Statistics Exercise 2

Madhukara S Holla

13th Octoner 2023

Question 1

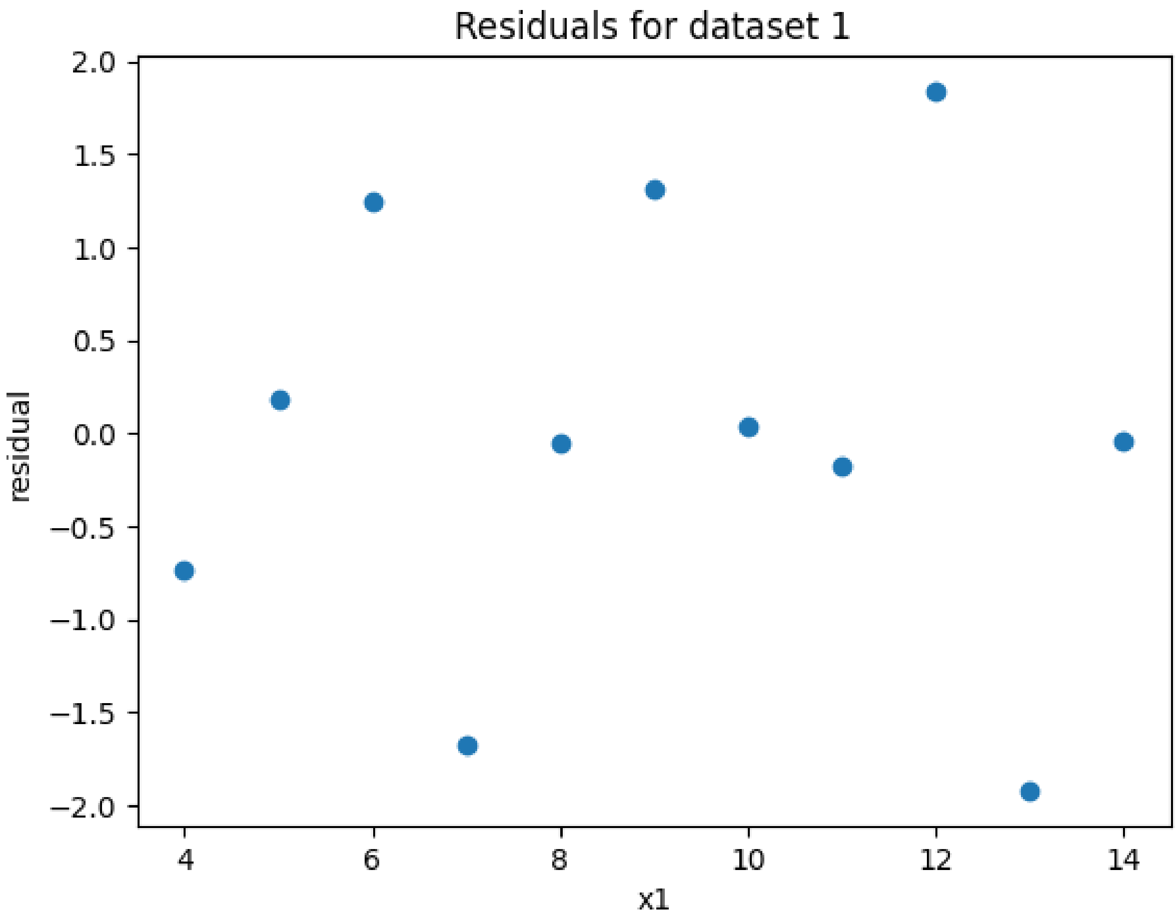
Project members: Madhukara S Holla (Master of Science in Computer Science, 1st Year)
Project description: NA

Question 2

2.a

Anscombe's Dataset 1

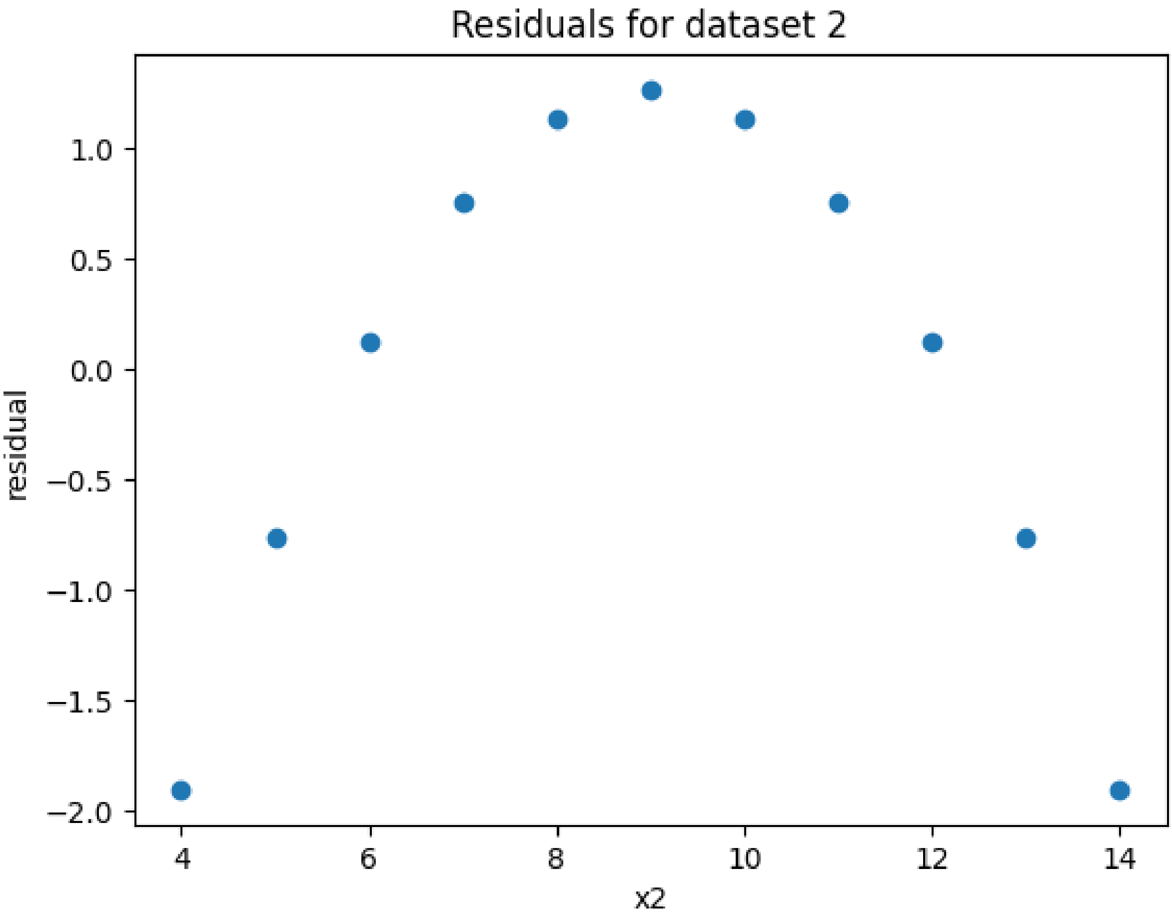
- Observed pearson co-efficient: 0.8164
- Observed co-efficient of multiple determination: 0.6665



2.a

Anscombe's Dataset 2

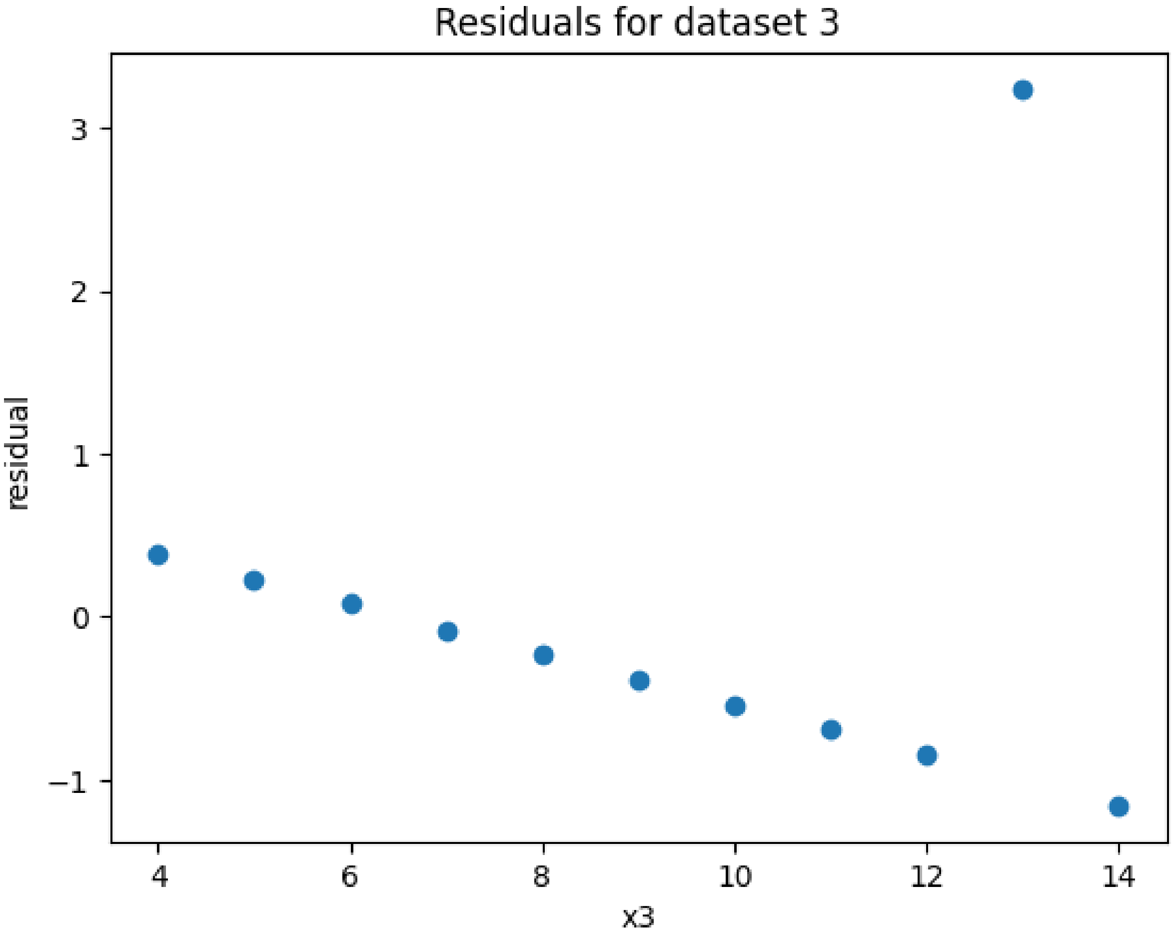
- Observed pearson co-efficient: 0.8162
- Observed co-efficient of multiple determination: 0.6662



2.a

Anscombe's Dataset 3

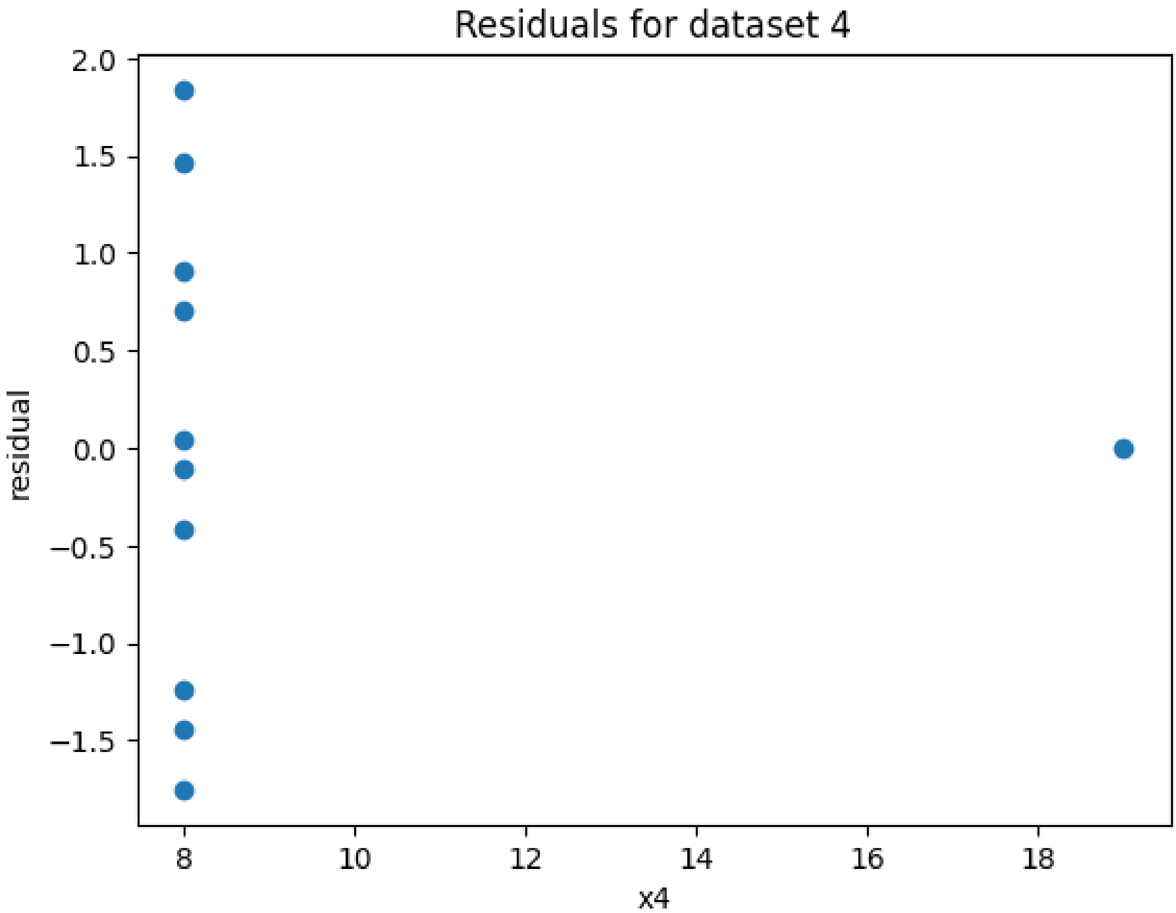
- Observed pearson co-efficient: 0.8163
- Observed co-efficient of multiple determination: 0.6663



2.a

Anscombe's Dataset 4

- Observed pearson co-efficient: 0.8165
- Observed co-efficient of multiple determination: 0.6667



2.b

Lack of fit tests for Anscombe's datasets.
Datasets grouped by x values (2 values per group)

Null Hypothesis H_0: A simple linear model is adequate to explain the systematic variations in the data.

Alternate Hypothesis H_a: A linear model is not adequate and a nonlinear model is required to capture the systematic variations in the data.

Anscombe's Dataset 1

P value: 0.86 - Fail to reject H_0.
No significant lack of fit.The dataset appears to be a simple linear relationship, and a linear regression model seems appropriate for this dataset.

Anscombe's Dataset 2

P value: 0.03 - Reject H_0 in favor of H_a.
Significant lack of fit. The data clearly follows a non-linear (quadratic) relationship, indicating that the linear model does not capture all systematic variations in the dataset.

Anscombe's Dataset 3

P value: 0.83 - Fail to reject H_0.
The outlier is ignored when we group the data by x values in pairs of 2.
No significant lack of fit. Since the influence of the outlier is ignored while calculating lack of fit, the dataset appears to be a simple linear relationship.

Anscombe's Dataset 4

P value: 0.04 - Reject H_0 in favor of H_a.
The outlier is ignored when we group the data by x values in pairs of 2.
Significant lack of fit. Test is not be appropriate due to the nature of the data. If forced, likely a significant lack of fit.

The lack of fit test assumes that there's some variation in the independent variable (x) that corresponds to variation in the dependent variable (y). In Dataset 4, for all but one observation, there's no variation in x. This goes against the fundamental premise of regression that we're trying to understand how y changes as x changes.

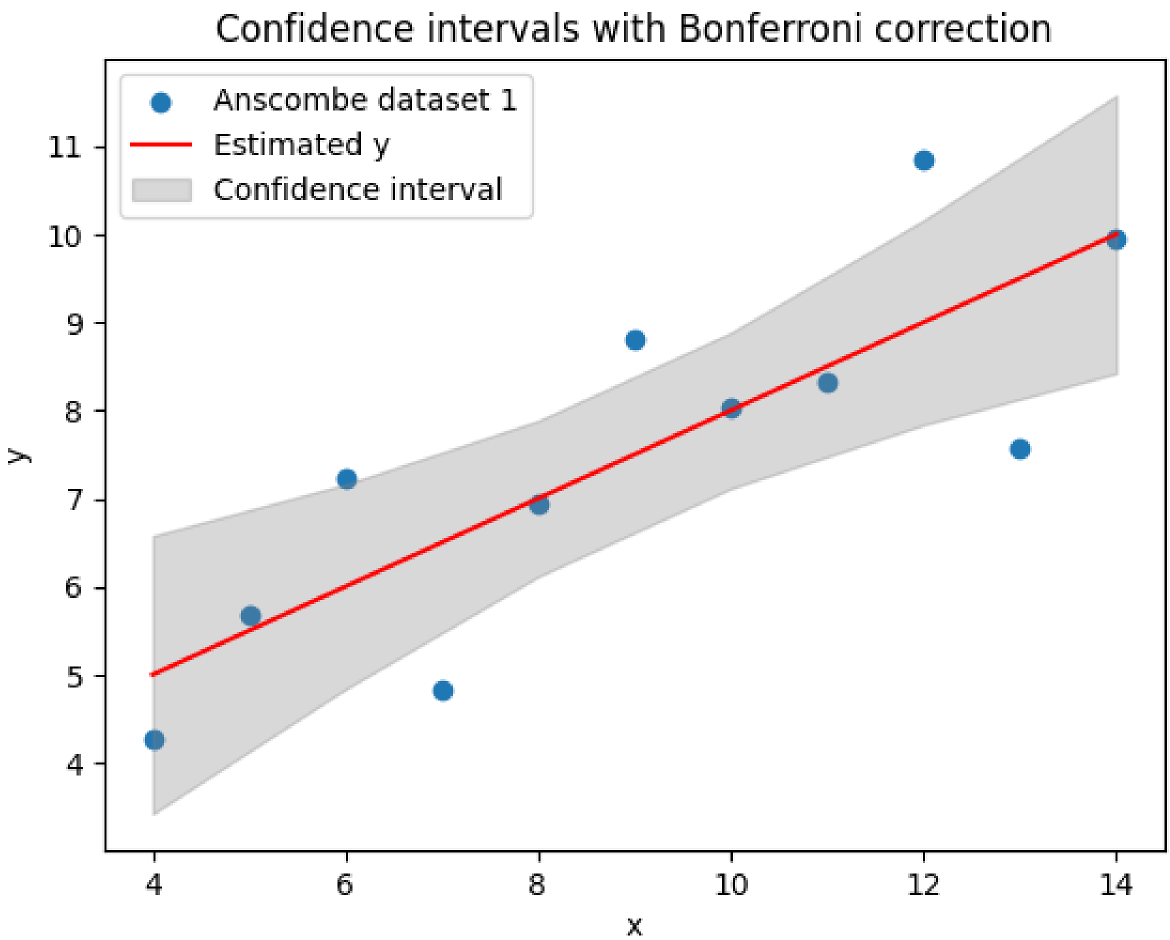
2.c

- Patterns in residual plots can indicate non-linearity and outliers, helping us to identify problems with the model.
- Lack of fit tests provide a formal statistical test to validate the model assumptions.
- But the lack of fit tests require replicate observations in the data which may not be available, or it may be inappropriate to run on datasets like Dataset 4.
- Both pearson coefficient and co-efficient of multiple determination do not clearly indicate the goodness of fit of the model. They just indicate the strength of the linear relationship and proportion of variance explained by the model respectively.

In conclusion, we need to use multiple methods such as visualization, lack of fit tests, and co-efficient determinations to validate the model assumptions.

Question 3

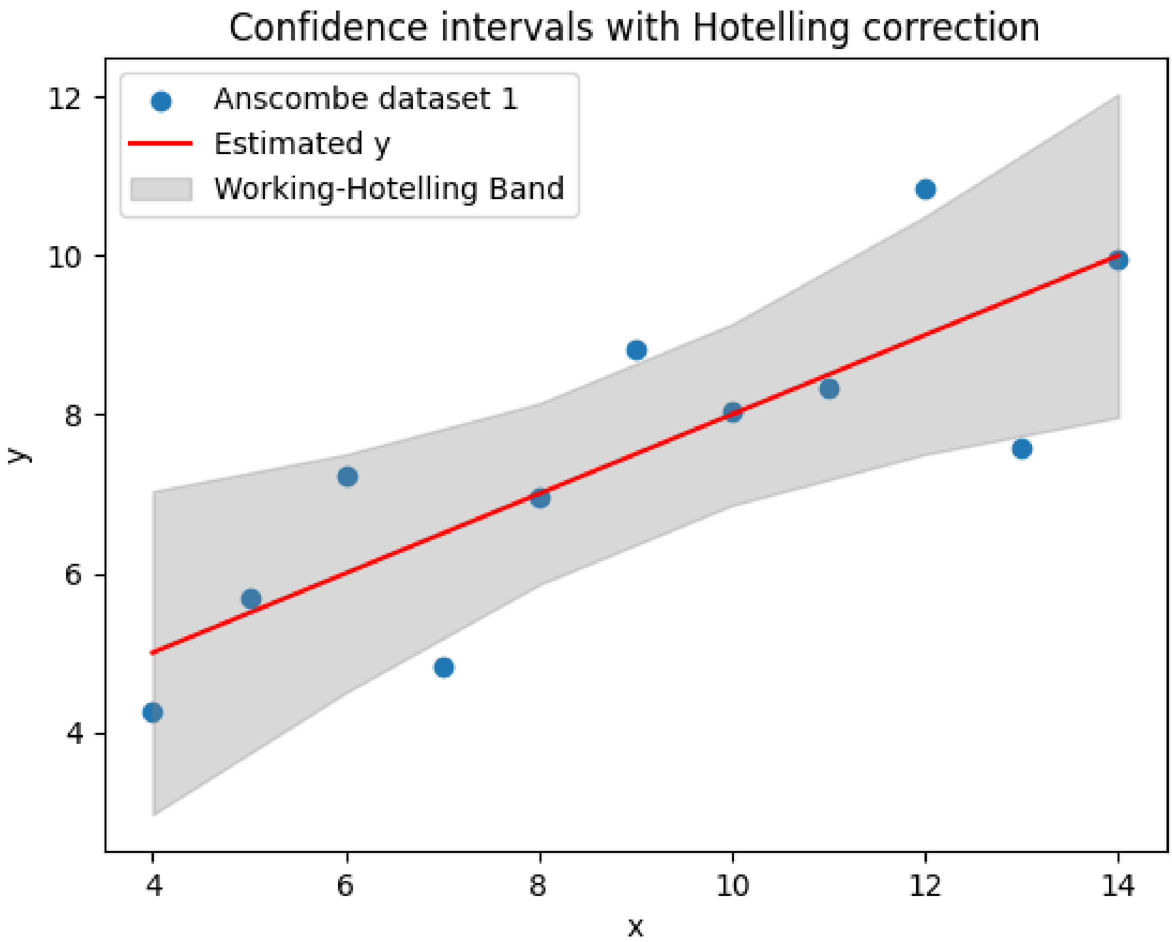
3.a



The Bonferroni adjustment controls the Familywise Error Rate (FWER). This is the probability of making at least one Type I error when performing multiple statistical tests.

By adjusting the significance level with the Bonferroni method, we ensure that our overall chance of incorrectly rejecting a true null hypothesis remains at the specified level (like 5%).

3.b



The Working-Hotelling confidence method is used to construct confidence intervals for the difference between the means of two multivariate data sets.
It controls for the precision of the estimate of the difference between the means. It provides a range of values within which we can be reasonably confident that the true difference between the means lies.

3.c

For a simple linear regression model using least squares estimation,

$$Y = \beta_0 + \beta_1.X + \epsilon$$

The estimates for β_0 and β_1 are given by:

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1.X$$

and ϵ is the error term.

$$cov(b_0, b_1) = E[(b_0 - E[b_0])(b_1 - E[b_1])]$$
$$cov(b_0, b_1) = E[b_0 b_1] - E[b_0].E[b_1]$$

Substituting for b_0,

$$E[b_0 b_1] = E[(\bar{Y} - b_1\bar{X})(b_1)]$$
$$= E[\bar{Y} b_1 - b_1^2 \bar{X}]$$
$$= \bar{Y} E[b_1] - \bar{X} E[b_1^2]$$

We know $E[b_1] = \beta_1$ and

$$E[b_0] = E[\bar{Y} - b_1\bar{X}]$$
$$= \bar{Y} - \bar{X}E[b_1] = \bar{Y} - \bar{X}.\beta_1$$

Now we have,

$$E[b_0 b_1] = \beta_1 \bar{Y} - \bar{X} E[b_1^2]$$
$$E[b_0].E[b_1] = (\bar{Y} - \bar{X}.\beta_1)\beta_1$$

Substituting in $cov(b_0, b_1)$ we have,

$$cov(b_0, b_1) = \beta_1 \bar{Y} - \bar{X} E[b_1^2] - (\bar{Y} - \bar{X}.\beta_1)\beta_1$$
$$= \beta_1 \bar{Y} - \bar{X} E[b_1^2] - \beta_1 \bar{Y} + \bar{X}.\beta_1^2$$
$$= \bar{X}.\beta_1^2 - \bar{X} E[b_1^2]$$

We need a formula for $E[b_1^2]$

$$Var(b1) = E[b_1^2] - (E[b_1])^2 = E[b_1^2] - \beta_1^2$$

From linear regression,

$$Var(b_1) = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$$

$$E[b_1^2] - \beta_1^2 = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$$

$$E[b_1^2] = \beta_1^2 + \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$$

Substituting into $cov(b_0, b_1)$ we have,

$$cov(b_0, b_1) = \bar{X}.\beta_1^2 - \bar{X}E[b_1^2]$$

$$= \bar{X}.\beta_1^2 - \bar{X}(\beta_1^2 + \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2})$$

$$= -\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}\bar{X}$$

$$cov(b_0, b_1) = -\bar{X}.Var(b_1)$$

Question 4

4.a

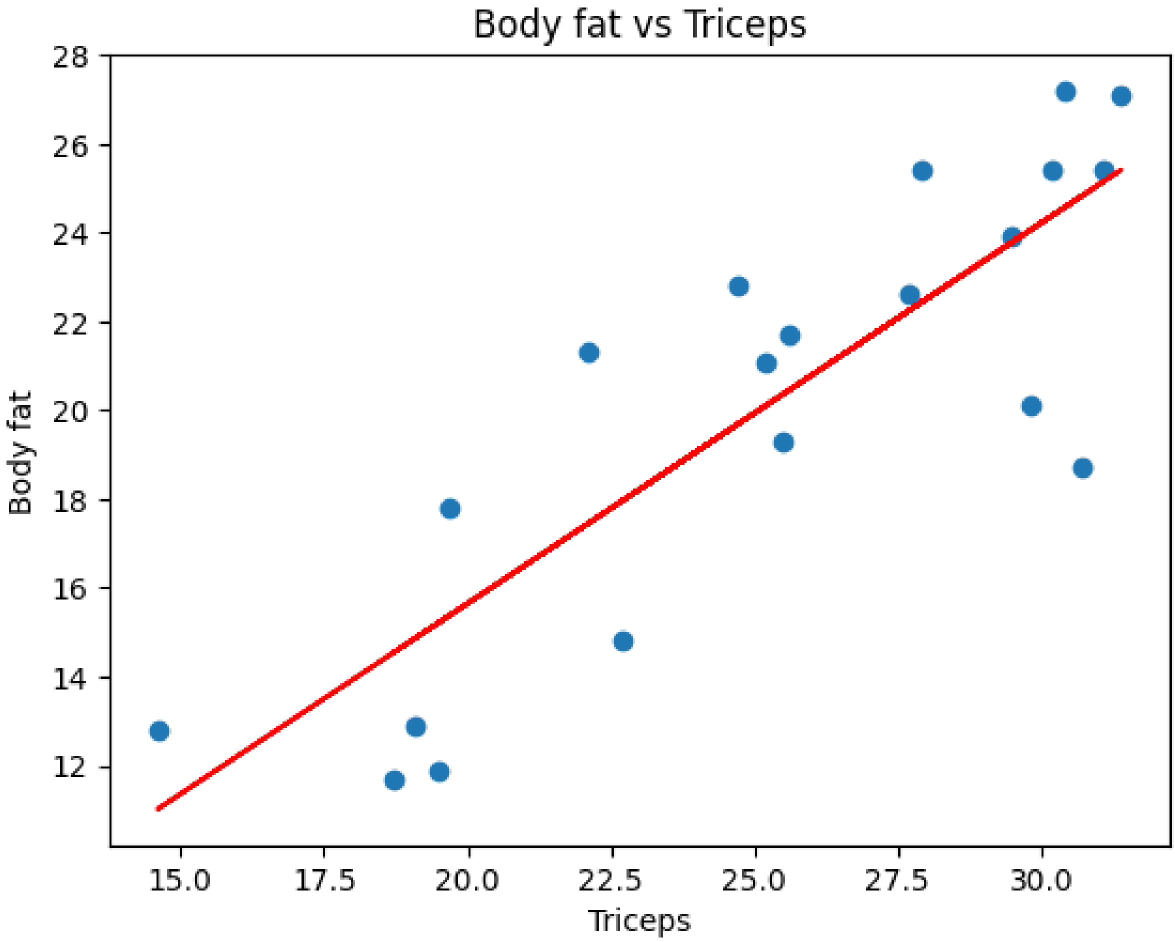
Covariance matrix for the data:

$$\begin{bmatrix} 25.23 & 24.29 & 8.39 & 21.63 \\ 24.29 & 27.4 & 1.62 & 23.47 \\ 8.39 & 1.62 & 13.3 & 2.65 \\ 21.63 & 23.47 & 2.65 & 26.07 \end{bmatrix}$$

Row 1: Covariance of Triceps with Triceps, Thigh, Midarm, Bodyfat
Row 2: Covariance of Thigh with Triceps, Thigh, Midarm, Bodyfat
Row 3: Covariance of Midarm with Triceps, Thigh, Midarm, Bodyfat
Row 4: Covariance of Bodyfat with Triceps, Thigh, Midarm, Bodyfat

4.b

Triceps as a linear predictor of Bodyfat

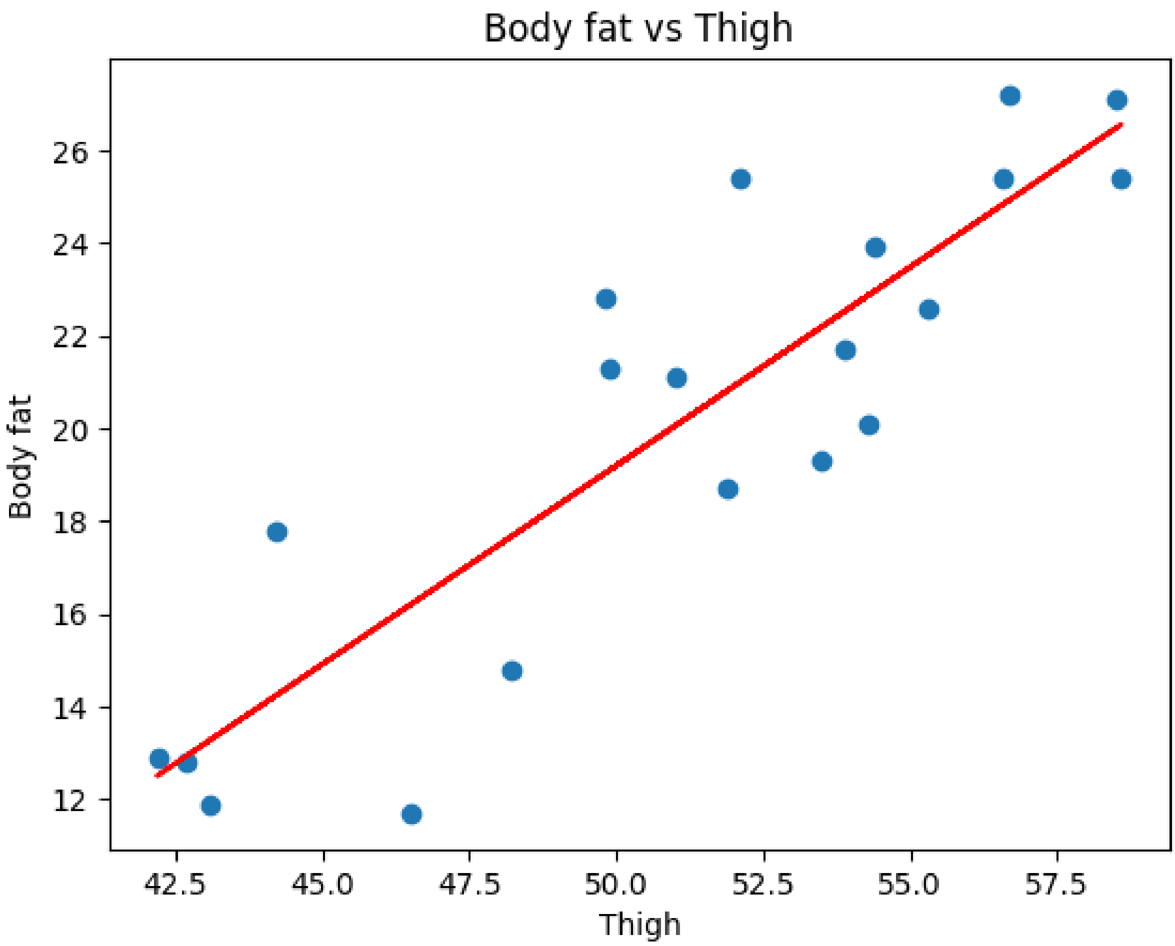


Summary of the data:

- Slope: 0.8572
- Intercept: -1.4961
- Mean squared error: 7.9511
- R^2: 0.7111
- The positive value of the slope signifies a positive linear association between Triceps and Body fat.
- An MSE of 7.9511 indicates there is a significant amount of error in the estimation of Bodyfat from Triceps.
- R^2 value of 0.71 indicates that $\approx 70\%$ of the variability in Bodyfat can be explained by Triceps.

From the graph and summary, Triceps show a strong linear association with Bodyfat.

Thigh as a linear predictor of Bodyfat

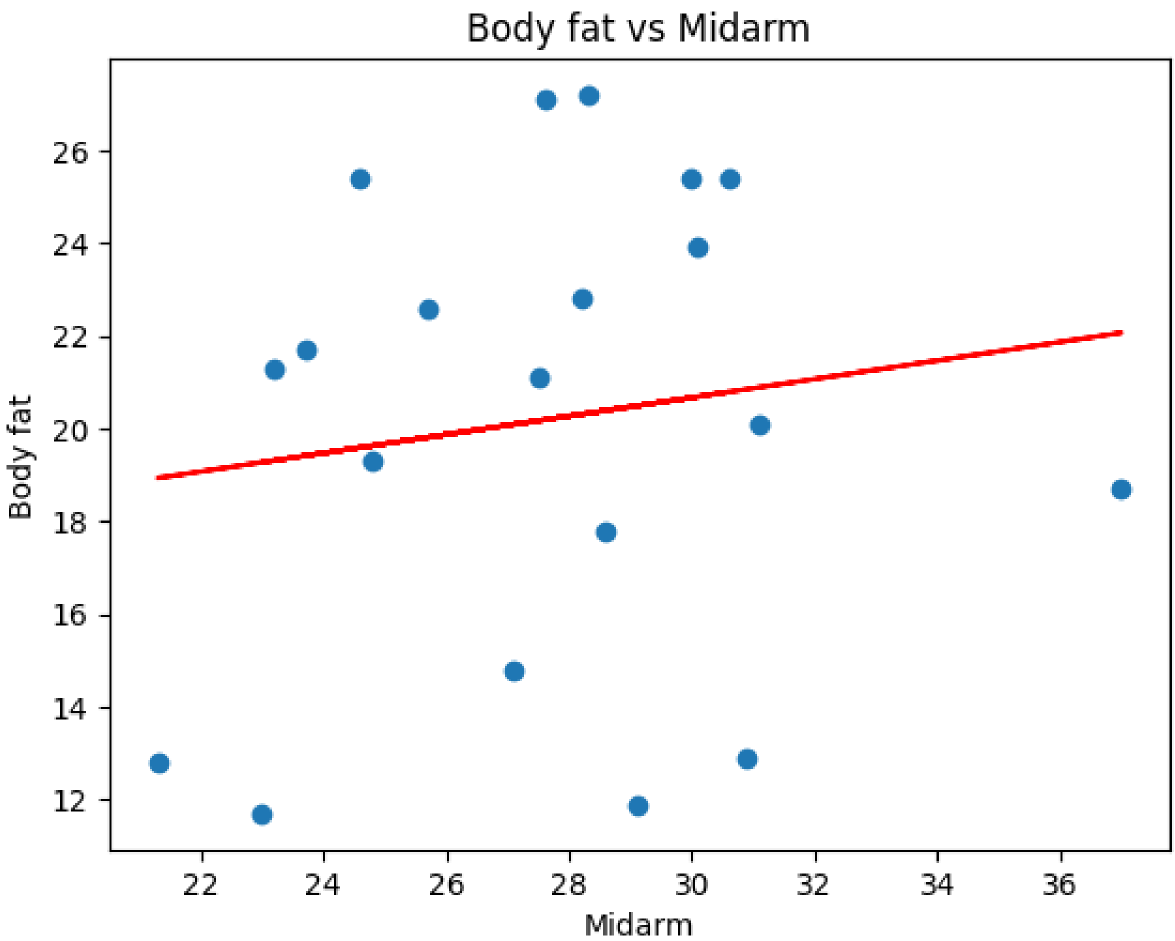


Summary of the data:

- Slope: 0.8565
- Intercept: -23.6345
- Mean squared error: 6.3013
- R^2: 0.7710
- The positive value of the slope signifies a positive linear association between Thigh and Body fat.
- An MSE of 6.3013 indicates there is still error in the estimation of body fat, but the error is less than that of Triceps.
- R^2 value of 0.77 indicates that $\approx 77\%$ of the variability in Bodyfat can be explained by Thigh. This association is even stronger than that of Triceps.

From the graph and summary, Thigh show a strong linear association with Bodyfat.F From a lower MSE and higher R^2 value, we can conclude that Thigh is a better predictor of Bodyfat than Triceps.

Midarm as a linear predictor of Bodyfat



Summary of the data:

- Slope: 0.8565
- Intercept: -23.6345
- Mean squared error: 6.3013
- R^2 value: 0.0203
- The positive value of the slope signifies a positive linear relationship with the Bodyfat. But from the graph we can see that the model has several outliers and does not fit the data well.
- MSE term indicates a similar variance to that of Triceps and Thigh, but given the low R^2 value, this can be misleading.
- R^2 value of 0.02 is considerably low and indicates that Midarm is not a good predictor of Bodyfat.

Despite the slope and MSE values for midarm being similar to that of Triceps and Thigh, the graph and R^2 value indicate that Midarm does not have a strong linear association with Bodyfat.

4.c

Why we may want to fit a multi-variate model with $Tricep$ and $Tricep^2$:

- To capture non-linearity: By including term $Tricep^2$, we are exploring the possibility of a non-linear relationship between Tricep and Bodyfat.
- Model flexibility: Adding polynomial terms offers more flexibility in capturing patterns in the data. This can lead to better predictions if underlying relationship is non-linear.

Summary of multivariate linear regression model with $Tricep$ and $Tricep^2$:

- Intercept: -6.1523
- Coefficient for $Tricep$: 1.2612
- Coefficient for $Tricep^2$: -0.0083
- Mean squared error: 8.3777
- R^2 value: 0.7125

From the summary, we can see that the co-efficient for $Tricep^2$ is ≈ 0. This means that the relationship between Tricep and Bodyfat is linear.
Due to this, the value of R^2 is similar to that of the simple linear model with Tricep as a predictor. This indicates that the multivariate model does not offer any significant improvement over the simple linear model.

4.d

Why we may want to fit a multi-variate model with $Tricep$ and $Thigh$:

- To capture combined influence: By including both $Tricep$ and $Thigh$ as predictors, we are exploring the possibility of a joint influence of both variables on Bodyfat.
- Increase predictive power: A multivariate model may capture more variance in the data and lead to better predictions.
- Reduce omitted variable bias: If $Tricep$ and $Thigh$ are correlated, then omitting one of them from the model can lead to omitted variable bias.

Summary of multivariate linear regression model with $Tricep$ and $Thigh$:

- Intercept: -19.1742
- Coefficient for $Tricep$: 0.2223
- Coefficient for $Thigh$: 0.6594
- Mean squared error: 6.4676
- R^2 value: 0.7780
- The coefficient for Tricep drops significantly from the simple linear model. This indicates potential multicollinearity, suggesting that the individual effects of Triceps and Thigh on Body fat are not as strong when considered together.
- Value of R^2 is marginally higher than that of the simple linear model with Thigh as a predictor. This indicates that the additional predictive power of adding Tricep to a model that already contains Thigh is minimal.

4.e

Summary of multivariate linear regression model with $Tricep$, $Thigh$ and $Midarm$:

- Intercept: 117.0847
- Coefficient for $Tricep$: 4.3341
- Coefficient for $Thigh$: -2.8568
- Coefficient for $Midarm$: -2.1861
- Mean squared error: 6.1503
- R^2 value: 0.8013
- Value of R^2 is higher than that of previous models, indicating that adding Midarm provides additional predictive power.
- The change in R^2 is only marginal compared to the model with only $Tricep$ and $Thigh$.
- The coefficients have changed notably from the bivariate regression. The negative coefficient for thigh and midarm, indicates that as these measurements increase, the predicted body fat decreases, holding other variables constant.

In conclusion, adding all the predictors has improved the predictive power of the model, but at the cost of complexity. The substantial change in coefficients when adding another predictor suggests possible multicollinearity.

4.f

If predictors are orthogonal in multivariate regression:

- Coefficients will be stable: Adding or removing a predictor does not change the coefficients of other predictors.
- Clear interpretability: coefficients will directly indicate the effect of each predictor on the response variable (unaffected by other predictors).
- In our specific case,the coefficient for $Tricep$ in three predictor model would remain close to the value in the bivariate model. (similarly for thigh).
- Coefficients would not flip signs or show significant changes in magnitude.

In conclusion, if the predictors were orthogonal: the results would be stable and easier to interpret.