Exam 1

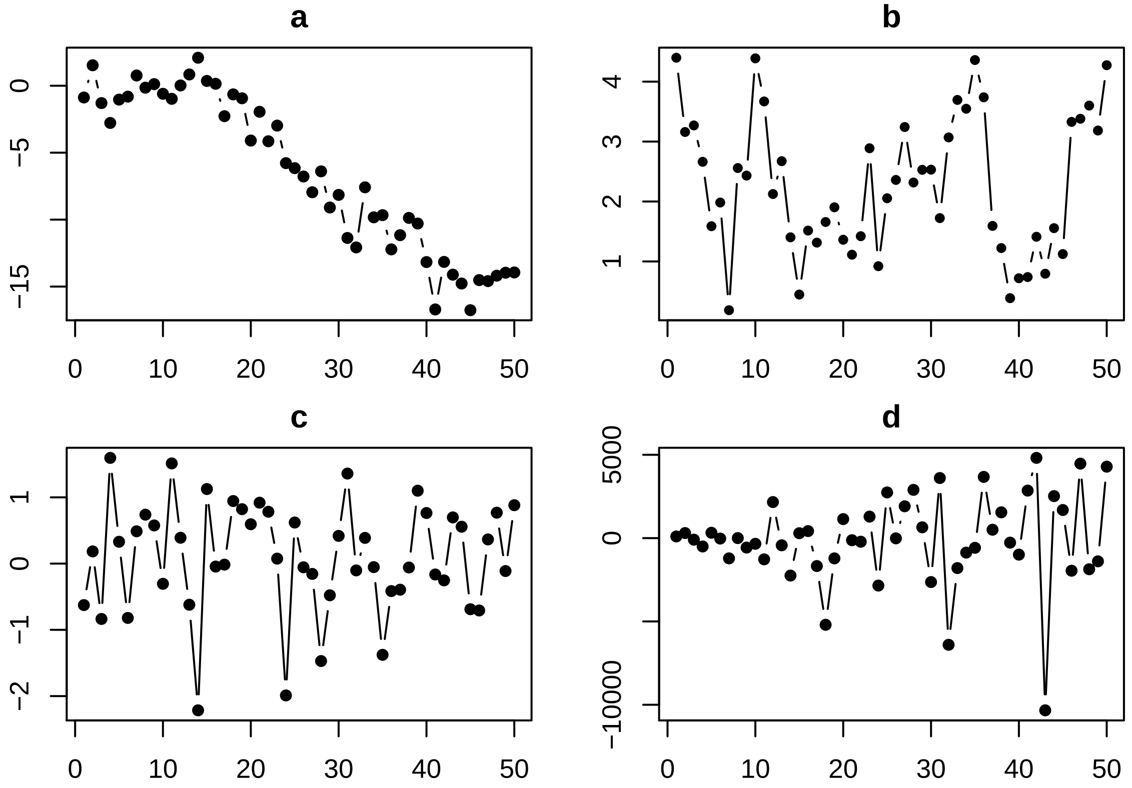
2/13/19

There are three questions, each of which has several parts, as well as a bonus question. Neither the questions nor the parts are necessarily in order from easiest to most difficult. Make sure you have taken a look at and attempted all of the questions in the allotted time. Stop working and immediately turn in your exam when time has been called.

Name:	________________________________
Question	Points
1	16
2	36
3	14
Bonus	________

1. Stationarity and The Autocorrelation Function

Indicate (i) whether or not there is clear evidence of nonstationarity for each of the time series shown below and (ii) what you expect the sample autocorrelation function of each time series to look like, with reference to what you would expect if each time series had no dependence across time and constant mean. Feel free draw a picture as part of your answers to (ii).



(a)
   i. There is clear evidence of nonstationarity. Worth 2 points.
   ii. I would expect the sample autocorrelations to be slowly decreasing as the lag h increases, with several outside of the approximate 95% intervals for sample autocorrelations of a Gaussian white noise process. Worth 2 points.

(b)
   i. There is **not** clear evidence of nonstationarity. Worth 2 points.
   ii. I would expect the sample autocorrelations to be slowly decreasing as the lag h increases, with several outside of the approximate 95% intervals for sample autocorrelations of a Gaussian white noise process. Worth 2 points.

(c)
   i. There is **not** clear evidence of nonstationarity. Worth 2 points.
   ii. I would expect the sample autocorrelations to be randomly distributed about 0, with few outside of the approximate 95% intervals for sample autocorrelations of a Gaussian white noise process. Worth 2 points.

(d)
   i. There is clear evidence of nonstationarity. Worth 2 points.
   ii. I would expect the sample autocorrelations to be randomly distributed about 0, with few outside of the approximate 95% intervals for sample autocorrelations of a Gaussian white noise process. Worth 2 points.

2. Sample Autocovariance and Autocorrelations

Suppose that you observe a time series of length 3, with elements $y_1 = 3$, $y_2 = 2$, and $y_3 = -5$.

(a) Compute the sample mean, $\bar{y}$.

$$\bar{y} = \frac{1}{3}(3 + 2 - 5) = 0$$

Worth 2 points.

(b) Provide the formula for the sample variance $\hat{\gamma}_y(0)$ of a time series $\boldsymbol{y}$ of length n. Use the formula we discussed in class, and which is given in the notes.

$$\hat{\gamma}_y(0) = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$$

Worth 4 points.

(c) Is the sample covariance computed as described in (b) an unbiased estimator of the variance $\gamma_y(0)$? Yes or no.

No. Worth 2 points.

(d) Compute the sample variance, $\hat{\gamma}_y(0)$. Do not bother simplifying fractions.

$$\hat{\gamma}_y(0) = \frac{1}{3}\left(3^2 + 2^2 + (-5)^2\right) = \frac{9 + 4 + 25}{3} = \frac{38}{3}$$

Worth 2 points.

(e) Provide the formula for the lag-h sample autocovariance $\hat{\gamma}_y(h)$ of a time series $\boldsymbol{y}$ of length n.

$$\hat{\gamma}_y(h) = \frac{1}{n}\sum_{t=h+1}^{n}(y_t - \bar{y})(y_{t-h} - \bar{y})$$

Worth 2 points.

(f) Are the sample autocovariances computed as described in (e) unbiased estimators of the autocovariances $\gamma_y(h)$? Yes or no.

No. Worth 2 points.

(g) For which values of $h > 0$ can we compute an estimate of the sample autocovariance, $\hat{\gamma}_y(h)$?

We have enough data to compute the sample autocovariance for $h = 1$ and $h = 2$. Worth 2 points.

(h) Compute the sample autocovariances for the values of $h > 0$ identified in (g). Do not bother simplifying fractions.

$$\hat{\gamma}_y(1) = \frac{1}{3}(3 \times 2 + 2 \times (-5)) = \frac{6 - 10}{3} = -\frac{4}{3}$$
$$\hat{\gamma}_y(2) = \frac{1}{3}(3 \times (-5)) = -\frac{15}{3}$$

Worth 3 points.

(i) Provide the formula for the lag-h sample autocorrelation $\hat{\rho}_y(h)$ of a time series $\boldsymbol{y}$ of length n.

$$\hat{\rho}_y(h) = \frac{\hat{\gamma}_y(h)}{\hat{\gamma}_y(0)}$$

Worth 2 points.

(j) Are the sample autocorrelations computed as described in (i) unbiased estimators of the autocorrelations $\rho_y(h)$? Yes or no.

No. Worth 2 points.

(k) For which values of $h > 0$ can we compute an estimate of the sample autocorrelation, $\hat{\rho}_y(h)$?

We have enough data to compute the sample autocorrelation for $h = 1$ and $h = 2$. Worth 2 points.

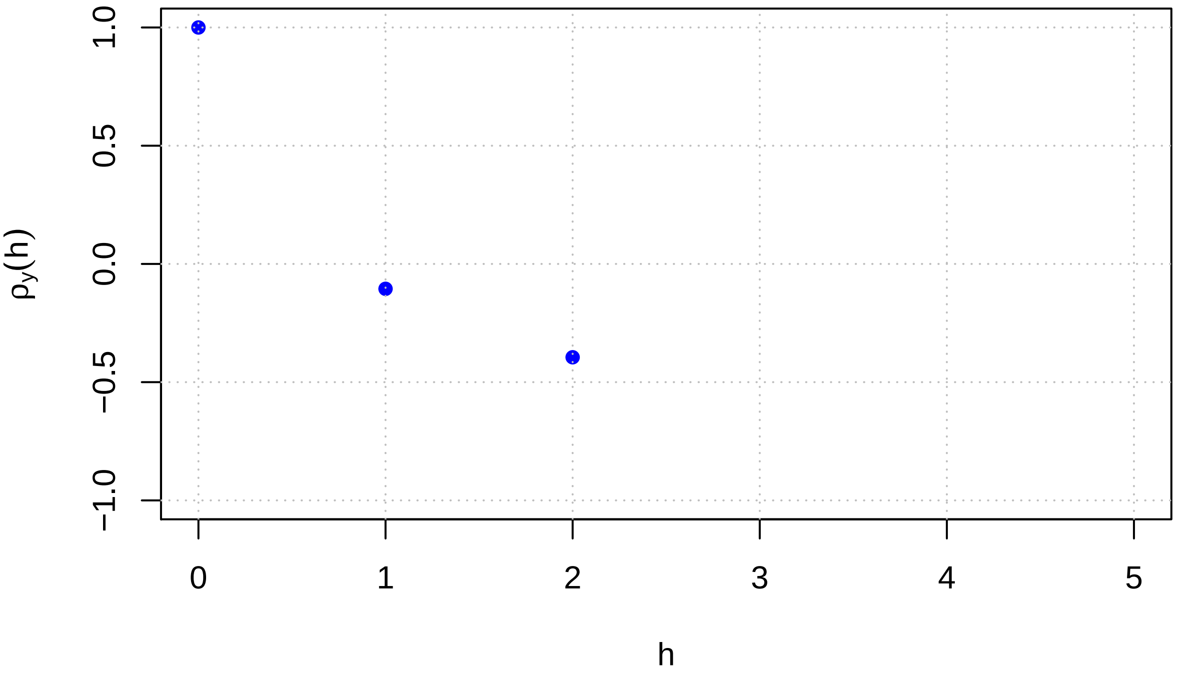
(l) Compute the sample autocorrelations for the values of $h > 0$ identified in (k). Do not bother simplifying fractions.

$$\hat{\rho}_y(1) = -\frac{4}{38}$$
$$\hat{\rho}_y(2) = -\frac{15}{38}$$

Worth 2 points.

(m) Plot the sample autocorrelations from (l) on the autocorrelation plot given below. You don't need to get the points exactly right, just try to make sure that they are placed correctly relative to each other and the gray dashed lines.

Worth 2 points.



(n) Add horizontal lines corresponding to the 2.5 and 97.5 quantiles of the distribution of each $\hat{\rho}_y(h)$ based on the asymptotic distribution of each $\hat{\rho}_y(h)$ as $n \to \infty$. Recall that the 2.5 and 97.5 percent quantiles for a standard normal distribution are -2 and 2. You don't need to get the lines exactly right, just try to make sure that they are placed correctly relative to the autocorrelations you plotted in (m).

The horizontal lines would be at $-2\sqrt{3}$ and $2/\sqrt{3}$ - these would be below -1 and above 1, which are beyond the y-axis on our plot! Full credit was given to anyone who noticed this, or to anyone who extended the y-axis on the plot by hand to include these values. Worth 3 points.

(o) In at most one sentence, state what you would conclude regarding the evidence of dependence across time in this data from (m) and (n).

I would conclude that we do not observe evidence of dependence across time based on the sample autocorrelations and the approximate 95 percent intervals of the sample autocorrelations of a white noise process as $n \to \infty$, because the two autocorrelations we calculated are within the approximate 95 percent intervals for the sample autocorrelations of a white noise process. However, I would also note that that one of the main reasons why we do not observe evidence of dependence is because we really don't have enough data. In fact, when we only have a time series of length $n = 3$, the approximate 95 percent intervals of the sample autocorrelations of a white noise process cover the entire range of possible values of the sample autocorrelations! So based the approximate 95 percent intervals of the sample autocorrelations of a white noise process as $n \to \infty$, we would never be able to conclude that there is evidence of dependence across time with a time series of this length. Worth 2 points.

(p) Rank the autocorrelations you computed in order of how much data you used to estimate them, from the autocorrelation you estimated using the most data to the autocorrelation you estimated using the least data.

1. $\hat{\rho}_y(1)$
2. $\hat{\rho}_y(2)$

I accepted answers in terms of $\hat{\rho}_y(h)$ and $\hat{\gamma}_y(h)$, and did not penalize people who wrote that we have the most data available to estimate $\hat{\rho}_y(0)$, even though it does not depend on the data at all. Worth 2 points.

3. Model Selection

3. This problem will refer back to the `broc` data posted on the course website, which contains the average price of one pound of broccoli in urban areas each month, from July 1995 through December 2019. Suppose we fit the following two regression models. Summaries of each model fit are given with the definition of each model.

- Model 1: $\texttt{price}_i = \mu + \sum_{j=1996}^{2019} \gamma_j(\texttt{year}_i = j) + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$.

```r
summary(linmod1)
```

```
## 
## Call:
## lm(formula = price ~ factor(year) + factor(month), data = broc)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.35438 -0.05841 -0.00637  0.04859  0.45224 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       0.94344    0.04733  19.935  < 2e-16 ***
## factor(year)1996  0.04469    0.05135   0.870 0.384999    
## factor(year)1997  0.11460    0.05135   2.232 0.026494 *  
## factor(year)1998  0.23094    0.05135   4.497 1.04e-05 ***
## factor(year)1999  0.13960    0.05135   2.719 0.007002 ** 
## factor(year)2000  0.27269    0.05135   5.310 2.37e-07 ***
## factor(year)2001  0.11919    0.05135   2.321 0.021070 *  
## factor(year)2002  0.32844    0.05135   6.396 7.46e-10 ***
## factor(year)2003  0.33410    0.05135   6.506 3.99e-10 ***
## factor(year)2004  0.32969    0.05135   6.420 6.50e-10 ***
## factor(year)2005  0.45302    0.05135   8.822  < 2e-16 ***
## factor(year)2006  0.58102    0.05135  11.314  < 2e-16 ***
## factor(year)2007  0.72044    0.05135  14.029  < 2e-16 ***
## factor(year)2008  0.80919    0.05135  15.757  < 2e-16 ***
## factor(year)2009  0.73277    0.05135  14.269  < 2e-16 ***
## factor(year)2010  0.71844    0.05135  13.990  < 2e-16 ***
## factor(year)2011  0.83477    0.05135  16.256  < 2e-16 ***
## factor(year)2012  0.73494    0.05135  14.311  < 2e-16 ***
## factor(year)2013  0.98619    0.05135  19.204  < 2e-16 ***
## factor(year)2014  0.89852    0.05135  17.497  < 2e-16 ***
## factor(year)2015  0.92577    0.05135  18.028  < 2e-16 ***
## factor(year)2016  0.76302    0.05135  14.858  < 2e-16 ***
## factor(year)2017  0.94694    0.05135  18.440  < 2e-16 ***
## factor(year)2018  0.98769    0.05135  19.233  < 2e-16 ***
## factor(year)2019  1.12969    0.05135  21.998  < 2e-16 ***
## factor(month)02  -0.04312    0.02944  -1.465 0.144248    
## factor(month)03  -0.07879    0.02944  -2.676 0.007930 ** 
## factor(month)04  -0.08812    0.02944  -2.993 0.003032 ** 
## factor(month)05  -0.12246    0.02944  -4.159 4.36e-05 ***
## factor(month)06  -0.10242    0.02944  -3.478 0.000592 ***
## factor(month)07  -0.10286    0.02920  -3.523 0.000505 ***
## factor(month)08  -0.13078    0.02920  -4.479 1.13e-05 ***
## factor(month)09  -0.09706    0.02920  -3.324 0.001016 ** 
## factor(month)10  -0.06346    0.02920  -2.173 0.030666 *  
## factor(month)11  -0.05722    0.02920  -1.960 0.051118 .
## factor(month)12  -0.05030    0.02920  -1.723 0.086156 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.102 on 258 degrees of freedom
## Multiple R-squared:  0.9249, Adjusted R-squared:  0.9148
## F-statistic: 90.83 on 35 and 258 DF,  p-value: < 2.2e-16
```

- Model 2: $\texttt{price}_i = \mu + \beta\texttt{year}_i + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$.

```
summary(linmod2)
```

```
##
## Call:
## lm(formula = price ~ year + factor(month), data = broc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42585 -0.08117 -0.00744  0.07183  0.44026
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -89.204084   2.270132 -39.295  < 2e-16 ***
## year              0.045199   0.001131  39.973  < 2e-16 ***
## factor(month)02  -0.043125   0.039576  -1.090  0.27679
## factor(month)03  -0.078792   0.039576  -1.991  0.04746 *
## factor(month)04  -0.088125   0.039576  -2.227  0.02676 *
## factor(month)05  -0.122458   0.039576  -3.094  0.00217 **
## factor(month)06  -0.102417   0.039576  -2.588  0.01016 *
## factor(month)07  -0.103819   0.039182  -2.650  0.00851 **
## factor(month)08  -0.131739   0.039182  -3.362  0.00088 ***
## factor(month)09  -0.098019   0.039182  -2.502  0.01293 *
## factor(month)10  -0.064419   0.039182  -1.644  0.10128
## factor(month)11  -0.058179   0.039182  -1.485  0.13871
## factor(month)12  -0.051259   0.039182  -1.308  0.19187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1371 on 281 degrees of freedom
## Multiple R-squared:  0.8523,   Adjusted R-squared:  0.846
## F-statistic: 135.1 on 12 and 281 DF,  p-value: < 2.2e-16
```

- Model 3: $\texttt{price}_i = \mu + \beta\texttt{days since start}_i + \sum_{j=1996}^{2019} \gamma_j(\texttt{year}_i = j) + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$.

```
summary(linmod3)
```

```
##
## Call:
## lm(formula = price ~ dayssincestart + factor(year) + factor(month),
##     data = broc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35513 -0.05845 -0.00622  0.04863  0.45149
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.59594    6.76385   0.236    0.814
## dayssincestart     0.00360    0.03732   0.096    0.923
## factor(year)1996  -1.27246   13.65342  -0.093    0.926
## factor(year)1997  -2.51715   27.28025  -0.092    0.927
## factor(year)1998  -3.71481   40.90090  -0.091    0.928
## factor(year)1999  -5.12015   54.52156  -0.094    0.925
## factor(year)2000  -6.30406   68.17332  -0.092    0.926
## factor(year)2001  -7.77216   81.80021  -0.095    0.924
## factor(year)2002  -8.87691   95.42087  -0.093    0.926
## factor(year)2003 -10.18525  109.04153  -0.093    0.926
## factor(year)2004 -11.50666  122.69330  -0.094    0.925
## factor(year)2005 -12.69793  136.32018  -0.093    0.926
## factor(year)2006 -13.88393  149.94085  -0.093    0.926
## factor(year)2007 -15.05851  163.56152  -0.092    0.927
## factor(year)2008 -16.28676  177.21328  -0.092    0.927
## factor(year)2009 -17.67778  190.84016  -0.093    0.926
## factor(year)2010 -19.00611  204.46083  -0.093    0.926
## factor(year)2011 -20.20378  218.08150  -0.093    0.926
## factor(year)2012 -21.62061  231.73326  -0.093    0.926
## factor(year)2013 -22.68396  245.36015  -0.092    0.926
## factor(year)2014 -24.08563  258.98081  -0.093    0.926
## factor(year)2015 -25.37238  272.60148  -0.093    0.926
## factor(year)2016 -26.85213  286.25324  -0.094    0.925
## factor(year)2017 -27.98281  299.88013  -0.093    0.926
## factor(year)2018 -29.25606  313.50080  -0.093    0.926
## factor(year)2019 -30.42806  327.12146  -0.093    0.926
## factor(month)02   -0.15473    1.15720  -0.134    0.894
## factor(month)03   -0.29209    2.21122  -0.132    0.895
## factor(month)04   -0.41302    3.36798  -0.123    0.902
## factor(month)05   -0.55536    4.48745  -0.124    0.902
## factor(month)06   -0.64692    5.64426  -0.115    0.909
## factor(month)07   -0.75536    6.76375  -0.112    0.911
## factor(month)08   -0.89488    7.92056  -0.113    0.910
## factor(month)09   -0.97276    9.07738  -0.107    0.915
## factor(month)10   -1.04716   10.19688  -0.103    0.918
## factor(month)11   -1.15252   11.35370  -0.102    0.919
## factor(month)12   -1.25360   12.47321  -0.101    0.920
##
## Residual standard error: 0.1022 on 257 degrees of freedom
## Multiple R-squared:  0.9249,  Adjusted R-squared:  0.9144
## F-statistic: 87.97 on 36 and 257 DF,  p-value: < 2.2e-16
```

- Model 4: $\texttt{price}_i = \mu + \beta \texttt{days since start}_i + \sum_{k=2}^{12} \alpha_k (\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$.

```
summary(linmod4)
```

```
##
## Call:
## lm(formula = price ~ dayssincestart + factor(month), data = broc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42589 -0.08115 -0.00745  0.07186  0.44025
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.899e-01  3.110e-02  31.824  < 2e-16 ***
## dayssincestart   1.237e-04  3.096e-06  39.971  < 2e-16 ***
## factor(month)02 -4.696e-02  3.958e-02  -1.187 0.236398
## factor(month)03 -8.612e-02  3.958e-02  -2.176 0.030382 *
## factor(month)04 -9.929e-02  3.958e-02  -2.509 0.012677 *
## factor(month)05 -1.373e-01  3.958e-02  -3.470 0.000602 ***
## factor(month)06 -1.211e-01  3.958e-02  -3.060 0.002423 **
## factor(month)07 -1.262e-01  3.918e-02  -3.222 0.001421 **
## factor(month)08 -1.580e-01  3.918e-02  -4.033 7.11e-05 ***
## factor(month)09 -1.281e-01  3.918e-02  -3.270 0.001209 **
## factor(month)10 -9.823e-02  3.918e-02  -2.507 0.012735 *
## factor(month)11 -9.583e-02  3.918e-02  -2.446 0.015068 *
## factor(month)12 -9.262e-02  3.918e-02  -2.364 0.018767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1371 on 281 degrees of freedom
## Multiple R-squared:  0.8523,   Adjusted R-squared:  0.846
## F-statistic: 135.1 on 12 and 281 DF,  p-value: < 2.2e-16
```

(a) Based on AIC or a z-test choose between the Model 1 and Model 2. Justify your choice in at most one sentence, and provide any relevant numerical evidence, e.g. AIC values, z-value and corresponding p-value, degrees of freedom and corresponding p-value. Indicate whether or not an F-test could have been used to make this choice as well.

Model 1 and Model 2 are not nested, so I would use AIC values to compare the two. Because the models are not nested, an F or z-test would not be an option.

The AIC of Model 1 would be given by:

$$\log\left(0.102^2 \times \left(\frac{258}{294}\right)\right) + \frac{294 + 2 \times 36}{294}$$

The AIC of Model 2 would be given by:

$$\log\left(0.1371^2 \times \left(\frac{281}{294}\right)\right) + \frac{294 + 2 \times 13}{294}$$

It is not feasible to compute these AIC's without a calculator, but I would choose whichever model had a lower AIC. Worth 3 points.

(b) Based on AIC or a z-test choose between the Model 1 and Model 3. Justify your choice in at most one sentence, and provide any relevant numerical evidence, e.g. AIC values, z-value and corresponding p-value, degrees of freedom and corresponding p-value. Indicate whether or not an F-test could have been used to make this choice as well.

Model 1 and Model 3 are nested, Model 3 contains one additional variable `dayssincestart` which is not contained in Model 1, but otherwise they are the same. It would be possible to choose between them using AIC, a z-test, or an F-test.

If I used AIC to compare the two, the AIC of Model 1 would be as given in (a).

$$\log\left(0.102^2 \times \left(\frac{258}{294}\right)\right) + \frac{294 + 2 \times 36}{294}$$

The AIC of Model 3 would be:

$$\log\left(0.1022^2 \times \left(\frac{257}{294}\right)\right) + \frac{294 + 2 \times 37}{294}$$

It is not feasible to compute these AIC's without a calculator, but I would choose whichever model had a lower AIC.

If I used a z-test to compare the two, I would compute the probability of observing a test statistic greater in absolute value than 0.096 under a t-distribution with 257 degrees of freedom. This probability would be 0.923, suggesting that a test statistic greater in absolute value than 0.096 is likely under the null that $\beta = 0$. This would lead me to choose Model 1.

Worth 3 points.

(c) Based on AIC or a z-test choose between the Model 3 and Model 4. Justify your choice in at most one sentence, and provide any relevant numerical evidence, e.g. AIC values, z-value and corresponding p-value, degrees of freedom and corresponding p-value. Indicate whether or not an F-test could have been used to make this choice as well.

Model 3 and Model 4 are nested, Model 3 contains eleven year indicators which is not contained in Model 4, but otherwise they are the same. It would be possible to choose between them using AIC or an F-test.

If I used AIC to compare the two, the AIC of Model 3 would be as given in (b).

$$\log\left(0.1022^2 \times \left(\frac{257}{294}\right)\right) + \frac{294 + 2 \times 37}{294}$$

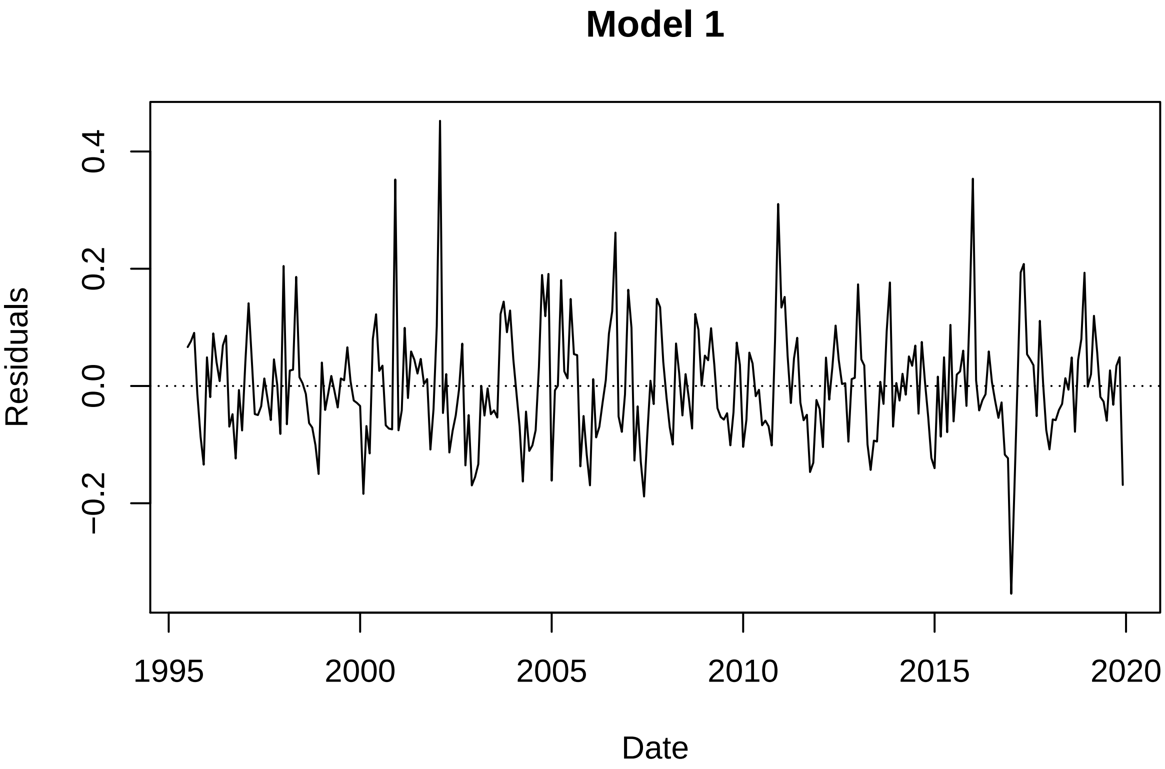
The AIC of Model 4 would be:

$$\log\left(0.1371^2 \times \left(\frac{281}{294}\right)\right) + \frac{294 + 2 \times 13}{294}$$

It is not feasible to compute these AIC's without a calculator, but I would choose whichever model had a lower AIC.

Worth 3 points.

(c) Examine the residuals from Model 1 in date order, from first to last. In at most one sentence, describe what (if any) evidence you observe for any remaining correlation across time.



The residuals from Model 1 still show some correlation across time - consecutive residuals tend to share the same sign. Worth 3 points.

(d) Suppose you wanted to forecast broccoli prices one month into the future, i.e. you wanted to compute $\mathbb{E}[y_{295}]$ under Model 1, Model 2, Model 3, or Model 4. Under which of the four models can you compute $\mathbb{E}[y_{295}]$ using the available data? Explain in at most one sentence.

We can compute $\mathbb{E}[y_{295}]$ under Models 2 and 4 only, because we will not be able to estimate a year effect for a future observation in 2020 from data that only includes observations from 2019 or earlier. Worth 2 points.

Bonus

In one sentence, describe the most surprising thing you've learned so far in this class. You will receive full credit for your answer as long as it is not something that is false.