Problem Set 5: More Regression Modeling

Linh Vu (Collab: Brice Laurent)

Due: October 27, 2023

This assignment is **due Friday, October 27 at 11:59pm**, handed in on Gradescope (remember, there are two separate submissions, one for your pdf, and another for you rmd file). Show your work and provide clear, explanations when asked. **Incorporate the relevant R output in this R markdown file**. Only the key output should be displayed for each problem and the relevant parts should be **highlighted** in some way. Make sure that you write-up any interpretation of R-code in your own words (don't just provide the output).

Collaboration policy (for this and all future homeworks): You are encouraged to discuss the problems with other students, but you must write up your solutions yourself and in your own words. Copying someone else's solution, or just making trivial changes is not acceptable.

Problem 1.

Let $X \sim \text{Unif}(a, b)$. Feel free to use results from the *Stat 110 Distribution Sheet* as seen on the midterm exam.

```
set.seed(139)
a = -0.5
b = 0.5
x = runif(10^6, a, b)
x2 = x^2
cov(x, x2)
```

```
## [1] -5.103786e-07
```

```
(a+b)*(a-b)^2/12
```

```
## [1] 0
```

(a) Determine the covariance between X and X^2.

Using the formula $Cov(X, Y) = E(XY) - E(X)E(Y)$, we get:

$$Cov(X, X^2) = E(X^3) - E(X)E(X^2)$$

Find $E(X^2)$ using LOTUS: $E(X^2) = \int_a^b X^2 \frac{1}{b-a} dX = \frac{b^3-a^3}{3(b-a)}$

Find $E(X^3)$ using LOTUS: $E(X^3) = \int_a^b X^3 \frac{1}{b-a} dX = \frac{b^4-a^4}{4(b-a)}$

Combining everything, we get:

$$Cov(X, X^2) = \frac{b^4 - a^4}{4(b-a)} - \frac{a+b}{2}\frac{b^3 - a^3}{3(b-a)}$$

$$= \frac{3(b^2 - a^2)(b^2 + a^2) - 2(a + b)(b^3 - a^3)}{12(b - a)}$$

Dividing both the numerator and denominator by $b - a$, we get:

$$= \frac{3(a + b)(b^2 + a^2) - 2(a + b)(a^2 + ab + b^2)}{12}$$

$$= \frac{(a + b)(3a^2 + 3b^2 - 2a^2 - 2ab - 2b^2)}{12}$$

$$= \frac{(a + b)(a - b)^2}{12}$$

(b) Assume $b - a = 1$ (so that the variability of X is fixed). For what values of a (and b) will this covariance be zero? When will this covariance be large (and positive)? When will it be negative (and large in magnitude)? What does this mean for where the distribution of X is centered in each case?

Using results from part (a), we know that the covariance in this case is $\frac{a+b}{12}$

The covariance is 0 when $a + b = 0$. And since $b - a = 1$, we find that $a = -0.5$, $b = 0.5$.

The covariance is large and positive when $a + b$ is large and positive (i.e. a and b are both large and 1 unit apart). Similarly, the covariance is negative and large in magnitude when $a + b$ is negative and large in magnitude (i.e. a and b are both negative and large in magnitude). When X is centered at a negative value far from 0, the covariance is very negative, and if the center is positive and far from 0, the covariance is large and positive.

(c) What are the implications of the results in (b) for a quadratic regression model (when X and X^2 are both used as predictors)? Is this phenomenon specific to the Uniform distribution?

*Note: do not be afraid to check your answers empirically using R.

Because X and X^2 very rarely have covariance of 0 (i.e. not at all correlated), when X and X^2 are both used as predictors, we face potential issues of multicollinearity. As a result, we need to be careful about including many polynomial terms. This phenomenon is not specific to the Uniform distirbution.???

Problem 2.

The file 'pregnancydata.csv' includes several variables to model the birthweight of babies (measured through an online survey). Those variables are defined below. Use this data set in R to answer the questions below:

id:	a unique identifier of the mother
weight:	birthweight of the newborn baby, in ounces
pregnancylength:	the length of the pregnancy, in days
country:	where the birth took place with categories United States (US), United Kingdom (UK), Canada (Can), and Other
motherage:	age of mother at childbirth, in years
multiples:	whether the baby was a 1=singleton or 2=twin
sex:	sex of the baby: girl or boy
induced:	a binary indicator for whether labor was induced with oxytocin
cesarean:	a binary indicator for whether a cesarean (c-section) was performed
previousbirths:	the number of births by the mother previous to this recorded one (from 0 to 10)

(a) Fit a regression model to predict weight from country and use the `relevel` command to make the "Other" group the reference group (call this **Model 1**). Interpret the results and provide a visual to support your conclusions.

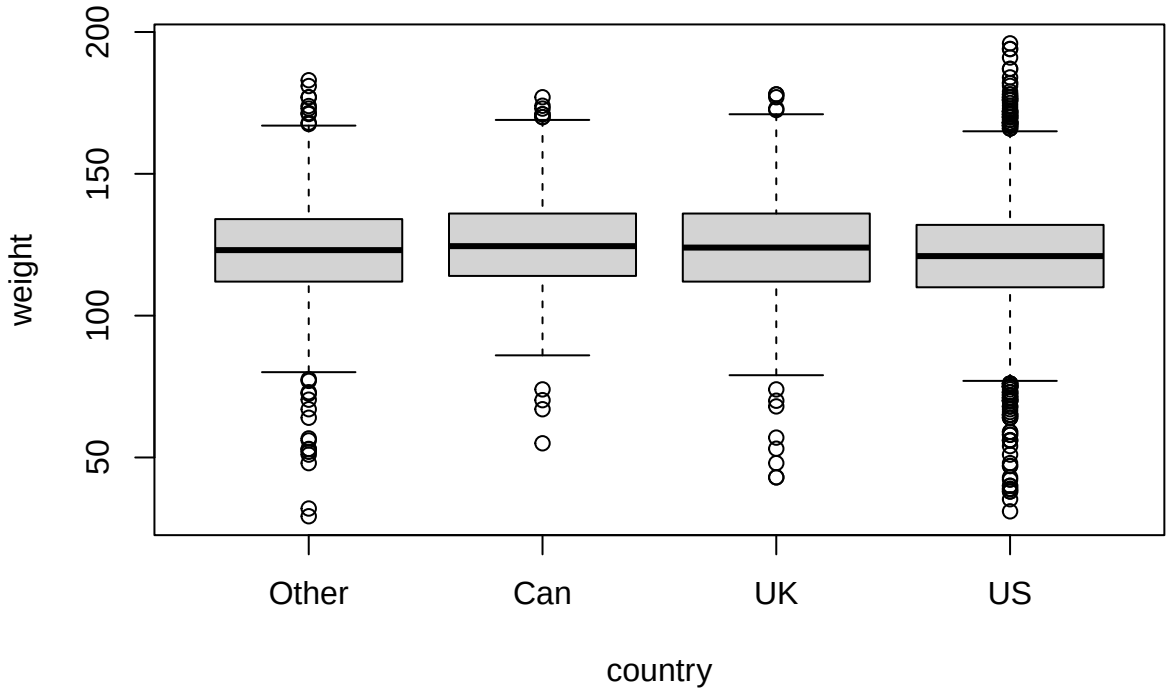
```
pregnancy <- read.csv("data/pregnancydata.csv")
pregnancy$country = relevel(as.factor(pregnancy$country), "Other")

mod1 <- lm(weight~country, pregnancy)
summary(mod1)
```

```
##
## Call:
## lm(formula = weight ~ country, data = pregnancy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.380 -11.311  -0.311  11.383  74.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 122.6570     0.6083 201.631   <2e-16 ***
## countryCan    2.2965     0.9712   2.365   0.0181 *
## countryUK     0.9596     0.7868   1.220   0.2227
## countryUS    -1.3458     0.6480  -2.077   0.0378 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 9061 degrees of freedom
## Multiple R-squared:  0.004002,   Adjusted R-squared:  0.003672
## F-statistic: 12.13 on 3 and 9061 DF,  p-value: 6.366e-08
```

```
x = c("Other", "Can", "UK", "US")
yhat = predict(mod1, new=data.frame(country=x))
plot(weight~country, pregnancy)
lines(yhat~x,col="red",lwd=3)
```

```
## Warning in xy.coords(x, y): NAs introduced by coercion
```




```
#???
```

The intercept means that babies born in other countries weigh 122.657 ounces on average. The other slope estimates mean the difference between average weigh of babies born in Canada, UK, US and those born in other countries. Specifically, compared to babies born in other countries, babies born in Canada weigh 2.297 ounce more; babies born in the UK weigh 0.96 ounce more; babies born in the US weigh 1.346 ounce less.

The side-by-side boxplot shows that babies in Canada weigh slightly more on average, and babies in the US weigh slightly less on average.

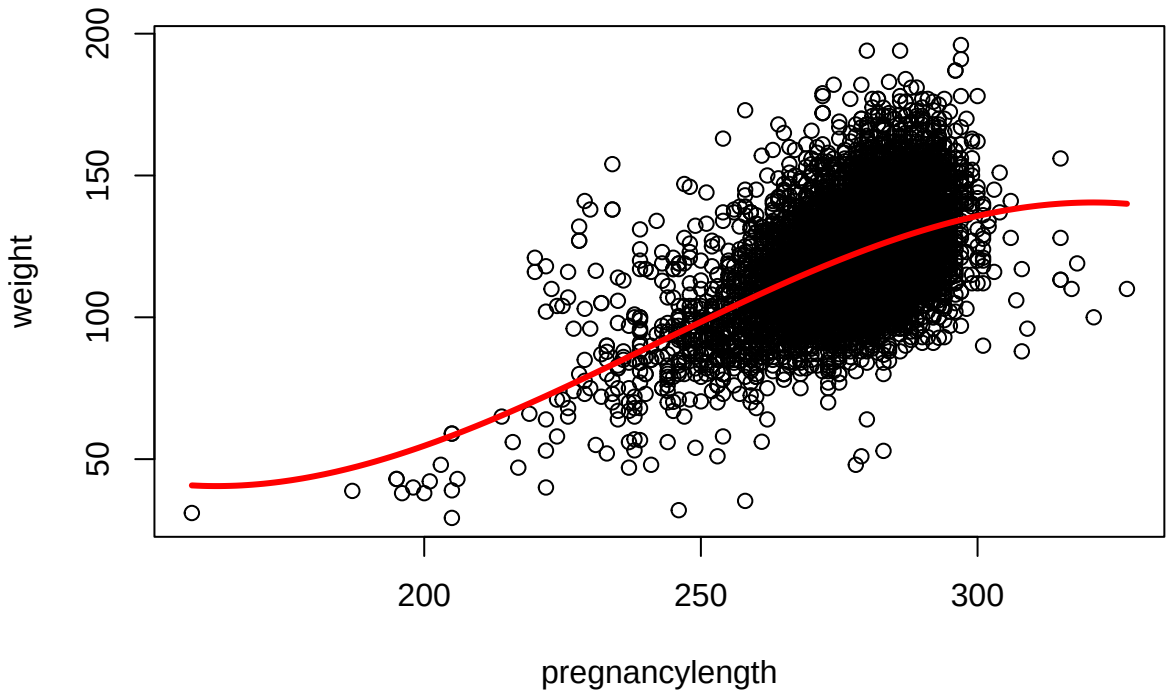
(b) Build a 3^{rd} order polynomial regression model to predict weight from `pregnancylength` (call this **Model 2**). Interpret the output and provide a visual to support the results of the model.

```
mod2 <- lm(weight~poly(pregnancylength, 3, raw=T), pregnancy)
summary(mod2)
```

```
##
## Call:
## lm(formula = weight ~ poly(pregnancylength, 3, raw = T), data = pregnancy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.486 -10.087  -0.761   9.364  70.727
##
## Coefficients:
```

```
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          5.712e+02  2.000e+02   2.856  0.00431 **
## poly(pregnancylength, 3, raw = T)1 -7.861e+00  2.353e+00  -3.341  0.00084 ***
## poly(pregnancylength, 3, raw = T)2  3.645e-02  9.195e-03   3.964 7.44e-05 ***
## poly(pregnancylength, 3, raw = T)3 -5.028e-05  1.193e-05  -4.213 2.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.33 on 9061 degrees of freedom
## Multiple R-squared:  0.2627, Adjusted R-squared:  0.2625
## F-statistic:  1076 on 3 and 9061 DF,  p-value: < 2.2e-16
```

```
x = min(pregnancy$pregnancylength):max(pregnancy$pregnancylength)
yhat = predict(mod2, new=data.frame(pregnancylength=x))
plot(weight~pregnancylength, pregnancy)
lines(yhat~x,col="red",lwd=3)
```



$\hat{\beta}_0$: when length of the pregnancy is 0, the baby weighs 571.16 ounces on average.

$\hat{\beta}_1$: when length of the pregnancy is 0, the baby's weight changes 571.16 ounces on average.???

(c) Use **Model 2** to estimate the probability that a baby will weigh less than 7 pounds (112 ounces) when born on day 280.

```
#???
new.data=data.frame(pregnancylength=280)
predict(mod2, type="response")
```

(d) It is of medical interest to determine at what gestational age a developing fetus is gaining weight the fastest. Use **Model 2** to estimate this *period of fastest growth*.

From model 2, we know that the line of best fit is

$$\hat{weight} = 517 - 7.86 \cdot length + 3.65(10)^{-2} \cdot length^2 - 5.02(10)^{-5} \cdot length^3$$

We take the 2nd derivative with respect to length to find the period of fastest growth

$$2(3.65)(10)^{-2} - 6(5.02)(10)^{-5} \cdot length = 0$$

Solving for length gives us $length = 242.36$, meaning that the period of fastest growth is estimated as day 242 according to model 2.

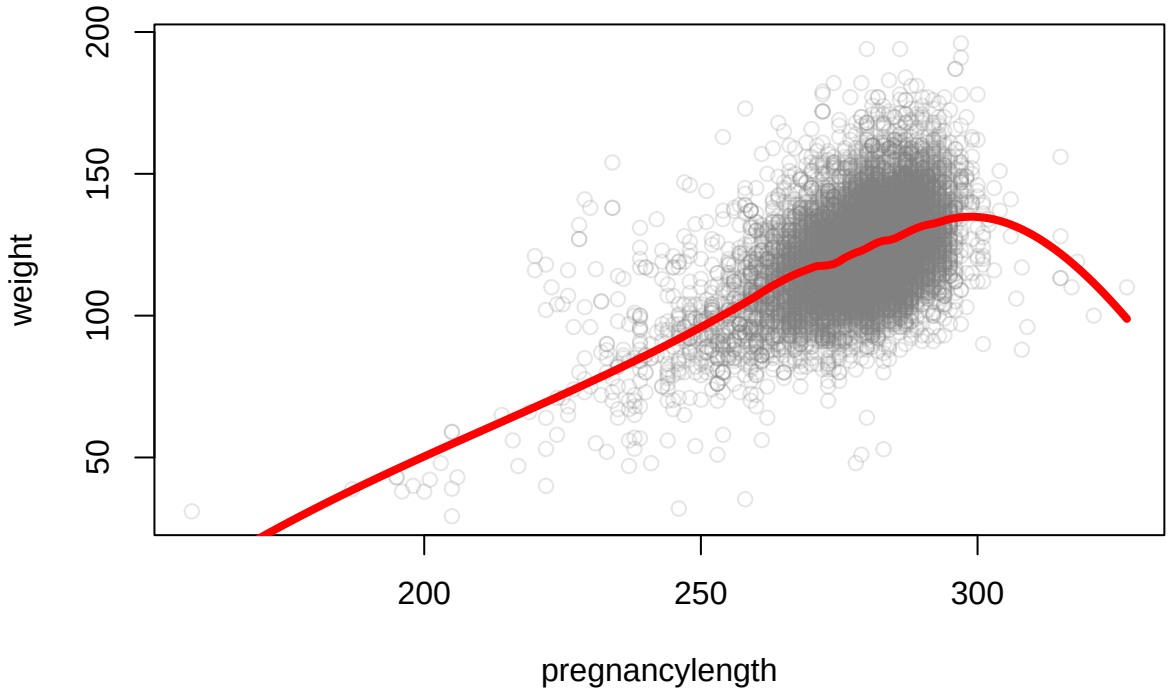
(e) Fit a LOESS model (call this **Model 3**) to predict weight from `pregnancylength` (use a smaller span of 0.3). Provide a visual to support the results of the model. How does this model compare to **Model 2** in its prediction accuracy?

The LOESS model seems to have higher accuracy because it captures the true trend of baby's weight: it increases as the pregnancy progresses and perhaps plateaus/decreases if the pregnancy is abnormally long. On the other hand, the polynomial regression model has a weird/unrealistic curve that starts high and decreases over time before increasing at around day 150 (in reality, babies born prematurely do not weigh that much, and baby's weight increases over time)

```
mod3 <- loess(weight~pregnancylength, pregnancy, span=0.3)

x=min(pregnancy$pregnancylength):max(pregnancy$pregnancylength)
yhat=predict(mod3, new=data.frame(pregnancylength=x))

plot(weight~pregnancylength, pregnancy, col=rgb(0.5, 0.5, 0.5, 0.2))
lines(yhat~x,col="red",lwd=4)
```


weight
50
100
150
200
200
250
300
pregnancylength

Problem 3.

In this problem, we will attempt to investigate whether the COVID-19 related restrictions imposed by the government had any effect on the reporting of criminal activity in the Boston Police Department (BPD). We will be using the same combined dataset from last time (now named 'bpd.csv') that includes the number of daily incident reports filed (`count`) and various weather indicators on those days (`maxtemp` is the only weather variable we will use in this problem). Note: we also used these data in Pset 3.

Note: a state of emergency was declared in Massachusetts on March 10, 2020, and restrictions on non-essential businesses, schools, and MBTA service were mainly put into effect on March 17,2020 (see this City of Boston article for the timeline).

The R chunk below reads in the data and includes some code to create a variable called `dayinyear` in the `bpd` data frame that counts the number of days into the year, starting with 0 for Jan 1 (similar to what was done on the previous pset).

```
bpd = read.csv('data/bpd.csv')

jan1_19 = as.Date("1/1/19",format="%m/%d/%y")
jan1_20 = as.Date("1/1/20",format="%m/%d/%y")
jan1_21 = as.Date("1/1/21",format="%m/%d/%y")

bpd$dayinyear = as.Date(bpd$date,format="%m/%d/%y") - jan1_19
bpd$dayinyear[bpd$year==2020] =
  as.Date(bpd$date,format="%m/%d/%y")[bpd$year==2020] - jan1_20
bpd$dayinyear[bpd$year==2021] =
  as.Date(bpd$date,format="%m/%d/%y")[bpd$year==2021] - jan1_21
```

(a) Create a binary/dummy variable (call it `restrictions`) to indicate whether that day falls under the time period of state of emergency or restricted business operations in the city of Boston (all dates between and including March 10, 2020 and Friday, May 28, 2020). How many days fall in this time period in the data set?

(b) Calculate the mean number of daily incident reports filed by the BPD during the restriction orders in Boston. Separately calculate the mean number of daily incident reports for a comparison group with the same calendar dates in the pre-pandemic portion of the data. Use these two groups to calculate a reasonable 95% confidence interval for the effect of COVID-19 restrictions on the reporting of crime in the BPD (based on a simple 2-group comparison method and not linear regression).

(c) Fit a linear regression model to predict `count` from `maxtemp` and `restrictions` (call it **model1**), and print out the `summary` results. Briefly interpret the coefficient estimates and use this model to estimate the effect of COVID-19 restrictions on the reporting of crime in the BPD (with 95% confidence).

(d) Fit a linear regression model to predict `count` from `maxtemp`, `restrictions`, `dayinyear` and all 2-way interactions between these 3 predictors (call it **model2**), and print out the `summary` results. Interpret what this model says about the relationship between crime reporting in the BPD and COVID-19 restrictions. Compute an estimate and 95% CI for the effect of restrictions on the 100th day of the year, assuming a `maxtemp` of 60 degrees. Also estimate `count` (and provide a 95% CI) on the 91st day of the year in 2020, assuming the temperature was 50 degrees. Do the same for 2019 and compare the difference.

(e) Perform a formal hypothesis test to determine whether **model2** performs significantly better at predicting `count` than **model1**.

(f) Investigate the assumptions for **model2**. Be sure to include and reference useful visuals.

(g) Determine which 4 dates **model2** did the worst job at predicting `count`. Can you think of a reason why any of these dates do not follow the relationships in this model? (all 4 are explainable with a little Google searching)

(h) Write a 200-300 word summary addressing whether there is evidence that COVID-19 reduced the amount of crime in Boston. Be sure to reference the results above (specifically, which approach you think was most reasonable) and mention any biases or confounders that may be present in this relationship.

Problem 4.

Perform a simulation study (with 1,000 iterations) where the data are **generated** from the following sin function:

$$Y_i = \sin(X_i) + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2 = 0.1^2)$ and independent, and X_i are sampled independently from a Unif$(a = 0, b = 6)$, for $n = 50$ observations.

For each iteration, fit 4 different polynomial models (use the raw form in R): (i) 3^{rd} order, (ii) 5^{th} order, (iii) 7^{th} order, and (iv) $= 9^{th}$ order. Save the β_1 coefficient (linear term) estimates for each of the 4 models for each of the 1,000 iteration (presumably a 1000x4 matrix) and either separately save all 10 β coefficient estimates for the 9^{th} order or the model objects themselves (in a list).

Evaluate which model is *best* in each iteration two ways: (i) based on sequential ESS F-tests (you will perform 3 of them in each iteration) and (ii) out-of-sample mean squared error (based on a single test set of $n_{test} = 1000$ generated from the same data generating process as the regular $n = 50$ set of observations... this does not need to be recreated in each iteration).

(a) Based on the ESS F-tests, how often is each of the 4 models considered the best? Based on out-of-sample mean squared error?

(b) Which metric is more conservative when it comes to overfitting? How do you know?

(c) Plot 10+ $\hat{\mu}_Y$ curves (the predicted curve) based on the estimated 9^{th} order polynomial model (for 10+ iterations): with at least 5 curves for when 9^{th} order model wins and at least 5 curves for when it does not win (based on out-of-sample mean square error). Be sure to color code these curves based on when this model wins vs. when it does not win. Interpret this plot: what does this say about how overfitting affects out-of-sample mean square error?

(d) Provide the boxplots of $\hat{\beta}_1$ estimates in each of the 4 models (should be a side-by-side boxplot with 4 boxplots based on 1000 estimates each). Interpret this plot in context of this situation. What does this illustrate? Why is this not surprising?