

STAT 435 Final Exam









Jaiden Atterbury

2024-05-30

Exercise 1: Luck and Fortune

Sign your name and date to affirm that you have not cheated on this exam:

On my integrity as a student at the University of Washington, I have neither given nor received unauthorized assistance on this midterm.

	
	
	
	
	Signature: Jaiden Atterbury
	Date: 05-30-24
	
	

Exercise 2: Regression

A ball is shot at three angles **a** (“low”/“med”/“high”) with varying speeds **v**. The reach **R** of each shot (horizontal flying distance until it hits the ground) is recorded. Below you see the R output of a model fit.

Note that $I(v^2)$ is R notation for including the square of predictor **v** in the regression. Additionally, $a:I(v^2)$ indicates the interaction between **a** and the square of the predictor **v**.

```
Call:
lm(formula = R ~ a + v + a:v + I(v^2) + a:I(v^2))

Residuals:
    Min       1Q   Median       3Q      Max
-0.43844 -0.15886 -0.02957  0.10579  0.52421

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.29480    0.28306  -1.041  0.303350
      amed       0.16155    0.36663   0.441  0.661632
      ahigh      0.11350    0.47887   0.237  0.813747
      v         0.21282    0.12349   1.723  0.091842 .
      I(v^2)     0.02478    0.01195   2.075  0.043889 *
      amed:v     -0.14725    0.15598  -0.944  0.350294
      ahigh:v    -0.13549    0.18130  -0.747  0.458838
      amed:I(v^2) 0.05847    0.01479   3.952  0.000277 ***
      ahigh:I(v^2) 0.07094    0.01599   4.438  6.02e-05 ***
---

Residual standard error: 0.222 on 44 degrees of freedom
Multiple R-squared:  0.9942, Adjusted R-squared:  0.9932
F-statistic: 949.9 on 8 and 44 DF,  p-value: < 2.2e-16
```

Part a

In this part of the exercise, we will mention what reference level of angle was used to fit the model. We will also mention how many data points were used to fit the model. In each case we will explain how we computed the answer. This is done below.

Since the predictor **a** has the levels **low**, **med** and **high**, based on the output from running the `summary()` function on the regression fit, the only level that is missing from the output is the level **low**. Therefore, it is apparent that **low** was the reference level of the angle variable in the above model fit.

Furthermore, since the degrees of freedom of this model are 44, and the degrees of freedom is calculated as $n - p - 1$ where n is the sample size, and p is the number of predictors, once we find p we can solve for n . As can be seen from the output from running the `summary()` function on the regression fit, there were 8 predictor coefficients that were estimated, which implies $p = 8$. Therefore, since $44 = n - p - 1$ and $p = 8$, it follows that $n = 44 + 8 + 1 = 53$ data points were used to fit the above model.

Part b

The above model fits three quadratic functions of **v** depending on the angle level. For instance, for **a=med**, and $v = v_1$, the fitted model satisfies the following

$$\hat{f}(a = med, v = v_1) = \hat{\alpha}_0 + \hat{\alpha}_1 v_1 + \hat{\alpha}_2 v_1^2$$

Subpart i

In this subpart of part b, we are tasked with identifying what the fitted values of $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ are in the fitted model shown above. This will be done below.

Since **a=med**, for the dummy variables corresponding to the **med** level, we can plug in the value of 1. For all of the other dummy variables (the ones corresponding to the **high** level), we will plug in 0. Furthermore, since

$v = v_1$, we will plug in v_1 for all of the terms involving the predictor v . Based on the problem description, after plugging in $a = \text{med}$ and $v = v_1$, we can deduce that \hat{f} takes the form

$$\begin{aligned}\hat{f}(a = \text{med}, v = v_1) &= \hat{\beta}_{int} + \hat{\beta}_{amed} + \hat{\beta}_v v_1 + \hat{\beta}_{v^2} v_1^2 + \hat{\beta}_{amed:v} v_1 + \hat{\beta}_{amed:v^2} v_1^2 \\ &= \hat{\beta}_{int} + \hat{\beta}_{amed} + \hat{\beta}_v v_1 + \hat{\beta}_{amed:v} v_1 + \hat{\beta}_{v^2} v_1^2 + \hat{\beta}_{amed:v^2} v_1^2 \\ &= (\hat{\beta}_{int} + \hat{\beta}_{amed}) + (\hat{\beta}_v + \hat{\beta}_{amed:v}) v_1 + (\hat{\beta}_{v^2} + \hat{\beta}_{amed:v^2}) v_1^2\end{aligned}$$

As calculated above, $\hat{\alpha}_0 = \hat{\beta}_{int} + \hat{\beta}_{amed}$, and since $\hat{\beta}_{int} = -0.29480$ and $\hat{\beta}_{amed} = 0.16155$, it follows that $\hat{\alpha}_0 = -0.13325$. Similarly, as calculated above, $\hat{\alpha}_1 = \hat{\beta}_v + \hat{\beta}_{amed:v}$, and since $\hat{\beta}_v = 0.21282$ and $\hat{\beta}_{amed:v} = -0.14725$, it follows that $\hat{\alpha}_1 = 0.06557$. Lastly, as calculated above, $\hat{\alpha}_2 = \hat{\beta}_{v^2} + \hat{\beta}_{amed:v^2}$, and since $\hat{\beta}_{v^2} = 0.02478$ and $\hat{\beta}_{amed:v^2} = 0.05847$, it follows that $\hat{\alpha}_2 = 0.08325$. Therefore, in summary, we have found that $\hat{\alpha}_0 = -0.13325$, $\hat{\alpha}_1 = 0.06557$, and $\hat{\alpha}_2 = 0.08325$.

Subpart ii

For $a = \text{low}$, and $v = v_1$, the fitted model satisfies the following

$$\hat{f}(a = \text{low}, v = v_1) = \hat{\gamma}_0 + \hat{\gamma}_1 v_1 + \hat{\gamma}_2 v_1^2$$

The goal of this subpart of part b is to see if any of the estimates $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ are equal to the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ from above. Since the reference level is **low**, it is obvious that the $\hat{\gamma}$ values will differ from the $\hat{\alpha}$ values from above. However, we will explicitly show this below.

Since $a = \text{low}$, for all of the dummy variables corresponding to the levels **med** and **high**, we will plug in 0. Furthermore, since $v = v_1$, we will plug in v_1 for all of the terms involving v . Based on the problem description, after plugging in $a = \text{low}$ and $v = v_1$, we can deduce that \hat{f} takes the form

$$\hat{f}(a = \text{low}, v = v_1) = \hat{\beta}_{int} + \hat{\beta}_v v_1 + \hat{\beta}_{v^2} v_1^2$$

As calculated above, $\hat{\gamma}_0 = \hat{\beta}_{int}$, and since $\hat{\beta}_{int} = -0.29480$, it follows that $\hat{\gamma}_0 = -0.29480$. Similarly, as calculated above, $\hat{\gamma}_1 = \hat{\beta}_v$, and since $\hat{\beta}_v = 0.21282$, it follows that $\hat{\gamma}_1 = 0.21282$. Lastly, as calculated above, $\hat{\gamma}_2 = \hat{\beta}_{v^2}$, and since $\hat{\beta}_{v^2} = 0.02478$, it follows that $\hat{\gamma}_2 = 0.02478$. Therefore, in summary, we have found that $\hat{\gamma}_0 = -0.29480$, $\hat{\gamma}_1 = 0.21282$, and $\hat{\gamma}_2 = 0.02478$. As found in subpart i, $\hat{\alpha}_0 = -0.13325$, $\hat{\alpha}_1 = 0.06557$, and $\hat{\alpha}_2 = 0.08325$. Thus we can conclude that none of the estimates $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ are equal to the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ from above.

Subpart iii

In this subpart of part b, we will explain if $\hat{\gamma}_1 = \hat{\alpha}_1$ if we decided to not include the interaction term $a:I(v^2)$. This will be explained below.

As can be seen from subpart i above, $\hat{\alpha}_1 = \hat{\beta}_v + \hat{\beta}_{amed:v}$. Similarly, as can be seen from subpart ii above, $\hat{\gamma}_1 = \hat{\beta}_v$. In both cases, the coefficients estimated do not depend on v_1^2 (instead $\hat{\alpha}_2$ and $\hat{\gamma}_2$ are the coefficients that depend on v_1^2). Therefore, even by dropping the interaction term $a:I(v^2)$, we would find that $\hat{\gamma}_1 \neq \hat{\alpha}_1$. The only way that these two estimated coefficients would equal each other, is if we dropped the interaction term $a:v$.

Subpart iv

In this subpart of part b, we determine, based only on the output, if we think that the interaction term $a:I(v^2)$ is significant at the 5% level. This will be explained below.

Based only on the output, it is likely that the interaction term $a:I(v^2)$ is significant at the 5% level. This is because the coefficients corresponding to the levels of the interaction term, **amed:I(v^2)** and **ahigh:I(v^2)**, have p-values of 0.000277 and 6.02×10^{-5} , respectively. However, it is important to note that this claim is a bit flawed because we are using the results from two different t-tests to make a statement about the significance of a single interaction term. In this case, the better thing to do would be to run an F-test between a model with and without this interaction term.

Subpart v

In this subpart of part b, we will state the null and the alternative hypotheses we would be testing in order

to make a statistical decision about dropping the interaction term $\mathbf{a}:\mathbf{I}(\mathbf{v}^2)$ from the model. This is done below.

As stated above, in order to make a statistical decision about dropping the interaction term $\mathbf{a}:\mathbf{I}(\mathbf{v}^2)$ from the model, we would need to run an F -test between a model with and without this interaction term. Therefore, the null and the alternative hypotheses corresponding to this F -test would be: $H_0 : (\beta_{amed:v^2}, \beta_{ahigh:v^2}) = (0, 0)$ versus $H_1 : \beta_{amed:v^2} \neq 0$ or $\beta_{ahigh:v^2} \neq 0$.

Subpart vi

In this subpart of part b, we will state the test statistic we would use for the above hypothesis test and specify the distribution this test statistic would follow under the null hypothesis. This is done below.

As can be seen from the optional reading corresponding to chapter 3 from the STAT 423 notes, given that a model was fit with q predictors, and a model was fit with p predictors, where $0 \leq q \leq p$ and the models are nested, it follows that the test statistic for an F -test is $F = \frac{(RSS_q - RSS_p)(n-p-1)}{(p-q)RSS_p}$. In the aforementioned test statistic, RSS_q is the residual sum of squares after fitting the model with q predictors, RSS_p is the residual sum of squares after fitting the model with p predictors, $n-p-1$ are the degrees of freedom of the denominator (and the model with p predictors), and $p-q$ are the numerator degrees of freedom. This test statistic, under the assumption that the null hypothesis is true, is $F_{p-q, n-p-1}$.

With that being said, in our case, fitting a model without the interaction term $\mathbf{a}:\mathbf{I}(\mathbf{v}^2)$ leads to $q = 6$ predictors. Therefore, RSS_6 corresponds to the residual sum of squares after fitting the model without the interaction term $\mathbf{a}:\mathbf{I}(\mathbf{v}^2)$. Furthermore, the full model leads to $p = 8$ predictors, and thus RSS_8 corresponds to the residual sum of squares after fitting the full model. Lastly, the numerator degrees of freedom is $p-q = 8-6 = 2$, and the denominator degrees of freedom is $n-p-1 = 53-8-1 = 44$. Therefore, the test statistic corresponding to the above hypothesis test is $F = \frac{44(RSS_6 - RSS_8)}{2RSS_8} = \frac{22(RSS_6 - RSS_8)}{RSS_8}$, and its distribution under the assumption that the null hypothesis is true is $F \sim F_{2, 44}$ (an F distribution with 2 numerator degrees of freedom and 44 denominator degrees of freedom).

Part c: (Extra credit)

An F -test can be used to determine whether a set of predictors have a significant effect on the response. In other words, given two models, an F -test can perform a hypothesis test to identify the better model. Furthermore, recall that cross-validation can also be used to identify which of two given models has better prediction error on the new data. In this extra credit exercise, we will compare and contrast these two methods for identifying a good model by providing at least one advantage and at least one disadvantage of each method. This is done below.

As hinted at in the problem description, although both the F -test and cross-validation both look to choose the best model and can be used for model selection, they do have some striking differences. One advantage of an F -test is that it tells you exactly if certain predictors are decreasing the residual sum of squares by a significant amount. However, some disadvantages of an F -test are that they require that the models they are comparing be nested, and they also require strong assumptions on the data (the fact that the F statistic follows an F distribution depends on the numerous assumptions that we made when using the linear regression framework). This means that, if the models aren't nested an F -test can't be used, and if the assumptions are violated and an F -test is ran, results could be inaccurate. On the other hand, some benefits of cross-validation are that there is no need for the models to be nested, there are no strong assumptions that the data must have in order to use cross-validation, and when there are more than two models, you don't need to run multiple tests, instead you just use the same algorithm on the third model. Although this method seems much better, cross-validation doesn't tell you if the differences in the error are significant or not, and running cross-validation on many models with large datasets can be computationally expensive.

Exercise 3: Classification

The following unrelated problems pertain to miscellaneous classification tasks.

Part a

In this part of the exercise, suppose we have 7 training data points with labels from one of $\{1, 2\}$. If we encounter a new data point and compute the distance between the new data point and the 7 training points as

#	Label of point	Distance from new point
1	2	11
2	1	20
3	2	12
4	1	7
5	2	13
6	1	6
7	2	10

then the goal of this exercise is, using the k -nearest neighbors approach, we must predict what the label of the new point will be for each of the following values of k , $k = 3, 4, 5$. In the k -nearest neighbors approach, in order to make a prediction one would simply find the k closest points, and out of these k nearest points, the most common label is declared our prediction.

Therefore, for $k = 3$, the 3 closest points are points 6, 4, and 7 with distances of 6, 7, and 10, respectively. The labels of these three points are 1, 1, and 2, respectively, leading to a prediction of the label of our new point being 1. Similarly, for $k = 4$, the 4 closest points are points 6, 4, 7, and 1, with distances of 6, 7, 10, and 11, respectively. The labels of these three points are 1, 1, 2, 2, respectively. In order to break the tie, as is done in R, one would randomly break this tie because the probability of choosing either label is the same, however we will instead use the class label of the point with the closest distance with the new point as the tiebreaker, this leads to a prediction of the label of our new point being 1. Lastly, for $k = 5$, the 5 closest points are points 6, 4, 7, 1, and 3, with distances of 6, 7, 10, 11, and 12, respectively. The labels of these three points are 1, 1, 2, 2, 2 respectively, leading to a prediction of the label of our new point being 2.

Part b

In this part of the exercise, we will again suppose we are working with the k -nearest neighbors approach. We will suppose there are n training data points, with 25% of them having label $Y = 1$; 35% of them having label $Y = 2$, and; 40% of them having label $Y = 3$. With this being said, we want to predict the label of a new data point x using $k = n$ neighbors (that is, using all of the training points in our neighborhood). The goal of this part of the exercise is to calculate $P(Y = 1|x)$, $P(Y = 2|x)$, and $P(Y = 3|x)$.

As explained in the previous part of the exercise, in the k -nearest neighbors approach, in order to make a prediction one would simply find the k closest points, and out of these k nearest points, the most common label is declared our prediction. Furthermore, if we define $\mathcal{N}_k(x)$ as the k closest points to x , then the probabilities used to make our prediction are $\hat{P}(Y = j|X = x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} I(y_i = j)$. Since we are using all n training points to make our prediction, the probability that the label of a new data point x is 1, 2, or 3 is simply the proportion of that label in the training data. Therefore, $P(Y = 1|x) = 0.25$, $P(Y = 2|x) = 0.35$, and $P(Y = 3|x) = 0.4$.

Part c

In this part of the exercise, we will consider the following 2×2 confusion matrix. In the confusion matrix, the horizontal axis contains the true labels and the vertical axis contains the predicted labels.

	Positive (true)	Negative (true)
Positive (predicted)	32	16
Negative (predicted)	7	45

The goal of this part of the exercise is to calculate the total error $\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$, where y is the true label and \hat{y} is the predicted label. We will also calculate the proportion of false positives and false negatives (these proportions are different from the false positive/negative rates). This is done below.

In a 2×2 confusion matrix, the off-diagonal entries are the misclassified observations, therefore the total error in whatever classification method was used on this data is $\frac{23}{100}$. Furthermore, the total number of false positives was 16, for a false positive proportion of $\frac{16}{100} = \frac{4}{25}$. Lastly, the total number of false negatives was 7, for a false negative proportion of $\frac{7}{100}$. If instead we wanted the false positive rate and the false negative rate, we would've gotten, $\frac{16}{61}$ and $\frac{7}{49} = \frac{1}{7}$, respectively.

Part d

In this part of the exercise, we assume that we built a binary classifier to classify whether a loan attempt was fraudulent (positive) or not fraudulent (negative). Based on three different classification thresholds, we obtain the following misclassification, false positive, and false negative rates:

#	Misclassification rate	False positive rate	False negative rate
1	0.25	0.15	0.10
2	0.20	0.05	0.15
3	0.20	0.15	0.05

The goal of this exercise is to choose, amongst the three different models, which one would we should prefer to detect fraudulent login attempts. We will also explain in terms of how relevant the three different metrics are in this scenario.

In terms of detecting fraudulent login attempts, and in any classification task for that matter, the misclassification rate, the total number of misclassified data points, is important for getting an overall sense of model performance. It is thus important to keep this as low as possible. However, in a high stakes classification task such as detecting fraud, the false positive rates and false negative rates are arguably more important to analyze (obviously there still needs to be balance when choosing the threshold). In terms of this classification task, a false positive corresponds to a login attempt that was non-fraudulent, but was classified as fraudulent. The consequences of a false positive in this setting is the inconvenience to true users who are just trying to use the website for its intended purpose. On the other hand, in this classification task, a false negative corresponds to a login attempt that was fraudulent, but was classified as non-fraudulent. The consequences of this error are much more severe, as this could lead to major security risks, which could be detrimental if the service provided had something to do with the identity or financial status of an individual.

Therefore, we want to minimize the overall misclassification rate, while at the same time keeping the false negative rate as low as we can (we must keep the false positive rate low as well, but it is not as important as the false negative rate). With all of this being said, since threshold 3 is tied with threshold 2 in terms of misclassification rate, but has a lower false negative rate than threshold 2 (while at the same time also having a relatively small false positive rate), threshold 3 would be the best choice for this classification task.

Part e

In this part of the exercise, we will recall that linear discriminant analysis (LDA) estimates the unknown means of a multivariate Gaussian and quadratic discriminant analysis (QDA) estimates the unknown means and the covariance matrix of a multivariate Gaussian. We will also assume that we have p predictors and k

classes. The goal of this exercise is to explain, for the k total classes, how many unknown parameters LDA estimates and how many unknown parameters QDA estimates. This is done below.

As stated on the Ed discussion board, as well as being indicated in the problem description, we will not include the prior probabilities $\hat{\pi}_k$ in our count of the total number of parameters estimated. Since LDA estimates the unknown means of a multivariate Gaussian in all k classes (as well as a single covariance matrix for all of the classes), and given the fact that we have p predictors, then LDA estimates kp unique parameters for the k mean vectors and $\frac{p(p+1)}{2}$ unique parameters for the single covariance matrix (since there are $\frac{p(p+1)}{2}$ unique terms in a covariance matrix). Therefore, in total, LDA estimates $kp + \frac{p(p+1)}{2}$ unique parameters.

On the other hand, since QDA estimates the unknown means and the covariance matrix of a multivariate Gaussian in all k classes, and given the fact that we have p predictors, then QDA estimates kp unique parameters for the k mean vectors and $\frac{kp(p+1)}{2}$ unique parameters for the k covariance matrices (since there are $\frac{p(p+1)}{2}$ unique terms in a covariance matrix). Therefore, in total, QDA estimates $kp + \frac{kp(p+1)}{2}$ unique parameters.

Lastly, suppose we instead constructed a modified QDA where we impose a constraint that all predictors are independent of each other, that is, for two predictors X_i and X_j , $Cov(X_i, X_j) = 0$ whenever $i \neq j$. We will now explain how many unknown parameters this constrained QDA estimates. Since all of the off-diagonal entries of the k class covariance matrices are now 0, there will now only be p unique entries in these matrices. Therefore, the constrained QDA estimates kp unique parameters for the k mean vectors and kp unique parameters for the k covariance matrices. Thus, in total, the constrained QDA estimates $kp + kp$ unique parameters. Note that, if we were to include the estimated prior probabilities into our total parameter count, we would simply add k to all of the above results.

Exercise 4: Bootstrap

In this exercise, we will work with a simple bootstrap procedure to estimate the mean and find the variance of our estimated mean. We will consider the following setup for the remainder of this problem:

We have n data points x_1, \dots, x_n that are fixed. In the bootstrap, we sample n points with replacement to obtain our bootstrap sample: X_1^*, \dots, X_n^* . The bootstrap sample points are independent and identically distributed with a probability distribution

$$P(X_j^* = x_1) = \dots = P(X_j^* = x_n) = \frac{1}{n}, \quad j = 1, \dots, n$$

Define the mean and the variance of the fixed data as

$$\begin{aligned} \hat{\mu} = \bar{x}_n &= \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\sigma}^2 = S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \end{aligned}$$

Define the bootstrap sample mean as

$$\bar{X}_n^* = \frac{1}{n} \sum_{i=1}^n X_i^*$$

Part a

In this part of the exercise, our goal is to show that \bar{X}_n^* is unbiased, that is $E[\bar{X}_n^*] = E[X_i^*] = \bar{x}_n$. This is done below.

Before we can delve into the proof, we must first show that $E[X_i^*] = \bar{x}_n$. Based on the definition of the expected value of a discrete random variable, we know that $E[X_i^*] = \sum_{j=1}^n x_j P(X_i^* = x_j)$. However, since the X_i^* are identically distributed, it follows that $E[X_i^*] = E[X^*]$ for all i . Based on the definition of the PDF of X^* , we can see that $P(X^* = x_j) = \frac{1}{n}$ (the probability that a bootstrapped sample equals x_j is equally likely for all x_j where $j \in \{1, \dots, n\}$). Therefore, $E[X_i^*] = E[X^*] = \frac{1}{n} \sum_{j=1}^n x_j = \bar{x}_n$. Putting this altogether we obtain the following

$$\begin{aligned} E[\bar{X}_n^*] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i^*\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i^*] \quad (\text{Linearity of expectation}) \\ &= \frac{1}{n} \sum_{i=1}^n E[X^*] \quad (\text{Since the } X_i^* \text{ are identically distributed}) \\ &= \frac{1}{n} \sum_{i=1}^n \bar{x}_n \quad (\text{Since } E[X^*] = \bar{x}_n) \\ &= \bar{x}_n \frac{1}{n} \sum_{i=1}^n 1 \\ &= \bar{x}_n \frac{n}{n} \\ &= \bar{x}_n \end{aligned}$$

Part b

In this part of the exercise, our goal is to show that $E[(\bar{X}_n^*)^2] = \frac{n-1}{n}(\bar{x}_n)^2 + \frac{1}{n^2} \sum_{i=1}^n x_i^2$. To do this we will use the following hints: $(a_1 + a_2 + \dots + a_n)^2 = \sum_{i=1}^n a_i^2 + 2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j$ and $\sum_{i=1}^n \sum_{j=i+1}^n ab = \frac{n(n-1)}{2} ab$.

Before we move onto the main proof we will first find $E[(X^*)^2]$. Based on the definition of the expected value of a discrete random variable, we know that $E[(X^*)^2] = \sum_{i=1}^n x_i^2 P(X^* = x_i)$. Based on the definition of the PDF of X^* , we can see that $P(X^* = x_i) = \frac{1}{n}$ (the probability that a bootstrapped sample equals x_i is equally likely for all x_i with $i \in \{1, \dots, n\}$). Therefore, $E[(X^*)^2] = \frac{1}{n} \sum_{i=1}^n x_i^2$. Putting this altogether we obtain the following

$$\begin{aligned}
E[(\bar{X}_n^*)^2] &= E\left[\left(\frac{1}{n} \sum_{i=1}^n X_i^*\right)^2\right] \\
&= E\left[\frac{1}{n^2} \left(\sum_{i=1}^n X_i^*\right)^2\right] \\
&= \frac{1}{n^2} E\left[\left(\sum_{i=1}^n X_i^*\right)^2\right] \\
&= \frac{1}{n^2} E\left[\sum_{i=1}^n (X_i^*)^2 + 2 \sum_{i=1}^n \sum_{j=i+1}^n X_i^* X_j^*\right] \quad (\text{Apply hint 1}) \\
&= \frac{1}{n^2} E\left[\sum_{i=1}^n (X_i^*)^2\right] + \frac{1}{n^2} E\left[2 \sum_{i=1}^n \sum_{j=i+1}^n X_i^* X_j^*\right] \quad (\text{Linearity of expectation}) \\
&= \frac{1}{n^2} \sum_{i=1}^n E[(X_i^*)^2] + \frac{2}{n^2} \sum_{i=1}^n \sum_{j=i+1}^n E[X_i^* X_j^*] \quad (\text{Linearity of expectation}) \\
&= \frac{1}{n^2} \sum_{i=1}^n E[(X_i^*)^2] + \frac{2}{n^2} \sum_{i=1}^n \sum_{j=i+1}^n E[X_i^*] E[X_j^*] \quad (\text{Since } X_i^* \text{ are independent for } i \neq j) \\
&= \frac{1}{n^2} \sum_{i=1}^n E[(X^*)^2] + \frac{2}{n^2} \sum_{i=1}^n \sum_{j=i+1}^n E[X^*] E[X^*] \quad (\text{Since the } X_i^* \text{ are identically distributed}) \\
&= \frac{n}{n^2} E[(X^*)^2] + \frac{2}{n^2} \frac{n(n-1)}{2} (E[X^*])^2 \quad (\text{Apply hint 2}) \\
&= \frac{1}{n} E[(X^*)^2] + \frac{(n-1)}{n} (E[X^*])^2 \\
&= \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n x_i^2\right) + \frac{n-1}{n} (\bar{x}_n)^2 \quad (\text{Since } E[(X^*)^2] = \frac{1}{n} \sum_{i=1}^n x_i^2 \text{ and } E[X^*] = \bar{x}_n) \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \frac{n-1}{n} (\bar{x}_n)^2 \\
&= \frac{n-1}{n} (\bar{x}_n)^2 + \frac{1}{n^2} \sum_{i=1}^n x_i^2
\end{aligned}$$

Part c

In this part of the exercise, we will use results from part (a) and (b) to show that $\text{Var}(\bar{X}_n^*) = \frac{n-1}{n^2} S_n^2$. This

is done below.

$$\begin{aligned}
\text{Var}(\bar{X}_n^*) &= E[(\bar{X}_n^*)^2] - (E[\bar{X}_n^*])^2 \quad (\text{Alternate definition of the variance}) \\
&= \frac{n-1}{n}(\bar{x}_n)^2 + \frac{1}{n^2} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2 \quad (\text{Using part a and b}) \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \frac{n-1}{n}(\bar{x}_n)^2 - (\bar{x}_n)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \left(\frac{n-1}{n} - 1\right)(\bar{x}_n)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \frac{-1}{n}(\bar{x}_n)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \left(\frac{1}{n} - \frac{2}{n}\right)(\bar{x}_n)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \frac{1}{n}(\bar{x}_n)^2 - \frac{2}{n}(\bar{x}_n)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 - \frac{2}{n}(\bar{x}_n)^2 + \frac{1}{n}(\bar{x}_n)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 - \frac{2}{n} \left(\frac{1}{n} \sum_{i=1}^n x_i\right) \bar{x}_n + \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n x_i\right) (\bar{x}_n)^2 \quad (\text{Since } \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } 1 = \frac{n}{n} = \frac{1}{n} \sum_{i=1}^n) \\
&= \frac{1}{n^2} \sum_{i=1}^n x_i^2 - \frac{2}{n^2} \sum_{i=1}^n x_i \bar{x}_n + \frac{1}{n^2} \sum_{i=1}^n (\bar{x}_n)^2 \\
&= \frac{1}{n^2} \left(\sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i \bar{x}_n + \sum_{i=1}^n (\bar{x}_n)^2 \right) \\
&= \frac{1}{n^2} \left(\sum_{i=1}^n x_i^2 - 2x_i \bar{x}_n + (\bar{x}_n)^2 \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \\
&= \frac{n-1}{n-1} \frac{1}{n^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \\
&= \frac{n-1}{n^2} \left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right) \\
&= \frac{n-1}{n^2} S_n^2
\end{aligned}$$

Exercise 5: Model Selection

The coefficient of determination for the multiple linear regression model $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}$ is given by $R^2 = 1 - \frac{RSS}{TSS}$ where $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$, and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Here there are p different predictors for each datapoint i : x_{i1}, \dots, x_{ip} that have corresponding coefficients $\beta_0, \beta_1, \dots, \beta_p$. Least squares gives us the following estimates for those coefficients: $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$.

Now, if we consider adding a new $(p+1)$ -th variable $x_{i,p+1}$, then this gives us a different set of least squares estimates for the $p+1$ coefficients $\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_p, \tilde{\beta}_{p+1}$ (in general, we are not guaranteed to have $\hat{\beta}_i = \tilde{\beta}_i$ for $i = 0, 1, \dots, p$). Our new coefficient of determination is $R_2^2 = 1 - \frac{RSS_2}{TSS_2}$ (where TSS_2 is simply TSS), where the subscript $()_2$ implies that we are referring to the expanded model with $x_{i,p+1}$. In the following problems, we will assume that $n > p+1$.

Part a

In this part of the exercise, we must show that $R^2 \leq 1$. We will assume that $TSS \neq 0$ because if $TSS = 0$, R^2 is undefined (there would also be no point in adding extra predictors if $TSS = 0$). First off, since $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, the square term indicates that $RSS \geq 0$. Furthermore, since we assumed that $TSS \neq 0$, we can see that $\frac{RSS}{TSS} \geq 0$. Multiplying both sides of the inequality by -1 we obtain $-\frac{RSS}{TSS} \leq 0$. Adding 1 to both sides of the inequality we obtain $1 - \frac{RSS}{TSS} \leq 1$. Since $R^2 = 1 - \frac{RSS}{TSS}$, we have shown that $R^2 \leq 1$.

Part b

In this part of the exercise, we must show that $R^2 \geq 0$. in order to do this we are told that the following might be helpful

$$\begin{aligned} RSS &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{ij})^2 = \min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\ \min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 &\leq \min_{\beta_0, 0, \dots, 0} \sum_{i=1}^n (y_i - \beta_0)^2 \\ \min_{\beta_0} \sum_{i=1}^n (y_i - \beta_0)^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 = TSS \end{aligned}$$

The above inequalities show us that $RSS \leq TSS$. We will also assume that $TSS \neq 0$. First off, since $RSS \leq TSS$ and $0 \leq RSS$, it follows that $0 \leq RSS \leq TSS$. Since we assumed that $TSS \neq 0$, we can see that $0 \leq \frac{RSS}{TSS} \leq 1$. Subtracting by $\frac{RSS}{TSS}$ on all sides we see that $0 \leq 1 - \frac{RSS}{TSS}$. Since $R^2 = 1 - \frac{RSS}{TSS}$, we have shown that $0 \leq R^2$.

Part c

In this part of the exercise, we must show that $R^2 \leq R_2^2$. in order to do this, we will adjust the hint from part b and notice that the following might be helpful

$$\begin{aligned} RSS_2 &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{ij} - \hat{\beta}_{p+1} x_{i,p+1})^2 = \min_{\beta_0, \beta_1, \dots, \beta_p, \beta_{p+1}} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} - \beta_{p+1} x_{i,p+1})^2 \\ \min_{\beta_0, \beta_1, \dots, \beta_p, \beta_{p+1}} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} - \beta_{p+1} x_{i,p+1})^2 &\leq \min_{\beta_0, \beta_1, \dots, \beta_p, 0} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\ \min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{ij})^2 = RSS \end{aligned}$$

The above inequalities show us that $RSS_2 \leq RSS$. We will also assume that $TSS \neq 0$. First off, multiplying by -1 on all sides we see that $-RSS_2 \geq -RSS$. Since we assumed that $TSS \neq 0$, we can see that $-\frac{RSS_2}{TSS} \geq -\frac{RSS}{TSS}$. Adding 1 on all sides we see that $1 - \frac{RSS_2}{TSS} \geq 1 - \frac{RSS}{TSS}$. Since $R_2^2 = 1 - \frac{RSS_2}{TSS}$ and $R^2 = 1 - \frac{RSS}{TSS}$, we have shown that $R^2 \leq R_2^2$.

Part d

In this part of the exercise, based on the inequality in part (c), we will give a reason for why maximizing R^2 is not a good heuristic when considering whether to include an additional variable in our model. This is done below.

Using R^2 is not a good heuristic when considering whether to include an additional variable in our model because, as shown in part (c), whenever we add an additional predictor to our model we are guaranteed that the R^2 will not decrease. Therefore, using R^2 as a model heuristic will always choose the largest model (except in the extremely rare case that the R^2 value stays the same when a new predictor is added). This is not a good thing as if the R^2 doesn't increase significantly (that is, if the new predictor doesn't do a good job at explaining the response variable and doesn't decrease the RSS by that much) then we'd be adding unnecessary complexity that will eventually lead to overfitting.