
UNIVERSITY OF EDINBURGH
Business School

# Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

**Dr Xuefei Lu**
The University of Edinburgh Business School


# Quiz

**Q1. Imagine that you want to predict house prices in Edinburgh based on the size. Which of the following statements is FALSE?**

- A. This is a supervised learning problem, since you have an outcome variable (house prices)
- B. If we regressed house prices on the size, we would have a continuous response variable and a discrete predictor.
- C. The relationship between house prices and size will be approximated by a straight line.
- D. The slope of the regression line will represent the effect of 1 unit increase in size on the house price.


# Quiz

**Q1. Imagine that you want to predict house prices in Edinburgh based on the size. Which of the following statements is FALSE?**

- A. This is a supervised learning problem, since you have an outcome variable (house prices)
- B. If we regressed house prices on the size, we would have a continuous response variable and a discrete predictor.
- C. The relationship between house prices and size will be approximated by a straight line.
- D. The slope of the regression line will represent the effect of 1 unit increase in size on the house price.


# Quiz

**Q2. A large retailer in the UK has asked you to regress CLV on the frequency of visiting the store. They have provided you with 5 observations: CLV = {10,5,20,15,5} and frequency = {10,3,6,7,4}. The value for $\widehat{\beta_1}$ has already been calculated and is equal to 1. Which of the following statements is FALSE?**

- A. The value for $\widehat{\beta_0}$ is equal to 5.
- B. The total variance in the response is equal to 153.
- C. The model has 1 degree of freedom.
- D. If you have visited the store 3 times, the model predicts CLV to be 8.


# Quiz

**Q2. A large retailer in the UK has asked you to regress CLV on the frequency of visiting the store. They have provided you with 5 observations: CLV = {10,5,20,15,5} and frequency = {10,3,6,7,4}. The value for $\widehat{\beta_1}$ has already been calculated and is equal to 1. Which of the following statements is FALSE?**

- A. The value for $\widehat{\beta_0}$ is equal to 5.
- B. The total variance in the response is equal to 153.
- C. The model has 1 degree of freedom.
- D. If you have visited the store 3 times, the model predicts CLV to be 8.

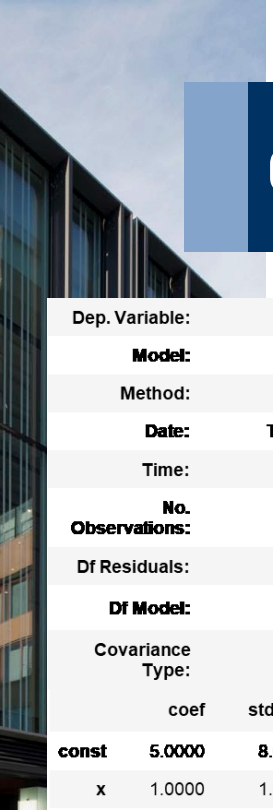
$\bar{y} = 11, \bar{x} = 6 \rightarrow \widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \cdot \bar{x} = 11 - 1 \cdot 6 = 5$

$\sum(y_i - \bar{y})^2 = (-1)^2 + (-6)^2 + (9)^2 + (4)^2 + (-6)^2 = 170$

We have 1 predictor, so 1 degree of freedom

The final equation is: $\hat{y} = 5 + 1 \cdot X = 5 + 1 \cdot 3 = 8$

# Quiz



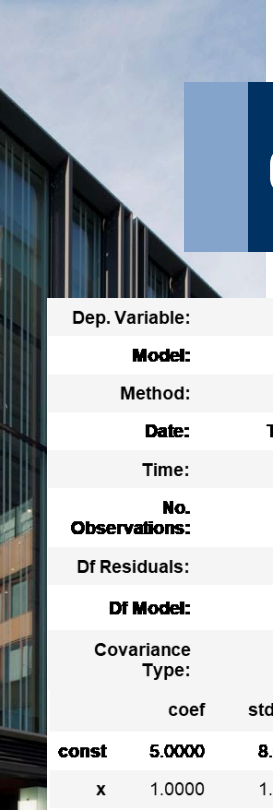
| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.176 |
| Model: | OLS | Adj. R-squared: | -0.098 |
| Method: | Least Squares | F-statistic: | 0.6429 |
| Date: | Thu, 11 Apr 2019 | Prob (F-statistic): | 0.481 |
| Time: | 15:34:16 | Log-Likelihood: | -15.425 |
| No. Observations: | 5 | AIC: | 34.85 |
| Df Residuals: | 3 | BIC: | 34.07 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.0000 | 8.083 | 0.619 | 0.580 | -20.723 | 30.723 |
| x | 1.0000 | 1.247 | 0.802 | 0.481 | -2.969 | 4.969 |

**Q3. You have built a simple linear regression model on the CLV example in question 2. What is the TRUE statement about the linear regression output? You can assume that Y is CLV and X is frequency.**

- A. Since the t-statistics are very small, we can reject the null hypothesis that frequency has no significant influence on CLV.
- B. A p-value of 0.580 means that the probability at the intercept is significant is 58%.
- C. The R2 is 17.6%, which means that the correlation between the CLV and frequency is $\sqrt{0.176}$.
- D. The regression line will cross the x-axis at 5.

# Quiz




| | | | | | | |
|---|---|---|---|---|---|---|
| Dep. Variable: | | | y | R-squared: | | 0.176 |
| Model: | | | OLS | Adj. R-squared: | | -0.098 |
| Method: | | | Least Squares | F-statistic: | | 0.6429 |
| Date: | | | Thu, 11 Apr 2019 | Prob (F-statistic): | | 0.481 |
| Time: | | | 15:34:16 | Log-Likelihood: | | -15.425 |
| No. Observations: | | | 5 | AIC: | | 34.85 |
| Df Residuals: | | | 3 | BIC: | | 34.07 |
| Df Model: | | | 1 | | | |
| Covariance Type: | | | nonrobust | | | |
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
| const | 5.0000 | 8.083 | 0.619 | 0.580 | -20.723 | 30.723 |
| x | 1.0000 | 1.247 | 0.802 | 0.481 | -2.969 | 4.969 |

**Q3. You have built a simple linear regression model on the CLV example in question 2. What is the TRUE statement about the linear regression output? You can assume that Y is CLV and X is frequency.**

- A. Since the t-statistics are very small, we can reject the null hypothesis that frequency has no significant influence on CLV.
- B. A p-value of 0.580 means that the probability at the intercept is significant is 58%.
- C. The R2 is 17.6%, which means that the correlation between the CLV and frequency is $\sqrt{0.176}$.
- D. The regression line will cross the x-axis at 5.


UNIVERSITY OF EDINBURGH
Business School