# Analytical Report

Yulin Zhou
i6336858

October 2, 2023

# 1 Linear Regression Model Analysis

## 1.1 Gender Bias Analysis

**Question:** Using the linear regression model given above, investigate the role of gender in determining the starting salary after graduation. Specifically, analyze whether the model indicates a gender bias in starting salaries. Explain your reasoning based on the coefficients $\hat{\beta}_0$-$\hat{\beta}_5$

**Conclusion:**

Yes, there is a gender bias in starting salaries based on the given information.

**Brief Explanation:**

For the females, although their function has a larger intercept(10 more than the males), the coefficient corresponding to GPA is 6 less than the males.

Therefore, on average, the males who get GPAs higher than $\frac{5}{3} \approx 1.67$, which is quite easy to get, could have higher salaries than the females with the same values of the other variables(i.e., GPA and IQ).

Obviously, there is a gender bias.

**Detailed Explanation:**

Although we could find that $\hat{\beta}_3 > 0$ at first glance, things become different when $\hat{\beta}_5 < 0$ is noticed. If $X_3 = 1$ for female and $X_3 = 0$ for male are put into the function, we can get two following regression functions:

$$Y_{Female} = 60 + 17 * GPA + 0.07 * IQ + 0.01 * GPA * IQ$$

$$Y_{Male} = 50 + 23 * GPA + 0.07 * IQ + 0.01 * GPA * IQ$$

Then, we will find that

$$Y_{Female} - Y_{Male} = 10 - 6 * GPA$$

Therefore, for the males and females with the same variables(i.e., IQ and GPA) and GPA is $\geq \frac{5}{3}$, the males can get higher salaries on average. By the way, the higher the GPA, the greater the average salary difference between males and females, even if they have the same IQ and GPA. For example, for females and males who have the same IQ and same GPA=3, males' salaries are 8000 euros higher than females on average.

Frankly, getting a GPA higher than $\frac{5}{3} \approx 1.67$ is really easy. It is even difficult for students with such low GPAs to obtain a diploma in lots of universities. To be honest, on average, we can say females don't have any advantageous conditions based on the given information.

In conclusion, there is an obvious gender bias between females and males in starting salaries. Specifically, males are much easier to get higher salaries than females on average.

### 1.2 Model Conversion

Based on the explanation mentioned above, if $X_3 = 1$ is put into the original function, then we can get the linear regression function for females. Also for the $X_3 = 0$ case, we could get the function for males.

Therefore, we obtain a model tree with two different conditions, which is shown in Fig.1.

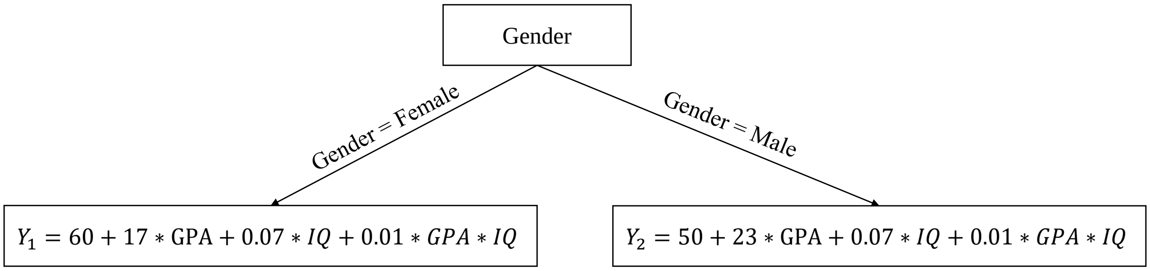
Gender
Gender = Female
Gender = Male
$Y_1 = 60 + 17 * \mathrm{GPA} + 0.07 * IQ + 0.01 * GPA * IQ$
$Y_2 = 50 + 23 * \mathrm{GPA} + 0.07 * IQ + 0.01 * GPA * IQ$

Figure 1: Converted model tree from the linear regression model

## 2 Data Generation and Model Fitting

### 2.1 First Data Generation

According to the requirement, I used the following codes to generate the vector $x$ and $eps$, and create the vector $y$ by the given function: $y = -0.5 + 0.75 * x + eps$

```
# create a vector x containing 100 observations (mean 0, variance 1)

# set the random seed to ensure the same experimental results
np.random.seed(42)

mean_x = 0; var_x = 1
x0 = np.random.normal(mean_x, var_x**0.5, 100)

# create a vector eps containing 100 observations (mean 0, variance 0.25)
mean_eps = 0; var_eps = 0.25
eps = np.random.normal(mean_eps, var_eps**0.5, 100)

# generate a vector y according to the model:
# y = -0.5 +  0.75x + eps
y0 = -0.5 + 0.75*x0 + eps
```

**Question a:** What is the length of the vector y?

**Answer:**

```
print(’The length of the vector y is’, len(y0))
## output: The length of the vector y is 100
```

Therefore, the length of the vector y is 100.

**Question b:** What are the values of $\beta_0$ and $\beta_1$ in this linear model?

**Answer:** The value of $\beta_0$ is the expectation of the variables or values that are independent of $x$. As we can see, the consistent and variable $eps$ are independent of $x$, therefore, $\beta_0 = E(-0.5 + eps)$.

For the value of $\beta_1$, it's the expectation of the coefficient of x linear term, which is $E(0.75)$.

The specific calculations are below:

$$\beta_0 = E(-0.5 + eps) = E(-0.5) + E(eps) = -0.5 + 0 = -0.5$$

$$\beta_1 = E(0.75) = 0.75$$

## 2.2 First Data Visualization

**Requirement:** Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

**Answer:**

The scatter plot of x and y is shown in Fig.2. As we can we, the range of x is around -2.5 to 2, and the range of y is roughly -2.5 to 1.5. In general, when x becomes larger, y will also be larger, showing a positive correlation. From the naked eye, it's a roughly linear relationship.

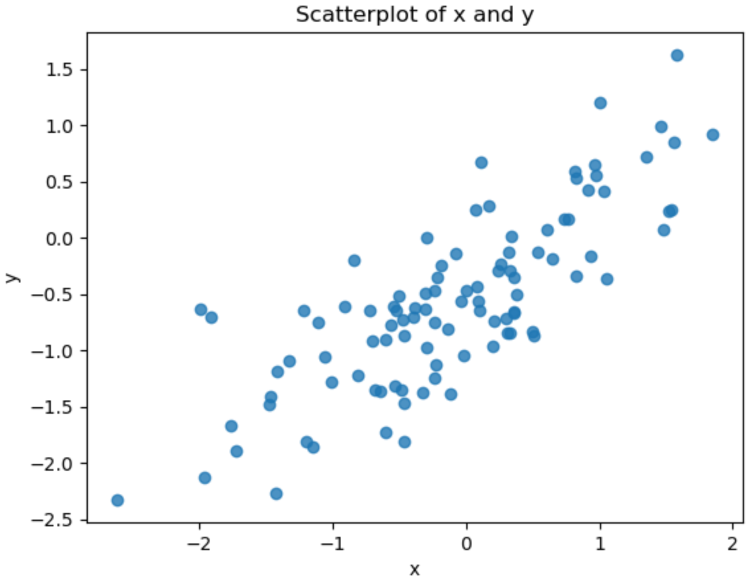

Figure 2: Scatter plot of x and y

## 2.3 Fitting First Linear Regression

**Question a:** How do the estimations of $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to $\beta_0$ and $\beta_0$?

**Answer a:**

After fitting the first linear regression model, we can know that $\hat{\beta}_0 = -0.4963$, $\hat{\beta}_1 = 0.6784$.

Compared with $\beta_0 = -0.5$ and $\beta_1 = 0.75$, it's obvious that they are really close to each other, which means the experimental results confirm the theoretical outcome to some extent.

**Question b:** Display the least squares line on the scatterplot obtained in Subsection 2.2.

**Answer b:**

Based on the first linear model, I predict $\hat{y}$ (i.e., y_m1 in the codes) on possible $x$ values. Then, I draw a red line according to $(x, \hat{y})$, which is shown in Fig.3.

```
# generate x values and predict on x_step to draw the line of model 1
x_step = np.arange(min(x), max(x), 0.02).reshape(-1,1)
y_m1 = model1.predict(x_step)

# draw the figure
scatter_fit_plot(x, y, x_step, y_m1,
                 'Line of model 1',
                 "Scatterplot of x and y with Line of model 1")
```

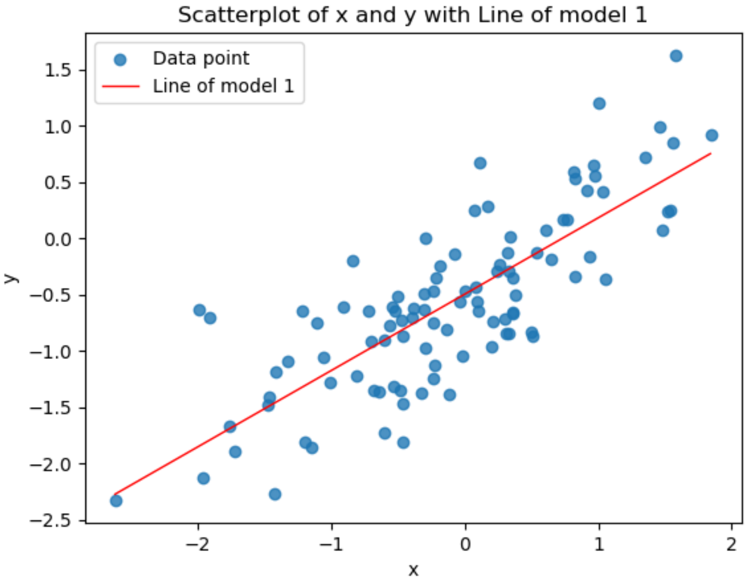

Figure 3: Scatter plot of x and y with the line of model 1

**Question c:** Compute $R^2$ statistics (using function r2 score from the sklearn.metrics module).

**Answer c:**

From the following codes, we can know that $R^2 = 0.6298$, meaning that the prediction model is reasonable.

```
from sklearn.metrics import r2_score

m1_r2 = cal_r2(model1, x, y)
m1_r2
## output: 0.6297598193059208
```

## 2.4 Fitting Second Linear Regression

Now fit a polynomial regression model that predicts $y$ using $x$ and $x^2$. Comment on the model obtained:

**Question a:** What is the estimated value for $\hat{\beta}_2$?

**Answer a:**

The estimated value for $\hat{\beta}_2 = 0.0922$

**Question b:** How do the estimations of $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to $\beta_0$ and $\beta_1$?

**Answer b:**

The estimated value for $\hat{\beta_0} \approx -0.5691$, $\hat{\beta_1} \approx 0.7121$, where $\beta_0 = -0.5$ and $\beta_1 = 0.75$. As we can see, they are still roughly equal to each other. Their differences are not so large that the interpretation of intercepts and one-time coefficients could be different in Model 1 and Model 2.

Therefore, in these two models, the effects of variable x are almost the same.

**Question c:** Display the least squares line on the scatter plot obtained in Subsection 2.2.

**Answer c:**

Similarly, I generated plenty of possible $x$ and corresponding $x^2$ values, which are the input data to draw the curve of model 2.

By applying the following codes, I got Fig.4.

```
# generate input data to draw the curve
df_x_step = pd.DataFrame({’x’: list(x_step), ’x^2’: list(np.square(x_step))},
                         columns=[’x’, ’x^2’])
y_step_m2 = model2.predict(df_x_step)

scatter_fit_plot(x, y, x_step, y_step_m2,
                 ’Model 2 curve’,
                 "Scatterplot of x and y with model 2 curve")
```

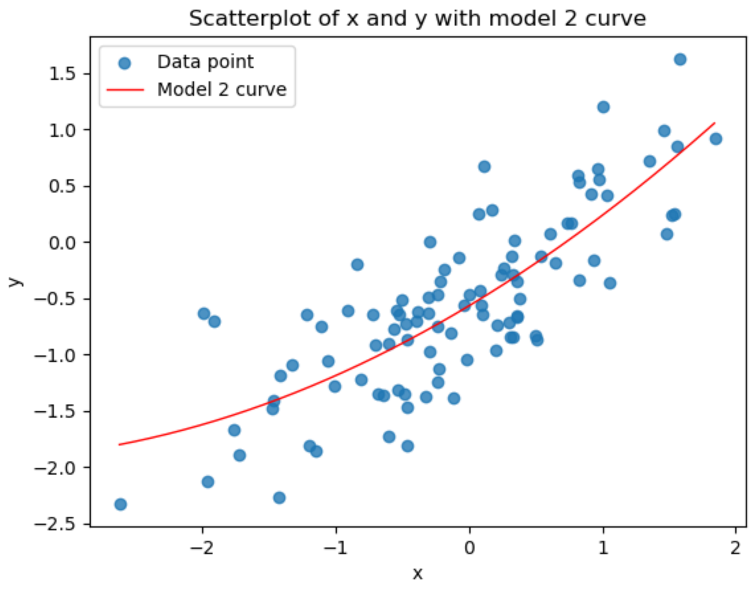

Figure 4: Scatter plot of x and y with the curve of model 2

**Question d:** Compute $R^2$ statistics.

**Answer d:** The statistic of $R^2 \approx 0.6470$.

```
m2_r2 = cal_r2(model2, df_quad_x, y)
m2_r2
## output: 0.6469951045504286
```

**Question e:** Is there evidence that the quadratic term improves the model fit? Explain your answer.

**Answer e:**

**Conclusion:** From the following three perspectives, there is no evidence that the quadratic term improves the model fit.

From the perspective of the change of $R^2$, it's increased from 62.98% for model 1 to 64.70% for model 2. The change is just around 1.7%. However, if we put more variables into the model, regardless of whether these variables are related to the dependent variables or not, the $R^2$ statistic will be increased definitely. Therefore, the $R^2$ improvement of 1.7% is not convincing enough to show that the quadratic term improves the model fit.

Also, I calculated the decrease of the mean squared error(i.e., MSE) of models 1 and 2. The outcome is that the MSE is only decreased by 0.010, which is really tiny. From the perspective of MSE, there is no improvement in model fit.

Besides, to further prove the above argument, I used the first 70% data to train the linear and quadratic models, and then I used the last 30% data to test the prediction accuracy. As a result, the mean absolute error(i.e., MAE) of the linear model is 0.329, while 0.336 for the quadratic model. This result is interesting because it shows that if we use these two models to predict unknown data, the linear model even could perform slightly better.

### 2.5 Second Data Generation

By the following codes, I generated the new vector y.

```
# Using x and eps, generate a vector y according to the model:
# y = -0.5 + 0.75x + x2 + eps

y = -0.5 + 0.75 * x0 + x_2 + eps
```

### 2.6 Second Data Visualization

**Question:** Create a new scatterplot displaying the relationship between x and y. Comment on what you observe.

**Answer:**

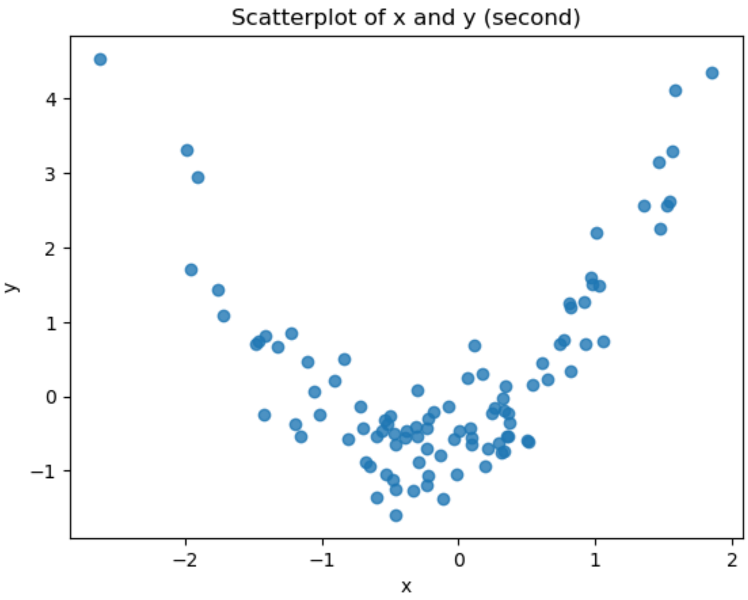

Figure 5: Scatter plot of x and y(second)

The scatter plot of x and y is shown in the Fig.5. As we can see, when x becomes larger, the values of y decrease at first until get around -0.3, then values of y increase when x is increasing. From the naked eye, the relationship between x and y is quadratic.

## 2.7 Fitting Third Linear Regression

Fit a least squares linear model LinearRegression() from the module linear model to predict y using x. Comment on the model obtained:

**Question a:** How do the estimations of $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to $\beta_0$ and $\beta_0$?

**Answer a:**

$$E(\beta_0) = E(-0.5 + eps) = -0.5 + E(eps) = -0.5$$
$$E(\beta_1) = E(0.75) = 0.75$$

After building the model 3, I found $\hat{\beta}_0 \approx 0.2930$ and $\hat{\beta}_1 \approx 0.3123$, which are quite different from the $\beta_0$ and $\beta_1$.

As we can see, the intercept even changes the sign. The intercept means the expected value of y when x is equal to 0. According to Fig.5, when values of x are around 0, the corresponding values of y are smaller than zero. Therefore, $\hat{\beta}_0 \approx 0.2930$ is really unreasonable, meaning the model 3 is quite bad.

Besides, it's clear that y goes down and then goes up in Fig.5. But, $\hat{\beta}_1 \approx 0.3123$ means that the expected value of y will always increase when x becomes larger. Thus, the obtained $\hat{\beta}_1$ is also unrealistic.

**Question b:** Display the least squares line on the scatterplot obtained in Subsection 2.6.

**Answer b:**

I used the possible x values as the input data, and then I used Model 3 to predict the corresponding y values. The result is shown in Fig.6.

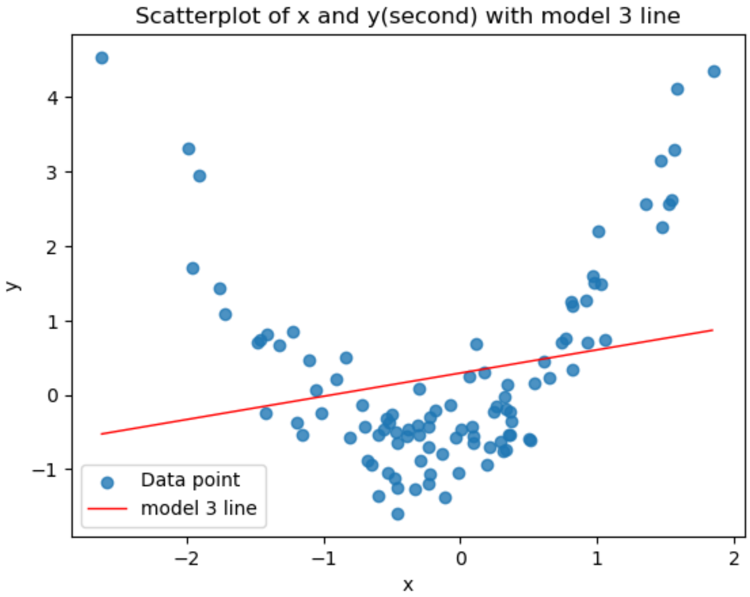

Figure 6: Scatter plot of x and y(second) with model 3 line

**Question c:** Compute $R^2$ statistics.

**Answer c:**

The value of $R^2$ of the model 3 is roughly 0.046, which is really low. In other words, model 3 can only explain the 4.6% of the variance of y.

```
m3_r2 = cal_r2(model3, x, y)
m3_r2
## output: 0.045956423052825435
```

## 2.8 Fitting Fourth Linear Regression

Now fit a polynomial regression model that predicts y using x and $x^2$. Comment on the model obtained:

**Question a:** How do the estimations of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ compare to $\beta_0$, $\beta_1$, and $\beta_2$ ?

**Answer a:**

According to the way of generating y, we can know that $\beta_0 = E(-0.5 + eps) = 0.5$, $\beta_1 = E(0.75) = 0.75$, $\beta_2 = E(1) = 1$.

As for the model 4, $\hat{\beta}_0 \approx -0.5691$, $\hat{\beta}_1 \approx 0.7121$, and $\hat{\beta}_2 \approx 1.0922$.

As we can see, they are really close to each other, meaning Model 3 is pretty good.

**Question b:** Display the least squares line on the scatterplot obtained in Subsection 2.6.

**Answer b:**

I used the possible x values and their corresponding $x^2$ as the input data, and then I used Model 4 to predict the corresponding y values. The result is shown in Fig.7, which seems to fit the data quite well.

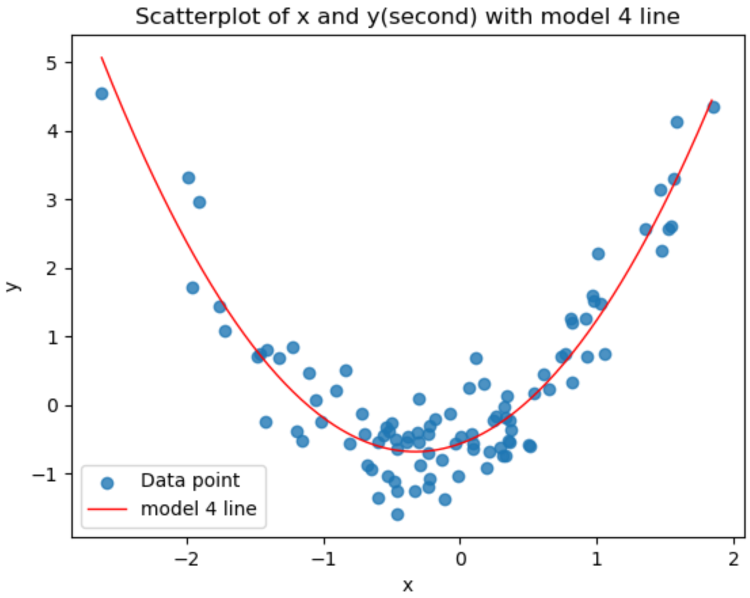

Figure 7: Scatter plot of x and y(second) with model 4 curve

**Question c:** Compute $R^2$ statistics.

**Answer c:**

The $R^2$ value is 0.8784, meaning model 4 can explain the most variance of the vector y. In other words, the model 4 is quite good.

**Question d:** Is there evidence that the quadratic term improves the model fit? Explain your answer.

**Answer d:**

**Conclusion:** There is enough evidence showing that the quadratic term improves the model fit.

Similar to section 2.4, I also used three perspectives to explain the conclusion.

From the perspective of the change of $R^2$, it's increased from 4.6% for model 3 to 87.8% for model 4. Thus, the increase is around 83.2%, from nearly can NOT explain the variance of y to almost can explain. This is a huge improvement.

Also, the value of MSE is almost decreased by 1.44, which is not ignorable. As for the prediction accuracy, MAE decreased from 1.133 for model 3 to 0.3359 for model 4. Thus, the MAE is deduced by around 0.796, which accounts for about 18% of the range of vector x. In other words, this deduction is valuable when applying the model.

Besides, from the answer (a) in the section2.7, we also can know that the estimated parameters in model 3 do NOT make sense, while the parameters in model 4 are quite reasonable.

All in all, there is enough evidence to show that the quadratic term improves the model fit.

## Academic Integrity Declaration

I, Yulin Zhou, 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.