Statistics Exercise 2

Madhukara S Holla

13th Octoner 2023

Question 1

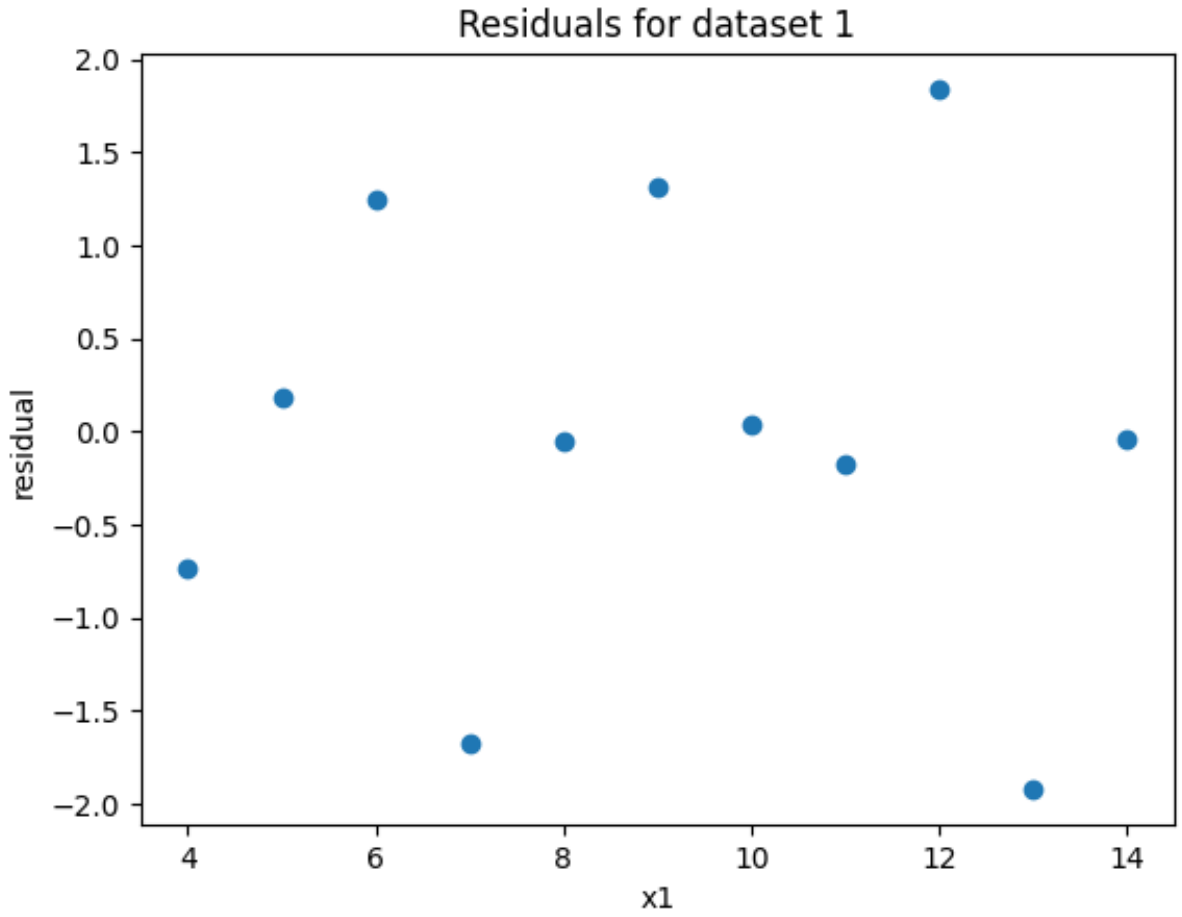
Project members: Madhukara S Holla (Master of Science in Computer Science, 1st Year)
Project description: Not decided yet.

Question 2

2.a

Anscombe's Dataset 1

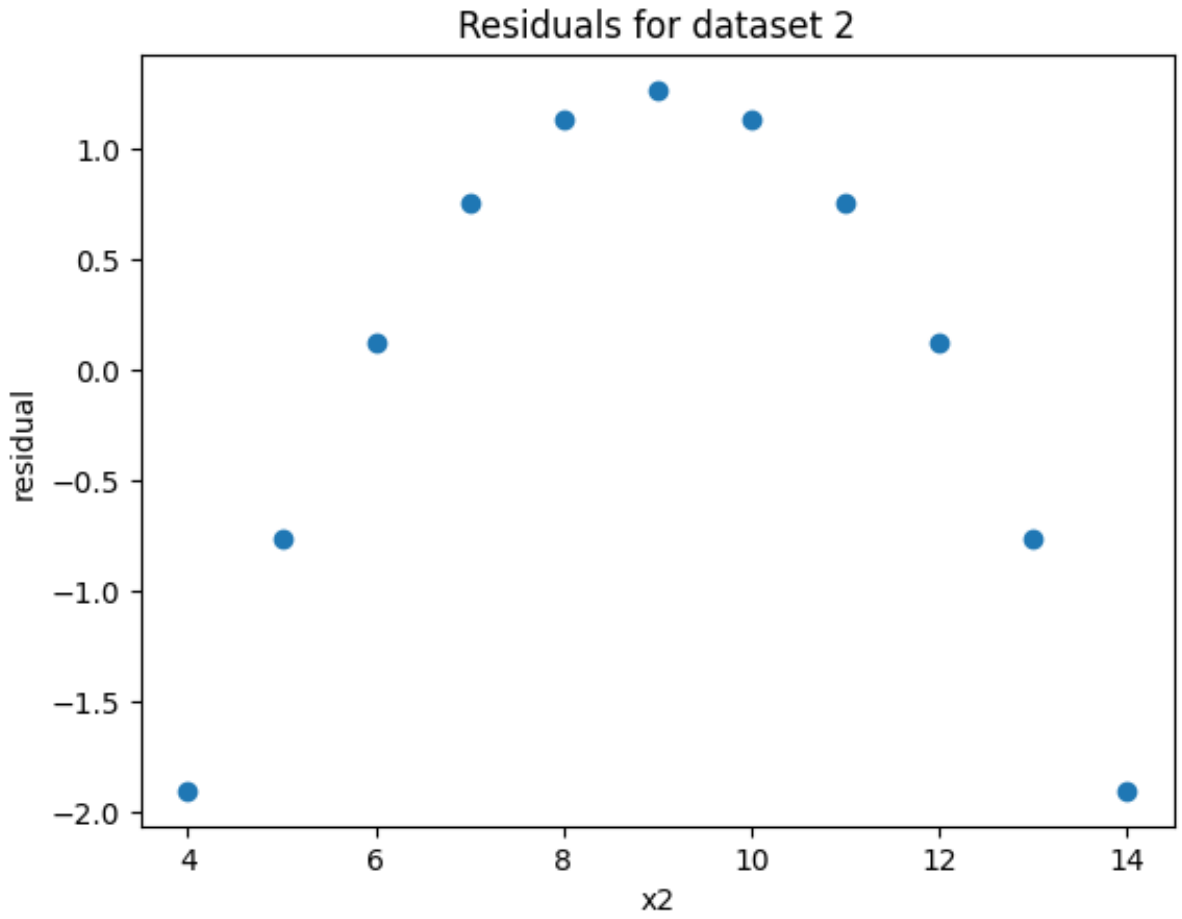
- Observed pearson co-efficient: 0.8164
- Observed co-efficient of multiple determination: 0.6665



2.a

Anscombe's Dataset 2

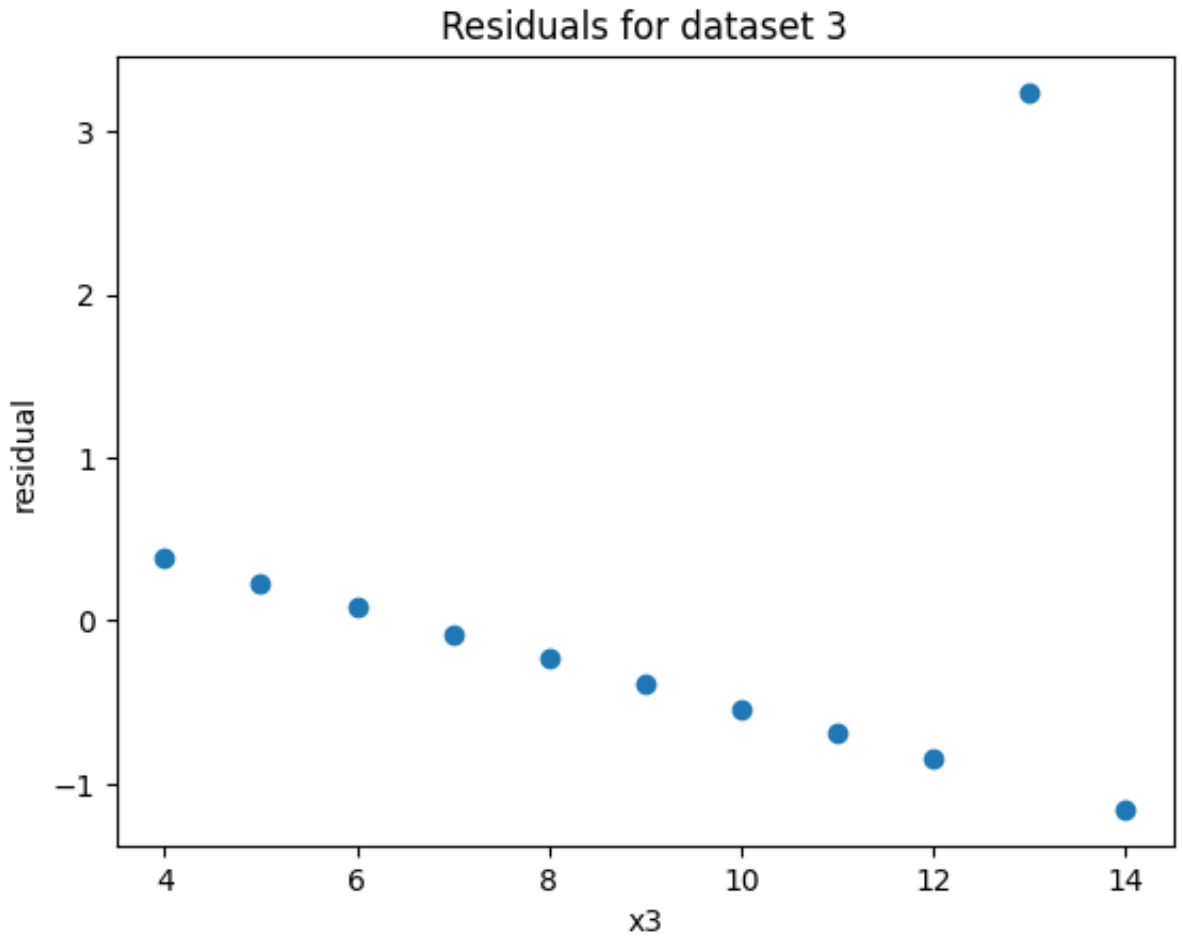
- Observed pearson co-efficient: 0.8162
- Observed co-efficient of multiple determination: 0.6662



2.a

Anscombe's Dataset 3

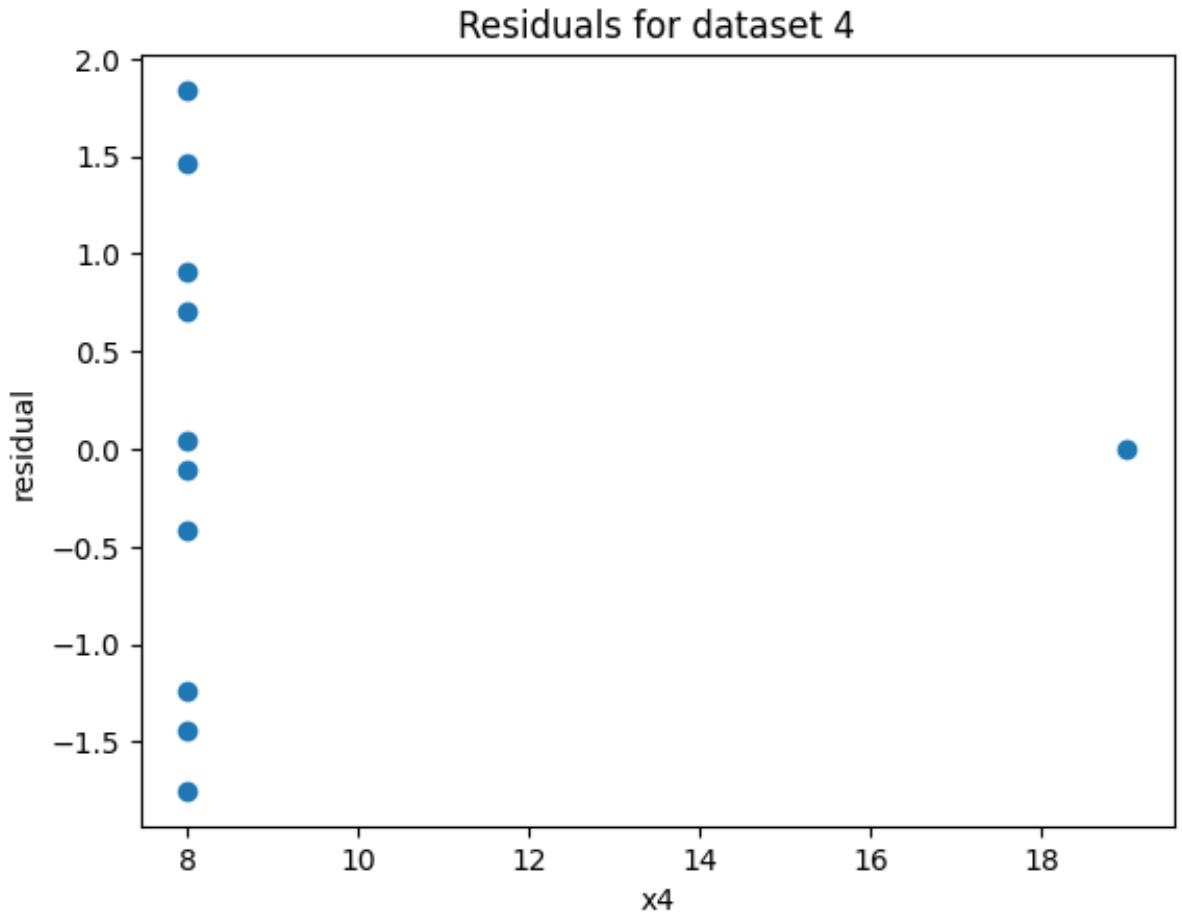
- Observed pearson co-efficient: 0.8163
- Observed co-efficient of multiple determination: 0.6663



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.

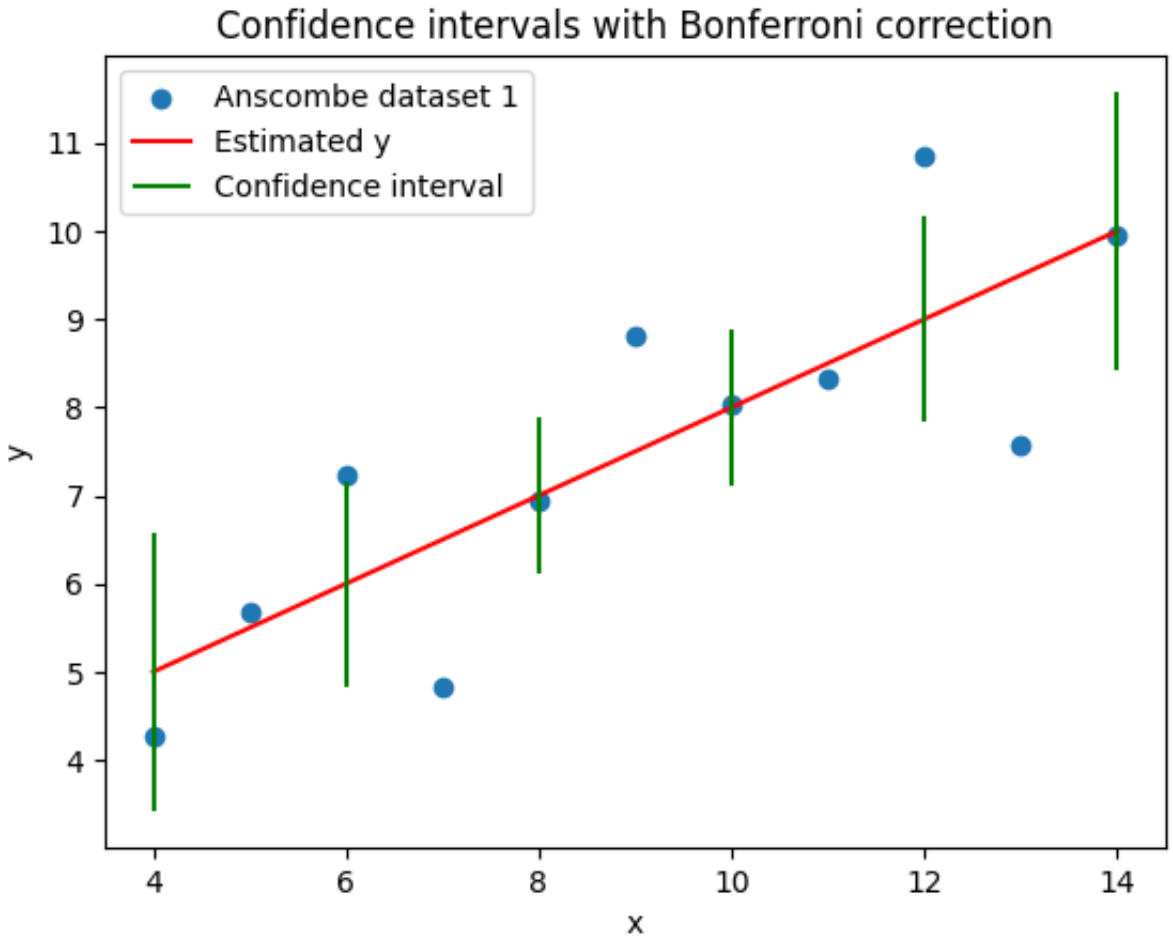
2.c

- 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.

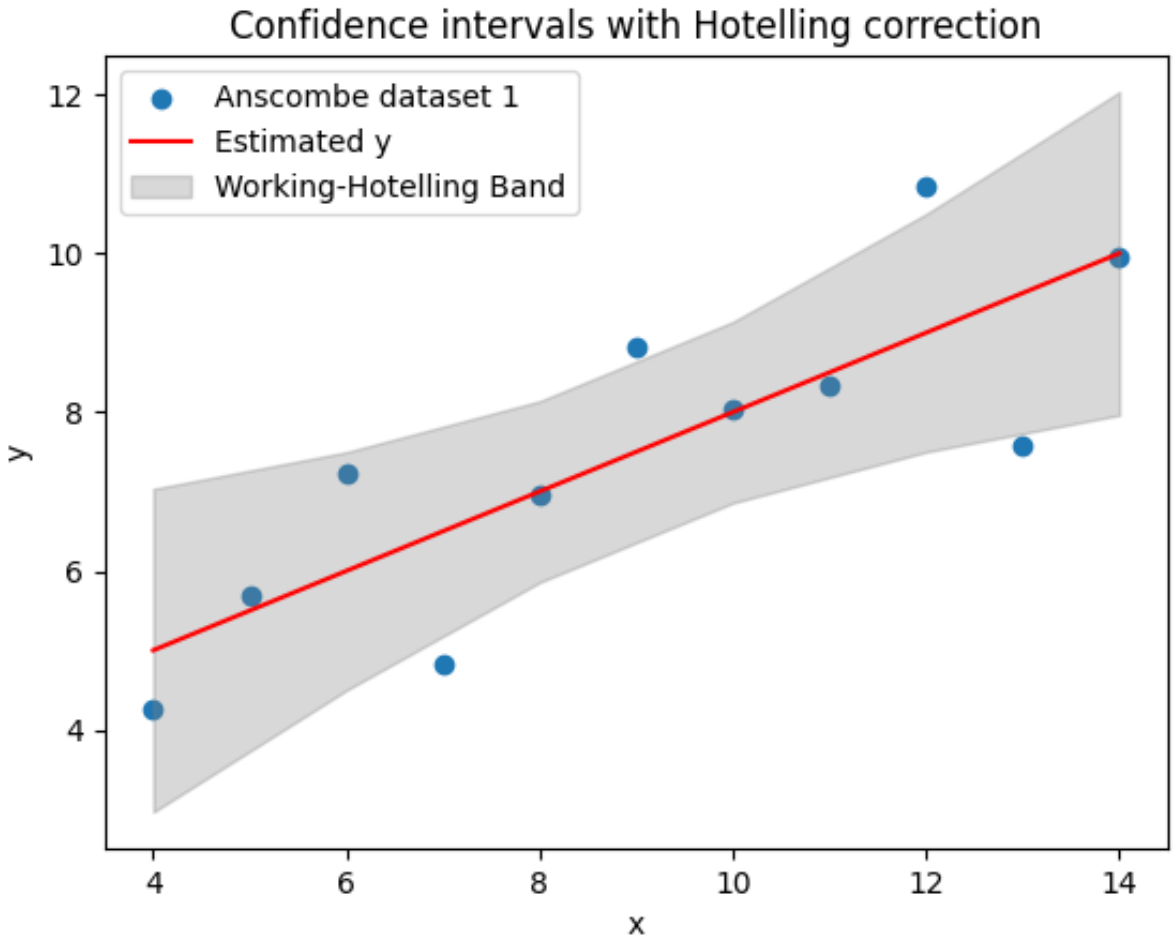
Question 3

3.a



The Bonferroni adjustment controls the Familywise Error Rate (FWER). To control FWER when making multiple comparisons, Bonferroni correction adjusts the significance level of each individual test.
So we are controlling the error rate in individual tests, to control the error rate in the family of tests.

3.b



In this method we are controlling the error rate of the entire regression curve.
The Working-Hotelling confidence method is used to construct confidence intervals for the difference between the means of two multivariate data sets. It provides a range of values within which we can be reasonably confident that the true difference between the means lies.

The efficiency of the procedure depends on the purpose: If we are making inferences at specific values of X1, then individual confidence intervals (Bonferroni) might be narrower and therefore more "efficient" in the sense of precision.

If we want to make inferences about the entire regression line, then the Working-Hotelling band is more efficient because it provides simultaneous coverage for the entire line. Individual confidence intervals wouldn't be appropriate for this broader type of inference.

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,

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

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

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

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

Importance for multiple testing:

- If we are simultaneously estimating multiple parameters, understanding covariance between the parameters is important.
- If the estimates are highly correlated or anti-correlated, then testing the hypotheses without accounting for this correlation can lead to incorrect inferences.

If the mean of predictor $\bar{X} = 0$, then $Cov(b_0, b_1) = 0$. This means that the estimates for β_0 and β_1 are uncorrelated.
This simplifies the analysis and interpretations as we can interpret or test each coefficient without considering the other.

Question 4

4.a

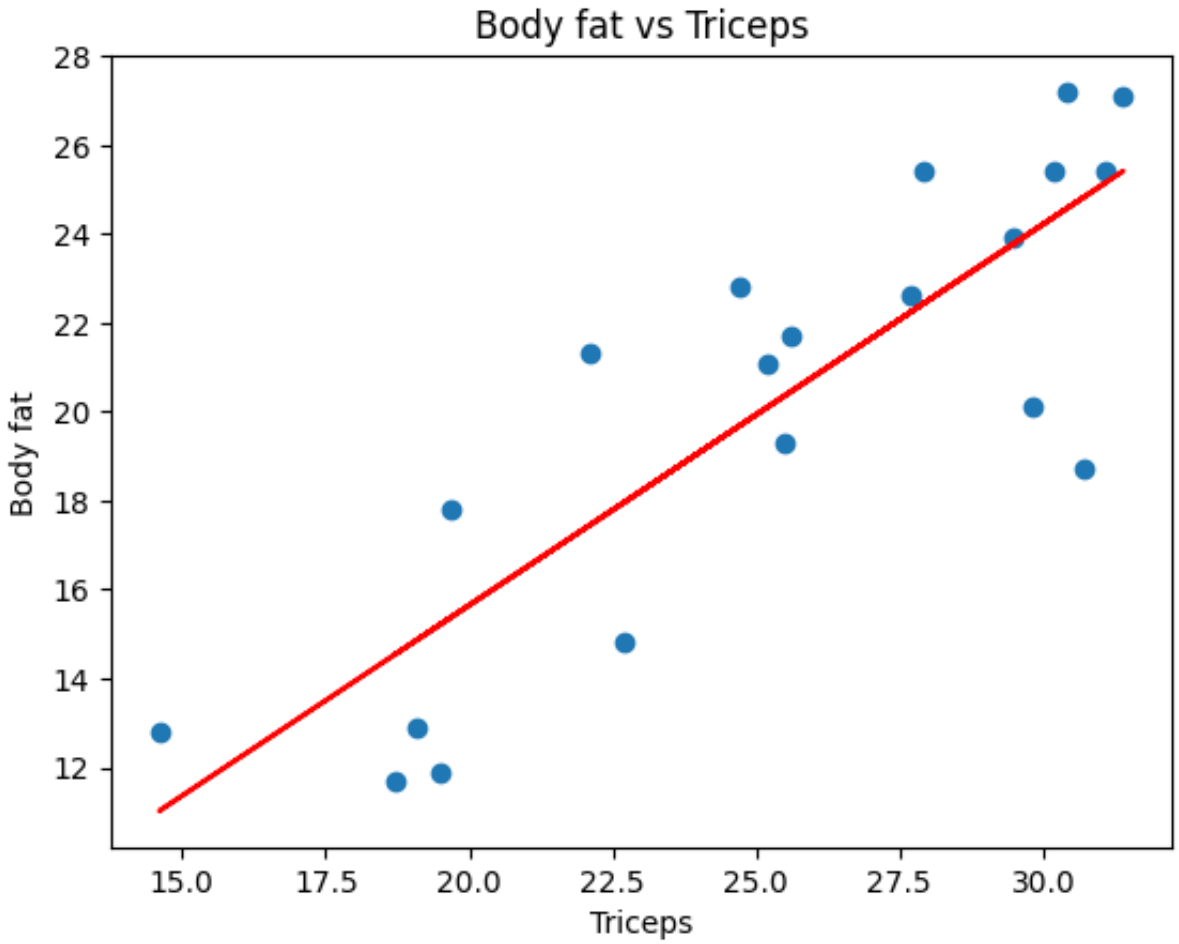
Covariance matrix for the data:

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

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

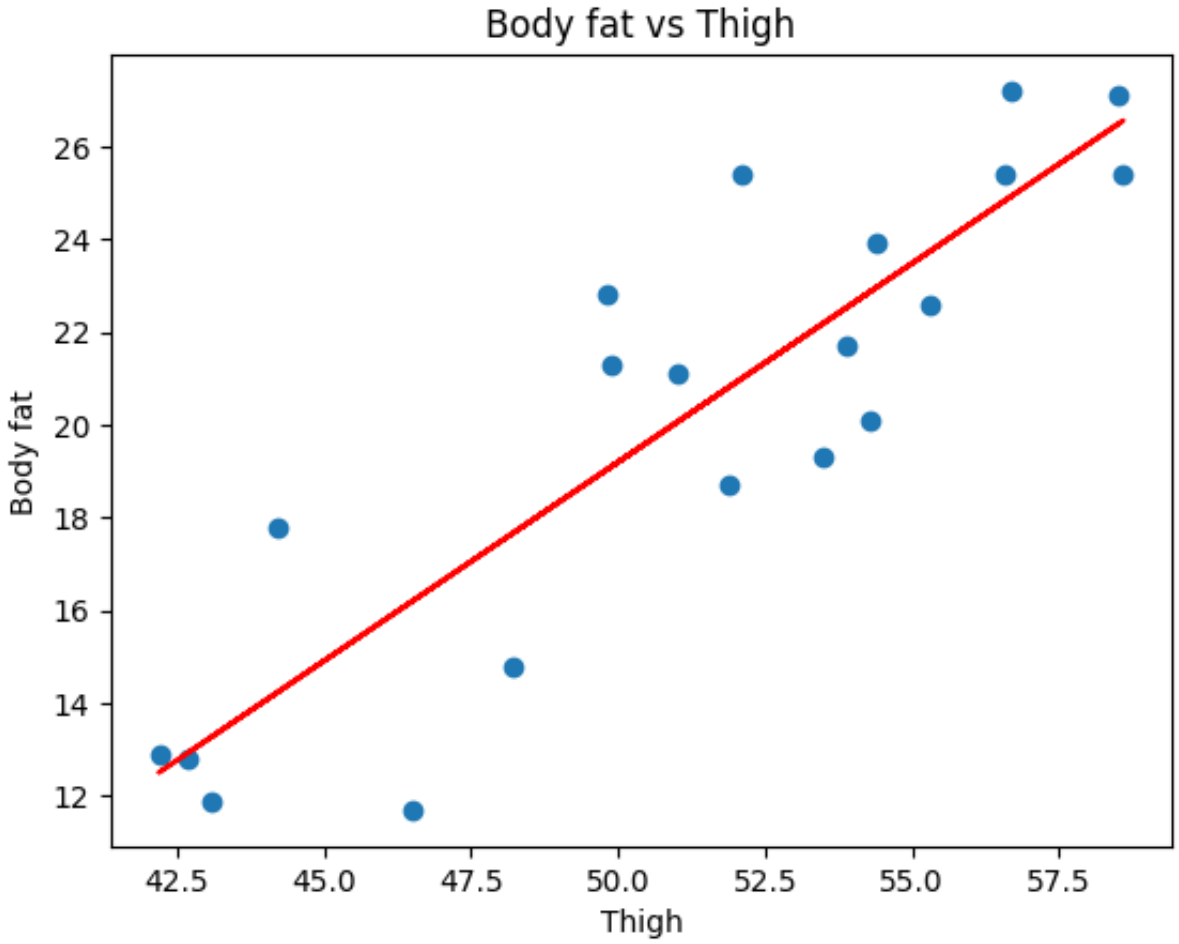


Summary of the data:

- 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

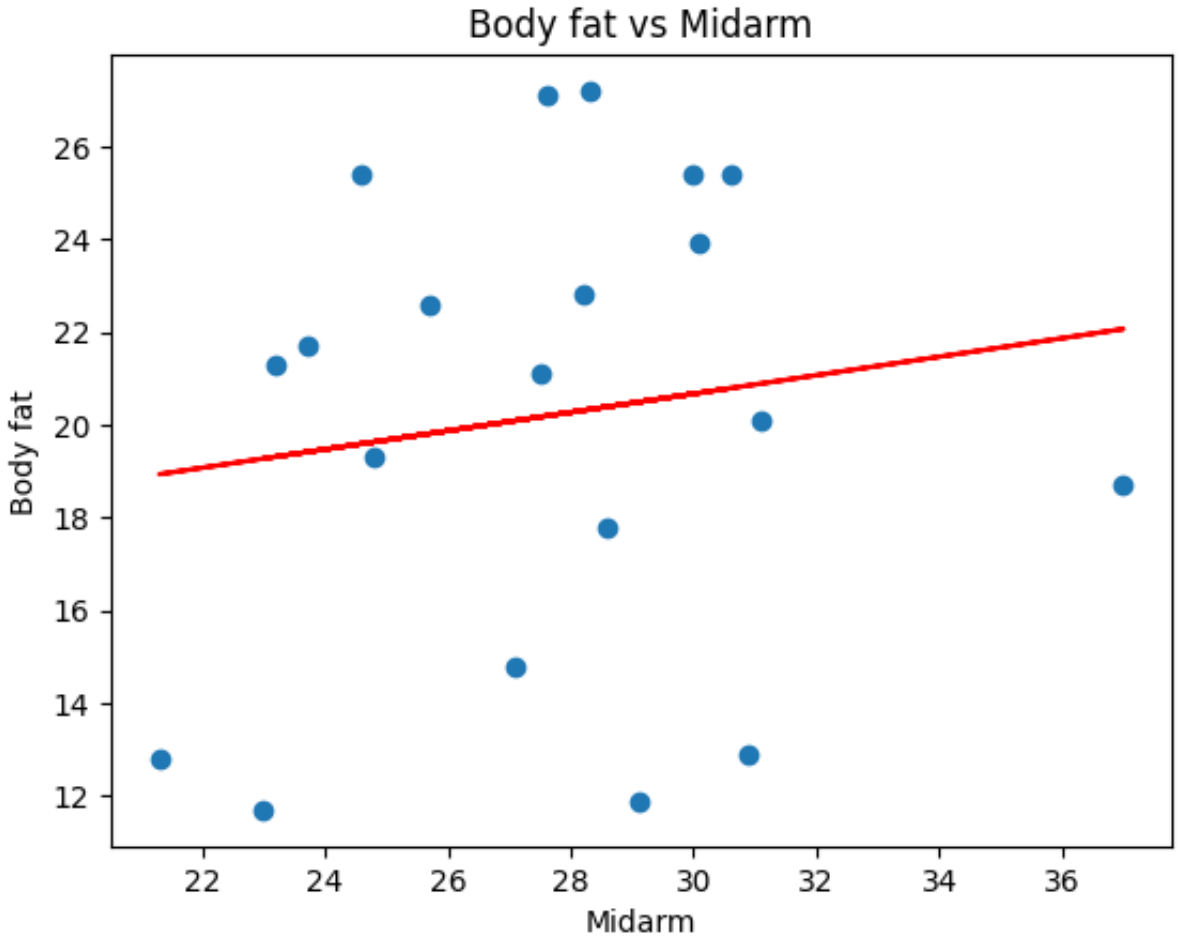


Summary of the data:

- 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: 0.1994
- Intercept: 14.6867
- Mean squared error: 26.9632
- R^2 value: 0.0203
- Low value of slope indicates a weak linear association between Midarm and Bodyfat.
- A high MSE term indicates that there is a significant amount of error in the estimation of Bodyfat from Midarm.
- R^2 value of 0.02 is considerably low and indicates that Midarm is not a good predictor of Bodyfat.

From the graph and summary, Midarm shows a weak linear association with Bodyfat. From a high MSE and low R^2 value, we can conclude that Midarm is not a good predictor of 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 marginally higher than that of higher than that of previous models, indicating that midarm is not a significant predictor of body fat.
- 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.

Question 5

5.a

Including the number of older siblings as a quantitative predictor

We can propose a linear regression model expressed as:

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \epsilon$$

where,

- Y is the maturation of the child
- X_1 represents the age of the child
- X_2 represents the number of older siblings
- β_0 is the y-intercept
- β_1 is the regression coefficient for Age.
- β_2 is the regression coefficient for number of older siblings.
- ϵ is the error term.

Assumptions regarding the effect of number of older siblings on the response variable:

- Linearity: The relationship between the number of older siblings and maturation is linear. For every additional older sibling, the change in maturation is constant β_2
- No multicollinearity: Age and number of siblings are not be highly correlated.
- Additivity: The combined effect of Age and number of siblings is the sum of their individual effects.
- Independence: The observations are independent of each other. The maturation of one child does not affect the maturation of another.
- Variance of residuals is constant across all values of the predictor.
- Normality: The residuals are normally distributed.

5.b

Including the number of older siblings as two indicator variables

By treating number of older siblings as indicator variables, we are essentially breaking down number of siblings into categories and representing each with a separate variable. Since number of older siblings can range from 0 to 2, we can create two indicator variables and represent the regression model as:

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.I_1 + \beta_3.I_3 + \epsilon$$

where,

- I_1 is an indicator variable: 1 if number of older siblings is 1, 0 otherwise.
- I_2 is an indicator variable: 1 if number of older siblings is 2, 0 otherwise.
- β_2 is the regression coefficient for I_1.
- β_3 is the regression coefficient for I_2.
- $Y, X_1, \beta_0, \beta_1, \epsilon$ are the same as in the previous case.

By considering number of older siblings as indicator variables, we are effectively creating multiple models within a single equation.

- Child with 0 siblings: $Y = \beta_0 + \beta_1.X_1 + \epsilon$
- Child with 1 sibling: $Y = \beta_0 + \beta_2 + \beta_1.X_1 + \epsilon$
- Child with 2 siblings: $Y = \beta_0 + \beta_3 + \beta_1.X_1 + \epsilon$

Each of these models represent different scenarios of number of older siblings.

By using indicator variables, we can model different intercepts for each level of the categorical variable (number of siblings) while keeping the same slope for age across all the levels.

Assumptions regarding the effect of number of older siblings on the response variable:

- Linearity: The relationship between the presence or absence of a specific number of older siblings (as denoted by indicator variables) and maturation is linear.
- Additivity: The combined effect of Age, having 1 or 2 older siblings is the sum of their individual effects.
- No multicollinearity: I_1 and I_2 are not highly correlated. In this case, they are mutually exclusive.
- In addition, Independence, constant variance and normality assumptions are the same as in the previous case.

5.c

Interactions along with indicator variables

Without interaction term, we are assuming that the effect of age and number of siblings is independent of each other. This would mean the number of siblings has the same effect on maturity irrespective of the age of the child.
This can be solved by adding an interaction term between age and number of siblings.

We can represent the regression model as:

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.I_1 + \beta_3.I_2 + \beta_4.(X_1.I_1) + \beta_5(X_1.I_2) + \epsilon$$

where,

- $X_1.I_1$ is the interaction between age and indicator for 1 older sibling
- $X_1.I_2$ is the interaction between age and indicator for 2 older siblings
- β_4 is the regression coefficient for $X_1.I_1$
- β_5 is the regression coefficient for $X_1.I_2$
- $Y, X_1, \beta_0, \beta_1, \beta_2, \beta_3, I_1, I_2, \epsilon$ are the same as in the previous case.

With this model, we can determine whether relationship between age and maturation changes depending on the number of older siblings. If the interaction terms are significant, it suggests that there is a differential effect.

Assumptions regarding the effect of number of older siblings on the response variable:

- Linearity: The relationship between the predictors (age, indicator variables, interaction terms) and maturation is linear.
- Additivity: This assumption is relaxed because of the inclusion of interaction terms. The model implies that the effect of one predictor may not be strictly additive with the effect of another predictor.
- No multicollinearity: The interaction term by nature can add some degree of multicollinearity. But individual predictors are not highly correlated.
- In addition, Independence, constant variance and normality assumptions are the same as in the previous case.