UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Quiz

I have modelled the average house prices in Boston (TARGET) based on the proportion of non-business retail acres per town (INDUS), the average number of rooms per dwelling (RM) and the proportion of lower status of the population (LSTAT). The output of the coefficients estimates look like this:

Which of the following statements about the interpretation of the coefficients is TRUE

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.4606	3.171	-0.461	0.645	-7.691	4.770
ZN	0.0158	0.012	1.359	0.175	-0.007	0.039
RM	5.0455	0.446	11.324	0.000	4.170	5.921
LSTAT	-0.6240	0.046	-13.644	0.000	-0.714	-0.534

- A. The LSTAT says that when the percentage of lower status population rises with 1%, the house prices will decrease by $624 on average, keeping everything else constant.
- B. The coefficient for ZN is significant on the 10% significance level, but not on the 5% significance level.

C. If the average number of rooms per dwelling rises with 1 unit, then the average house price will rise by $500 on average.

D. The LSTAT says that when the percentage of lower status population rises with 1%-point, the house prices will decrease by $624 on average, keeping everything else constant.

Quiz

I have modelled the average house prices in Boston (TARGET, * $1000) based on the proportion of non-business retail acres per town (INDUS), the average number of rooms per dwelling (RM) and the proportion of lower status of the population (LSTAT). The output of the coefficients estimates look like this:

Which of the following statements about the interpretation of the coefficients is TRUE

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.4606	3.171	-0.461	0.645	-7.691	4.770
ZN	0.0158	0.012	1.359	0.175	-0.007	0.039
RM	5.0455	0.446	11.324	0.000	4.170	5.921
LSTAT	-0.6240	0.046	-13.644	0.000	-0.714	-0.534

- A. The LSTAT says that when the percentage of lower status population rises with 1%, the house prices will decrease by $624 on average, keeping everything else constant.
- B. The coefficient for ZN is significant on the 10% significance level, but not on the 5% significance level.

C. If the average number of rooms per dwelling rises with 1 unit, then the average house price will rise by $500 on average.

D. The LSTAT says that when the percentage of lower status population rises with 1%-point, the house prices will decrease by $624 on average, keeping everything else constant.

Quiz

Which of the following statements about the F-statistic is correct?

A: If the dataset has 506 observations, the degrees of freedom of the F-statistic are equal to 3 and 502.

B: The p-value of the F-statistic is smaller than 0.0001. This means that all the variables are related to the response.

C: If the F-statistic is close to 1, we reject the null hypothesis.

D: Instead of looking at the F-statistic, you can just have a look at the p-values of the predictors.

Quiz

Which of the following statements about the F-statistic is correct?

A: If the dataset has 506 observations, the degrees of freedom of the F-statistic are equal to 3 and 502.

B: The p-value of the F-statistic is smaller than 0.0001. This means that all the variables are related to the response.

C: If the F-statistic is close to 1, we reject the null hypothesis.

D: Instead of looking at the F-statistic, you can just have a look at the p-values of the predictors.

Quiz

The R-squared of the model is 0.640 and the adjusted R-squared is 0.638. What can you tell about the model performance?

A: The total variation in house prices explained by all the predictors is 0.638.

B: The formula for the adjusted R-squared is given by the following formula. The denominator for RSS is 503, and 505 for TSS.

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}}$$

C: The adjusted R-squared and the R-squared are almost equal, so we can conclude that the model is not overfitting.

D: R-squared and adjusted R-squared both assume that all predictors explain variation in the data.

Quiz

The R-squared of the model is 0.640 and the adjusted R-squared is 0.638. What can you tell about the model performance?

A: The total variation in house prices explained by all the predictors is 0.638.

B: The formula for the adjusted R-squared is given by the following formula. The denominator for RSS is 503, and 505 for TSS.

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}}$$

C: The adjusted R-squared and the R-squared are almost equal, so we can conclude that the model is not overfitting.

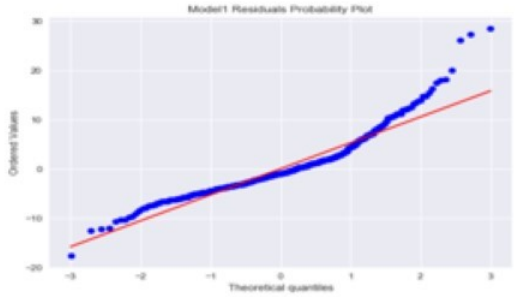
D: R-squared and adjusted R-squared both assume that all predictors explain variation in the data.

Quiz

You are testing the assumptions and you are confronted with the following correlation matrix and QQ-plot. What can you say about the assumptions?

A. The multivariate normality assumption is not violated.
B. There is no severe multicollinearity problem. However, you should look at the VIFs to be sure.
C. The residuals have a clear normal distribution.
D. The correlation LSTAT and RM is too high. As a result, one of the two variables should be removed.

	ZN	RM	LSTAT
ZN	1.000000	0.311991	-0.412995
RM	0.311991	1.000000	-0.613808
LSTAT	-0.412995	-0.613808	1.000000



Quiz

You are testing the assumptions and you are confronted with the following correlation matrix and QQ-plot. What can you say about the assumptions?

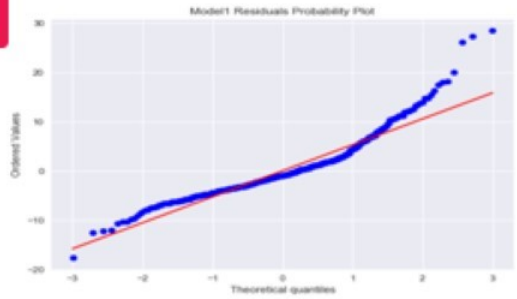
A. The multivariate normality assumption is not violated.

B. There is no severe multicollinearity problem. However, you should look at the VIFs to be sure.

C. The residuals have a clear normal distribution.

D. The correlation LSTAT and RM is too high. As a result, one of the two variables should be removed.

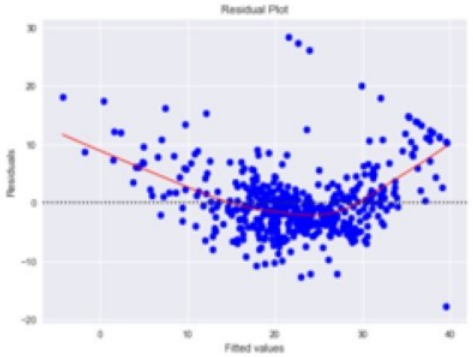
	ZN	RM	LSTAT
ZN	1.000000	0.311991	-0.412995
RM	0.311991	1.000000	-0.613808
LSTAT	-0.412995	-0.613808	1.000000



Quiz

Looking at this residual plot, what can you tell about the assumption of the model?

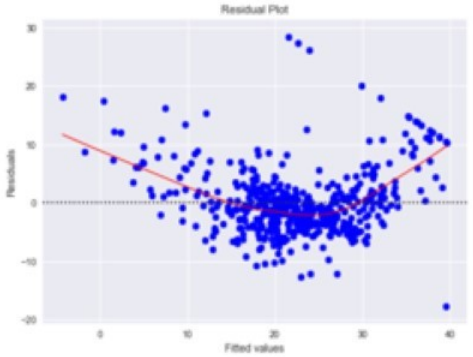
A. We don't notice a funnel shape in the residual plot so there is no clear evidence of heteroscedasticity.
B. Homoscedasticity means that the predictors have a constant variance.
C. The residual plot does not inform us about the linearity of the data.
D. There is evidence of heteroscedasticity since there is a clear pattern in the residual plot.



Quiz

Looking at this residual plot, what can you tell about the assumption of the model?

A. We don't notice a funnel shape in the residual plot so there is no clear evidence of heteroscedasticity.
B. Homoscedasticity means that the predictors have a constant variance.
C. The residual plot does not inform us about the linearity of the data.
D. There is evidence of heteroscedasticity since there is a clear pattern in the residual plot.



UNIVERSITY OF EDINBURGH
Business School