JAMES COOK UNIVERSITY AUSTRALIA

19-B-MA5821-ONL-EXT-SP85

Advanced Statistical Methods for Data Scientists

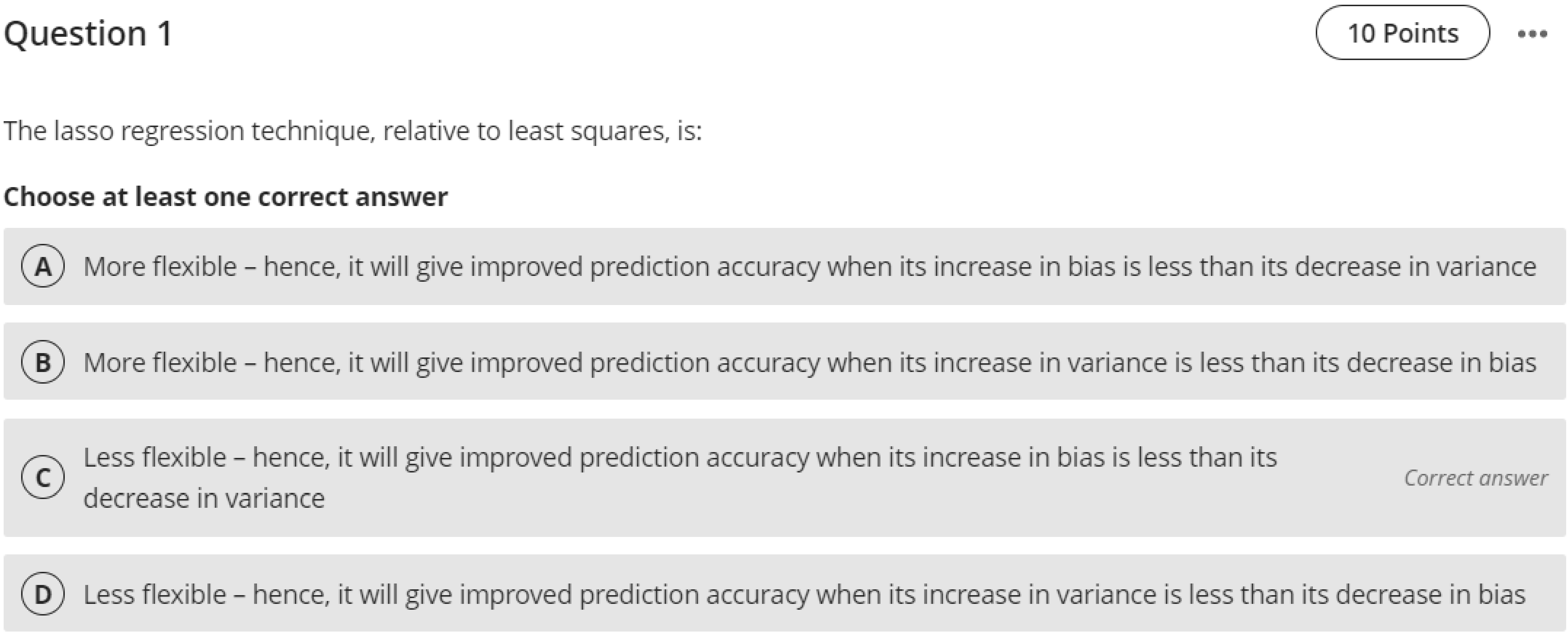
Week-4

Presented by
Zilani (JCU)
mohammed.zilani@jcu.edu.au
mgzilani15@gmail.com

Main Focus:

- Quiz from week3
- Logistic Model

Week3 quiz:



Question 1

10 Points •••

The lasso regression technique, relative to least squares, is:

Choose at least one correct answer

Ⓐ More flexible – hence, it will give improved prediction accuracy when its increase in bias is less than its decrease in variance

Ⓑ More flexible – hence, it will give improved prediction accuracy when its increase in variance is less than its decrease in bias

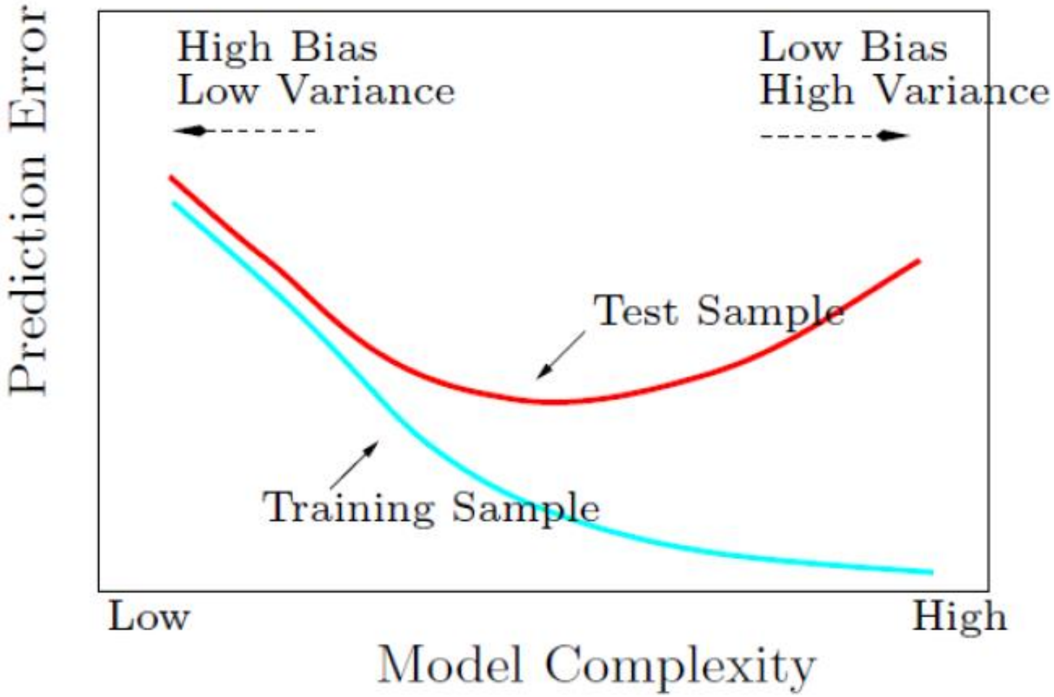
Ⓒ Less flexible – hence, it will give improved prediction accuracy when its increase in bias is less than its decrease in variance *Correct answer*

Ⓓ Less flexible – hence, it will give improved prediction accuracy when its increase in variance is less than its decrease in bias

Week3 quiz:

- Explanation of Q1 answer

  Bias Variance trade off.



Things to remember:

- Increasing model complexity brings greater flexibility and, therefore, lower bias. However, this comes at a cost of higher variance. Overfitting can be a problem.

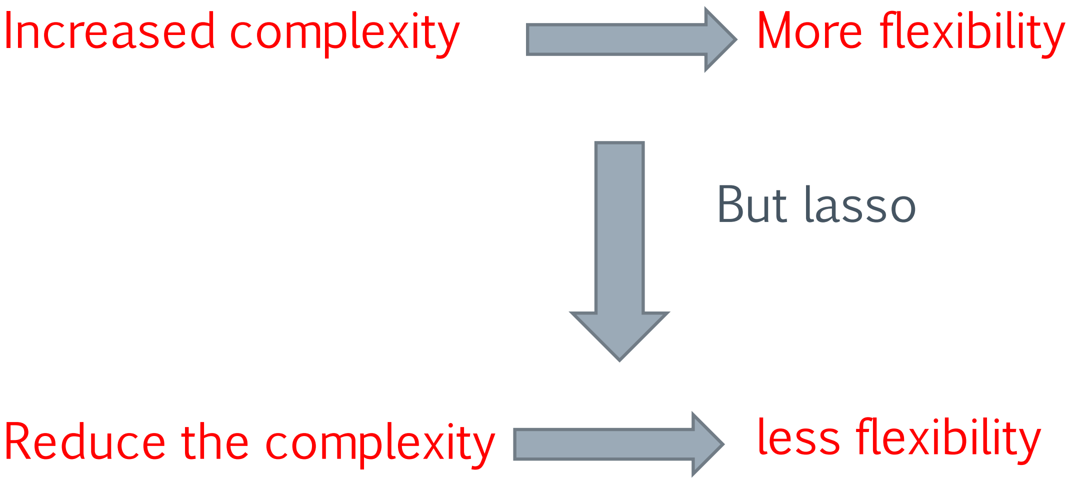
- Decreasing model complexity leads to lower variance. However, simpler models may not be sufficiently flexible to capture the underlying patterns in the data, leading to higher bias.

- **Lasso or L1:** in lasso some of the coefficient of the variable we make exactly to 0, which mean that variable will drop automatically. We consider these variables as less important for the model.

Week3 quiz:

- Explanation of Q1 answer

Bias Variance trade off.

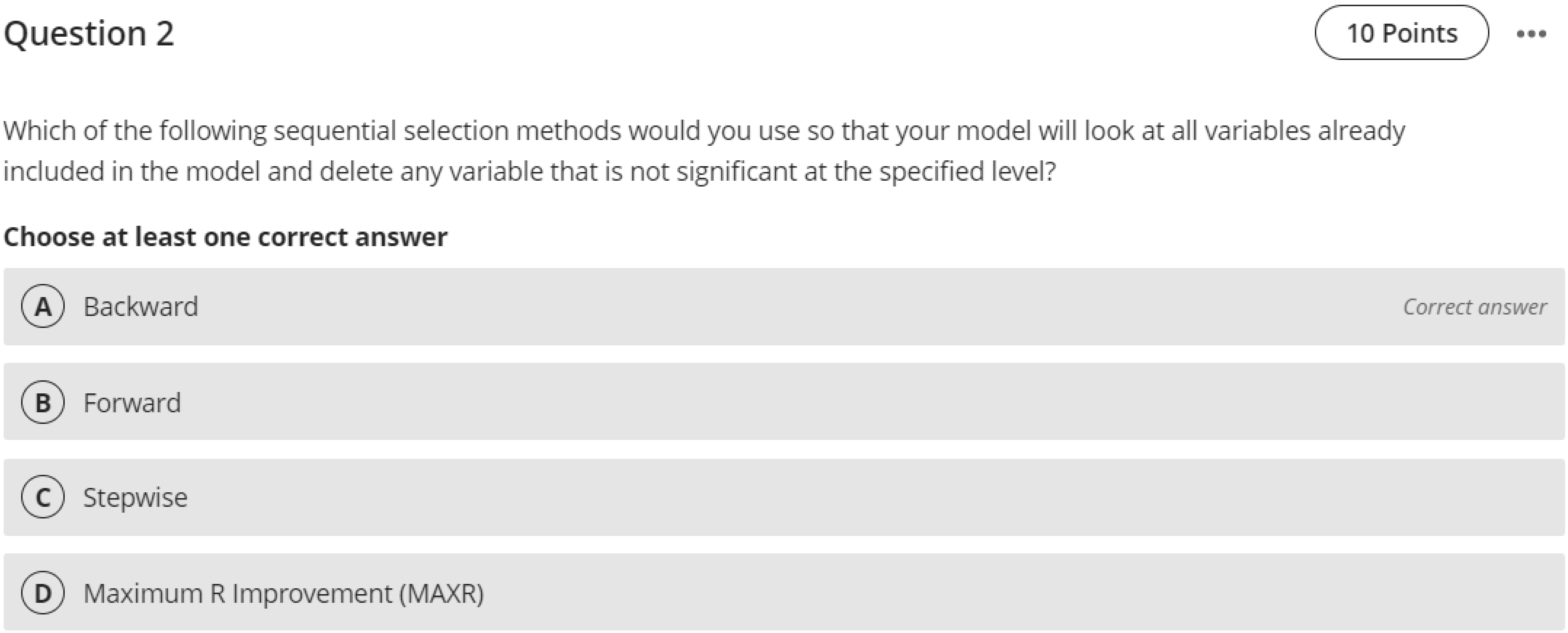


Increased complexity → More flexibility

But lasso

Reduce the complexity → less flexibility

Week3 quiz:



Question 2

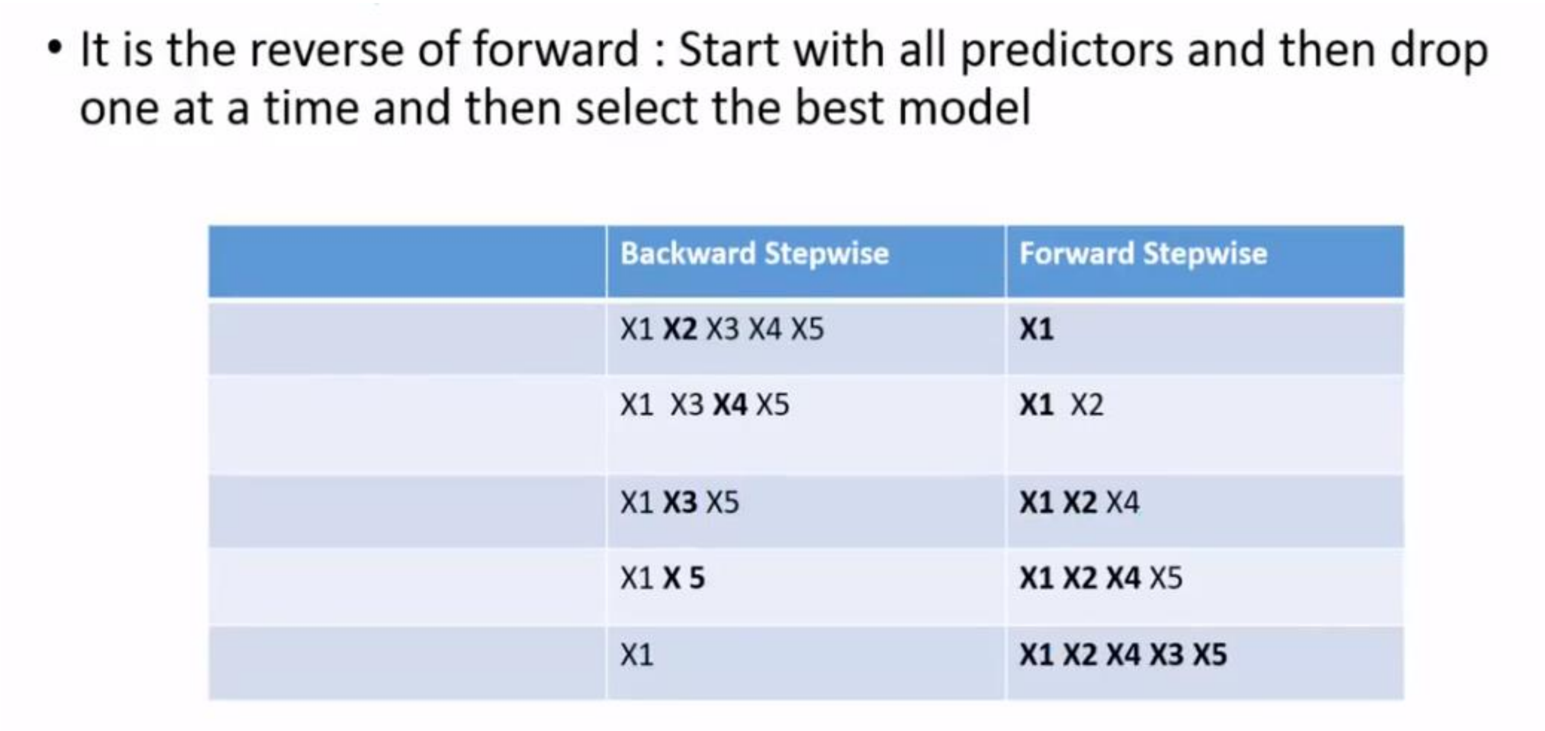
Which of the following sequential selection methods would you use so that your model will look at all variables already included in the model and delete any variable that is not significant at the specified level?

Choose at least one correct answer

- (A) Backward *Correct answer*
- (B) Forward
- (C) Stepwise
- (D) Maximum R Improvement (MAXR)

Week3 quiz:

- Explanation of Q2 answer

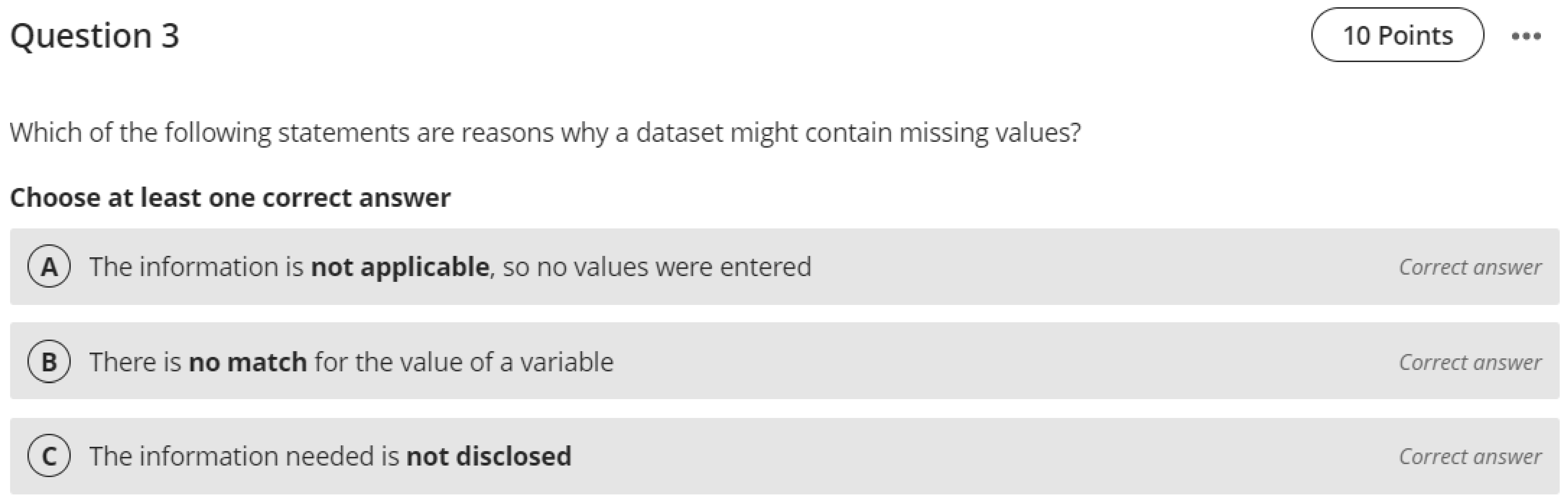


- It is the reverse of forward : Start with all predictors and then drop one at a time and then select the best model

	Backward Stepwise	Forward Stepwise
	X1 **X2** X3 X4 X5	**X1**
	X1 X3 **X4** X5	**X1** X2
	X1 **X3** X5	**X1 X2** X4
	X1 **X 5**	**X1 X2 X4** X5
	X1	**X1 X2 X4 X3 X5**

- The variable which drop once never been added again.

Week3 quiz:



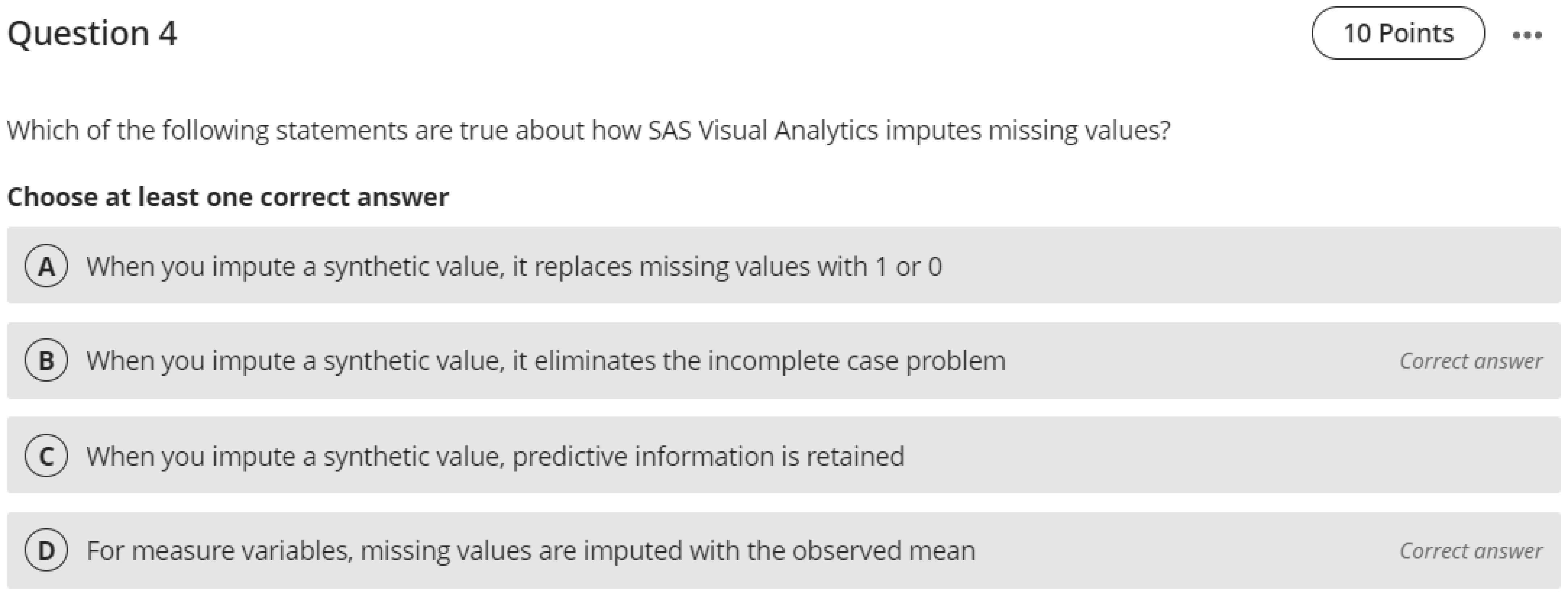
Question 3

Which of the following statements are reasons why a dataset might contain missing values?

Choose at least one correct answer

- Ⓐ The information is **not applicable**, so no values were entered *Correct answer*
- Ⓑ There is **no match** for the value of a variable *Correct answer*
- Ⓒ The information needed is **not disclosed** *Correct answer*

Week3 quiz:



Question 4

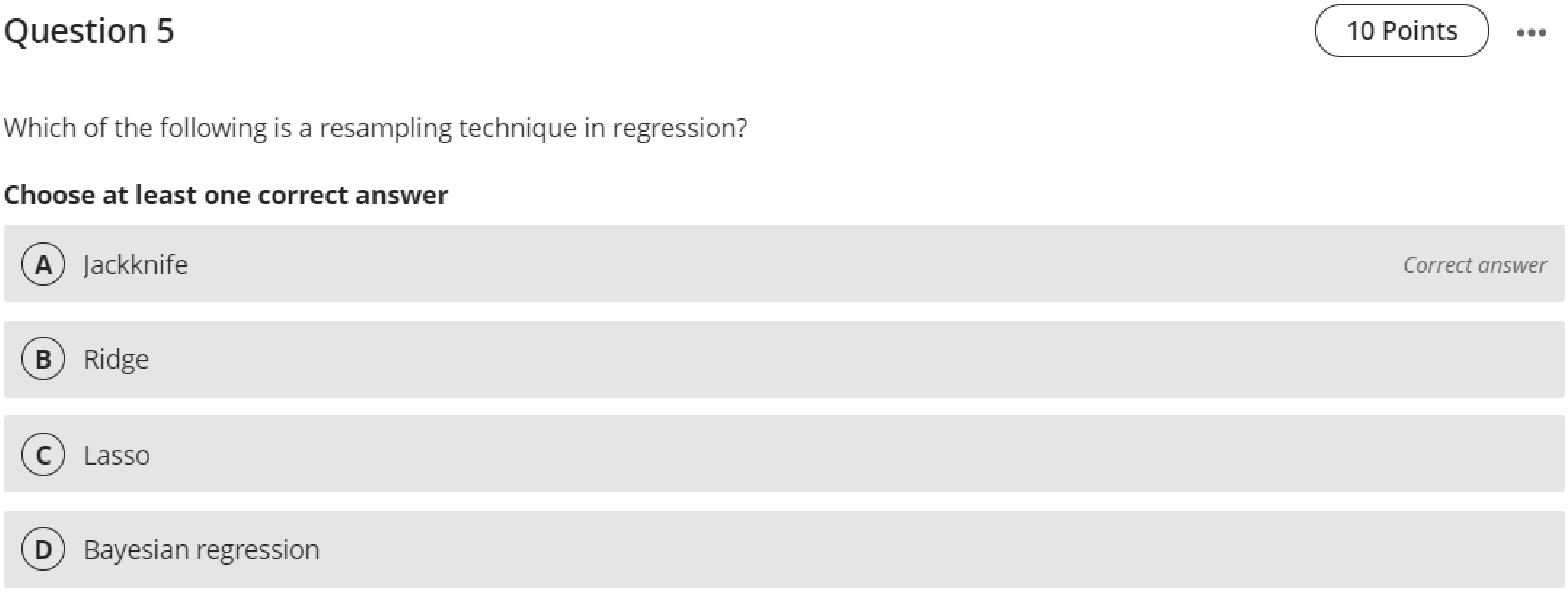
10 Points •••

Which of the following statements are true about how SAS Visual Analytics imputes missing values?

Choose at least one correct answer

- A When you impute a synthetic value, it replaces missing values with 1 or 0
- B When you impute a synthetic value, it eliminates the incomplete case problem *Correct answer*
- C When you impute a synthetic value, predictive information is retained
- D For measure variables, missing values are imputed with the observed mean *Correct answer*

Week3 quiz:



Question 5

Which of the following is a resampling technique in regression?

Choose at least one correct answer

- (A) Jackknife *Correct answer*
- (B) Ridge
- (C) Lasso
- (D) Bayesian regression

Logistic Regression: Recap

Simple Linear Regression Model:

We call it as a linear regression model because we use the linear equation to describe the model.

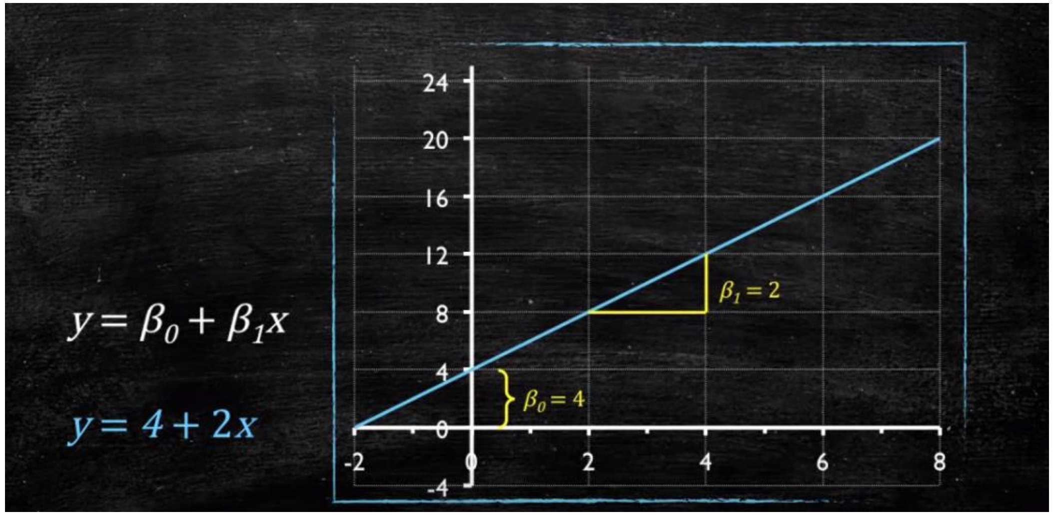
where we denote it as: $y = \beta_0 + \beta_1 x$

In here β_0= Intercept

β_1= slop

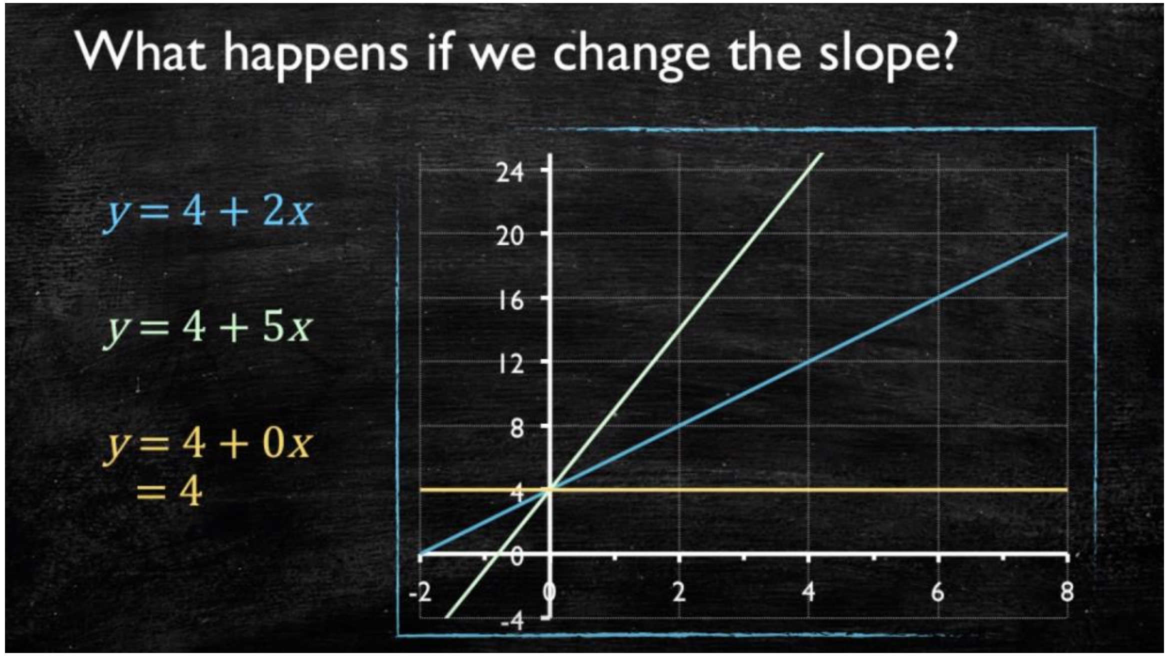
y=Dependent variable/Response var

x=Independent variable/predictor



Logistic Regression: Recap

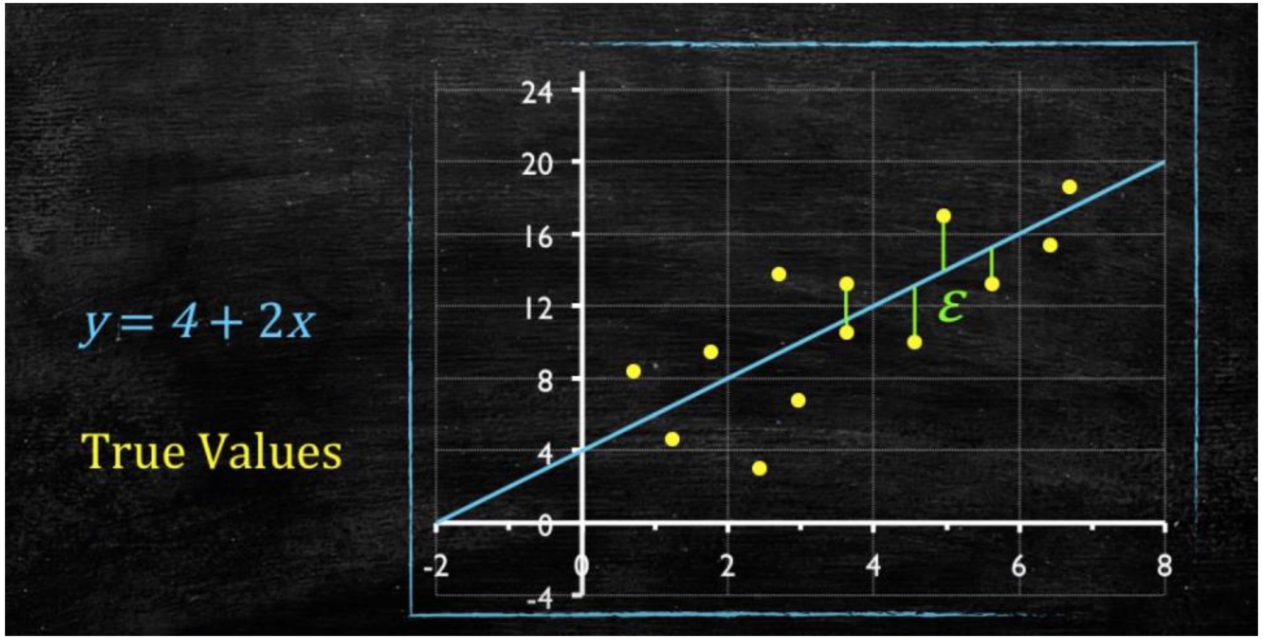
Simple Linear Regression Model:



- In here if slop is increase or decrease the value of y is change, if the slop is 0, it doesn't matter what value x have y will be remain always the same.
- But all the data in the real world is not stay in the striate line. They may have a linear pattern that's why we use linear model to represent the linear data.
- Where each data point may have a x and y value.

Logistic Regression: Recap

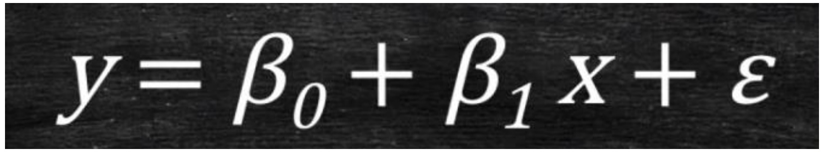
Simple Linear Regression Model:



- If we draw the line which may suits the most of the trend of the data, we can explain the model.
- In this case each data point is having the distance from the fitted line which we call the error term.
- Which we have to consider in the linear model. However we have to minimise the error term to get the best fit of our model.

Logistic Regression: Recap

Simple Linear Regression Model:



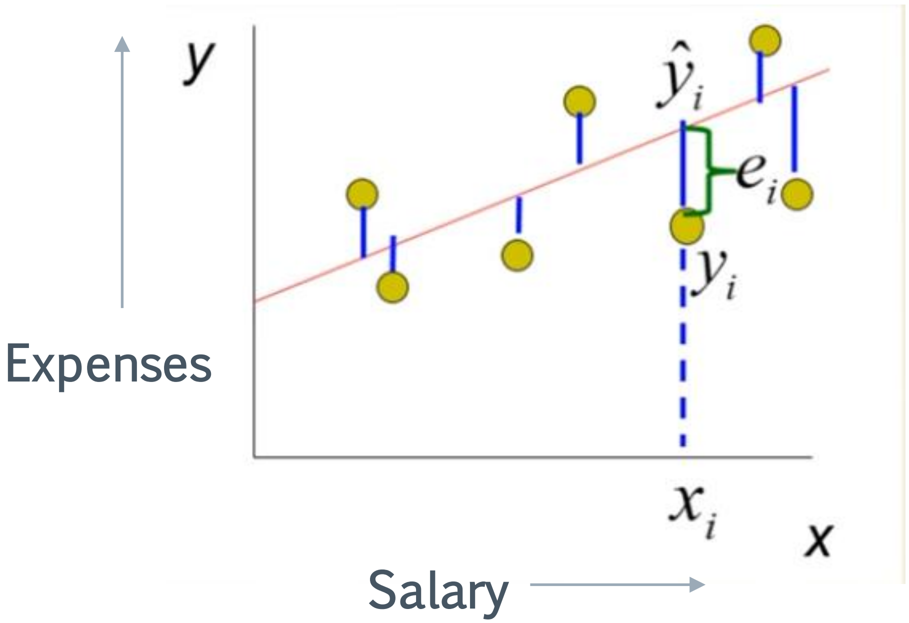
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- So this is our linear regression model.
- In this case each data point is having the distance from the fitted line which we call the error term.
- Which we have to consider in the linear model. However we have to minimise the error term to get the best fit of our model.
- We call the error term as Greek letter epsilon .

Logistic Regression: Recap (Linear regression modelling)

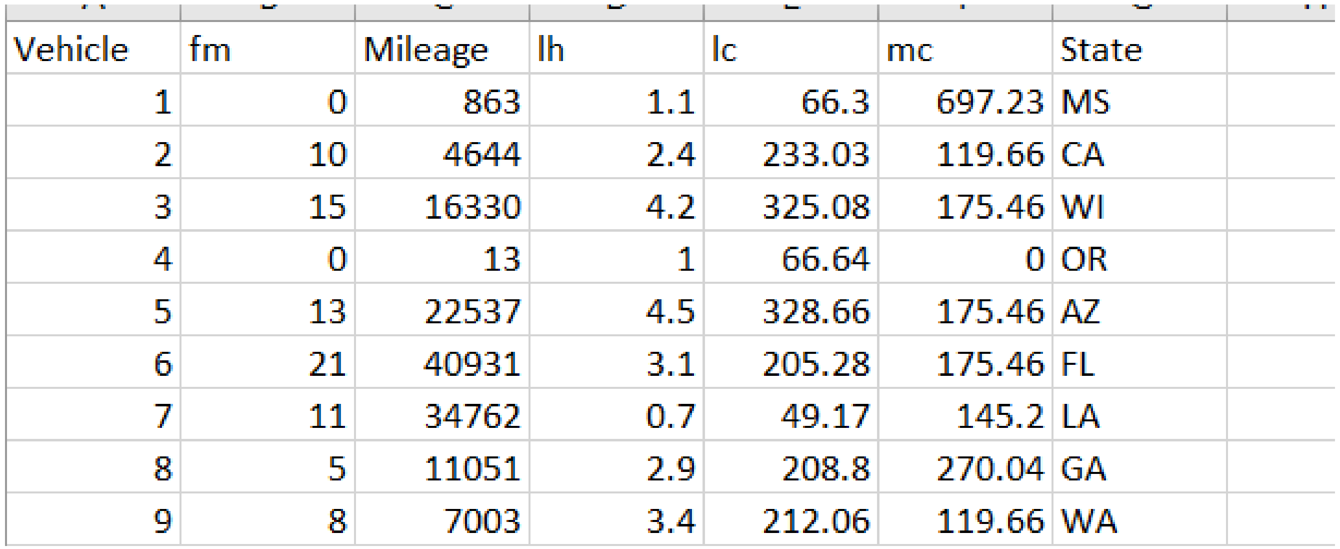
- Linear regression model try to fit a best fitted straight line which describe the relationship between x and y.
- Line fitting can be done on the basis of ordinary least squire (OLS)
- H0=No relation between the variables, slop=0
- Error/residual=observe value(y)-predicted value(y hat)

$$Min\sum^{n} e_i^2$$



Example: Linear regression (vehicle data set)

- Vehicle:



Vehicle	fm	Mileage	lh	lc	mc	State
1	0	863	1.1	66.3	697.23	MS
2	10	4644	2.4	233.03	119.66	CA
3	15	16330	4.2	325.08	175.46	WI
4	0	13	1	66.64	0	OR
5	13	22537	4.5	328.66	175.46	AZ
6	21	40931	3.1	205.28	175.46	FL
7	11	34762	0.7	49.17	145.2	LA
8	5	11051	2.9	208.8	270.04	GA
9	8	7003	3.4	212.06	119.66	WA

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Lc=β_0 + β_1 lh + β2 mileage + ε

↓

Lc=1.375+73.55 lh-.0000847 Mileage + ε

↓

Lc= - 0.236 + 73.51 lh + ε

Example: Linear regression (vehicle data set)

- Vehicle:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$Lc = \beta_0 + \beta_1 lh + \beta 2\ mileage + \varepsilon$

$Lc = 1.375 + 73.55\ lh - 0.0000847\ Mileage + \varepsilon$

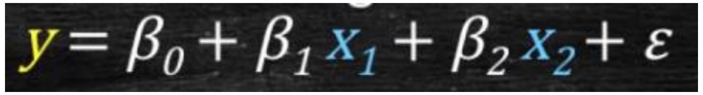
$Lc = -\ 0.236 + 73.51\ lh + \varepsilon$

Model Interpretation:

- In the first model mileage is not playing any significant roll , so we can avoid this variable from the model and in the final model only the independent variable will be lh.
- In the final model if the lh is 0 the lc will be -0.236
- Generally most cases intercept doesn't have any intuitive interpretation.
- If the lh increase by 1 unit, the lc will be increase by 73.51$ approximately.

Logistic Regression:

- In a general regression model if there are more than one independent variables or the predictors general regression model can be define as:



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Independent variables (x) can be:
  - Continuous: age, income height (numerical values)
  - Categorical: gender, good/bad etc (use of dummy variables)

- Dependent variable(y) also ca be
  - Continuous:
  - Categorical:

Logistic Regression:

If the dependent variable(y) can be a nature like:

- should a bank give a person loan or not?
- Is an individual transection is fraud or not?
- A person likely vote against the law or not?
- A person will back to shop or not?

  - ➢ This kind of problem will bring the binary outcome.
  - ➢ The answer can be yes/no
  - ➢ Using the dummy variables we may define the outcome as
    - ✓ Yes=1
    - ✓ No=0

Logistic Regression:

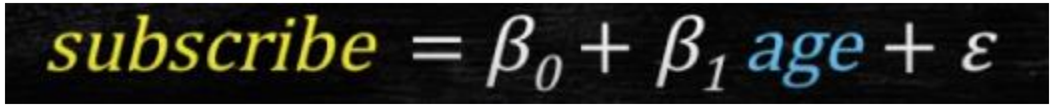
- We have the data of 1000 random customer, from a given city we want to know whether they will make a decision to subscribe a magazine or not:

  - In this case subscribe can be our dependent variable, which can get the value of 0 or 1:
    - 0=don’t subscribe
    - 1=will subscribe

    So the outcome is binary in nature.

  - The independent variable we have is age.

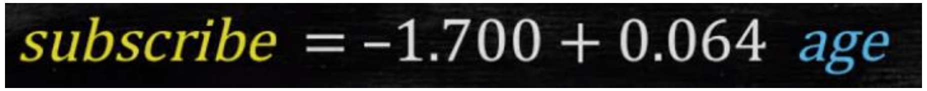
- Using the simple linear regression model we can design the model:



$$subscribe = \beta_0 + \beta_1\, age + \varepsilon$$

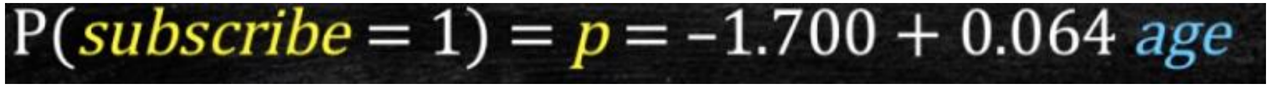
Logistic Regression:

- Using the real time data from the statistical package give us the following result: what does this result mean to us.



$$subscribe = -1.700 + 0.064\ age$$

- If our dependent variable is binary, than we want to see what make it change from 0 to 1.
- This also can be interpreted as what increase the likelihood of subscription (p=1), or the probability of buying the subscription.
- The result can be read as:

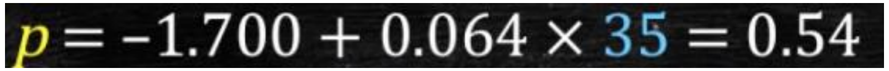


$$P(subscribe = 1) = p = -1.700 + 0.064\ age$$

- Which mean that every additional year of age increase the probability of buying the subscription increased by 6.4%

Logistic Regression:

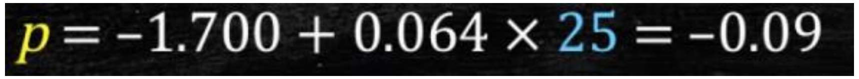
- Since the probability is always bounded by: 0<=p <=1 (between 0 to 1)
- The age range on the data set was 20 to 55
- So the probability of 35 years one is :



$$p = -1.700 + 0.064 \times 35 = 0.54$$

  - ➢ Which is a good outcome.

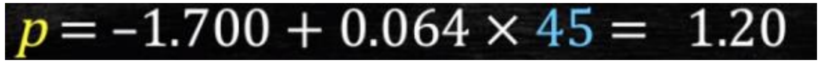
- However what is the probability of 25 years of age buying the subscription: using the same model



$$p = -1.700 + 0.064 \times 25 = -0.09$$

  - ➢ Which is not possible, since the probability can’t be negative.

- What about of 45 years of age buying the subscription:

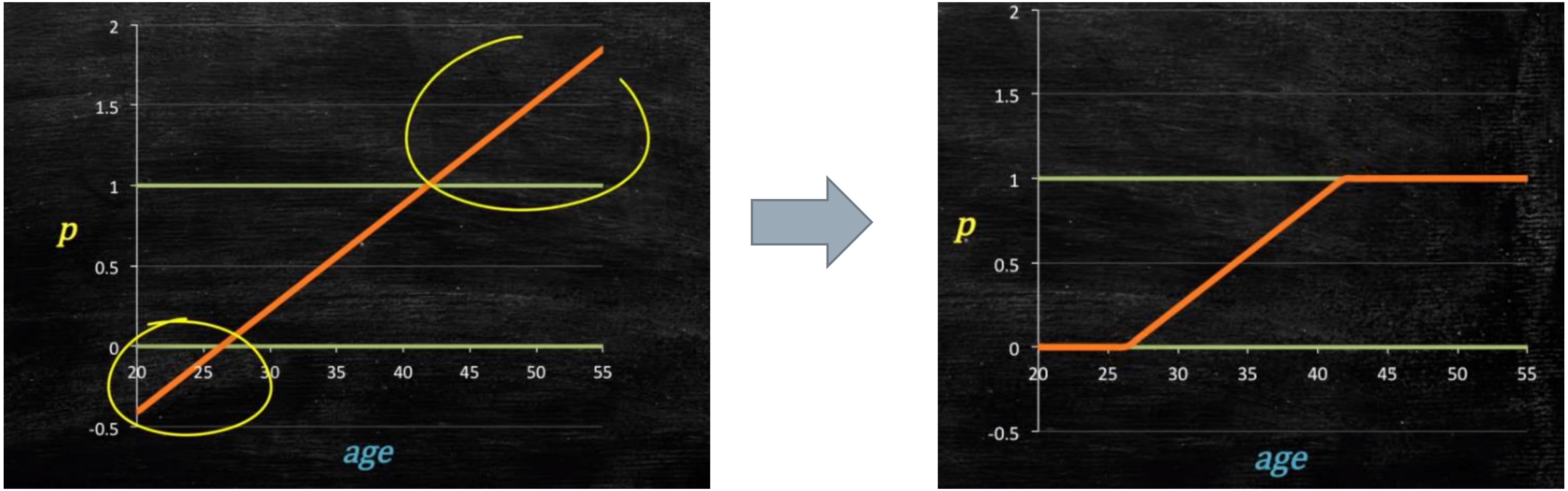


$$p = -1.700 + 0.064 \times 45 = \;\; 1.20$$

  - ➢ Which is also impossible since the probability is over 1

Logistic Regression:

- If we plot this situation:



Logistic Regression:

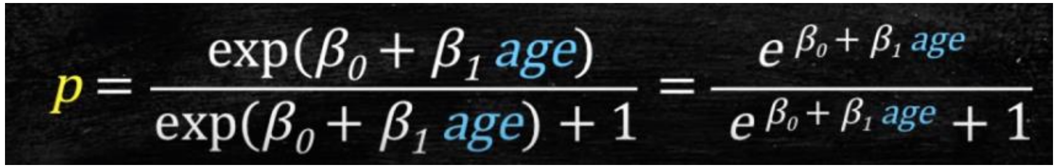
❑ Fixing the prior approach:

- We need to constrain p such a way that $0 \le p \le 1.$
- In here p is a function of age p=f(age)
- F(…) must satisfy two thing:
  - ➢ It must always be positive
  - ➢ Must be less than 1

- To make it must be positive we can either do any one of the approach:
  - 1. f(x)=abs(x)=lxl
  - 2. f(x)=x^2
  - 3. $p = \exp(\beta_0 + \beta_1\, age) = e^{\beta_0 + \beta_1\, age}$

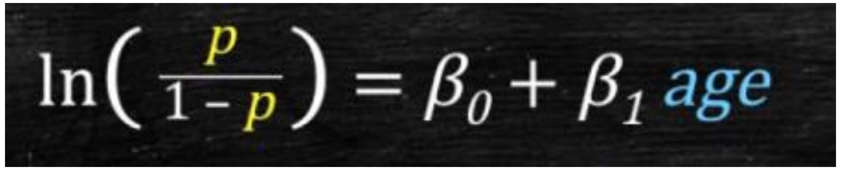
Logistic Regression:

- Since the previous exp function can be more than 1 to avoid that problem we can divide that with the same value and adding with one.



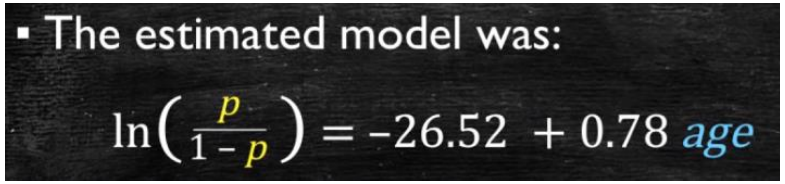
$$p = \frac{\exp(\beta_0 + \beta_1\, age)}{\exp(\beta_0 + \beta_1\, age) + 1} = \frac{e^{\beta_0 + \beta_1\, age}}{e^{\beta_0 + \beta_1\, age} + 1}$$

- Using some algebra the above equation can be rewrite as follows:



$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\, age$$

- So the estimated model for this function is:

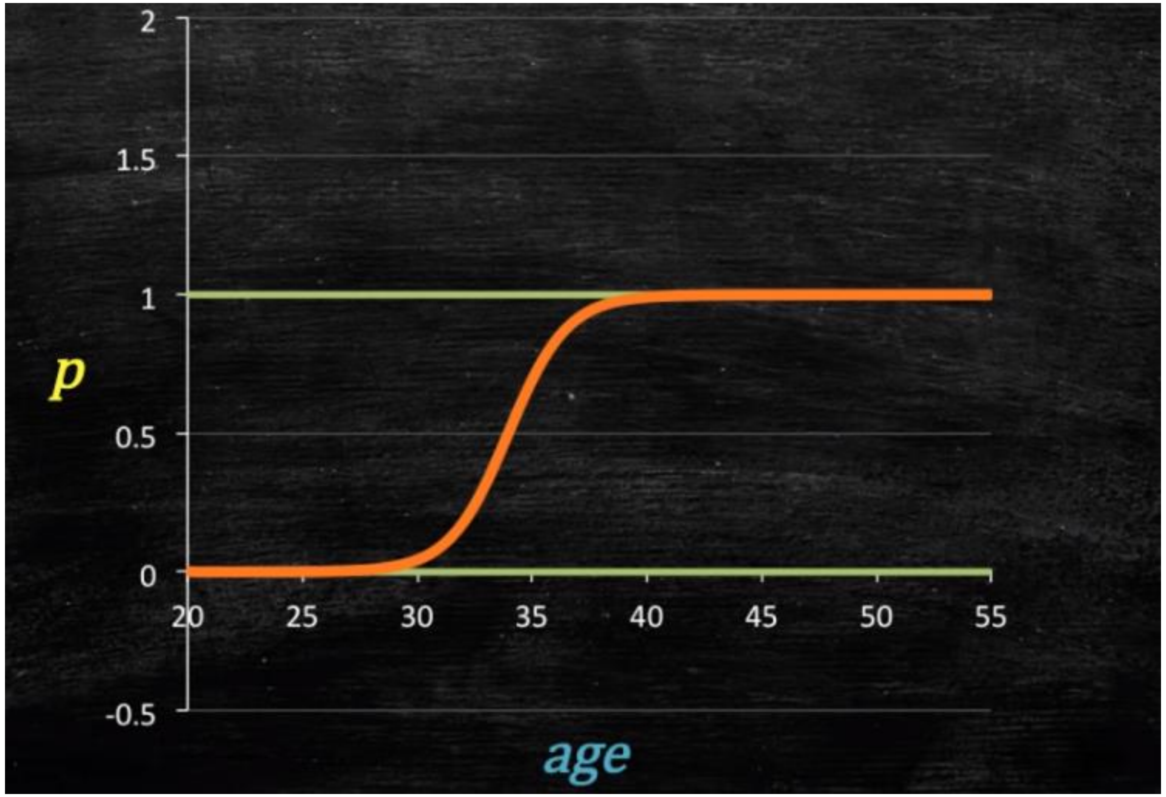


- The estimated model was:

$$\ln\left(\frac{p}{1-p}\right) = -26.52 + 0.78\, age$$

Logistic Regression:

So the final model of our logistic regression is given below:



Logistic Regression:

- Say we replace age by x for simplicity age=x:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Work book solution

Question 1

What is the AIC value for the model? **408,433.17**

Question 2

How many parameters ended up in the model? **30**

Question 3

What is the AIC value for the GLM – Gamma model? **332,364.36**

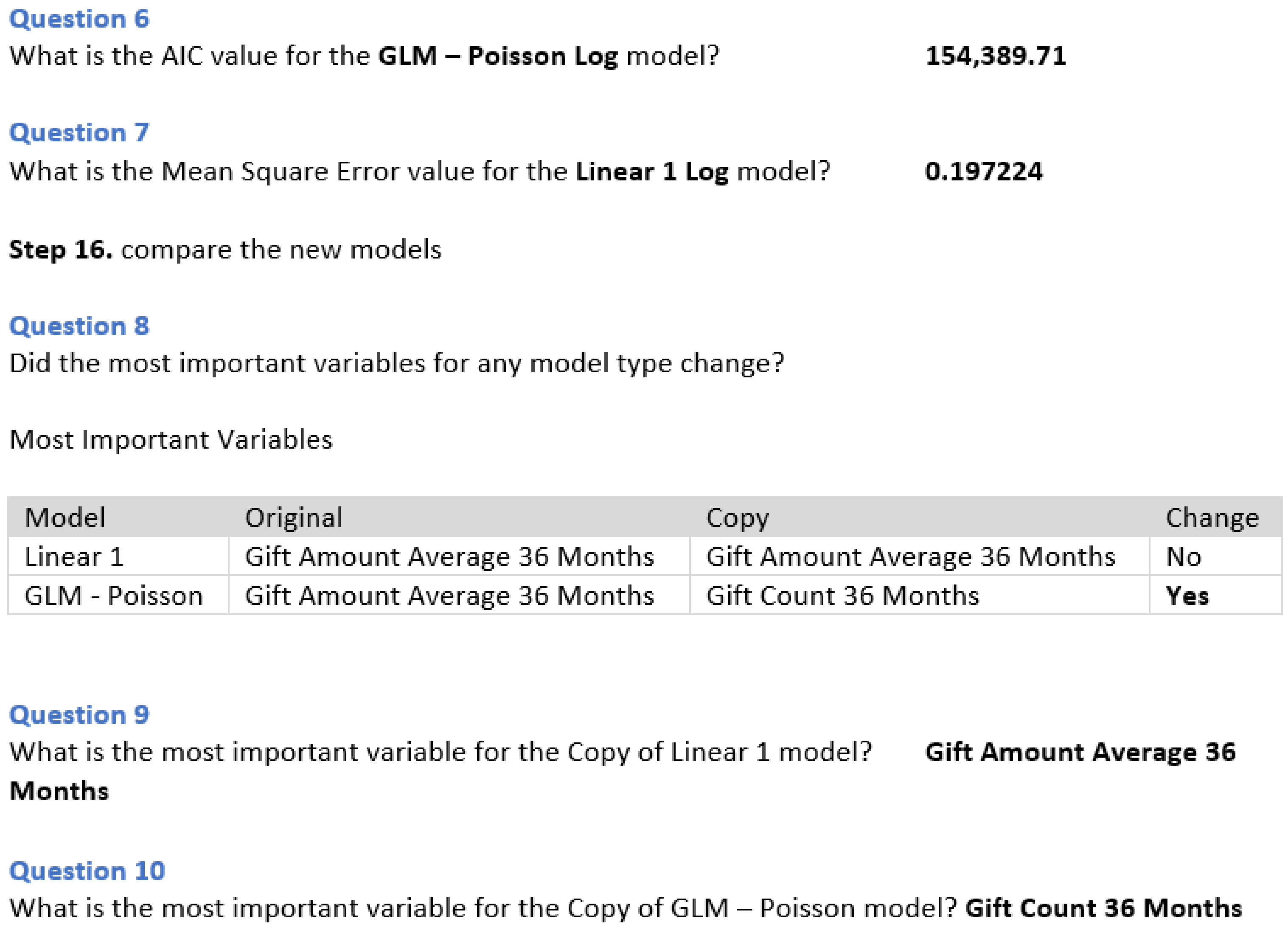
Question 4

Which model is selected as the best model? **Linear 1**

Question 5

What is the most important variable for the Linear 1 model? **Gift Amount Average 36 Months**

Work book solution



Question 6

What is the AIC value for the **GLM – Poisson Log** model? **154,389.71**

Question 7

What is the Mean Square Error value for the **Linear 1 Log** model? **0.197224**

Step 16. compare the new models

Question 8

Did the most important variables for any model type change?

Most Important Variables

Model	Original	Copy	Change
Linear 1	Gift Amount Average 36 Months	Gift Amount Average 36 Months	No
GLM - Poisson	Gift Amount Average 36 Months	Gift Count 36 Months	**Yes**

Question 9

What is the most important variable for the Copy of Linear 1 model? **Gift Amount Average 36 Months**

Question 10

What is the most important variable for the Copy of GLM – Poisson model? **Gift Count 36 Months**

Thank You