1. Linear Regression Model Analysis

1.1 Gender Bias Analysis

Yes there exists a gender bias. This bias is created by the value of:

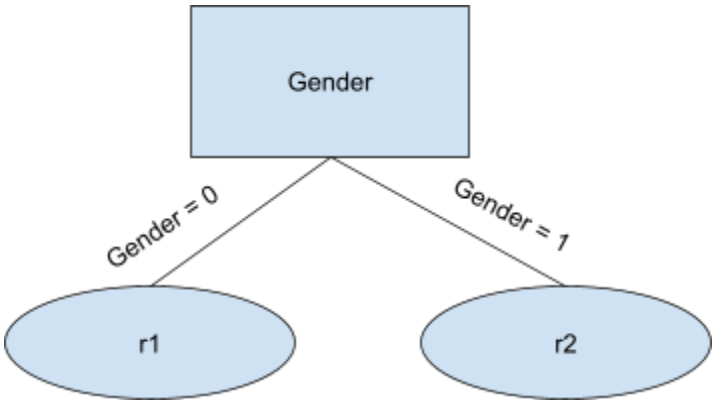
β^3= 10 → coefficient of gender

β^5 = -3 → coefficient for interaction between GPA and gender

These 2 coefficients are only influencing the result when the participant is female, because otherwise the value for gender is 0 and all coefficients are gonna be multiplied by 0. However if the candidate is female, she gets a value of 1 for gender variable and that is multiply by β^3 would add 10 to the final result. β^5 influence the salary depending on the GPA and gender. So with higher GPA this value gets smaller after being multiplied by -3 and reduces the salary, if the person is female.
Basically female candidates with lower GPA have higher salaries than males with the same features, however as the GPA of the female gets higher they have less salary compared to a male with the same characteristics, because of the influence of β^5 .

1.2 Model Conversion

Because we don't have any data we don't know the right threshold to divide the variables with a range, like IQ and GPA, because of this the 2 other features for the interactions are not useful either. So the only feature that can be used to get divided is gender, which make the tree below:



where:

$r1 = \beta^0 + (\beta^1 * GPA) + (\beta^2 * IQ) + (\beta^4 * (GPA * IQ)) \rightarrow$

$r1 = 50 + (20 * GPA) + (0.07 * IQ) + (0.01 * (GPA * IQ))$

(Gender and the coefficient got deleted here because gender is 0.)

$r2 = \beta^0 + (\beta^1 * GPA) + (\beta^2 * IQ) + \beta^3 + (\beta^4 * (GPA * IQ)) + (\beta^5 * GPA)$

$r2 = 50 + (20 * GPA) + (0.07 * IQ) + 10 + (0.01 * (GPA * IQ)) + (-3 * GPA)$

(I didn't write gender here because its 1 so i just wrote the coefficient.)

2. Data Generation and Model Fitting

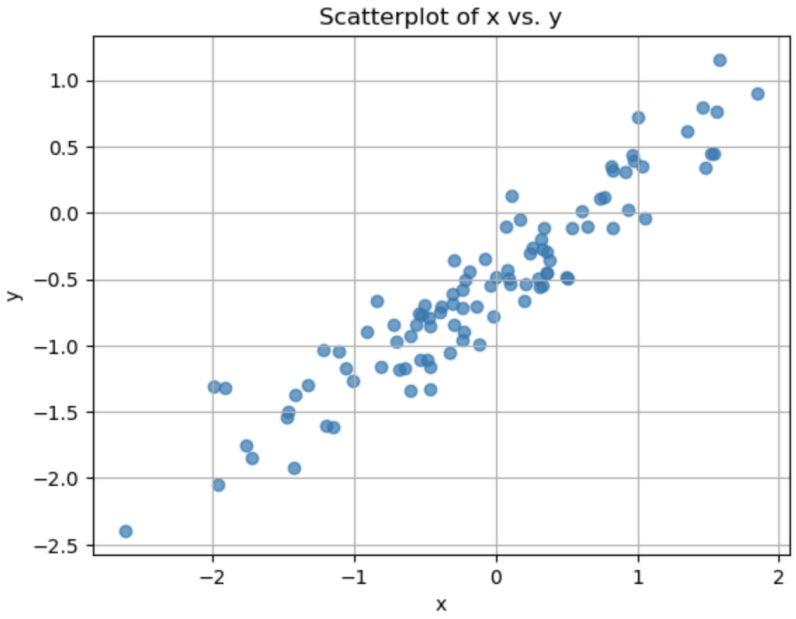
2.1 First Data Generation

a What is the length of the vector y? 100

b What are the values of β0 and β1 in this linear model?

β0 = - 0.5 β1 = 0.75

2.2 First Data Visualization



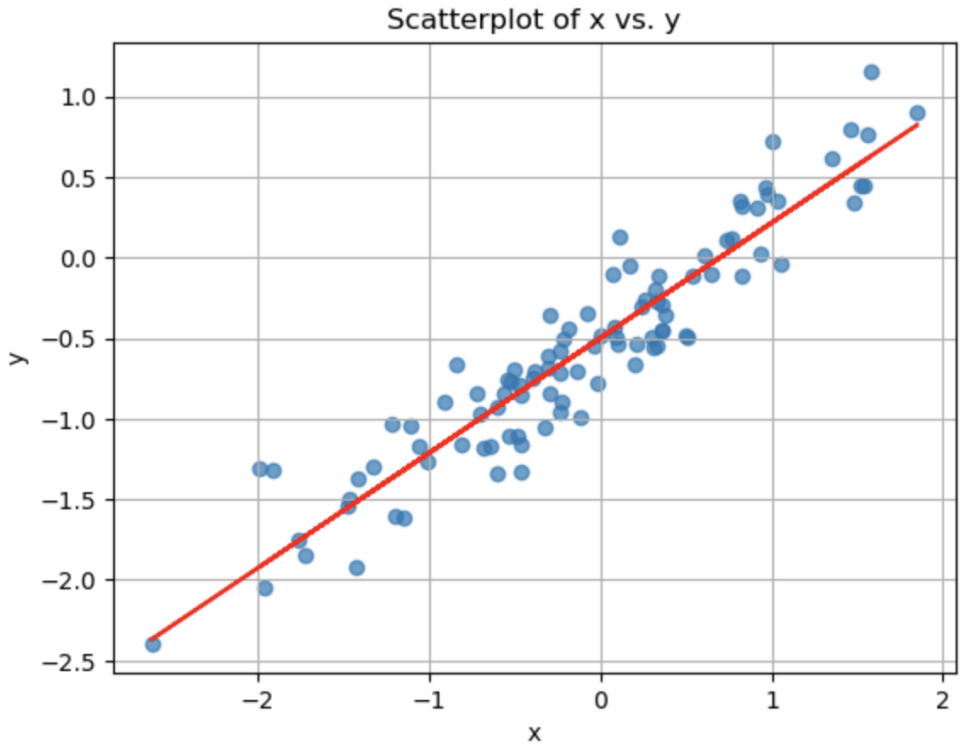
2.3 Fitting First Linear Regression

(a) How do the estimations of βˆ0 and βˆ1 compare to β0 and β1 ?

βˆ0 = - 0.4981430425340084

βˆ1 = 0.7141857099321391
Both are close to the real betas but slope or βˆ1 has a bigger difference to the original β1 than βˆ0 to β0.

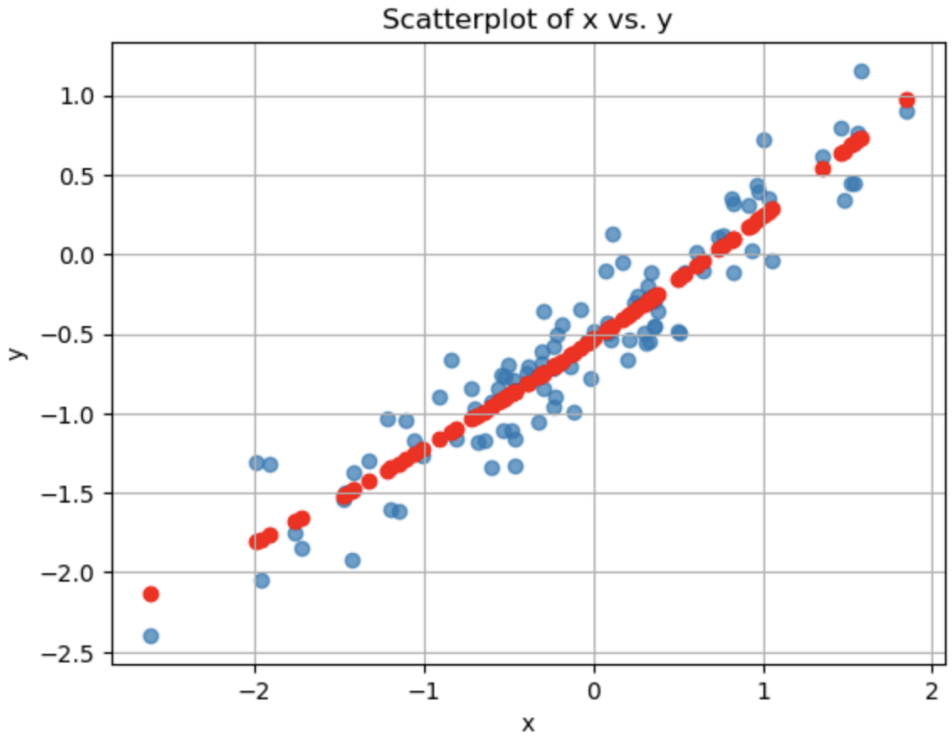
(b) Display the least squares line on the scatterplot obtained in Subsection 2.2.



(c) Compute R2 statistics (using function r2 score from the sklearn.metrics module).
R2 statistics = 0.8829199943498208

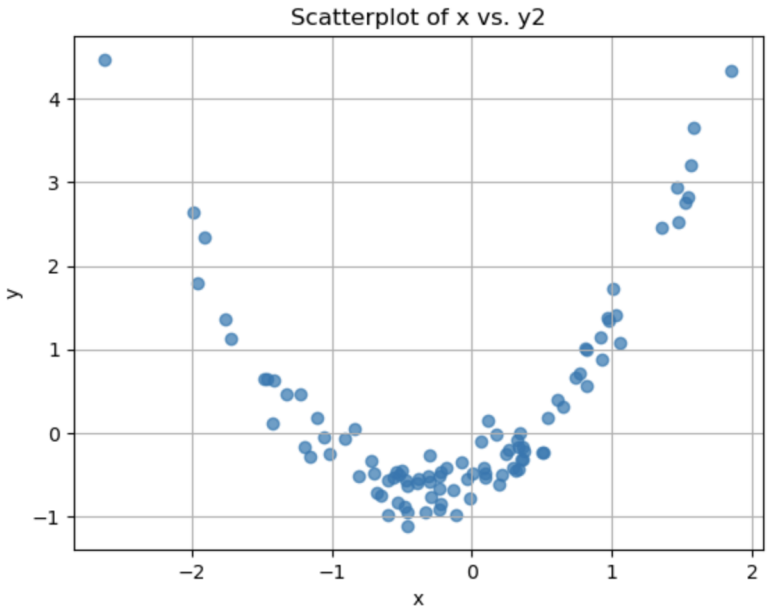
2.4 Fitting Second Linear Regression
(a) What is the estimated value for βˆ2? 0.04610749
(b) How do the estimations of βˆ0 and βˆ1 compare to β0 and β1 ?
βˆ0 = - 0.5345352870115676
βˆ1 = 0.73106418

(b) Display the least squares line on the scatterplot obtained in Subsection 2.2.



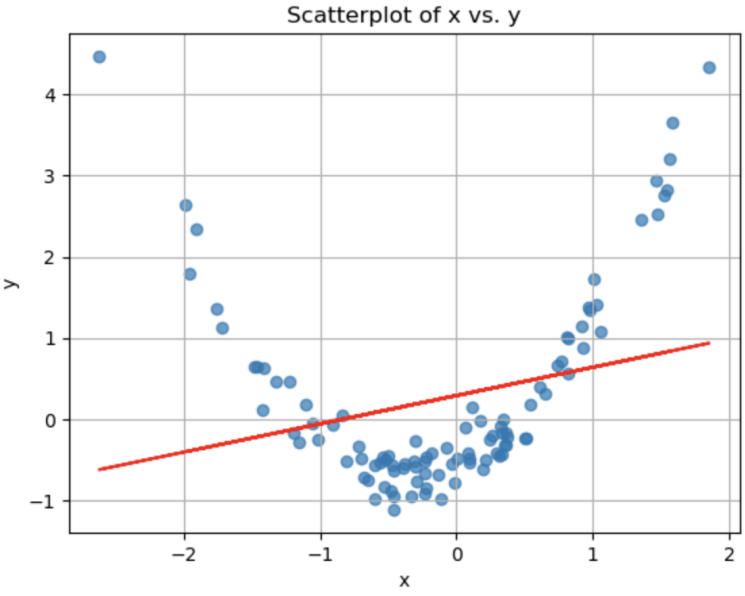
(d) Compute R2 statistics. 0.8883702598775292
(e) Is there evidence that the quadratic term improves the model fit? Explain your answer.
Yes. we can see that the quadratic model has a higher R2 score. This shows that the quadratic term improved the model fit. However from the value of R2 and the plots we can see that the difference is not that big. Another interesting thing we can see is that the βˆ0 and βˆ1 in the second linear regression have higher difference to the original β0 and β1than the first linear regression, however because of the quadratic term added still the R2 statics of the second linear regression is higher.

2.6 Second Data Visualization



2.7 Fitting Third Linear Regression

(a) How do the estimations of β^0 and β^1 compare to β0 and β1 ?
β^0 = 0.2911483870392203
β^1 = 0.3481179278875428
Really really bad, non of them are not even slightly close to the original β0 and β1 and this can be seen better in the plot below.
(b) Display the least squares line on the scatterplot obtained in Subsection 2.6.



(c) Compute R2 statistics. 0.06708446395233303

2.8 Fitting Fourth Linear Regression

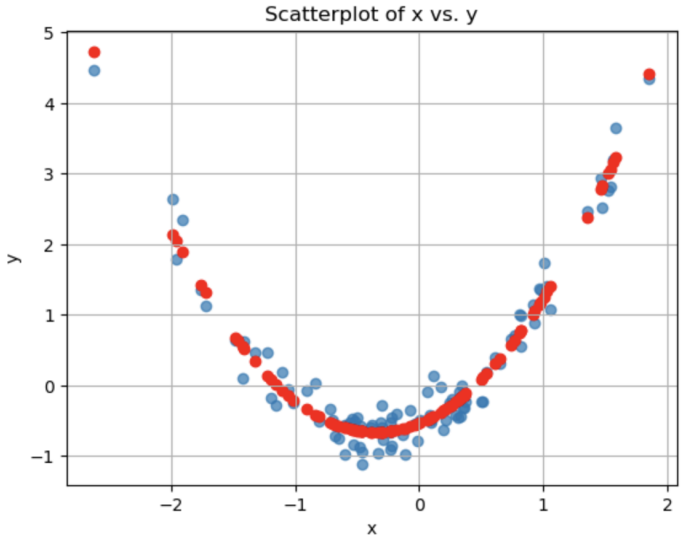
(a) How do the estimations of βˆ0, βˆ1, amd βˆ2 compare to β0, β1, and β2 ?

βˆ0 = -0.5345352870115672

βˆ1 = 0.73106418

βˆ2 = 1.04610749

(b) Display the least squares line on the scatterplot obtained in Subsection



2.6.

(c) Compute R2 statistics. 0.9643014933895998

(d) Is there evidence that the quadratic term improves the model fit? Explain your answer.

Yes, as we can see from the R2 score and the plots the quadratic term really improved the model fit. This is because the data is not linear so fitting it using a linear model is not gonna give good results as we can see from the 2.7(b) plot.

3. LASSO Regression Model Analysis

In this section I'm doing the same thing as before but using Lasso instead of normal linear regression. The difference here is that Lasso uses L1 regularization and the strength of regularization is controlled by value alpha. So basically if we set the value of alpha = 0 its the same model as a normal linear regression. I did the testing with different value of alpha but the smaller the alpha the better the result. For this PDF I'm leaving alpha =1 so the really bad results are shown.

There are many reasons why regularization is not useful in this case. Here we only have 1 feature and not many data samples. Regularization is usually useful for cases where the number of features are more than samples, which is not true in this example.

The result are shown in the jupyter notebook.

I, Parand Mohri, hereby declare that I have not used large language models or any automated tools for generating the written answers and interpretations in this Analytical Report.