Homework 1 Solutions

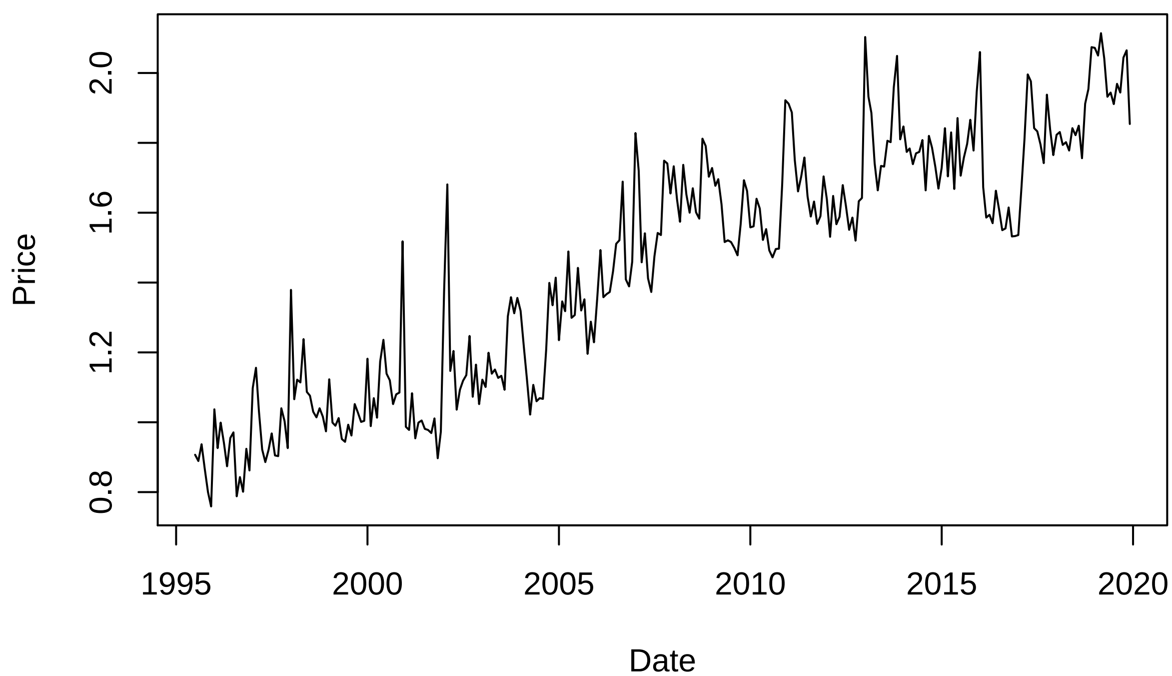
Due: Tuesday 1/28/20 by 10:00am

Clearly, 1. and 2. do not have right or wrong answers - you will be given full credit for each as long as your responses indicate that you made an honest effort.

Regression and R Review

3. This problem will require that you work with `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.
   (a) Working with time series data often involves working with dates, which can be tricky to manipulate because they are often provided as characters. For instance, in the `broc` data the first value of the `date` variable is a string `"1995-07-01"`. This makes things like plotting difficult or extracting the year, month, or day difficult. Fortunately, it is easy to convert a string variable to something `R` calls a `Date` variable! See the Quick-R tutorial on on converting strings to `Date` variables: https://statistics.berkeley.edu/computing/r-dates-times. Create a new variable `fdate` in the `broc` data frame that is a `Date` object, and line plot of `broc$fdate` against `broc$price`.

```
load("~/Dropbox/Teaching/TimeSeries2020/stat697/content/data/broc.RData")
broc$fdate <- as.Date(broc$date, "%Y-%m-%d")
plot(broc$fdate, broc$price, type = "l", xlab = "Date", ylab = "Price")
```



   (b) Fit the following two regression models. Model 1 is given by $\texttt{price}_i = \mu + \sum_{j=1996}^{2019} \gamma_j(\texttt{year}_i = j) + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$. Model 2 is given by $\texttt{price}_i = \mu + \beta\texttt{year}_i +$

$\sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim}$ normal$(0, \sigma^2)$. Revisit the tutorial (https://statistics.berkeley.edu/computing/r-dates-times) for help extracting the month or year from a `Date` object in `R`. Using AIC, a z-test, or an F-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, or F-value, degrees of freedom and corresponding p-value.

```
broc$year <- as.numeric(format(broc$fdate, "%Y"))
broc$month <- format(broc$fdate, "%m")
linmod1 <- lm(price~factor(year)+factor(month), data = broc)
linmod2 <- lm(price~year+factor(month), data = broc)
n <- nrow(broc)
k.fit1 <- length(coef(linmod1))
k.fit2 <- length(coef(linmod2))
ss.fit1 <- mean((broc$price - linmod1$fitted.values)^2)
ss.fit2 <- mean((broc$price - linmod2$fitted.values)^2)
aic.fit1 <- log(ss.fit1) + (n + 2*k.fit1)/n
aic.fit2 <- log(ss.fit2) + (n + 2*k.fit2)/n
```

Because these models are not nested, we will compare them using AIC. The AIC of Model 1 is -3.45, which is lower than the AIC of Model 2, -2.93, so we would choose Model 1.

(c) Fit an additional model, which we'll call Model 3: $\texttt{price}_i = \mu + \beta\texttt{days since start}_i + \sum_{j=1996}^{2019} \gamma_j(\texttt{year}_i = j) + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim}$ normal$(0, \sigma^2)$. Using AIC, a z-test, or an F-test, choose between 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, or F-value, degrees of freedom and corresponding p-value.

```
broc$dayssincestart <- as.numeric(broc$fdate) - min(as.numeric(broc$fdate))
linmod3 <- lm(price~dayssincestart+factor(year)+factor(month), data = broc)
```

Models 1 and 3 are nested, so we can compare them using a z/t-test or an F-test of the null hypothesis that $\beta = 0$. The t-statistic is 0.096, and the corresponding p-value (probability of a t-random variable with 257 degrees of freedom exceeding 0.096) in absolute value is 0.923. This would lead us to choose Model 1.

Note that alternatively, we could use an F-test.

```
p <-  length(linmod3$coef)
p1 <- length(linmod1$coef)
ssr1 <- sum(linmod1$residuals^2)
ssr <- sum(linmod3$residuals^2)
f.stat <- ((ssr1 - ssr)/ssr)*
  (n - p)/(p - p1)
```

The F-statistic is 0.009, and the corresponding p-value (probability of an F-random variable with 1 and 257 degrees of freedom exceeding 0.009) is 0.923. This would also lead us to choose Model 1.

Last, we could use AIC.

```
k.fit3 <- length(coef(linmod3))
ss.fit3 <- mean((broc$price - linmod3$fitted.values)^2)
aic.fit3 <- log(ss.fit3) + (n + 2*k.fit3)/n
```

The AIC of Model 1 is -3.45, which is lower than the AIC of Model 3, -3.44, so we would choose Model 1 if we used AIC as our model selection criteria.

(d) Fit one more additional model, which we'll call Model 4: $\texttt{price}_i = \mu + \beta\texttt{days since start}_i + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim}$ normal$(0, \sigma^2)$. Revisit the tutorial (https://statistics.berkeley.edu/computing/r-dates-times) for help computing the number of days since the start, July 1995. Using

AIC, a z-test, or an F-test, choose between 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, or F-value, degrees of freedom and corresponding p-value.

```
linmod4 <- lm(price~dayssincestart+factor(month), data = broc)
```

Models 3 and 4 are nested but Model 3 includes more than one covariate that is absent from Model 4, so we could use an F-test.

```
p <-  length(linmod3$coef)
p1 <- length(linmod4$coef)
ssr1 <- sum(linmod4$residuals^2)
ssr <- sum(linmod3$residuals^2)
f.stat <- ((ssr1 - ssr)/ssr)*
  (n - p)/(p - p1)
```

The F-statistic is 10.363, and the corresponding p-value (probability of an F-random variable with 24 and 257 degrees of freedom exceeding 10.363) is 0. This would also lead us to choose Model 3.

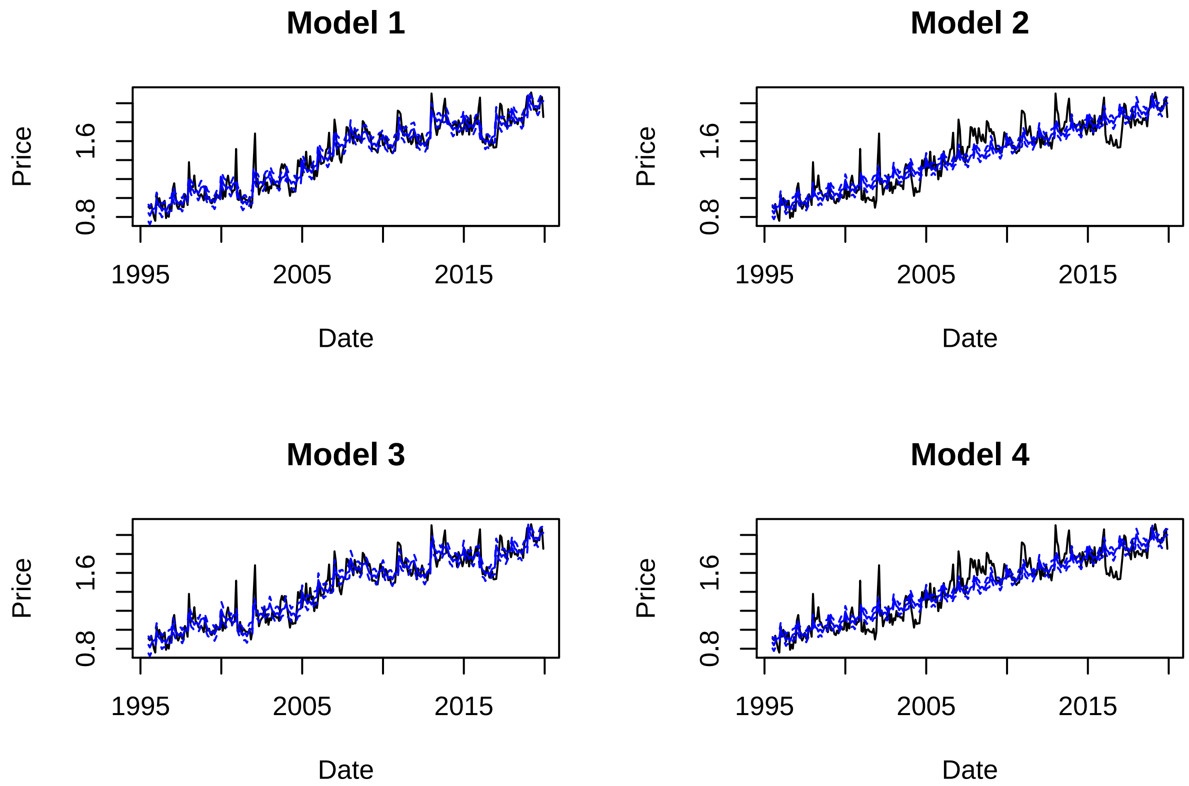
Alternatively, we could use AIC.

```
k.fit4 <- length(coef(linmod4))
ss.fit4 <- mean((broc$price - linmod4$fitted.values)^2)
aic.fit4 <- log(ss.fit4) + (n + 2*k.fit4)/n
```

The AIC of Model 3 is -3.44, which is lower than the AIC of Model 4, -2.93, so we would choose Model 3 if we used AIC as our model selection criteria.

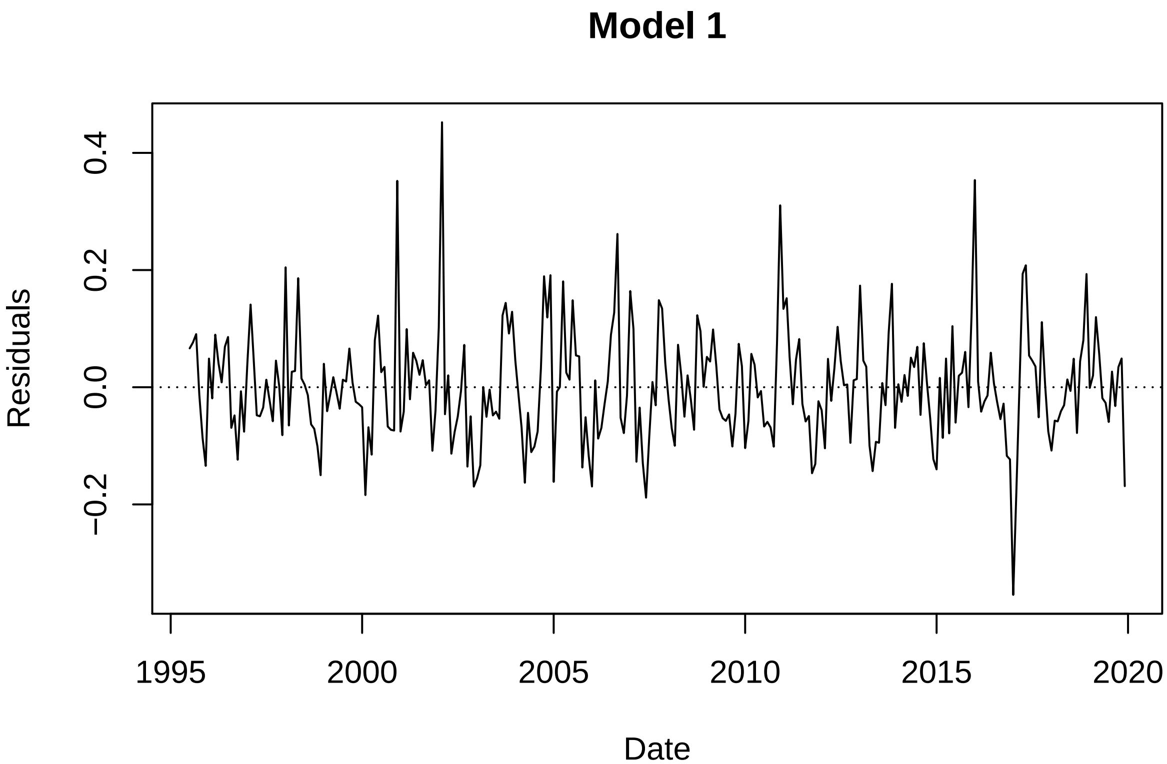
(e) Make one plot with four panels. In each panel, plot the prices in date order, from first to last. In the first panel, add fitted values from Model 1 along with 95% confidence intervals for each fitted value. In the second panel, add fitted values from Model 2 along with 95% confidence intervals for each fitted value. In the third panel, add fitted values from Model 3 along with 95% confidence intervals for each fitted value. In the fourth and last panel, add fitted values from Model 4 along with 95% confidence intervals for each fitted value.

```
par(mfrow = c(2, 2))
for (i in 1:4) {
  plot(broc$fdate, broc$price, type = "l", xlab = "Date", ylab = "Price",
       main = paste("Model ", i, sep = ""))
  pred <- predict(get(paste("linmod", i, sep = "")), se.fit = TRUE)
  lines(broc$fdate, pred$fit, col = "blue")
  lines(broc$fdate, pred$fit + qnorm(0.975)*pred$se.fit, col = "blue", lty = 2)
  lines(broc$fdate, pred$fit + qnorm(0.025)*pred$se.fit, col = "blue", lty = 2)
}
```



(f) Plot 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.

```
plot(broc$fdate, linmod1$residuals, type = "l", xlab = "Date", ylab = "Residuals",
       main = "Model 1")
abline(h = 0, lty = 3)
```



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

(g) 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.