Statistical Inference HW#4 programming sections

Student Name:
Pouya Haji Mohammadi Gohari

SID:810102113

Date of deadline
Monday 6th February, 2023

Dept. of Computer Engineering

University of Tehran

Contents

Attention: Note that we have fully solved 11 first questions in the previous submission and for last submission we solve the last 2 remaining questions.

1 Problem 12

From statsmodels we had fit a regression line and as you can see in the table 1: And we can see the paramters(B_0, B_1) as you can see in table 2:

Table 1: Result of ordinary least square

Dep. Variable:	lpsa	R-squared:	0.539
Model:	OLS	Adj. R-squared:	0.535
Method:	Least Squares	F-statistic:	111.3
Date:	Fri, 02 Feb 2024	Prob (F-statistic):	1.12e-17
Time:	02:32:08	Log-Likelihood:	-113.45
No. Observations:	97	AIC:	230.9
Df Residuals:	95	BIC:	236.1
Df Model:	1		
Covariance Type:	nonrobust		

Table 2: Fited line with regression using least square method

	coef	std err	t	$P > \|t\|$	$[0.025, 0.975]$
const	1.5073	0.122	12.361	0.000	1.265 , 1.749
lcavol	0.7193	0.068	10.548	0.000	0.584 , 0.855

The residuals can be calucalted with the model.scale which it is 0.620155621631307. Cramer-Rao Lower Bounds for the variances of the estimators: $[0.0148686, 0.00465027]$.
Empirical standard errors of the estimates:0.121937. From the Cramer-Rao lower bound we know that we must calculate the fisher information for $\hat{\beta}_1$:
First we must calculate the Likelihood:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \tag{1}$$

Log Likelihood:

$$\ln(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} \tag{2}$$

First deviation of Log Likelihood with respect to β_1:

$$\frac{\partial \ln(L)}{\partial \beta_1} = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)(-x_i)}{\sigma^2} \tag{3}$$

Second deviation:

$$\frac{\partial^2 \ln(L)}{\partial \beta_1^2} = \sum_{i=1}^{n} \frac{x_i^2}{\sigma^2} \tag{4}$$

Fisher information:

$$I(\beta_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})2}{\sigma^2} \tag{5}$$

From Cramer-Rao if the estimated parameter is unbiased(where we know that):

$$\text{Var}(\hat{\beta}_1) \geq \frac{1}{I(\beta_1)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{6}$$

The Empirical value is 0.004650270803149708 and the Cramer-Rao using formula is 0.004650270803149714 where it is same from the code provided in first place.

2 Problem 13

2.1 Which explanatory variable do you guess is the more significant predictor and why?

For this purpose we can fit a line on this datas and see the p-value of each, The one have smaller p-value is generally considered to be the more statistically significant predictor. And as you can see in table 3:

Table 3: Fited line with regression using least square method

	coef	std err	t	$P > \|t\|$	[0.025 , 0.975]
const	2.2359	0.328	6.814	0.000	1.584 , 2.887
age	0.0163	0.005	3.150	0.002	0.006 , 0.027
lpsa	0.1430	0.033	4.296	0.000	0.077 , 0.209

Since the p-value for lpsa is smaller therefore it is considered to be more statistically significant predictor.

2.2 For each explanatory variable

2.2.1 Investigate the linearity of data points using scatter plot of residuals.

We can use our model in previous part and obtain the residuals of the fitted regression line with OLS and see each scatter plot of both explanatory variable (age, lpsa) and as you can see in figures 1 and 2:

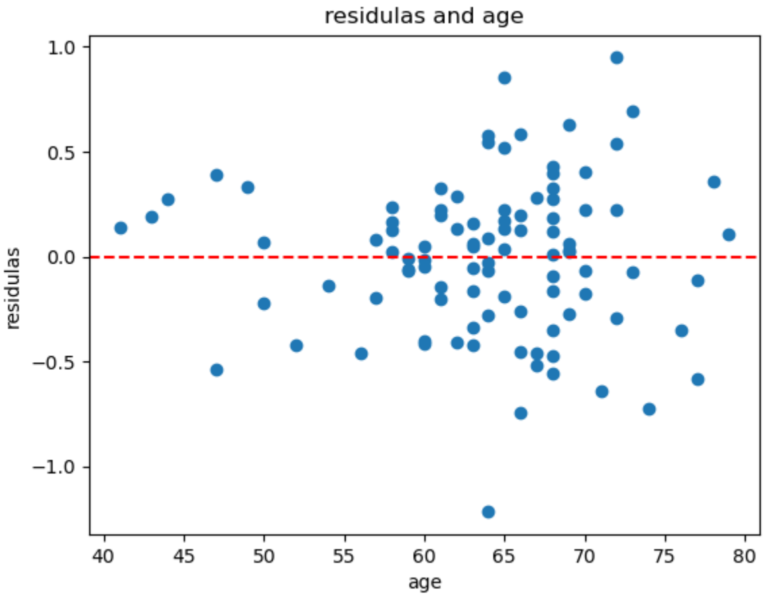

Figure 1: scatter plot for residuals and age

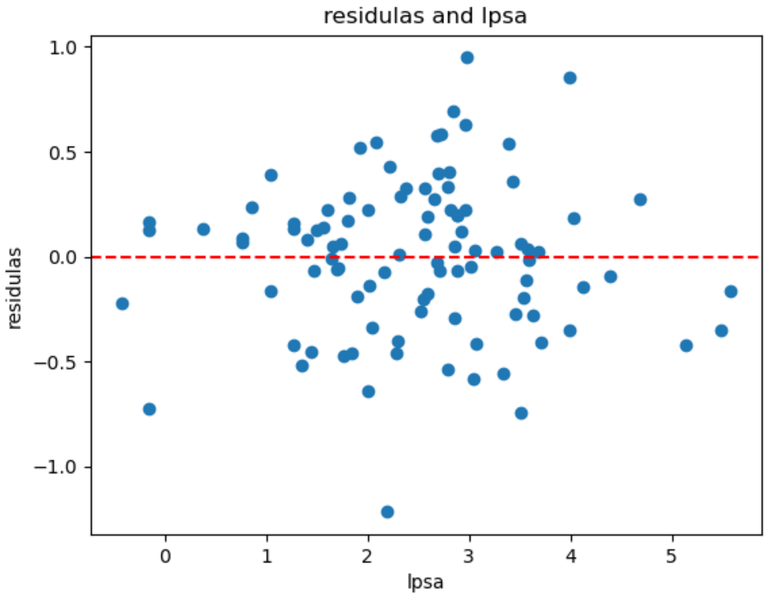

Figure 2: scatter plot for residuals and lpsa

In order to investigate the linearity of the relationship between the explanatory variables and the dependent variables using residuals we should consider the following aspcets:

- Randomness of residuals: Residuals must be scattered randomly around the horizontal line at zero.
- Pattern of symmetric structures: If there are clear patterns this indicates non-linearity.
- Homoscedasticity: The spread of the residuals should be consistent across all values of the explanatory variables.

So from both figures 1 and 2 we can see no pattern and also see a good scattering of randomness of residuals over horizontal line at zero this reasons indicates that the relationship might be linear.

2.2.2 Compute the least squares regression.

Approach one to see both explanatory variable in one equation: The general framework of least square regression is as follow:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, 2, \cdots, n \tag{7}$$

We can rewrite it as matrix notation as follow:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Now we can esimate with linear square method in order to estimate the parameters of fitted line:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{8}$$

The result of calculating this parameters using python is:

$$\hat{\beta} = \begin{bmatrix} 2.2358606 \\ 0.01626208 \\ 0.14303179 \end{bmatrix}$$

As you can see it is very close to OLS result in python.
Second approach see each explanatory variable once:
For this purpose we have fitted 2 regression lines with age and lpsa variable(using them one at a time): The result of OLS for age is in table4:

Table 4: Fitted line with regression using least square method for age and weight

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	2.3502	0.356	6.604	0.000	1.644 , 3.057
age	0.0200	0.006	3.618	0.000	0.009 , 0.031

The predictive line using same formula as before:

$$\text{weight} = 2.35015161 + 0.02002304\text{age} \tag{9}$$

The result of OLS for lpsa is in table ??:

Table 5: Fitted line with regression using least square method for lpsa and weight

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	3.2304	0.094	34.462	0.000	3.044 , 3.416
lpsa	0.1608	0.034	4.686	0.000	0.093 , 0.229

The predictive line is:

$$\text{weight} = 3.23036915 + 0.16081973\text{lpsa} \tag{10}$$

2.2.3 Write the predictive equation for the response variable and interpret its parameters.

From OLS the preditive equation is:

$$\text{weight} = 2.24 + 0.02\text{age} + 0.14\text{lpsa} \tag{11}$$

From my own calculation the predictive equation is:

$$\text{weight} = 2.2358 + 0.0162\text{age} + 0.143\text{lpsa} \tag{12}$$

The result are close to each other either it is from calculation or OLS(built-in function). The interpretation:

- Tntercept: This is the expected value of weight when both age and lpsa are zero.
- Slope of age: This parameter represents the change in the dependent variable weight for a one-unit change in the independent variable age.
- Slope of lpsa: This parameter represents the change in the dependent variable weight for a on-unit change in the independent variable lpsa.

Second approach fiting single line for each explanatory variables:
From OLS(age) results the line is:

$$\text{width} = 2.35 + 0.02\text{age} \tag{13}$$

From OLS(lpsa) results the line is:

$$\text{width} = 3.23 + 0.16 * lpsa \tag{14}$$

Both regression line from OLS results are very close to the calculations in the previous section!

2.2.4 Draw a scatter plot of the relation between these two variables overlaid with the least-squares fit as a dashed line.

Scatter plot of the relation between these two variables with the least-squares fit is in figures 3 and 4:

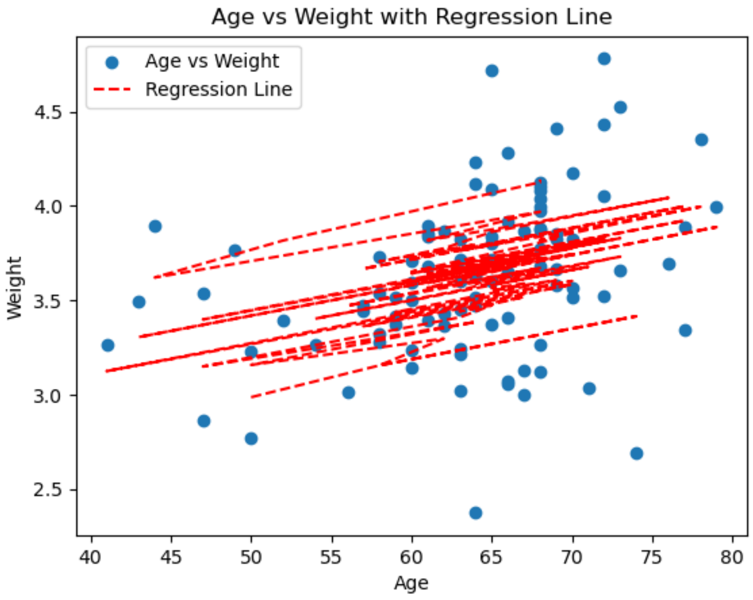

Figure 3: Age and fitted line

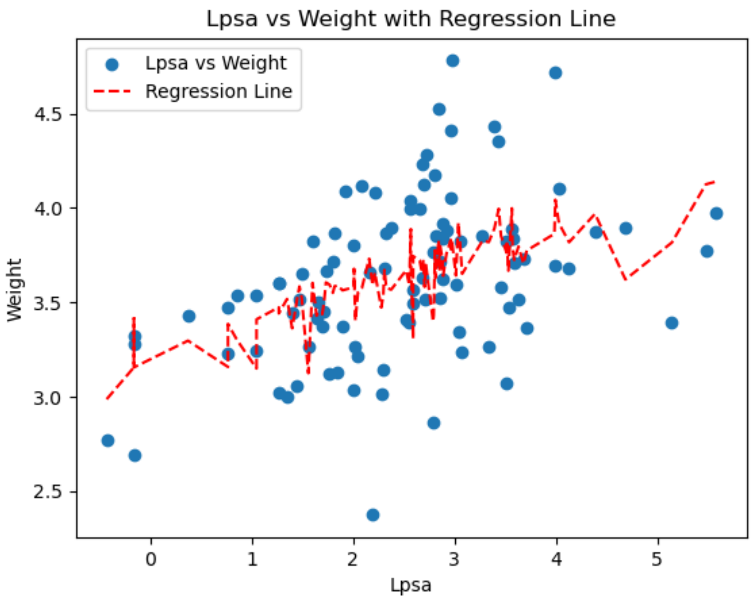

Figure 4: Lpsa and fitted line

As you can see in the figures above regression has beeen shown by red dashed line. In the first scatter plot the data points are fairly scattered around the regression line without a clear pattern of increasing or decreasing residuals which could indicate a linear relationship between age and weight.

From the second scatter plot(Lpsa and weight), we can conclude that the points do not follow the regression line like the first scatter plot.(There seems to be more variablity in weight for given lpsa values).

So we must consider the transformations of the lpsa variables.

Attention: We have seen the a lower p-value indicates a statistically significant relationship, but assesing linearity requires looking at the pattern of the residuals and the fit of the regression line to the data points.It's possibel for a variable to have a significant relationship with the dependent variable but not a prefectly linear one.From the scatter plots provided, if the regression line for lpsa vs weight appears ro be less straight or shows more variablity around the line compared to the plot for age vs weight, this suggest that while there is a significant relationship, it might not be prefectly linear, or it might be other factors influencing the variabltiy in weight that are not captured by lpsa alone. Second approach when for each explanatory variable we fit a regression line:

For age as a single explanatory variabl and weight the scatter plot between these two variables and and the regression line is in figure 6:

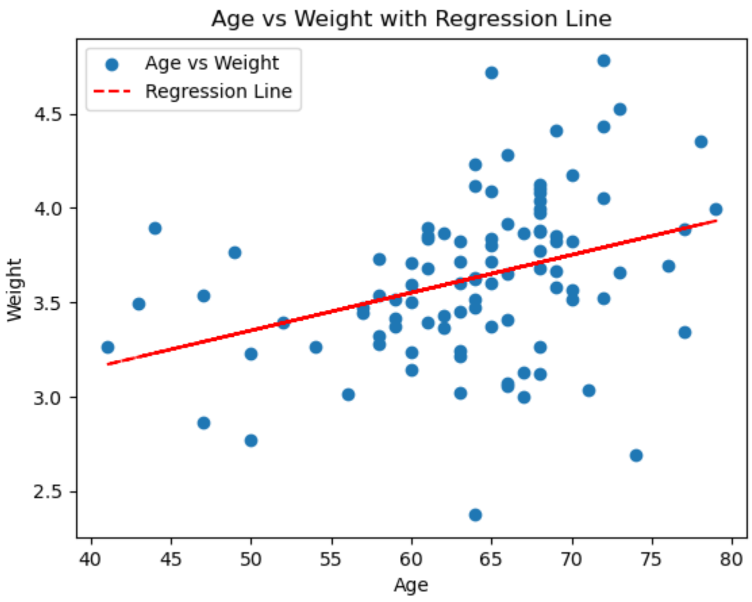

Figure 5: Age and fitted line

For lpsa and weight:

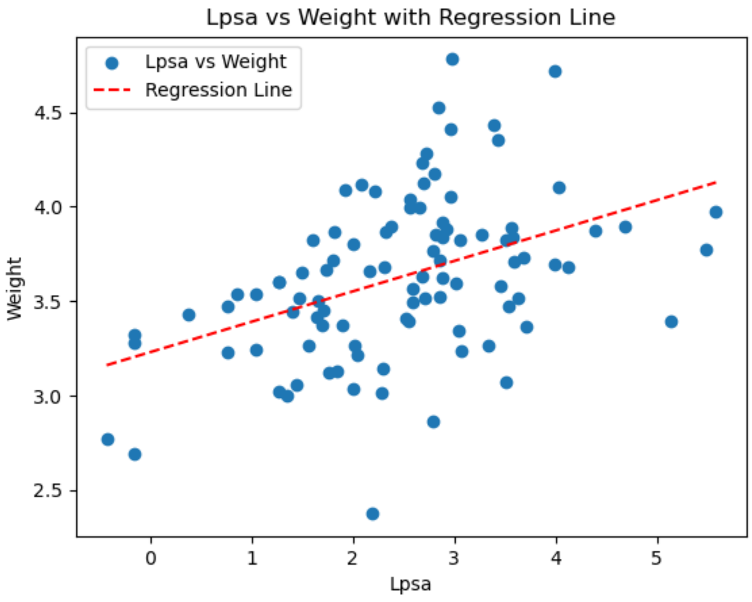

Figure 6: Lpsa and fitted line

2.3 Using the previous part results, try to explain which variable is the more significant predictor.

From table 3 since we have 0.14 coeffient for lpsa and lower p-value therefore lpsa is more significant predictive.

2.4 Choose a random sample of 50 data points from the dataset.

We have generated 50 sample from real data for following purposes:

2.4.1 By 90 percent of data, build two Linear Regression models and design hypothesis tests to see if these explanatory variables are a significant predictor of the response variable or not.

In this section we use split test train function to split our 50 data into 90% train and 10% test size. OLS regression results for age is in table 6:

Table 6: Fitted line with regression using least square method for age and weight

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	2.3804	0.595	4.000	0.000	1.180 , 3.580
age	0.0179	0.009	1.901	0.064	-0.001 , 0.037

The p-value for const is aprroximately zero and for age is 0.064(where is not significantly for 0.05 significant level) and the predictive equation is:

$$\text{weight} = 2.3804 + 0.0179\text{age} \tag{15}$$

The R-squared is 0.056 suggest that 5.6% of the variation of weight is accounted for by linear regression on age where the relationship between the two is weakly linear with a positive slope.
OLS regression results for lpsa is in table 7:

Table 7: Fitted line with regression using least square method for lpsa and weight

	coef	std err	t	P>|t|	[0.025 0.975]
const	3.1612	0.115	27.460	0.000	2.929 3.393
lpsa	0.1595	0.046	3.454	0.001	0.066 0.253

Since the p-value is of lpsa is 0.001 it is significant predictor(for 0.05 significant level) and predictive equation is:

$$\text{width} = 3.1612 + 0.1595\text{lpsa} \tag{16}$$

The R-squared is 0.217 suggest that 21.7% of the variation of weight is accounted for by linear regression on lpsa where the relationship between the two is not strongly linear with a positive slope. But compared to age it has more linear relationship with weight.
Result: lpsa has higher impact on weight since has higher R-squared but significant predictor is lpsa and not age since the p-value for age is higher than 0.05.

2.4.2 Calculate the 95% confidence interval for the slope of the relationship between response variable and explanatory variables. Interpret these CIs.

We have conduct a confidence interval for both slope of predicted equations where for age is :

$$CI = [-0.001, 0.03692720085038906] \tag{17}$$

The interpret of CI for slop corresponding to the age is we are 95% confindent that the real slop is in this range and since this CI include zero therefore weight might has not linear relationship with age.
confidence interval for slope of lpsa:

$$CI = [0.066, 0.2527] \tag{18}$$

The interpret of lpsa CI is we are 95% confindent that real slop is in this range and since this CI does not include zero therfore we can say that there is a linear relationship between lpsa and weight with positive slope.

2.4.3 Use your models to predict the values of the response variable for the remaining percent of samples.

We have predict remain data(0.1 percentage of sample data) where you can see the actual values for weight and predicted with different explanatory variables in table 8:

Actual values	Predict with age	Predict with lpsa
3.974998	3.598828	4.051898
3.764682	3.258379	3.607009
3.539509	3.222542	3.328320
4.524502	3.688420	3.614630
3.825375	3.616746	3.648827

Table 8: Actual values vs predicted values with age and lpsa

2.4.4 Compare the predicted values with actuals. Report the success rate.

In here we don't have labelize data where we can set an accuracy for our method, Instead we can use different metrics to see the success rate, generally R^2 and MSE are two common metrics used to evaluate the performance of regression models:

- R^2 score:The R^2 score is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. In simpler terms, it measures how well the regression predictions approximate the real data points. An R^2 score of 1 indicates that the regression predictions perfectly fit the data. Mathematically, R2R2 is defined as $1 - \frac{SSres}{SStot}$ where SSres is the sum of squares of residuals (the differences between the observed and predicted values) and SStot is the total sum of squares (the differences between the observed values and the mean of observed values). R^2 can sometimes be negative in models that do not intercept, indicating that the model is worse than simply predicting the mean of the dependent variable.

- MSE: Is a measure of the average squared difference between the actual and predicted values, giving insight into the error of a model. It's used to assess how close a fitted line is to the actual data points. The smaller the MSE, the closer the fit is to the data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

And we can see these metrics in table 9: where R^2 scores shows that the model has significant issues, possibly due to overfitting or underfitting or using features that do not have a predictive relation with the target value.But the MSE value provides a measure of the average error magnitude, but without more context such as range or distribution of dependent variable, it is hard to assess whether this values are high or low.Neverthelss, obtained the negeative R^2 score, improving the model is neccessary.

Table 9: Some metrics for success rates

metrics	R^2	MSE
age samples	-1.2721269110345732	0.24817508603508687
lpsa samples	-0.7109715689039531	0.18688239387252475