DS 3010 Homework 4 Solutions

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code in your solutions.** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework *as a separate file* (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Concept Review

(a) You fit a multiple linear regression model with 3 predictors. The output is seen below:

```
 > summary(lm(y~x1+x2+x3))

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82903 -0.55227  0.07452  0.46413  2.39613

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.87750    0.08782  21.380   <2e-16 ***
x1           2.95472    0.02991  98.774   <2e-16 ***
x2           1.04783    0.02907  36.042   <2e-16 ***
x3          -1.4306     1.47332  -0.971    0.334
```

To gain insight into the model, you propose testing whether the regression coefficient related to X_3 is statistically significantly different from 0. Your collaborator (who knows nothing about data science) says that is a waste of time. They state that it's obvious that $\hat{\beta}_3$ is different from 0 because the estimate is not 0! So why bother doing any hypothesis testing? Provide a clear and compelling explanation for why hypothesis testing is valuable here and why data scientists rely on hypothesis test to draw meaningful conclusions about the data.

Our dataset is just a random sample from the population. Therefore, what we observe from our sample is not the absolute truth, but simply an estimate or approximation of the truth. How do we account for the fact that what we have is just an estimate? Inference procedures, such

as hypothesis test, help us to do that in a rigorous and statistically sound way. Hypothesis test tell us whether or the estimate we observe is statistically different from 0. What does 'statistically different' mean? It answer the question of whether or not we observed -1.4306 *just by chance OR is this estimate truly different from 0. If we got this estimate just by chance, then if we had another random sample, we would expect that estimate to be very close to 0. Therefore, hypothesis testing allows us to* ***generalize*** *our conclusions about a random sample to the entire population. This allows us to draw meaningful conclusions about the true relationship between* X_3 *and our response* Y*.*

(b) This same colleague claims that if you knew the true $f(X)$, then we could obtain perfect prediction. Do you agree or disagree? Explain clearly.

I disagree. We know that the true model for Y *involves both* $f(X)$ *and the irreducible error. We do not know the irreducible error, so we can never obtain perfect prediction even if we knew* $f(X)$*.*

For the following statement, evaluate whether they are True or False and **justify your answer**.

(c) The expected test MSE is defined as: $E(y_0 - \hat{f}(x_0))^2$. Here y_0 is from our training set and $\hat{f}()$ is the model we built from our training set. We evaluate $\hat{f}(x_0)$ on the x_0 values from our test set.

False. y_0 *should be from our test set.*

(d) The bias-variance decomposition tells us that sometimes reducing the complexity of our model (for example, removing a predictor), can actually improve our expected test MSE.

True. The bias-variance decomposition tells us that using a simpler model can sometimes reduce variance (at the price of a little bias), which can result in a smaller expected test MSE.

(e) The expected test MSE can be smaller than the irreducible error.

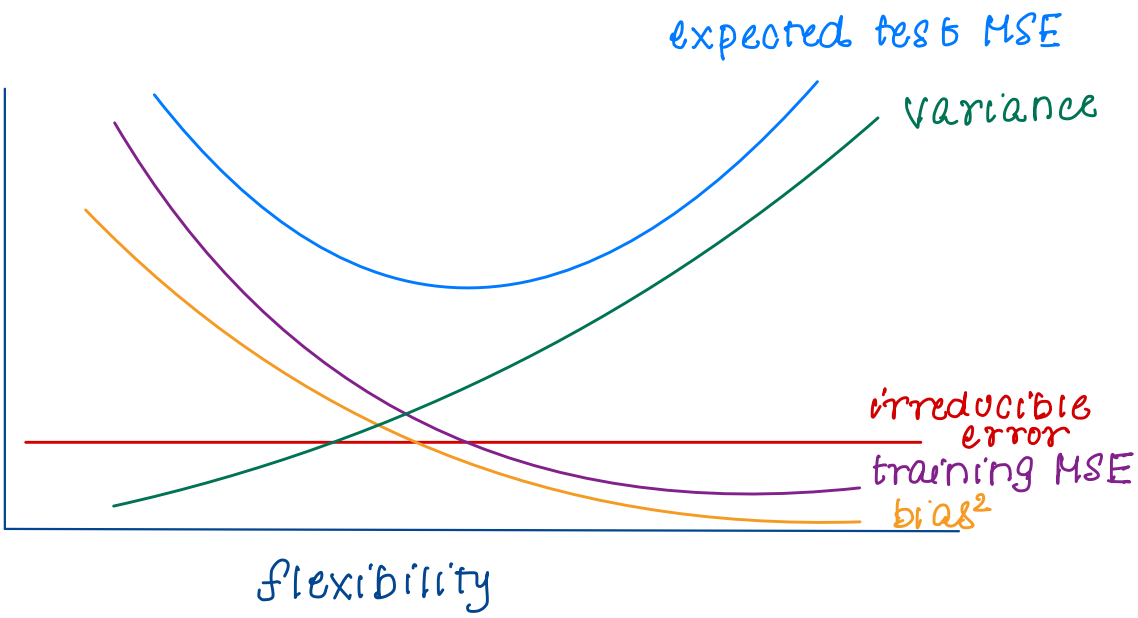
False. The expected test MSE is bounded by the irreducible error.

(f) The training MSE can be smaller than the irreducible error.

True. The training MSE can go to zero if you set $\hat{y}_i = y_i$*.*

Problem 2: Bias-variance decomposition

(a) On a single plot, provide a sketch of typical curves for (squared) bias, variance, expected test MSE, training MSE, and the irreducible error as we go from less flexible statistical learning methods towards more flexible methods. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be 5 curves. Make sure to label each one.



(b) Define in plain language (so that a non-data scientist can understand) what the quantities expected test MSE, training MSE, bias, variance and irreducible error mean.

- ***MSE*** *stands for mean-squared error and it measures the average (squared) difference between what we observe and what we predict.*
- ***Training MSE*** *is computed using the training data: it evaluates the discrepancy between our observed response and predicted value in the training set.*
- *On the other hand,* ***expected test MSE*** *can be conceptualized as the 'true average test MSE' that we would obtain if could repeatedly estimate f using a large number of training sets and evaluate each of them at x_0. The expected test MSE can be decomposed into three quantities: bias, variance, and irreducible error.*
- ***Bias*** *refers to the error that is introduced by approximating our real-life phenomenon.*
- ***Variance*** *refers to the amount by which our model would change if we had estimated it using a different raining set.*
- ***Irreducible error*** *refers to the inherent noise and randomness that we observe in Y. No matter how well we estimate our model f, we cannot reduce the error introduced by this irreducible error.*

(c) Explain why each of the five curves has the shape displayed in part (a).

Students should go through each of the curves systematically. From the definition of variance, it is clear that models that tend to be more flexible/complex will have higher variance because they tend to 'fit' the data better. Therefore, any change in the training set could lead to

substantial changes in the fitted model. On the other hand, as flexibility increases and the model 'fits' the data better, the bias (our approximation error from estimating f) will decrease. Irreducible error is a constant - it is a source of randomness in our dataset and is independent of the model we use. Since expected test MSE can be broken down into bias (squared), variance, and irreducible error, the U-shaped curve results from the trade-off between the bias and variance. Finally, training MSE will tend to decrease as flexibility of the model increases since the trained model will be a better approximation of the training set.

(d) Suppose I collect a data set of ($n = 100$ observations) containing a single predictor and a quantitative response Y. I fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 + \beta_2 X^2 + \beta_3 X^3 + \epsilon$. Suppose that the true relationship between X and Y is linear. Consider the training MSE for the linear regression and also the training MSE for the cubic regression. Would we expect one to be lower than other, or is there not enough information to tell? Justify your answer.

We would expect that the training MSE would be smaller for the cubic regression model. The cubic regression model is more flexible than the standard linear regression model; we know that as the flexibility of our model increases, the training MSE will decrease (see above plot).

(e) Same setup as (d). Consider the test MSE for the linear regression and also the test MSE for the cubic regression. Would we expect one to be lower than other, or is there not enough information to tell? Justify your answer.

Since the true relationship between X and Y is linear, we expect that the linear regression model will have a lower the test MSE - it will achieve the sweet spot of the bias variance tradeoff. We would expect the cubic regression model to overfit and suffer from high variance.

Problem 3: Optimal degree for Boston dataset

We'll use the `Boston` dataset. Our response is `medv` and our predictor is `lstat`. The setup is the same as our in-class activity (`bv_example.R`): we want to determine the optimal level of flexibility for our model. But now instead of just doing a one time split into a training/test set, use 5-fold CV to determine the optimal degree d for a polynomial regression model. **Write your own code to do so** - that means you should not be using other another package's functions to implement this (other than `caret` to create your folds). You can set $d = 1$ through 9. Report your 5-fold CV error for each model you considered and show a plot here. What degree did you choose?

```
library(caret)
library(ISLR2)

flds = createFolds(Boston$medv, k = 5, list = TRUE, returnTrain = FALSE)
k = 5

testMSE = matrix(NA,nrow=5,ncol=9)
for(i in 1:k){
  test_index = flds[[i]]
```

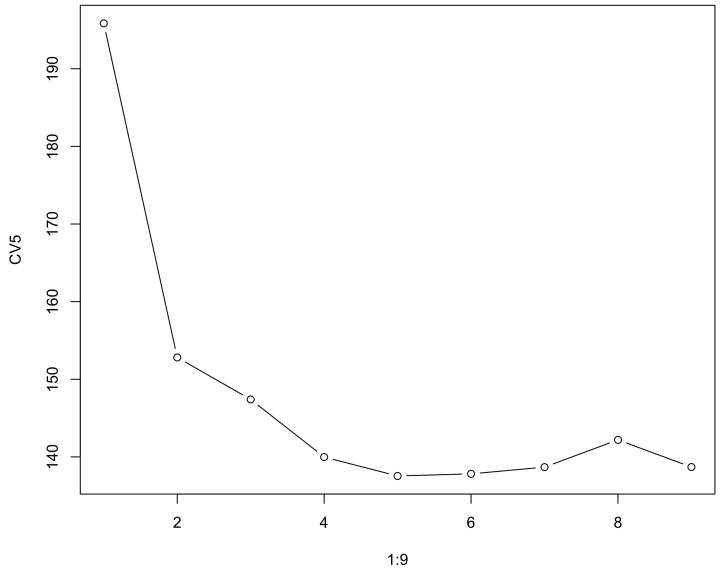
```
  test = Boston[test_index,]
  train = Boston[-test_index,]

  for(j in 1:9){
    model = lm(medv~poly(lstat,j,raw=TRUE),data=train)

    Yhat_test = predict(model,test)
    testMSE[i,j] = mean((test$medv-Yhat_test)^2)
  }
}
CV5 = colSums(testMSE)
plot(1:9,CV5,type='b')
```

```
 CV5
[1] 195.8410 152.8118 147.4013 139.9819 137.5441 137.8180
[7] 138.6833 142.1965 138.6956
```

Students answers may vary. My 5-fold CV error was minimized when $d = 5$.



Problem 4: Cross-validation

(a) Explain how k-fold cross-validation is implemented.

The dataset is divided into k roughly equal folds. A fold is chosen to be the test set and the remaining observations are the training set. This process is repeated so that each fold gets the chance to be the test set. The average of the k-test MSEs from this process is our k-fold cross-validation error.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i. The validation set approach?

Advantage of k-fold is we get to utilize more data for training/testing purposes. The disadvantage is that it is not as easy to implement and can be more computationally intensive to implement.

ii. LOOCV?

Advantage is that k-fold CV is less computationally intensive than LOOCV (the exception is for linear models). Disadvantage is that k-fold utilizes less data and it still has randomness.

(c) For the following questions, we will perform cross-validation on a simulated data set. Generate a simulated data set such that $Y = X - 2X^2 + \epsilon$, with $\epsilon \sim N(0, 1^2)$. Fill in the following code:

```
set.seed(1)
x = rnorm(100)
error = ??
y = ??
```

Solution:

```
set.seed(1)
x = rnorm(100)
error = rnorm(100,0,1)
y = x - 2*(x^2) + error
data = data.frame(x,y)
```

(d) Set a random seed, and then compute the LOOCV errors that result from fitting the following 4 models using `lm` and `poly`:

$M1$: a linear model with X
$M2$: a polynomial regression model with degree 2
$M3$: a polynomial regression model with degree 3
$M4$: a polynomial regression model with degree 4

You may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

```
> set.seed(1)
> MSE_M1 = MSE_M2 = MSE_M3 = MSE_M4 = rep(0,100)
> for(i in 1:100){
+   test = data[i,]
+   train = data[-i,]
```

```
+
+   M1 = lm(y~x,data=train)
+   M2 = lm(y~poly(x,2),data=train)
+   M3 = lm(y~poly(x,3),data=train)
+   M4 = lm(y~poly(x,4),data=train)
+
+   M1_y = predict(M1,newdata=test)
+   M2_y = predict(M2,newdata=test)
+   M3_y = predict(M3,newdata=test)
+   M4_y = predict(M4,newdata=test)
+
+   MSE_M1[i] = (test$y - M1_y)^2
+   MSE_M2[i] = (test$y - M2_y)^2
+   MSE_M3[i] = (test$y - M3_y)^2
+   MSE_M4[i] = (test$y - M4_y)^2
+ }
> mean(MSE_M1)
[1] 7.288162
> mean(MSE_M2)
[1] 0.9374236
> mean(MSE_M3)
[1] 0.9566218
> mean(MSE_M4)
[1] 0.9539049
```

(e) Repeat the above step using another random seed, and report your results. Are your results the same as what you got in (d). Why?

Yes. The results should be exactly the same. There is no randomness in LOOCV because each observations gets a chance to be left out as the test set.

(f) Which of the models in (d) had the smallest LOOCV error? Is this what you expected? Explain your answer.

M_2: the polynomial regression with degree 2. This what we expected since our true model has an X^2 in it. If students say another model happened to have the smallest LOOCV, as long as their code is correct, you can give them full credit.

(g) LOOCV will provide approximately unbiased estimates of the test error. Provide some intuition as to why.

You're taking the average of n test MSEs (where n is the number of observations in the dataset). And each training set contains $n-1$ observations, which is almost as many as the number of observations in the full dataset. Therefore, LOOCV will provide approximately unbiased estimates of the test MSE.

(h) There is bias-variance trade-off associated with the choice of k in k-fold cross-validation. As k increases in k-fold cross-validation, does the bias of the test error estimate increase, decrease, or stay the same? Explain why.

The bias decreases as k increases since k-fold starts to approach LOOCV. And based on the answer above, LOOCV will give us approximately unbiased estimates.

(i) As k increases in k-fold cross-validation, does the variance of the test error estimate increase, decrease, or stay the same? Explain why.

The variance will increase. Consider the most extreme version of this: $k = n$, so we are implementing LOOCV. When we perform LOOCV, we are in effect averaging the output of n fitted models, each of which is trained on an almost identical set of observations. Therefore, these outputs are highly positively correlated with each other. The average of many highly correlated quantities has average variance than does the mean of many quantities that are not highly correlated. Therefore, the test error estimates resulting from LOOCV will have higher variance than the test error estimates from k-fold CV. Applying this logic, as k increases, the variance of the test error estimates will increase.

End of assignment.