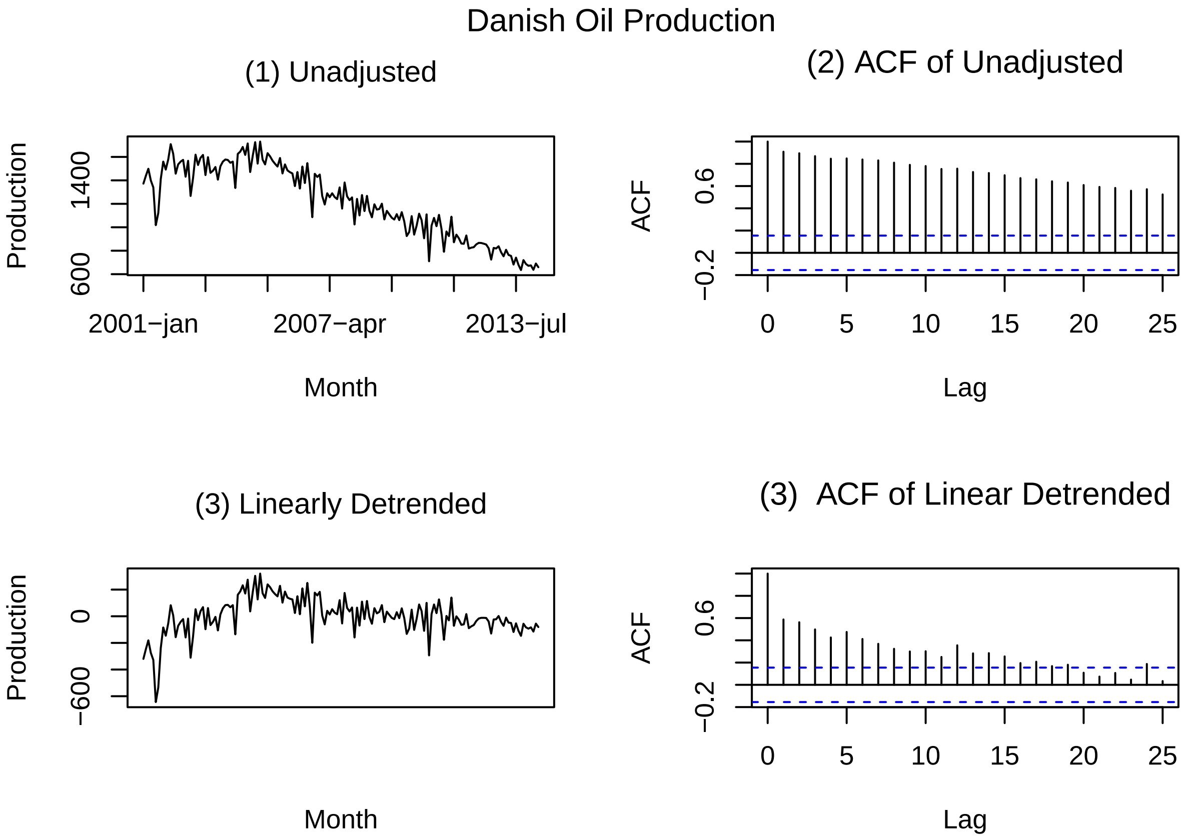
Problem Set 3

[redacted]

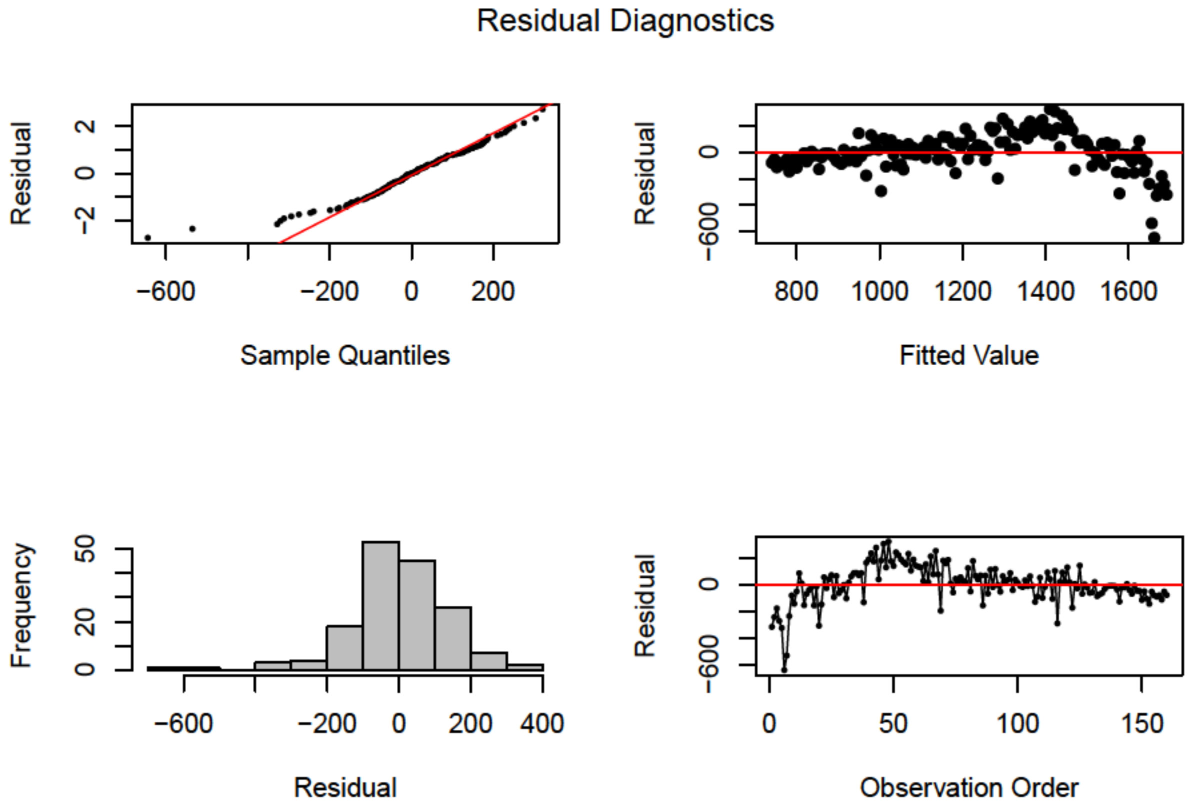
2/25/2022

2.26 Table B.22 contains data from the Danish Energy Agency on Danish crude oil production. Plot the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF. Interpret the graph.



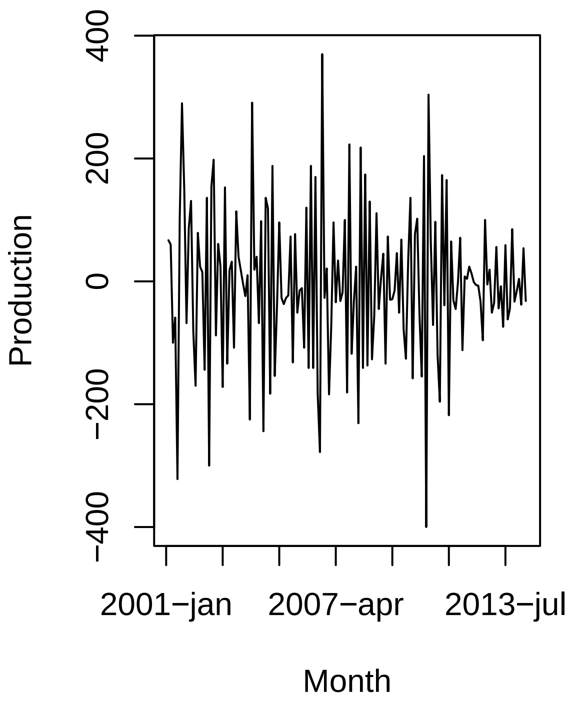
Let's take a look at the plot of the crude oil production from the Danish Energy Agency between the months of January, 2021 and April, 2014. As we can see from the first plot titled "Unadjusted Oil Production", there appears to be an overall downward trend in the time series regarding the production of oil in Denmark. In the mean time, the apparent fluctuations also indicate that certain seasonality embedded in the data. Since there's neither constant mean nor variance, the oil production during this time period is most likely non-stationary. Therefore, we have also attached the plot titled "Linearly Detrended Oil Production" (3). In comparison to the first plot, the linearly detrended has closed in the mean to 0. Notice that the downward sloping trend has been removed, which is one step closer to what we want to achieve. However, the fluctuations are still signs that there might be underlying seasonality.

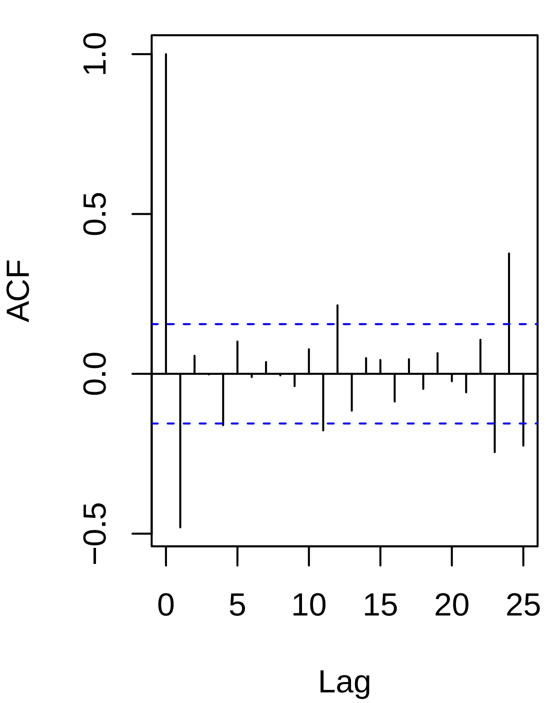
Now let's shift our attention to the ACF graphs on the right. It's clear according to the ACF of the unadjusted oil production (2) that the Danish oil production in any month is highly autocorrelated with the oil production in the past. In the ACF for the adjusted oil production (4), we can see that even though the data has been transformed in the way that productions in different periods are less autocorrelated to the past production, the ACF decreased too slowly, which means that the linear detrend is not a sufficient transformation on the time series we have.



By running the residual diagnostics, we can see from the Residual vs. Sample Quantiles plot on the top left that, while most of the observations' residuals are normally distributed, there are some significant deviations on the left side. The histogram that describes the spread of residuals confirms our previous observation that the residuals after removing the linear trend are relatively normally distributed, while also having a long tail on the left. There is no obvious pattern in the Residual vs. Fitted Value plot on the top right, which is an indication that the residuals are quite likely uncorrelated. Furthermore, there also does not appear to be any discernible pattern in the residuals with respect to time. Therefore, we can say that the linear detrend has not provided us with a stationary series.

2.27 Reconsider the Danish crude oil production data from Exercise 2.26. Plot the first difference of the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF for the differenced data. Interpret the graph. What impact did differencing have?

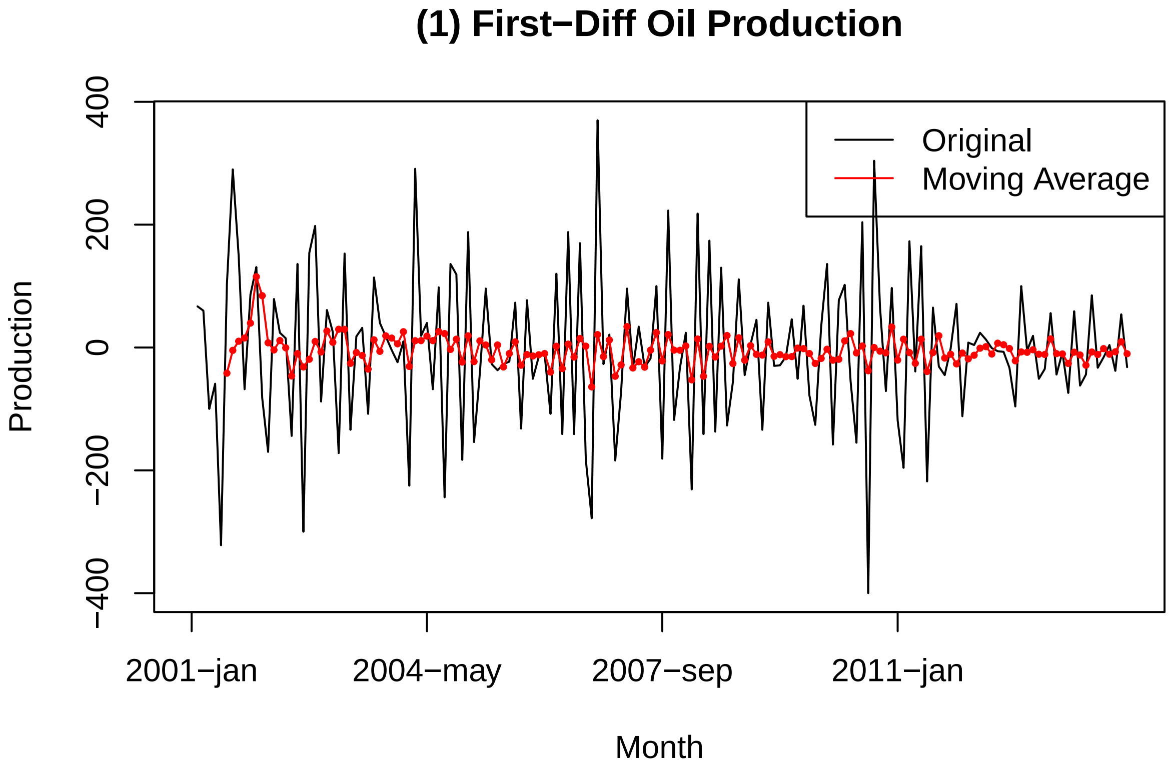

(1) First−Differenced Production

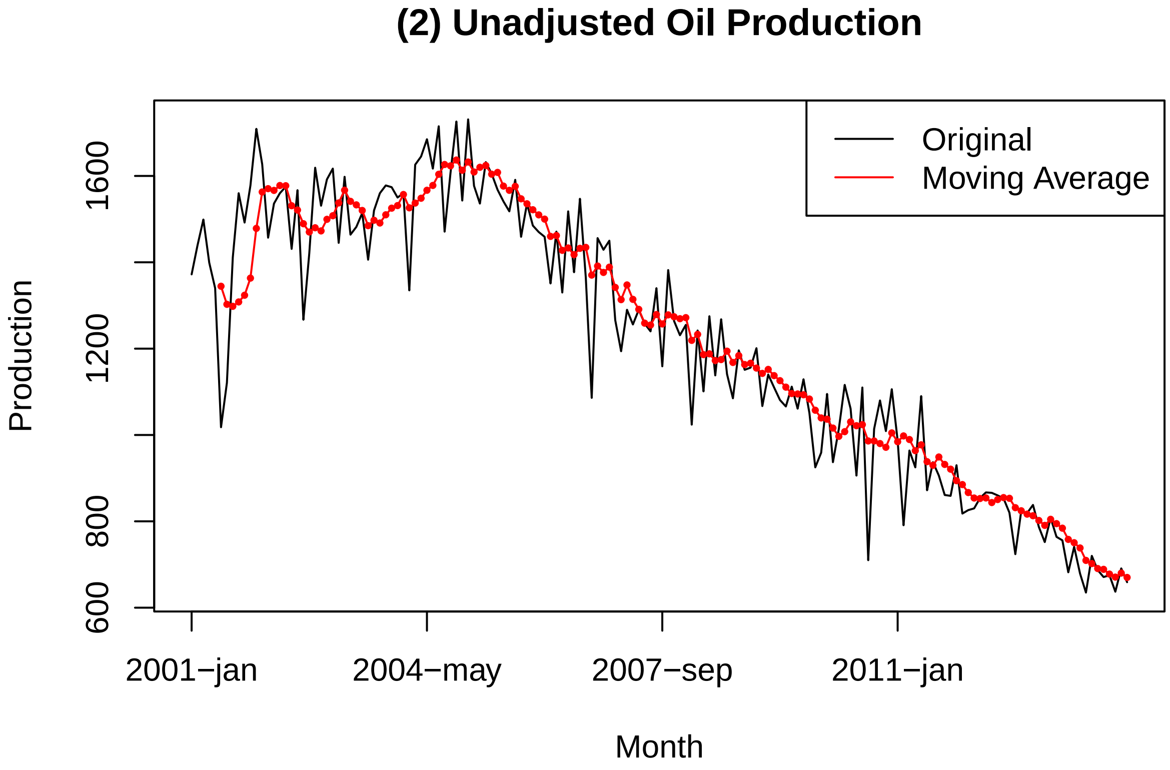

(2) First−Differenced ACF

With first-differencing, we can see clearly from the first plot that the data seems to have a closer to constant mean around 0 comparing to its original form, which suggests that this transformed data is more stationary than before. However, it is also true that its variance does not appear to be constant over time, with patterned fluctuations indicating seasonality.

The ACF for the first-difference adjusted production also presents to be more stationary in comparison to the original unadjusted data's ACF. However, we can still detect a seasonal pattern in the first-differenced autocorrelation plot: every multiple of 12-month lag, we see a spike in the autocorrelation. Therefore, even though first-differencing was helpful in removing a portion of the trend, the seasonal trend still remains to be removed in order for us to obtain a truly trend-stationary data to perform forecast.

2.28 Use a six-period moving average to smooth the first difference of the Danish crude oil production data that you computed in Exercise 2.27. Plot both the smoothed data and the original data on the same axes. What has the moving average done? Does the moving average look like a reasonable forecasting technique for the differenced data?

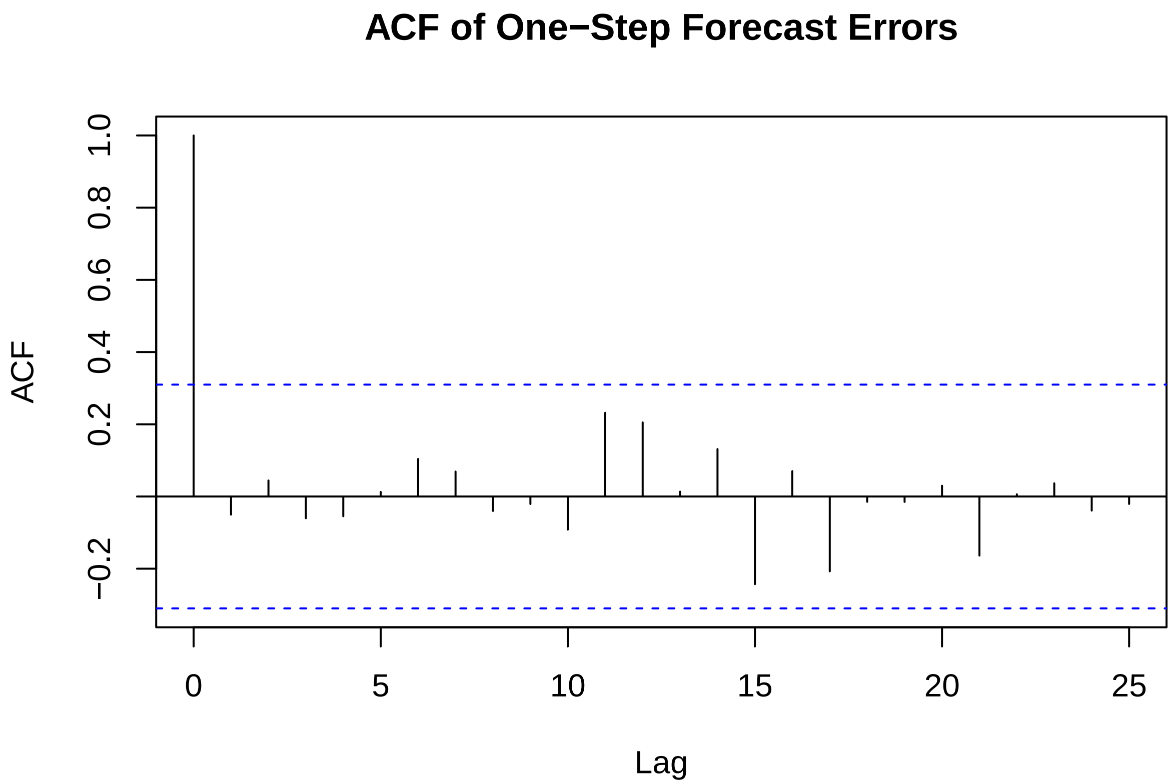

(1) First−Diff Oil Production


(2) Unadjusted Oil Production

By comparing the application of moving averages on the two different plots (first-differenced vs. original), we can clearly see that the moving average of the first-differenced data approaches a constant mean and has no drastic changes in variances over time, which signals stationality. Therefore, we can say that the moving average on first-differenced data has a smoothing effect on the data. On the other hand, while the moving average on the original data did diminish the variance in the time series over time, it did not facilitate with stablising to a constant mean, leaving it utterly unstationary.

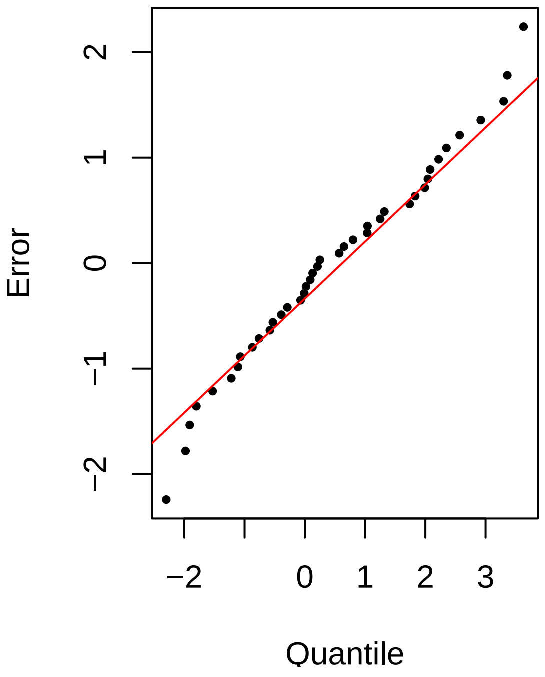
2.44 Table E2.1 contains 40 one-step-ahead forecast errors from a forecasting model

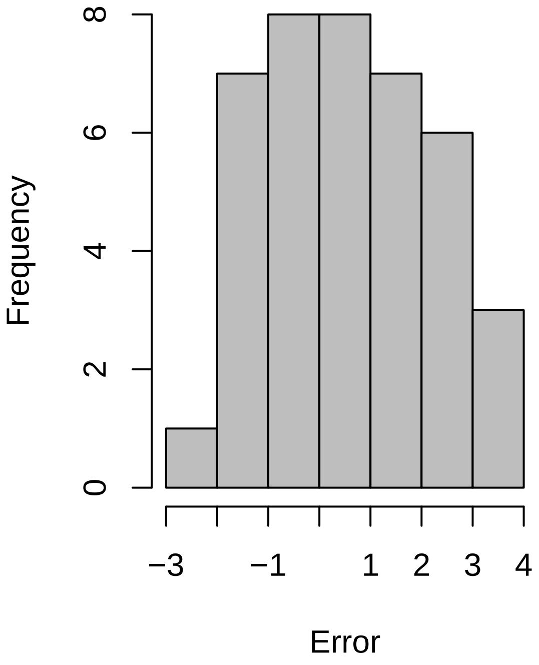
a. **Find the sample ACF of the forecast errors. Interpret the results.**



Please see above for the ACF for the one-step forecast errors provided in table E2.1. Recall that the one-step forecast error describes the difference between the forecasted value one period before and the actual value. With the quickly declined autocorrelation after the first lag, we can deduce with confidence that errors are at large not autocorrelated with the past values. Moreover, there does not appear to be any other outliers and no perceivable indication for unaccounted seasonality.

b. **Construct a normal probability plot of the forecast errors. Is there evidence to support a claim that the forecast errors are normally distributed?**





Observe the normal probability (histogram, on the right). It appears that, even though it has a large portion of errors distributed around the median quantile, there appear to be some outliers that affect the overall distribution of the errors. As we may observe from the quantiel-quantile plot on the left side, the errors on both ends as well as in the middle deviate relatively substantially from the quantile-quantile line, which suggests that the errors are not normally distributed.

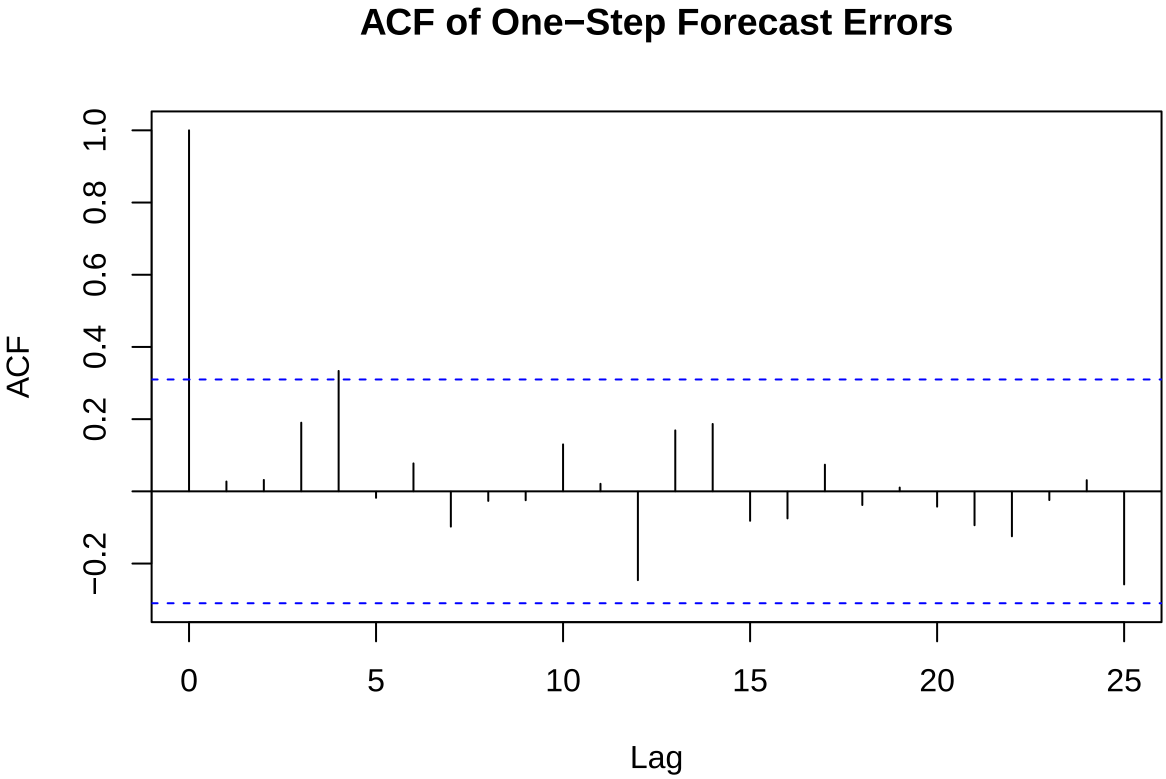
c. **Find the mean error, the mean square error, and the mean absolute deviation. Is it likely that the forecasting technique produces unbiased forecasts?**

Table 1: ME, MSE, & MAD

	Mean Error	Mean Squared Error	Mean Absolute Deviation
Value	0.52445	2.781211	1.34545

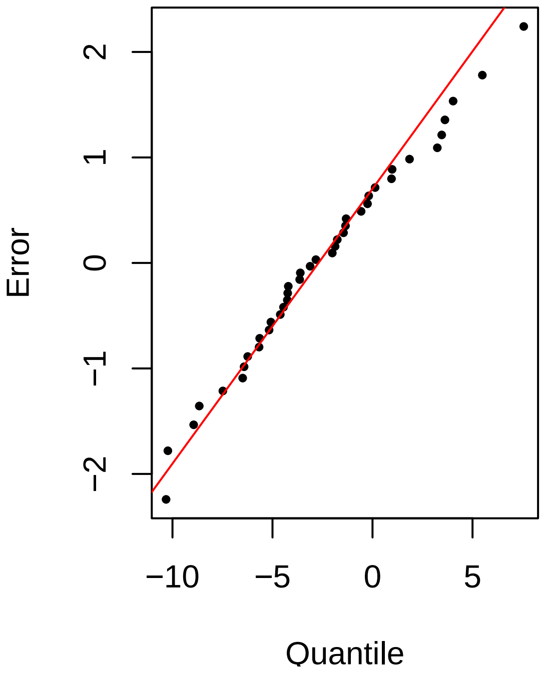
See Table 1 for Mean Error, Mean Squared Error, and Mean Absolute Deviation. Even though the mean error is pretty small (as well as the two other numbers, in ccomparison to Exercise 2.45), a nonzero mean implies that the overall estimation will be biased.

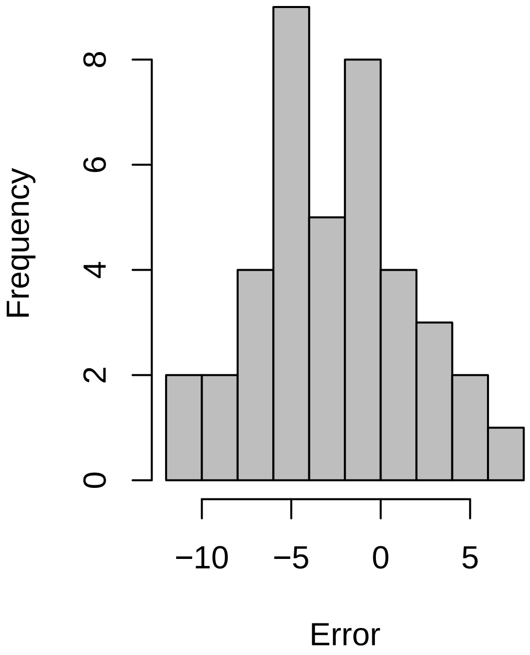
2.45 Table E2.2 contains 40 one-step-ahead forecast errors from a forecasting model a. **Find the sample ACF of the forecast errors. Interpret the results.**



Please see above for the ACF for the one-step forecast errors provided in table E2.2. Recall that the one-step forecast error describes the difference between the forecasted value one period before and the actual value. With the quickly declined autocorrelation after the first lag, we can deduce with confidence that errors are at large not autocorrelated with the past values. Moreover, there does not appear to be any other outliers and no perceivable indication for unaccounted seasonality. One caveat as suggested by the plot is that the autocorrelation at lag 4 (errors four lags apart) extends beyond the 95% confidence interval for 0 autocorrelation, which means that observations 4 lags apart are somewhat correlated.

b. **Construct a normal probability plot of the forecast errors. Is there evidence to support a claim that the forecast errors are normally distributed?**





Notice that the normal probability plot does not have the usual appearance of a normal probability plot: it has relatively longer right tail and it does not peak in the center. Therefore, it does not look like a normal distribution, meaning that the errors might not be normally distributed.

c. **Find the mean error, the mean square error, and the mean absolute deviation. Is it likely that the forecasting technique produces unbiased forecasts?**

Table 2: ME, MSE, & MAD

	Mean Error	Mean Squared Error	Mean Absolute Deviation
Value	-2.51925	24.02144	4.08475

In contrast to the previous exercise, what we have here is much larger of a deviation from the true value since the mean error and other relevant statistics are larger. Therefore, it's less likely to obtain unbiased forecast from the forecasting technique employed in this question.

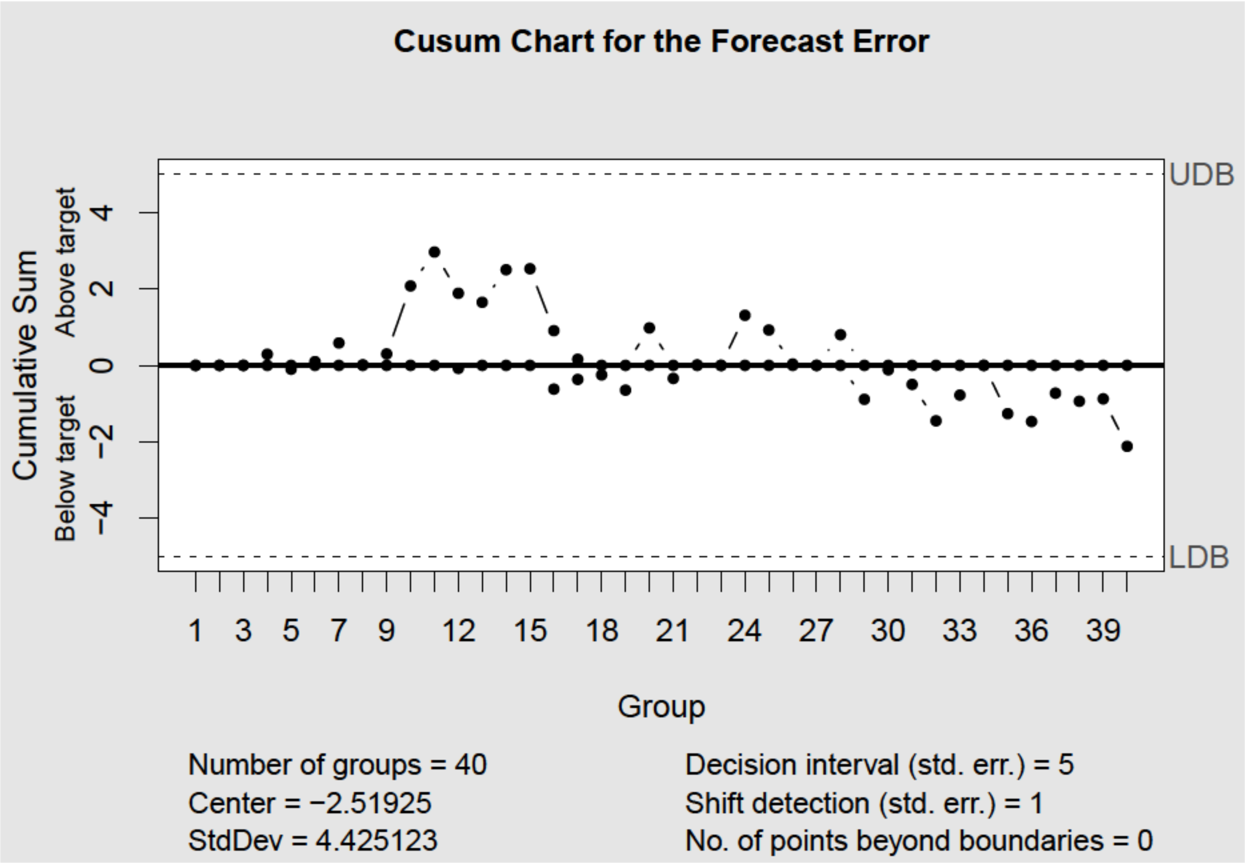
2.46 Exercise 2.44 and 2.45 present information on forecast errors. Suppose that these two sets of forecast errors come from two different forecasting methods applied to the same time series. Which of these two forecasting methods would you recommend for use? Why?

I would recommend the method incorporated in Exercise 2.44 for a couple of reasons. First, just basing off of the two ACF plots of the one-step ahead forecast errors, we can see clearly that, by the first method in Exercise 2.44, none of the autocorrelation after lag 0 exceeds the 95% confidence interval for 0 correlation. However, as previously noted, the ACF for the second method in Exercise 2.45 at lag=4 extends beyond the confidence interval, meaning that there is autocorrelation for observations 4 lags apart. Hence, the forecasting method in the Exercise 2.44 does a better job to obtain a stationary time series comparing to in E2.45 just at first glance.

We can further substantiate this claim by looking at the normal probability plots of the forecast errors. In Exercise 2.44, the histogram of forecast errors points to a normal-adjacent distribution. On the other hand, we can observe from the histogram for the forecast errors in Exercise 2.45 a left skew

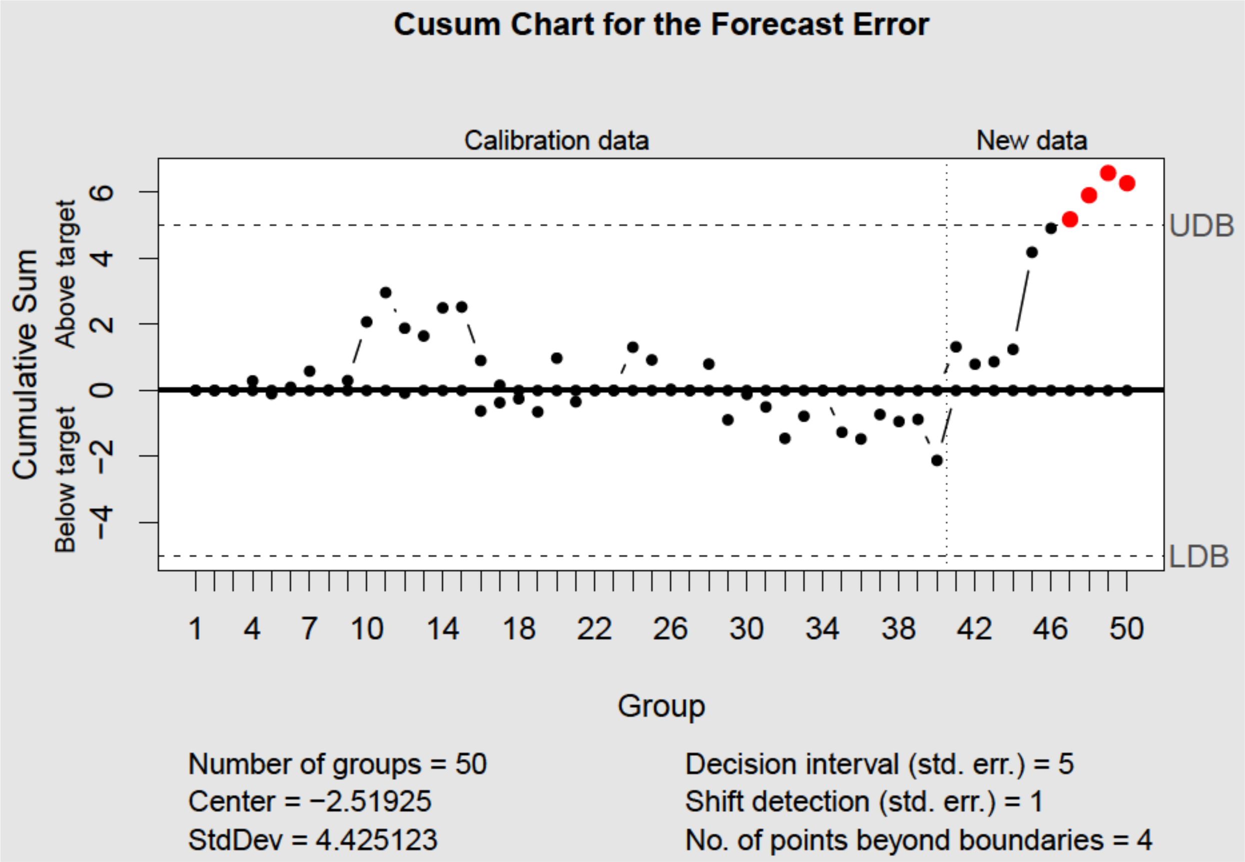
And of course, our most reliable tool is the plot residual analysis. Since all the data we have are composed of errors(residuals), we can examine their respective outlooks.

2.50 Consider the forecast errors in Exercise 2.45. Construct a cumulative sum control chart for these forecast errors. Does the forecasting system exhibit stability over this time period?



It appears that forecast errors fluctuate relatively randomly (no apparent trend) around 0. Moreover, it does not appear at any point that our cumulative sum of deviations exceeds the upper or lower bound of the control chart, which is a signal that our forecasting model is working adequately so far.

2.51 Ten additional forecast erros for the forecasting model in Exercise 2.44 are as follows: 5.5358, -2.6183, 0.0130, 1.3543, 12.6980, 2.9007, 0.8985, 2.9240, 2.6663, and -1.6710. Plot these additional 10 forecast errors on the individuals and moving range control charts constructed in Exercise 2.47. Is the forecasting system still working satisfactorily?



As we can see from the cusum chart with newly added forecast errors, the cumulative sum of deviations from the target mean for this group of errors drift out of the upper limit. This is an indication that our forecasting system is no longer working satisfactorily.

3.7 The quality of Pinot Noir wine is thought to be related to the properties of clarity, aroma, body, flavor, and oakiness. Data for 38 wines are given in Table E3.4

a. **Fit a multiple linear regression model relating wine quality to these predictors. Do not include the "Region" variable in the model.**

b. **Test for significance of regression. What conclusions can you draw?**

Table 3: F-Test

	F-stats	Number of Variables	Degrees of Freedom
Value	16.506	5	32

As we can see from the F-test summary table, 16.506 is quite large. Recall that the F-stat is mathematically calculated as $\frac{SSR/k}{SSE/(n-p)}$, meaning that the SSR is much greater than SSE in proportion. Under normal circumstances, an F-stat of 3.95 would suffice to reject the null hypothesis that none of the parameters is significant.

c. **Use *t*-test to assess the contribution of each predictor to the model. Discuss your findings.**

Table 4: Coefficients & T-Tests

	coefficient	p-value	t-stats
clairty	2.339	0.187	1.349
aroma	0.483	0.086	1.771
body	0.273	0.418	0.821
flavor	1.168	0.001	3.837
oakiness	-0.684	0.017	-2.522

Table 5: Summary for Parameters

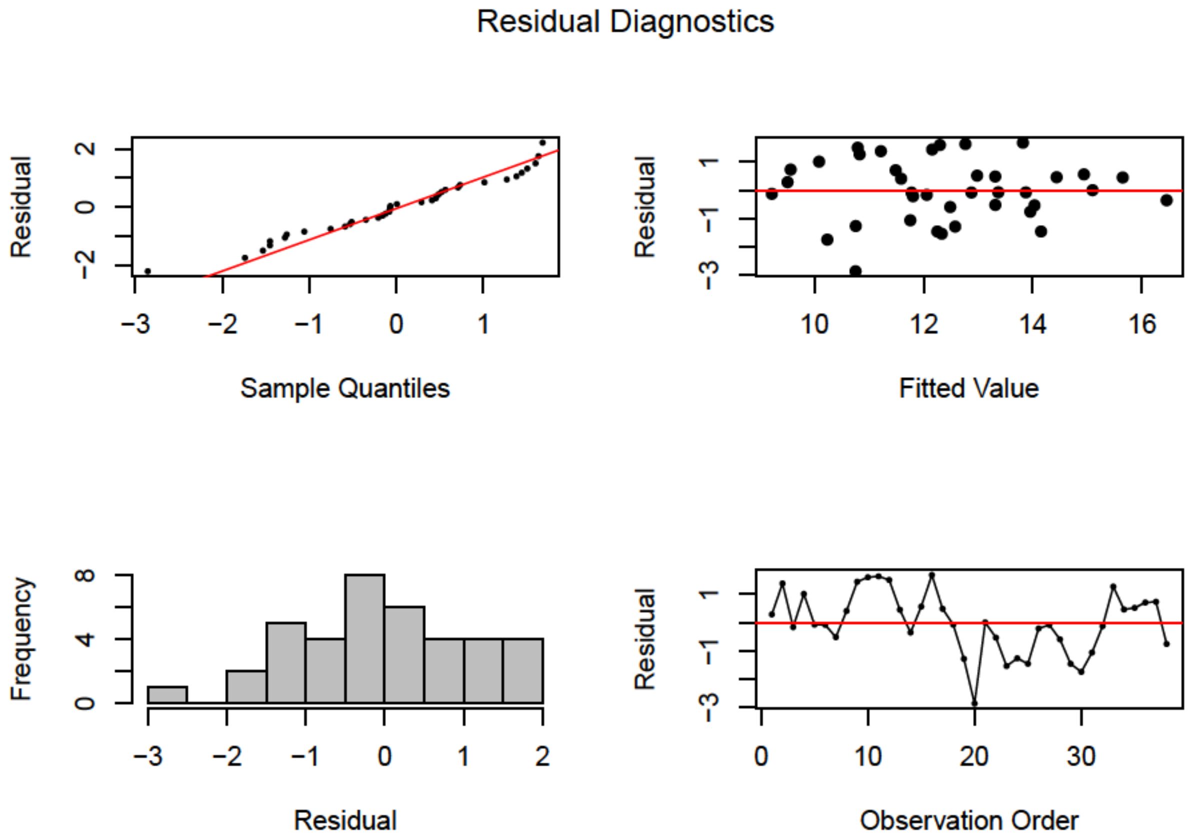
	Min	Max
clairty	0.5	1.0
aroma	3.3	7.7
body	2.6	6.6
flavor	2.9	7.0
oakiness	2.9	6.0

As we can see from Table 4 that sums up the coefficients, p-values, as well as t-statistics for our independent variables in the regression, it's clear that not all of our coefficients are significant. We can see both from the p-value as well as the t-stats. The only really significant results we have are falvor and oakiness. We then turn our attention to the column of coefficients, which summarize the contributions these different characteristics have on the quality of the Pinot Noir.

At first glance, it appears that clarity eclipses all other characteristics by the size of its regression coefficient (albeit insignificant). However, we should also take into consideration the scale upon which our coefficients operate. In Table 5, we can see that clarity ranges between 0 and 1, so a better way to interpret the result is that a 10% increase (0.1) in clarity corresponds to a 0.234 point increase in the quality of the Pinot Noir. On average, flavor seems to account for most of the contribution, with it ranging from 2.9 to 7.0 and a relatively large coefficient. One unit increase of flavor rating raises the overall quality point of the Pinot Noir by 1.168.

The oakiness of the Pinot Noir stands out as the only characteristic that works against the quality. Even though its impact isn't as large in comparison to flavor, the decently wide range (2.9-6.0) means that it could potentially affect the overall quality quite negatively.

d. Analyze the residuals from this model. Is the model adequate?



From the Residual vs. Fitted Value plot, we can see conclude that there is no apparent trend in the residual. Further more, there is no serial correlation within the residuals based on the residual vs time plot. It's also clear that the residual approaches a normal distribution with relatively long tails on both ends. The quantile-quantile residual-sample plot also seems to suggest that the distribution of residuals seems to have a larger left skew due to one extreme outlier.

e. **Calculate Rˆ2 and the adjusted R^2_{adj} for this model. Compare these values to the R^2 and R^2_{adj} for the linear regressin model relating wine quality to only the predictors "Aroma" and "Flavor". Discuss your results.**

Table 6: Full/Partial Model Rˆ2

	R-squared	Adjusted R-squared
Full Model	0.7205992	0.6769428
Partial Model	0.6585515	0.6390402

As shown in Table 6 above, we have the data for R^2 as well as R^2_{adj} for both the full regression model and the regression model on just flavor and aroma. From a linear regression model standpoint, it's quite normal to see the R^2 and R^2_{adj} values increase as we fit more variables into the regression because a larger portion of variances in the original data are now accounted for. Adjusted R^2 is naturally lower in both models than the regular R^2 for it penalizes us for adding new variables. Of course, all the variables added are quite correlated with the quality of said Pinot Noir, but it's also clear that there is still existing endogeneity in the omitted variable. For instance, region is not included, and we know for a fact that wine from different regions have different taste and make use of different material, which are highly correlated to the quality of the wine. Therefore, even though the full model is better than the partial model in the sense that it has included more pertinent variables, it's still not a great estimation.

Another important thing to keep in mind is that a high R^2 value, or, more precisely, adjusted R^2 value, is not the most important thing in time series data. If R^2 is high, our forecasting model is most likely useless.

f. **Find a 95% CI for the regression coefficient for "Flavor" for both models in part e. Discuss any differences.**

Recall that the calculationn for 95 confidence interval is given by the following:

$$\hat{\beta} - t_{\frac{\alpha}{2}} \cdot se(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{\frac{\alpha}{2}} \cdot se(\hat{\beta})$$

Notice that in this case, we have the following table:

Table 7: Mean & SE of Flavor

	Mean	Standard Error
Full Model	1.168324	0.3044807
Partial Model	1.170166	0.2905446

Hence we can have the 95% confidence interval for flavor in the original/full model:

$$\begin{gathered} 1.170166 - 1.96 \cdot 0.0.2905446 \leq \beta \leq 1.170166 + 1.96 \cdot 0.2905446 \\ 0.57154183 \leq \beta \leq 1.76510617 \end{gathered}$$

In the partial model, we can obtain the 95% confidence interval:

$$\begin{gathered} 1.168324 - 1.96 \cdot 0.3044807 \leq \beta \leq 1.168324 + 1.96 \cdot 0.3044807 \\ 0.60069858 \leq \beta \leq 1.73963342 \end{gathered}$$

3.8 Reconsider the wine quality data in Table E3.4. The "Region" predictor refers to three distinct geographical regions where the wine was produced. Note that this is a categorical variable.

a. **Fit the model using the "Region" variable as it is given in Table E3.4. What potential difficulties could be introduced by including this variable in the regression model using the three levels shown in Table E3.4?**

The new linear regression is performed R markdown file. There is no reason to regard this regression with the same seriousness as the other ones because, by including region as a variable directly, the baseline is unclear and there is no logical progression from going from one region to another. It could easily be the case that region 1 has much better quality of Pinot Noir on average than region 2, while region 3 also has better quality of wine than region 2. However, a negative/positive/0 coefficient would " attempt" to sort the three regions into logical order, which doesn't necessarily exists. Therefore it's better to create dummy variables to analyze the effect of region on wine quality.

b. **An alternative way to include the categorical variable "Region" would be to introduce two indicator variables x_1 and x_2. Why is this approach better than just using the codes 1, 2, and 3?**

Yes, this approach will provide a much better result because we will have a baseline and other regions will be pairwise compared to the baseline group. Each coefficient of the regions will be either positive or negative in contrast with the baseline.

c. **Rework Exercise 3.7 using the indicator variables defined in part b for "Region".**

Table 8: F-Test

	F-stats	Number of Variables	Degrees of Freedom
Value	22.102	7	30

As we can see from Table 8 that sums up the F-test statistics, a value of 22.102 is quite large, which indicates that it's very likely that at least one of the coefficients for the variables is statistically significant.

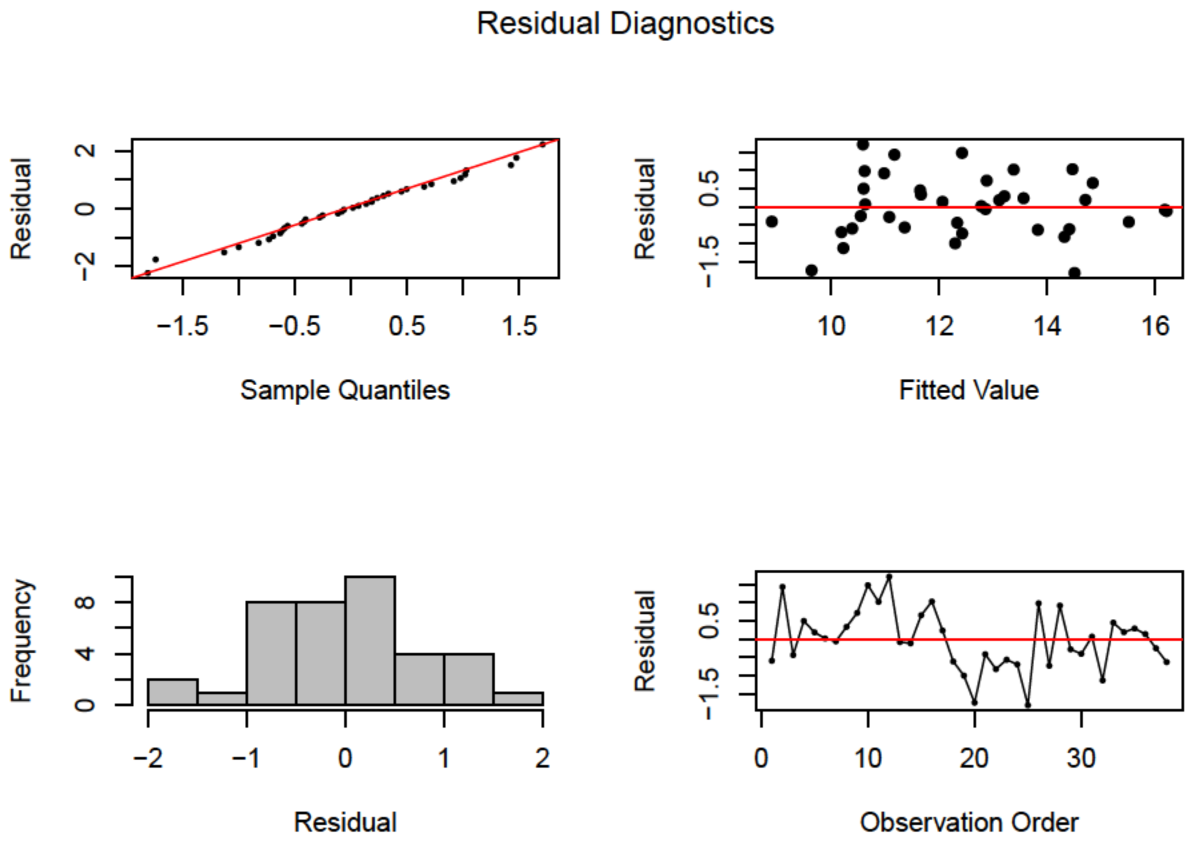
Table 9: Coefficients & T-Tests

	coefficient	p-value	t-stats
clairty	0.017	0.991	0.012
aroma	0.089	0.727	0.353
body	0.080	0.768	0.298
flavor	1.117	0.000	4.650
oakiness	-0.346	0.148	-1.487
region 1	-0.973	0.066	-1.906
region 2	-2.485	0.000	-4.222

Table 10: Summary for Parameters

	Min	Max
clairty	0.5	1.0
aroma	3.3	7.7
body	2.6	6.6
flavor	2.9	7.0
oakiness	2.9	6.0

As we can see from Table 9, out of the original regressors, only flavor still remains (and to a large extent enhances) its significance in the regression with respect to wine quality. As indicated, being produced in region 1 in comparison to region 3 lowers the quality on average by 1.906, while being produced in region 2 comparing to region 3 lowers the quality on average by 4.222. This indicates that, holding all else constant, we get the best quality Pinot Noir from region 3, then region 1, and lastly region 2.



From the residual diagnostics plots, we can see that there is no clear pattern in the residual vs. fitted value plot, and the residuals are pretty normally distributed along the qqline. Overall, these graphs suggest that the new model including the region dummies is a much better model than the original one.

Table 11: Full Model (with Region) Rˆ2

	R-squared	Adjusted R-squared
Original Model	0.7205992	0.6769428
Model with Region	0.8375838	0.7996867

We can see that, in comparison to the original regression model's R-squared and adjusted R-squared values, the newer model explains a wider range of variations in the time series. However, the caveat still stands in time series data – where the best fit in terms of R-squared values is not best forecasting model.

Table 12: Mean & SE of Flavor

	Mean	Standard Error
Full Model	1.11723	0.2402562

Observe the table above that summarizes the flavor coefficient's mean and standard error. Hence we can have the 95% confidence interval for flavor in the original/full model:

$$1.11723 - 1.96 \cdot 0.0.2402562 \leq \beta \leq 1.11723 + 1.96 \cdot 0.2402562$$
$$0.64632785 \leq \beta \leq 1.58813215$$