Homework 2

Andre Ye

3 April 2021

Problem 1

Context: Construct the line of best fit for each of the following sets of points, compute the R2 and adjusted R2 values, and comment on the quality of the fit. You don't need to apply an F-test.

Part A Problem: (0, 0), (1, 3), (2, 2), (4, 6)

Part A Solution: The formula for a linear regression model $f(x) = mx + b$ is given by:

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$

Let us calculate the means first; $x_i = \frac{0+1+2+4}{4} = \frac{7}{4} = 1.75$ and $y_i = \frac{0+3+2+6}{4} = \frac{11}{4} = 2.75$. We can then calculate m:

$$m = \frac{(0-1.75)(0-2.75) + (1-1.75)(3-2.75) + (2-1.75)(2-2.75) + (4-1.75)(6-2.75)}{(0-1.75)^2 + (1-1.75)^2 + (2-1.75)^2 + (4-1.75)^2} = \frac{11.75}{8.75} \approx 1.34285$$

We can find b as $\frac{11}{4} - \frac{11.75}{8.75}\left(\frac{7}{4}\right) = 2.75 - 2.35 = 0.4$. The line of best fit is thus approximately $y = 1.34285x + 0.4$.

The R2 score is given by $R^2 = 1 - \frac{\sum(y_i - f(x_i))^2}{\sum(y_i - \bar{y})^2}$. The denominator can be calculated as follows:

$$\sum(y_i - \bar{y})^2 = (2.75-0)^2 + (2.75-3)^2 + (2.75-2)^2 + (2.75-6)^2 = 18.75$$

The numerator can be calculated as follows:

$$\sum(y_i - f(x_i))^2 = (0 - f(0))^2 + (3 - f(1))^2 + (2 - f(2))^2 + (6 - f(4))^2$$

$$\sum(y_i - f(x_i))^2 = (0 - 0.4)^2 + (3 - 1.74285714)^2 + (2 - 3.08571429)^2 + (6 - 5.77142857)^2 \approx 2.97142$$

The denominator can be calculated as follows:

$$\sum(y_i - \bar{y})^2 = (0-2.75)^2 + (3-2.75)^2 + (2-2.75)^2 + (6-2.75)^2 = 18.75$$

Plugging these into the R2 score formula yields $1 - \frac{2.97142}{18.75} \approx 0.84152$. Using this R2 score in the adjusted R2 formula, $1 - (1 - R^2)\frac{n-1}{n-3}$, yields $1 - (1 - 0.84152)\,\frac{4-1}{4-3} = 0.52456$.

This line of fit performs pretty well on the data, although there is not enough data.

Part B Problem: (0, 1), (2, 1), (3, 2), (4, 1), (5, 3), (6, 4)

Part B Solution: Using the same formula and process as outlined above, we can calculate the line of best fit. m evaluates to $\frac{11}{23.3333} \approx 0.47142$. b is thus approximately 0.4286. The line of best fit is thus $y = 0.47142x + 0.4286$. The numerator of the R2 score evaluates to 2.8143; the denominator of the R2 score evaluates to 8 (referring to the second term of the R2 score, ignoring the "1−"). The R2 score is thus 0.64821. To calculate the adjusted R2, 1 − 0.64821 is multiplied by 1.6667; this quantity is subtracted from 1 to yield an adjusted R2 of 0.41369. This line's fit is mediocre.

Part C Problem: (0, 1), (1, 3), (2, 5), (3, 7), (4, 9)

Part C Solution: Using the same formula and process as outlined above, we can calculate the line of best fit. m evaluates to $\frac{20}{10} = 2$. b is thus equal to 1. The line of best fit is thus $y = 2x + 1$. The numerator of the R2 score evaluates to 0, meaning that the model predicts every point perfectly. The R2 score is thus 1. The adjusted R2 score is irrelevant when the R2 score is also 1, since the $(1 - R^2)\frac{n-1}{n-3}$ term evaluates to 0 when $R^2 = 1$. This line is a perfect fit for the data.

Part D Problem: (0, 2), (1, -1), (4, 3), (5, -6), (7, 0), (8, 2)

Part D Solution: Using the same formula and process as outlined above, we can calculate the line of best fit. m evaluates to $\frac{-3}{50.83333} \approx -0.05901$. b is thus approximately 0.245902. The line of best fit is thus $y = -0.05901x + 0.245902$. The numerator of the R2 score evaluates to 53.82295; the denominator evaluates to 54. The R2 is hence 0.003279. Multiplying $1 - 0.003279$ by 1.6667 and subtracting the result from 1 yields an adjusted R2 of -0.6612. This line performs very poorly on the dataset; it is barely better than a constant model that guesses the average. When the dataset size is taken into account with the R2, we see that this predictor negatively impacts the prediction so much that the adjusted-R2 becomes negative.

Part E Problem: (0, 2), (3, 5)

Part E Solution: Given only two points, the best line of fit is the one that passes through them. The line is thus $y = x + 2$. Since the line passes through all the data points in the dataset, the R2 and adjusted R2 are both equal to 1. This is to be expected, since a linear model will always perfectly model two points. As such, it does not make sense to use the R2 as a comparative metric.

Problem 2

Context: Suppose that the line of best fit for a certain set of data is $y = 2x - 1$, and the R2 for this model is 0.8.

Part A Problem: If 1 were added to all of the y-coordinates of the data, what would the new line of best fit be? Can you predict what the new R2 would be?

Part A Solution: All the data has shifted up one unit, so the line of best fit would be shifted up one unit. This yields a new best fit line of $y = 2x$.

The equation for R2 was $1 - \frac{\sum(y_i - 2x_i + 1))^2}{\sum(y_i - \bar{y})^2}$. Adding 1 to the y-coordinates of the data and updating the line of best fit would yield the following result:

$$1 - \frac{\sum(y_i + 1 - 2x_i)^2}{\sum(y_i + 1 - \bar{y} - 1)^2}$$

Note that $\bar{y}$ still represents the mean of the y-values prior to adding 1; we have simply expressed the new mean as $\bar{y} + 1$. We can prove the truth of this step as follows: $\bar{y} = \frac{\sum y_i}{N}$. Adding 1 to the elements yields $\frac{\sum y_i + 1}{N}$. Using the commutative property of addition, this can be distributed as $\frac{\sum y_i}{N} + \frac{\sum 1}{N}$. The latter term $\frac{\sum 1}{N}$ equals 1 because there are $\sum 1 = N$. The expression $\frac{\sum y_i}{N} + \frac{\sum 1}{N}$ simplifies to $\bar{y} + 1$.

Thus, we can prove that the R2 prior to and after adding 1 to each y_i term is the same:

$$\begin{aligned} 1 - \frac{\sum(y_i - 2x_i + 1))^2}{\sum(y_i - \bar{y})^2} &= 1 - \frac{\sum(y_i + 1 - 2x_i)^2}{\sum(y_i + 1 - \bar{y} - 1)^2} \\ &= 1 - \frac{\sum(y_i - 2x_i + 1)^2}{\sum(y_i - \bar{y})^2} \end{aligned}$$

Therefore, the R2 remains the same, at 0.8.

Part B Problem: If 1 were added to all of the x-coordinates of the data, what would the new line of best fit be? Can you predict what the new R2 would be?

Part B Solution: Since all the data shifts 1 unit to the right, the line of best fit also does; this yields $2(x - 1) - 1 = 2x - 3$. The line of best fit for this dataset is thus $y = 2x - 3$.

The equation for R2 was $1 - \frac{\sum(y_i - 2x_i + 1))^2}{\sum(y_i - \bar{y})^2}$. After adding 1 to each x-term, the R2 becomes $1 - \frac{\sum(y_i - 2(x_i + 1)))^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - 2(x_i + 1) + 3))^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - 2x_i + 1))^2}{\sum(y_i - \bar{y})^2}$. Since the R2 remains the same, the R2 remains the same, at 0.8.

Part C Problem: If all of the y-coordinates of the data were doubled, what would the new line of best fit be? Can you predict what the new R2 would be?

Part C Solution: Since all the y-coordinates have doubled, the model's output should have doubled too. Thus yields $2(2x-1)=4x-2$. The line of best fit for this dataset is thus $y=4x-2$.

The equation for R2 was $1-\frac{\sum(y_i-2x_i+1))^2}{\sum(y_i-\bar{y})^2}$. After doubling each y element, the R2 becomes $1-\frac{\sum(2y_i-4x_i+2))^2}{\sum(2y_i-2\bar{y})^2}$.

As in part A, $\bar{y}$ refers to the mean of the y-coordinates prior to being doubled. Doubling the coordinates is equivalent to doubling the mean. We can prove this as follows: $\bar{y}=\frac{\sum y_i}{N}$. Doubling each element yields $\frac{\sum 2y_i}{N}$. Per the distributive property, this can be rewritten as $2\frac{\sum y_i}{N}=2\bar{y}$.

We can prove that the R2 prior to and after doubling each y_i term is the same:

$$
\begin{aligned}
1-\frac{\sum(y_i-2x_i+1))^2}{\sum(y_i-\bar{y})^2} &= 1-\frac{\sum(2y_i-4x_i+2))^2}{\sum(2y_i-2\bar{y})^2} \\
&= 1-\frac{4\sum(y_i-2x_i+1))^2}{4\sum(y_i-\bar{y})^2} \\
&= 1-\frac{\sum(y_i-2x_i+1))^2}{\sum(y_i-\bar{y})^2}
\end{aligned}
$$

Since the R2 remains the same, the R2 remains the same, at 0.8.

Part D Problem: If all of the x-coordinates of the data were doubled, what would the new line of best fit be? Can you predict what the new R2 would be?

Part D Solution: Since all the x-coordinates have doubled, the model would need to be stretched horizontally by a factor of 2: $2\left(\frac{1}{2}x\right)-1=x-1$. The line of best fit for this model is thus $y=x-1$.

The equation for R2 was $1-\frac{\sum(y_i-2x_i+1))^2}{\sum(y_i-\bar{y})^2}$. After doubling each x element, the R2 becomes $1-\frac{\sum(y_i-(2x_i)+1))^2}{\sum(y_i-\bar{y})^2}$.

Since the R2 remains the same, the R2 remains the same, at 0.8.

Part E Problem: If all of the y-coordinates of the data were squared, can you predict what the new line of best fit would be? Can you predict what the new R2 would be?

Part E Solution: The slope of the line of best fit is given by $m=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$. That is, finding the line of best fit requires knowing the mean of the data after its elements are squared. This is impossible to do, given only the mean of the data prior to its elements being squared. Thus, we do not know the line of best fit. Correspondingly, we do not know the new R2 because the R2 is dependent on knowing the line of best fit.

Problem 3

Problem: Laverne, Gordon, and Tanya are trying to model the price of gas per gallon in their city. Each of them tries a slightly different approach.

- Laverne constructs a line of best fit for the data and performs an F-test. She gets an R2 of about 0.02 and an F-score of 2.04, and based on that she decides that we can't conclude anything with 90% confidence or better.
- Gordon takes Laverne's results to mean that we can't accurately model gas prices with just one line, so he tries a multipart linear model. He tries a whole bunch of different options, and eventually comes up with a five-part linear model which has an R2 of 0.99, and an F-score of nearly 10,000.
- Tanya recalls from the news that, at a certain point during the period of time they have data on, a toll was added to one of the major highways in the area. She hypothesizes that the toll may have changed people's driving habits and therefore changed the demand for gas, so she breaks up the data into "before the toll" and "after the toll" data sets, and constructs lines of best fit for each one, resulting in a two-part linear model. Her model has an R2 of 0.76, and the F-scores of the pieces are 27 and 660, respectively.

Rank these three approaches from "most trustworthy" to "least trustworthy", and from "most successful" to "least successful". As always, explain your reasoning.

Solution: In determining which of the three people are the most "trustworthy", we will need to define what "trustworthiness" entails. For the purposes of this question, we will define "trustworthy" as "well-represents the truth of the problem as a whole". For example, in an extreme case, we could have a multipart linear function with $N-1$ lines for a dataset with N points. The metrics would point to a great model, but because the model is high-variance, it is difficult to trust these as well-representing the truth of a model in predicting the phenomena of gas prices and not just the given set of data points. The trustworthiness is then simply a ranking of how variant the models are; Laverne's is one piece, Tanya's is two pieces, and Gordon's is five pieces. The approaches ranked from most trustworthy to least trustworthy are thus Laverne's, Tanya's, and Gordon's.

To make more concrete the notion of "success", we will define the "success" of an approach by how well it will perform in modelling the phenomenon. While some of this analysis will borrow from the metrics of how well the model fits the data, we also need to consider the variance of the model. While Laverne's model is very trustworthy, it likely wouldn't do well in modelling the price of gas. Gordon's model is very untrustworthy; it is high-variance and the R^2 is "too high" given the nature of the model; Gordon's model is likely over-fitting to the current data and will not perform well on new gas price data. On the other hand, Tanya's model seems to be complex enough to model the phenomena well without being so high-variance that it overfits to the data. Thus, Tanya clearly has had the most success. The ranking of least successful, then, depends on whether we consider Gordon or Laverne's endeavors to be more likely to model the phenomena successfully. To answer this question rigorously would require understanding the nature of the problem – assuming that it is a forecasting problem, it seems that Laverne's model would be more successful at modelling new data, since the line is the "product" of considering all the data, rather than only one subset, as Gordon's model does. Thus, the approaches ranked from most successful to least successful are Tanya's, Laverne's, and Gordon's.

Problem 4

Problem: According to the Washington Post, the total number of new COVID-19 cases reported in the US each day of the final week of March 2021 was as follows:

Date	Cases
3/25	66,740
3/26	76,242
3/27	66,826
3/28	45,728
3/29	66,429
3/30	61,907
3/31	69,216

By constructing a best-fit line for this data, can we conclude with 95% confidence that the number of cases was increasing, decreasing, or neither during that time?

Solution: Firstly, let us change the date into "Days since 3/25" so it can be modelled by a mathematical model:

Days Since 3/25	Cases
0	66,740
1	76,242
2	66,826
3	45,728
4	66,429
5	61,907
6	69,216

We can use the same process as outlined in Problem 1 to find the best line of fit. The numerator for m evaluates to about -21639; the denominator evaluates to 28. Therefore, the resulting slope of the best line of fit is about -772.8214. The y-intercept can be correspondingly calculated to be 67,240.958. The line of best fit is thus $y = -772.8214x + 67,240.958$.

Using the line of best fit, we can calculate the R2; the numerator evaluates to ≈ 516292437.9643 and the denominator evaluates to ≈ 533015520.8571. Dividing and subtracting from 1 yields an R2 score of 0.0314. The F-score is thus $\frac{0.0314}{1-0.0314}(7-2) = 0.16208$. The F-test value for 95% confidence with a sample size of 7 is 5.59; the F score of the line of best fit falls far from this. Thus, we cannot conclude whether the number of cases was decreasing or increasing – at least in a linear sense.

Problem 5

Problem: Again According to the Washington Post, the total number of new COVID-19 cases reported in the US each day from Jan. 13 2021 to Jan. 31 2021 was as follows: [table omitted]

By constructing a best-fit line for this data, can we conclude with 95% confidence that the number of cases was increasing, decreasing, or neither during that time?

Solution: Substituting the date as the independent variable for "number of days since 1/13", we can construct a line of best fit. The numerator for the slope evaluates to $-2,805,000$ and the denominator evaluates to 570; dividing yields a slope of -4921.0526. The y-intercept is correspondingly 219578.6316. The line of best fit is thus $y = -4921.0526x + 219578.6316$. To calculate the R2, we find the numerator to be 8212330914.9474 and the denominator to be 22015883546.5263; dividing and subtracting from 1 yields an R2 of about 0.6270. The F-score is thus $\frac{0.627}{1-0.627}(19-2) = 28.5764$. The F-test value for 95% confidence with a sample size of 20 is 4.35; the F-score for our model far surpasses this. Given that the slope is negative, we can conclude with 95% confidence that the number of cases was decreasing.

Problem 6

Problem: Take a look at the epidemic simulator linked in the description. Run it a few times to get a feel for how it works and what it does. Then, based on your observations and the data provided by the simulator, construct statistically justified models of the number of infected individuals and the number of recovered individuals over time. Finally, (rigorously) test your model against another run of the simulator, and analyze the model's performance. The simulator allows you to adjust particular parameters; feel free to leave it on the defaults, or to experiment with other options. If you experiment, choose a set of parameters with nontrivial results – the simulation should last at least 500 frames.

Solution: We will set the infection rate at 0.01, the infection strength at 90, and the infection persistence at 0.7. We collect the following data:

Frame	Infected	Recovered
0	1	0
100	29	0
200	143	0
300	439	2
400	1010	31
500	1615	161
600	2377	448
700	3064	1048
800	3899	1803
900	4638	2878
1000	5397	4164
1100	5994	5789
1200	6558	7592
1300	7188	9647
1400	8039	11844
1500	8060	14263
1600	6809	16958
1700	4707	19945
1800	2935	22315
1900	1728	23758
2000	910	24667
2100	343	25257
2200	116	25484
2300	21	25579
2400	1	25599

Interestingly, we can observe two patterns that may be useful: the number of infected people seems to be symmetric in that it rises and then falls. The number of recovered people seems to increase steadily, although faster in the middle and slower near the beginning and ending frames.

Let us first attempt to model the number of infected people. For the sake of simplicity, let our model consist of two lines modelling frames 0-1400 and 1500-2400. The former should model the up-sloping trend of number of infected individuals, and the latter should model the down-sloping trend. For the first line, we find m using process outlined in previous problems through the formula for slope of the line of the best fit; this yields $\frac{17,424,400}{2,800,000} = 6.223$. The y-intercept is -996.7. The line for the first few frames is thus $y_1 = 6.223x - 996.7$, where x is in the frame number (0, 100, 200, etc.). The R2 score for y_1 is $1 - \frac{2,234,114.4}{110,666,155.6} \approx 0.9798$, which is a great fit. Multiplying $1 - 0.9798$ by 1.166 and subtracting from 1 yields an adjusted R2 score of 0.97645, which is still very high. The small difference between the R2 and the adjusted R2 suggests that there is sufficient data to attain a trustworthy model.

For the second line, we find $m = \frac{-7{,}579{,}800}{825{,}000} \approx -9.1876$. The y-intercept comes out to $20,478.891$. The model for the second part of the data is thus $y_2 = -9.1876x + 20,478.891$. The R2 score for y_2 is $1 - \frac{10{,}711{,}649.891}{80{,}352{,}096} \approx 0.8667$, which is a good fit. Multiplying $1 - 0.8667$ by 1.2857 and subtracting from 1 yields an adjusted R2 score of 0.8286.

Our model for the number of infected people is thus the following piecewise linear model:

$$y = \begin{cases} 6.223x - 996.7 & \text{if } x < 1450 \\ -9.1876x + 20,478.891 & \text{if } x \geq 1450 \end{cases}$$

The following new data was generated with the same parameters from the simulation:

Frame	Infected	Recovered
0	1	0
100	38	0
200	241	0
300	621	3
400	1105	38
500	1717	249
600	2372	630
700	3166	1174
800	3890	2013
900	4619	3077
1000	5338	4389
1100	6057	5990
1200	6720	7789
1300	7502	9838
1400	7784	12121
1500	7437	14637
1600	6176	17418
1700	4620	20019
1800	3119	22117
1900	1853	23629
2000	944	24644
2100	370	25230
2200	119	25481
2300	12	25588
2400	1	25599

The R2 score for y_1 when evaluated on this dataset is $1 - \frac{2{,}005{,}131.8}{109{,}377{,}285.6} \approx 0.9817$. The root mean squared error is 365.6165. The R2 score for y_2 when evaluated on this new dataset is $1 - \frac{7{,}750{,}440.63}{68{,}233{,}276.89} \approx 0.8864$. The root mean squared error is 880.3657. Despite the simplicity of this two-part linear model, it has shown to model both the training and testing data well.

The number of recovered individuals seems to just be increasing; we can just fit one linear model to it. The slope is $\frac{177{,}263{,}900}{13{,}000{,}000} \approx 13.6357$. The y-intercept thus is -5593.5415. The R2 score comes out to be $1 - \frac{194{,}761{,}968.9469}{2{,}611{,}876{,}603.04} \approx 0.9254$, which indicates that this is a pretty good fit. Multiplying $1 - 0.9254$ by 1.0909 and subtracting from 1 yields an adjusted R2 of 0.9186. Testing on the test dataset yields an R2 of $1 - \frac{182{,}650{,}576.6438}{2{,}591{,}939{,}198.24} \approx 0.9295$ and a root mean squared error of ≈ 2702.9656. This simple linear regression model performs well both on the training and testing data sets, which suggests that it is a good model for the number of recovered people in any new simulation (run with the same parameters).