Homework 2 Solutions

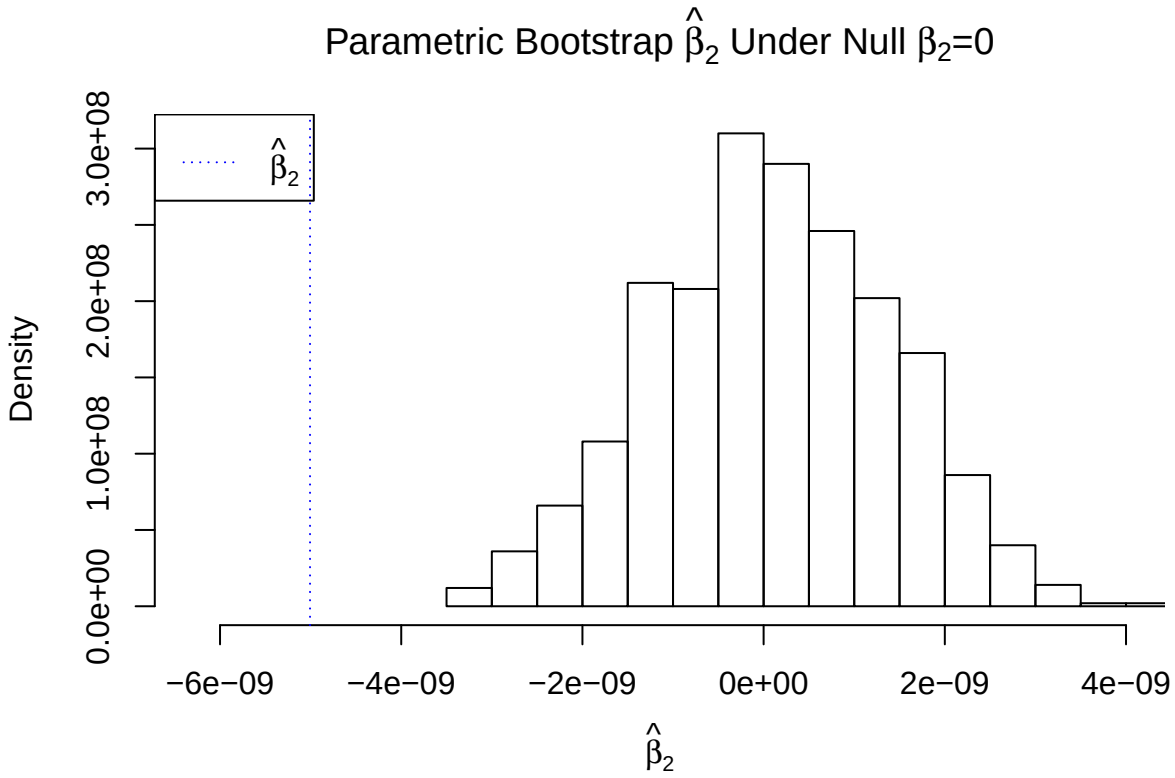
Due: Tuesday 2/4/20 by 10:00am

Continued Regression and R Review

1. This problem will require that you continue to 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. We will consider the following three regression models:
   1. $\texttt{price}_i = \mu + \beta_1 \texttt{days since start}_i + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$
   2. $\texttt{price}_i = \mu + \beta_1 \texttt{days since start}_i + \beta_2 \texttt{days since start}_i^2 + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$
   3. $\texttt{price}_i = \mu + \phi_1 \texttt{z}_{1i} + \phi_2 \texttt{z}_{2i}^2 + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$, where $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ correspond to the orthogonal polynomials of degree 1 and 2, respectively, over the set of points given by **`days since start`**. You can use the `poly` function to construct the orthogonal polynomials.

   (a) Using the parametric bootstrap, choose between Model 1 and Model 2. Justify your choice in at most one sentence, and provide any relevant numerical evidence.

```
load("~/Dropbox/Teaching/TimeSeries2020/stat697/content/data/broc.RData")
set.seed(1)
broc$fdate <- as.Date(broc$date, "%Y-%m-%d")
broc$month <- format(broc$fdate, "%m")
broc$dayssincestart <- as.numeric(broc$fdate) - min(as.numeric(broc$fdate))
linmod1 <- lm(price~dayssincestart+factor(month), data = broc)
linmod2 <- lm(price~dayssincestart+I(dayssincestart^2)+factor(month), data = broc)
nboot <- 1000
beta.vals <- rep(NA, nboot)
for (i in 1:nboot) {
  by <- rnorm(nrow(broc), mean = fitted.values(linmod1), sd = summary(linmod1)$sigma)
  blinmod2 <- lm(by~dayssincestart+I(dayssincestart^2)+factor(month), data = broc)
  beta.vals[i] <- blinmod2$coef[3]
}
hist(beta.vals, xlab = expression(hat(beta)[2]), freq = FALSE,
     main = expression(paste("Parametric Bootstrap ", hat(beta)[2],
                             " Under Null ", beta[2], "=0", sep = "")),
     xlim = c(min(linmod2$coef[3] - sd(beta.vals), beta.vals),
              max(beta.vals)))
abline(v = linmod2$coef[3], col = "blue", lty = 3)
legend("topleft", lty = 3, col = "blue",
       legend = expression(hat(beta)[2]))
```



The parametric bootstrap indicates that the estimate we obtain for $\hat{\beta}_2$ by fitting Model 2 is much smaller than we would expect if β_2 were equal to zero and Model 1 were true, so we would choose Model 2.

(b) In at most one sentence, explain whether or not it is appropriate to choose between Models 1 and 3 using a t/z-test, an F-test, or parametric bootstrap versions of either.

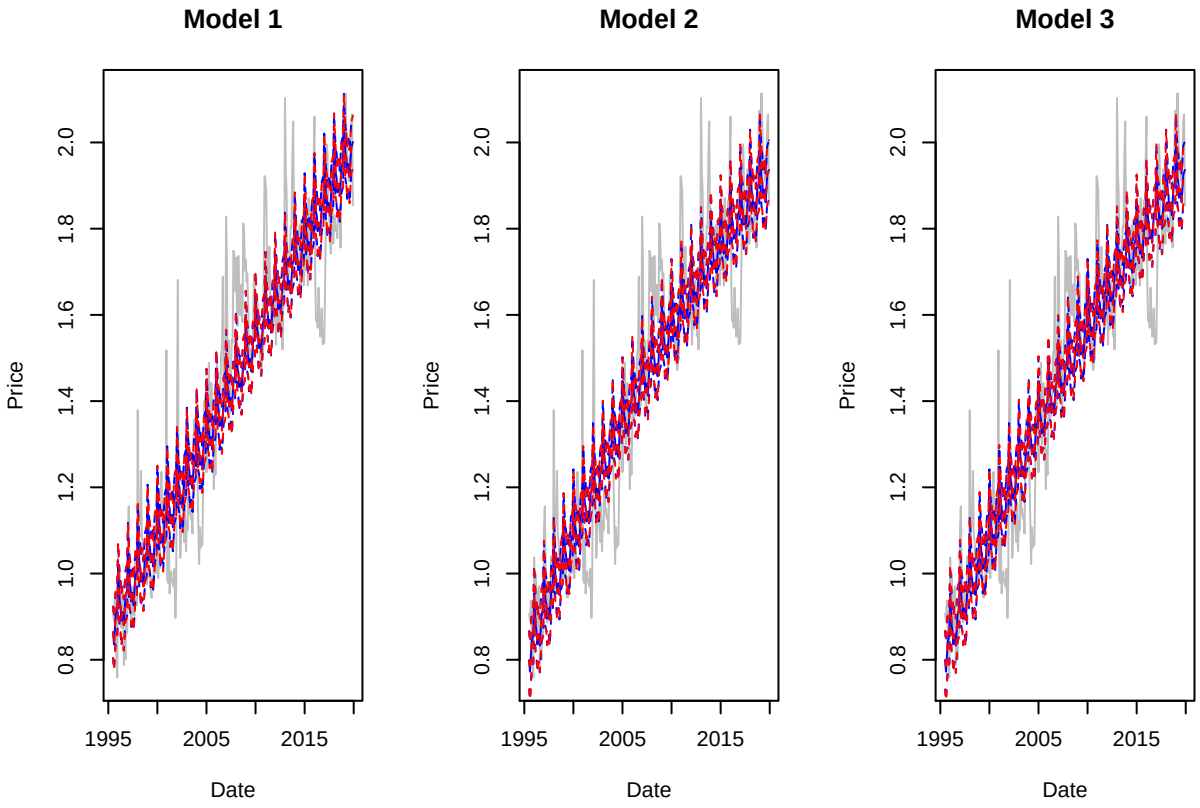
Models 1 and 3 are not nested, so it does not really make sense to choose between them based on a t-test, a z-test, an F-test, or parametric bootstrap versions of either.

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

```
linmod3 <- lm(price~poly(dayssincestart, 2)+factor(month), data = broc)

nboot <- 1000
par(mfrow = c(1, 3))
for (i in 1:3) {
    plot(broc$fdate, broc$price, type = "l", xlab = "Date", ylab = "Price",
         main = paste("Model ", i, sep = ""), col = "gray")
    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)
    fit.vals <- matrix(NA, nrow = nboot, ncol = nrow(broc))
    for (j in 1:nboot) {
        by <- rnorm(nrow(broc), mean = fitted.values(get(paste("linmod", i, sep = ""))),
                    sd = summary(get(paste("linmod", i, sep = "")))$sigma)
```

```
        if (i == 1) {
            blinmod <- lm(by~dayssincestart+factor(month), data = broc)
        } else if (i == 2) {
            blinmod <- lm(by~dayssincestart+I(dayssincestart^2)+factor(month), data = broc)
        } else {
            blinmod <- lm(by~poly(dayssincestart, 2)+factor(month), data = broc)
        }
        fit.vals[j, ] <- fitted.values(blinmod)
    }
        lines(broc$fdate, pred$fit + qnorm(0.975)*apply(fit.vals, 2, sd), col = "red",
              lty = 2)
        lines(broc$fdate, pred$fit + qnorm(0.025)*apply(fit.vals, 2, sd), col = "red",
              lty = 2)
        lines(broc$fdate, apply(fit.vals, 2, mean), col = "red",
              lty = 2)
}
```



(d) For Models 1, 2 and 3, compute predictions and standard errors for $\mathbb{E}[y_{295}]$.

```
broc.new <- data.frame("dayssincestart"=
                         broc$dayssincestart[length(broc$dayssincestart)] + 31,
                       "month"=
                         "01")
pred1 <- predict(linmod1, broc.new, se.fit = TRUE)
pred2 <- predict(linmod2, broc.new, se.fit = TRUE)
pred3 <- predict(linmod3, broc.new, se.fit = TRUE)
```

The predicted values for each model are 2.097, 2.029, and 2.029, respectively. The corresponding standard errors for each model are 0.031, 0.036, and 0.036, respectively.

(e) For Models 1, 2, and 3, compute a parametric bootstrap mean and standard error for $\mathbb{E}[y_{295}]$. In at most one sentence, justify your choice of the number of bootstrap samples. Compare to the results from (d), and comment on the differences in at most one sentence.

```
nboot <- 10000

preds <- matrix(nrow = nboot, ncol = 3)
for (j in 1:nboot) {

  for (i in 1:3) {
  by <- rnorm(nrow(broc), mean = fitted.values(get(paste("linmod", i, sep = ""))),
                    sd = summary(get(paste("linmod", i, sep = "")))$sigma)
        if (i == 1) {
            blinmod <- lm(by~dayssincestart+factor(month), data = broc)
        } else if (i == 2) {
            blinmod <- lm(by~dayssincestart+I(dayssincestart^2)+factor(month), data = broc)
        } else {
            blinmod <- lm(by~poly(dayssincestart, 2)+factor(month), data = broc)
        }
  preds[j, i] <- predict(blinmod, broc.new)
  }
}
pred.means <- colMeans(preds)
pred.ses <- apply(preds, 2, sd)
```

The parametric bootstrap mean predictions for each model are 2.098, 2.029, and 2.028, respectively. The corresponding standard errors for each model are 0.031, 0.036, and 0.035, respectively. I used 25,000 bootstrap samples, because I kept increasing the bootstrap size until the forecasts and standard errors from the two models that are identical (Models 2 and 3) differed by at most 0.001.

(f) Choose between Models 1, 2, and 3 using leave-one-out cross validation, using out-of-sample prediction errors to assess model fit. Justify your choice in at most one sentence.

```
y <- broc$price
X1 <- model.matrix(~dayssincestart+factor(month), data = broc)
X2 <- model.matrix(~dayssincestart+I(as.numeric(dayssincestart)^2)+
                     factor(month), data = broc)
X3 <- model.matrix(~poly(dayssincestart, 2)+
                     factor(month), data = broc)

test.mse1 <- matrix(nrow = nrow(broc), ncol = 3)

for (o in 1:nrow(broc)) {
   y.test <- y[o]
   y.train <- y[-o]

   X1.test <- X1[o, , drop = FALSE]
   X2.test <- X2[o, , drop = FALSE]
   X3.test <- X3[o, , drop = FALSE]

   X1.train <- X1[-o, , drop = FALSE]
   X2.train <- X2[-o, , drop = FALSE]
   X3.train <- X3[-o, , drop = FALSE]

   beta1 <- lm(y.train~X1.train-1)$coef
   beta2 <- lm(y.train~X2.train-1)$coef
```

```
   beta3 <- lm(y.train~X3.train-1)$coef

   fit.test1 <- X1.test%*%beta1
   fit.test2 <- X2.test%*%beta2
   fit.test3 <- X3.test%*%beta3

   test.mse1[o, 1] <- mean((y.test - fit.test1)^2)
   test.mse1[o, 2] <- mean((y.test - fit.test2)^2)
   test.mse1[o, 3] <- mean((y.test - fit.test3)^2)
}
```

Based on the results of leave-one-out cross validation, we would choose Model 2 or 3, which are in fact different representations of the same model.

(g) Choose between Models 1, 2, and 3 using leave-five-out cross validation, using out-of-sample prediction errors to assess model fit. Justify your choice in at most one sentence. In at most one additional sentence, describe any practical challenges you encounter.

```
test.mse5 <- matrix(nrow = nrow(broc), ncol = 3)

for (o in 1:nrow(broc)) {
   os <- sample(1:nrow(broc), 5, replace = FALSE)
   nos <- which(!1:nrow(broc) %in% os)
   y.test <- y[os]
   y.train <- y[nos]

   X1.test <- X1[os, , drop = FALSE]
   X2.test <- X2[os, , drop = FALSE]
   X3.test <- X3[os, , drop = FALSE]

   X1.train <- X1[nos, , drop = FALSE]
   X2.train <- X2[nos, , drop = FALSE]
   X3.train <- X3[nos, , drop = FALSE]

   beta1 <- lm(y.train~X1.train-1)$coef
   beta2 <- lm(y.train~X2.train-1)$coef
   beta3 <- lm(y.train~X3.train-1)$coef

   fit.test1 <- X1.test%*%beta1
   fit.test2 <- X2.test%*%beta2
   fit.test3 <- X3.test%*%beta3

   test.mse5[o, 1] <- mean((y.test - fit.test1)^2)
   test.mse5[o, 2] <- mean((y.test - fit.test2)^2)
   test.mse5[o, 3] <- mean((y.test - fit.test3)^2)
}
```

Based on the results of leave-five-out cross validation, we would choose Model 2 or 3, which are in fact different representations of the same model.

(h) Choose between Models 1, 2, and 3 using leave-ten-out cross validation, using out-of-sample prediction errors to assess model fit. Justify your choice in at most one sentence. In at most one additional sentence, describe any challenges you encounter.

```
test.mse10 <- matrix(nrow = nrow(broc), ncol = 3)
```

```
for (o in 1:nrow(broc)) {
   os <- sample(1:nrow(broc), 10, replace = FALSE)
   nos <- which(!1:nrow(broc) %in% os)
   y.test <- y[os]
   y.train <- y[nos]

   X1.test <- X1[os, , drop = FALSE]
   X2.test <- X2[os, , drop = FALSE]
   X3.test <- X3[os, , drop = FALSE]

   X1.train <- X1[nos, , drop = FALSE]
   X2.train <- X2[nos, , drop = FALSE]
   X3.train <- X3[nos, , drop = FALSE]

   beta1 <- lm(y.train~X1.train-1)$coef
   beta2 <- lm(y.train~X2.train-1)$coef
   beta3 <- lm(y.train~X3.train-1)$coef

   fit.test1 <- X1.test%*%beta1
   fit.test2 <- X2.test%*%beta2
   fit.test3 <- X3.test%*%beta3

   test.mse10[o, 1] <- mean((y.test - fit.test1)^2)
   test.mse10[o, 2] <- mean((y.test - fit.test2)^2)
   test.mse10[o, 3] <- mean((y.test - fit.test3)^2)
}
```

Based on the results of leave-twn-out cross validation, we would choose Model 2 or 3, which are in fact different representations of the same model.

(i) In at most one sentence, explain which procedure you would use to choose a model from those described in (f), (g), and (h) and justify your choice.

In this case, I would choose leave-ten-out cross validation because it is least robust to individual outliers, however in this case, all three procedures lead to the same conclusion - that we should choose Model 2 or Model 3, which are different represenations of the same model.