Homework 1

Due: Thursday 1/31/19 by 12:00pm

Grading Scheme:

- 1 point for some response to 1.
- 1 point for some response to 2.
- Maximum of 3 points for 3., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only
  - 1 point for an honest effort but very few correct answers;
  - 2 points for mostly correct answers but some major mistakes;
  - 3 points for one or fewer mistakes.

Solutions are given below in blue.

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.

Course and Project Preparation

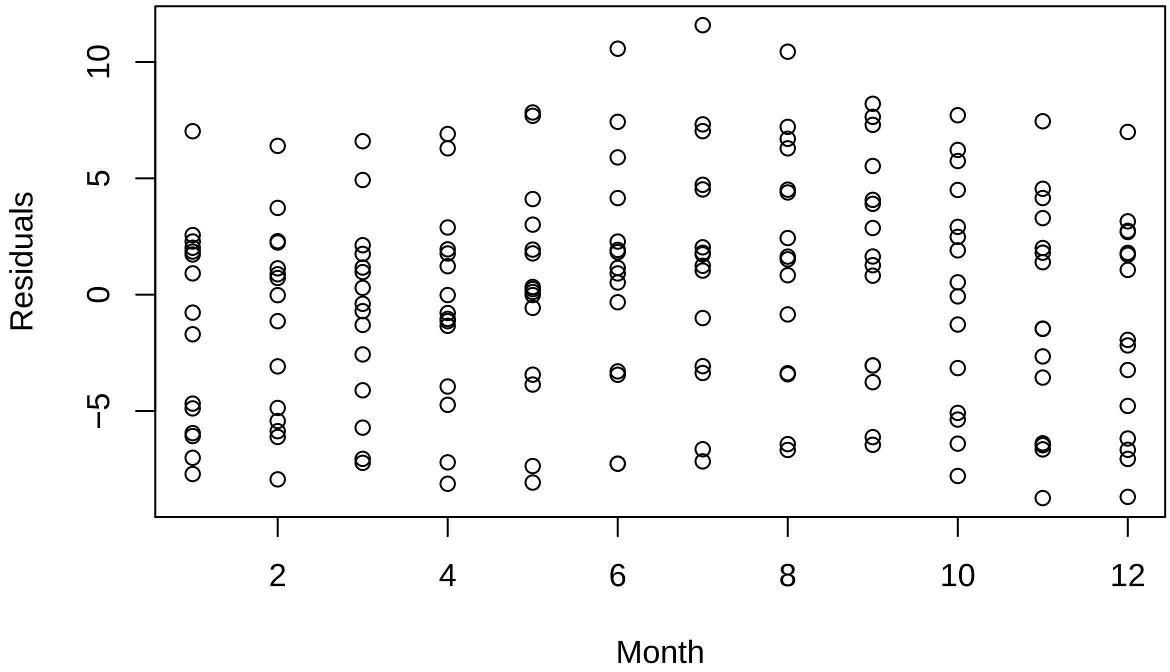
1. List up to three topics you hope we'll cover, from the most to least interesting to you. I will try to incorporate the most popular topics into the last few weeks of the course as time permits.
2. In a few sentences sentences, describe the kind of data you are interested in analyzing using the tools you gain in this course. The goal of this exercise is to get you started thinking about the kind of data you might use for your course project.

Regression Review and Basic Time Series Concepts

3. This problem will require that you work with the `chicken` data from the `astsa` package.

(a) In class, we regressed the the monthly price per pound of chicken on an intercept and the times the chicken prices were measured, $\boldsymbol{t} = (2001 + 7/12, \dots, 2015 + 11/12)$. Plot the residuals from this regression as a function of the month.

```
library(astsa)
data(chicken)
fit <- lm(chicken~time(chicken))
month <- round((time(chicken) - floor(time(chicken)))*12, 0) + 1
plot(month, fit$residuals, xlab = "Month", ylab = "Residuals")
```



(b) Using `lm`, regress monthly price per pound of chicken on an intercept, $\boldsymbol{t}$, and indicators for the month the price was recorded.

```
fit.new <- lm(chicken~time(chicken)+factor(month))
```

(c) The `lm` function used in part (b) returns an estimate of σ_w. Is this the unbiased estimate s_w^2, or the maximum likelihood estimate $\hat{\sigma}_w^2$?

The `lm` function returns the unbiased estimate. We can verify this by computing the unbiased estimate and comparing it to the output of `lm`.

```
lm.sig <- summary(fit.new)$sig
s.w <- sqrt(sum((chicken - fit.new$fitted.values)^2)/(length(chicken) - length(coef(fit.new))))
```

The output of `lm` is `4.712`, which is the same as the square root of the unbiased estimate `4.712`.

(d) Describe and interpret the results of an F-test of the model used in (a) versus the regression model used in (b).

```
ano <- anova(fit, fit.new)
```

A level-0.05 F-test of the null hypothesis that the regression coefficients for the month indicators are exactly equal to zero fails to reject the null, with p-value of `0.55`. This means that the F-test indicates that we should prefer the smaller model, with just an intercept and a linear function of time.

(e) Compute AIC values for the the model used in (a) and the model used in (b). Based on AIC, which model provides a better fit? Do AIC and the F-test agree?

There are two ways we could compute AIC values - the first would be to use the formula we discussed in class.

```
n <- length(chicken)
k.fit <- length(coef(fit))
k.fit.new <- length(coef(fit.new))
ss.fit <- mean((chicken - fit$fitted.values)^2)
ss.fit.new <- mean((chicken - fit.new$fitted.values)^2)
aic.fit <- log(ss.fit) + (n + 2*k.fit)/n
aic.fit.new <- log(ss.fit.new) + (n + 2*k.fit.new)/n
```

The smaller model has a lower AIC value of 4.1046, and the larger model has a higher AIC value of 4.1698. This suggests that we should choose the smaller model over the larger model with indicators for month included. This is consistent with what the F-test indicated.

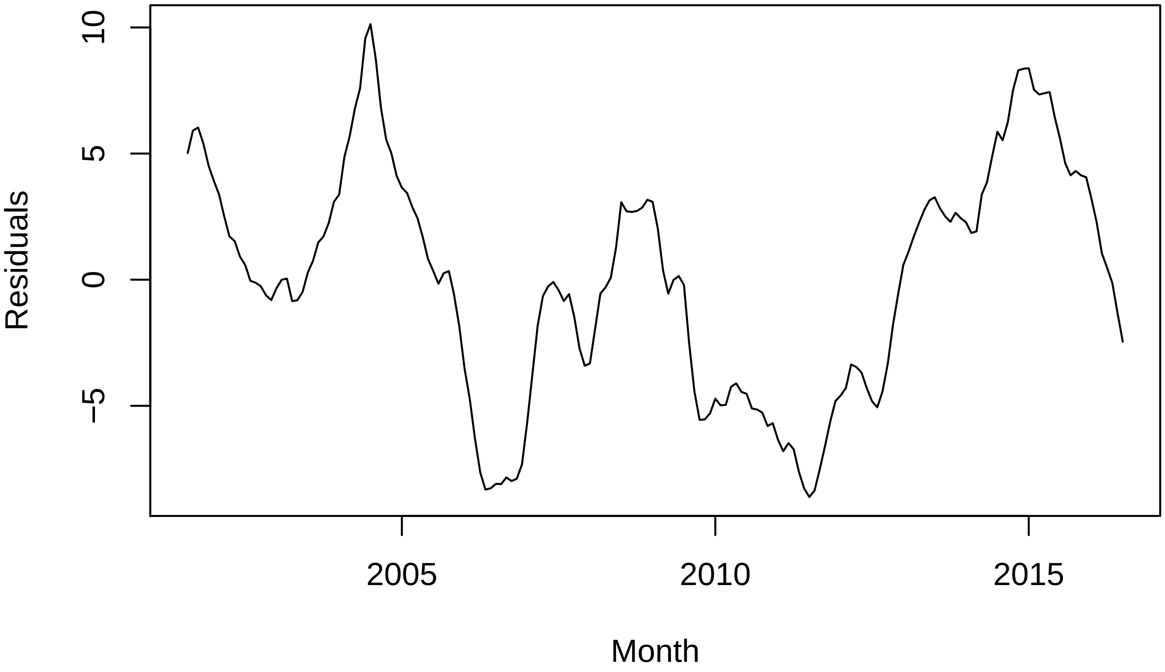
Alternatively, we could calculate the AIC values using the AIC function. Recall that AIC is only meaningful relatively. The AIC function gives different values, but the same overall conclusions because it returns AIC on a different scale than the formula we used in class. Specifically, the AIC function includes a constant and does not divide by n.

```
AIC.fit <- AIC(fit)
AIC.fit.new <- AIC(fit.new)
```

The smaller model has a lower AIC value of 1071.6507, and the larger model has a higher AIC value of 1083.3789. This suggests that we should choose the smaller model over the larger model with indicators for month included. This is consistent with what the F-test indicated.

(f) Plot the residuals from model (b) as a function of time.

```
plot(time(chicken), fit.new$residuals, xlab = "Month", ylab = "Residuals", type = "l")
```



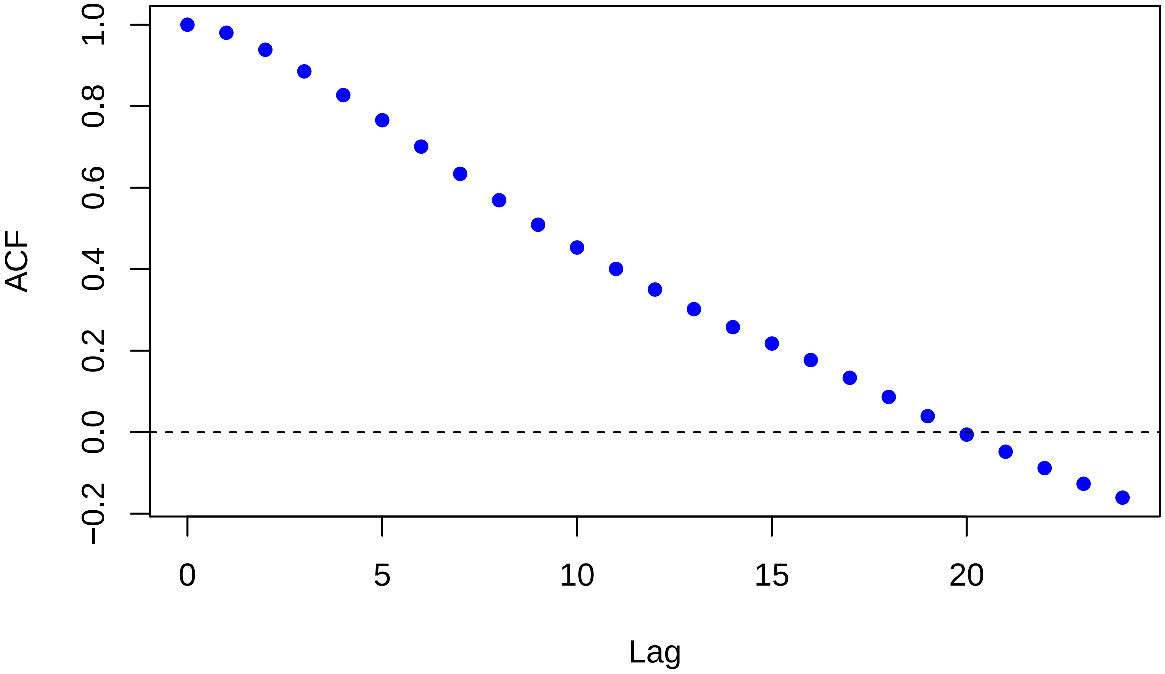
(g) Referring back to (f), do you see evidence of correlated residuals across time?

Yes - the residuals that are closer together in time are more similar to each other than residuals that are far apart in time.

(h) Plot the sample autocorrelation function of the residuals from model (b) for $h \leq 24$ without using the `acf` function or any other third party function that automatically computes the ACF. Include a dashed horizontal line at 0. Revisit (g) - do you see evidence of correlated residuals across time?

```
r <- fit.new$residuals
h <- seq(0, 24, by = 1)
acvf <- numeric(length(h))
time.diff <- outer(1:n, 1:n, "-")
rr <- outer(r, r, "*")
for (i in 1:length(h)) {
  acvf[i] <- acvf[i] + (sum(rr[which(time.diff == h[i])]))/n
}
```

```r
acrf <- acvf/acvf[1]
plot(h, acrf, ylab = "ACF", xlab = "Lag", pch = 16, col = "blue")
abline(h = 0, lty = 2)
```



We do see evidence of correlated residuals across time - many sample autocorrelation function estimates are much larger than 0.

You might notice that I did not subtract off the mean when I computed the sample autocorrelation function - this was because since our regression included an intercept, the residuals $\boldsymbol{r}$ will have mean zero, i.e. $\bar{r} = \left(\sum_{i=1}^{n} r_i\right)/n = 0$. You can verify this yourself by computing `mean(r)`.