UNIVERSITY OF EDINBURGH
Business School

# Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

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


# Multiple Linear Regression

# Multiple linear regression: Intro

*Multiple linear regression* enables you to predict a continuous response using two or more continuous or discrete predictors.

Recall that the simple linear regression:

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

The basic equation that defines **the multiple linear regression (MLR)** model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- The structural part of the model $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- The error part of the model $\epsilon$
- The j coefficients $\beta_j$ are called the **partial slopes** of the response variable with respect to the j-th predictor $x_j$.

# Multiple linear regression: Intro

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- Estimation of $\beta_j$'s is still performed using **least squares** and all the quantities we introduced in simple linear regression model are used in multiple regression in a similar way.
- Specifically, the principle to find the optimal coefficient estimates remains the same as in simple linear regression, namely minimising the **residual sum of squares (RSS)**:

$$\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = \sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_{i1} - \widehat{\beta_2}x_{i2} \ldots - \widehat{\beta_p}x_{ip}\right)^2$$

Once we have the estimates of the coefficient, we can make a prediction for a single instance $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ using the equation

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}x_{i1} + \widehat{\beta_2}x_{i2} \ldots + \widehat{\beta_p}x_{ip}$$

In multiple linear regression, the coefficient $\widehat{\beta_j}$ would represent the average effect of one unit increase in $x_j$ **while keeping all other predictors fixed**.

# Multiple linear regression: Assumptions

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

**Assumption 1: linear regression assumes that there is a linear relationship between predictor and response.**

To test whether or not there is a linear relationship between a predictor and a response, you can make **a scatterplot between each predictor and the response**. If the relationship is linear, the outcome variable will linearly increase or decrease in terms of the response.

Another option is to make **a residual plot**. This plot depicts the residuals against the fitted values. If there is a linear relationship, the points should be randomly distributed around the horizontal line.

**Assumption 2: Multiple linear regression assumes that there is NO multicollinearity in the data.**

This means that the predictors cannot be too closely related to each other or, in others words, that the correlation between the predictors is not too high. The easiest, and most effective, way to deal with multicollinearity is to delete highly correlated variables.

# Multiple linear regression: Assumptions

Two ways to check for multicollinearity:

1. calculate the Pearson correlation matrix among all predictors. This correlation matrix shows you the correlation between all predictors and a rule of thumb is that the correlation between two predictors should be smaller than 0.80.
2. A second way is to calculate the **Variance Inflation Factor (VIF)**. The VIF for the j-th predictor in a linear regression model with p predictors is defined as

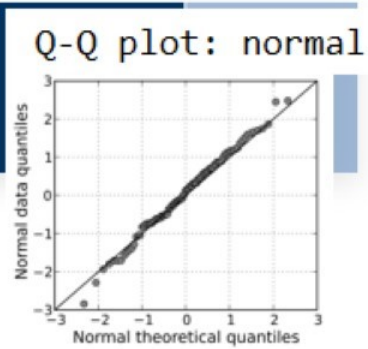
$$VIF_j = \frac{1}{1 - R_j^2}$$

where $R_j^2$ is the R-squared value for the (auxiliary) regression of the j-th predictor versus all the remaining ones.

The VIF provides a measure of **the reduction in the precision** of a coefficient estimate. VIF is a number greater than or equal to 1.

- When it is equal to 1, it means that the predictor is not affected by any collinearity problem.
- The larger it is, the stronger is the association of that predictor with all the other ones in the model.
- If the VIF of a predictor is larger than 10, we flag the predictor as affected by a severe collinearity problem.

# Multiple linear regression: Assumptions




**Assumption 3: Multivariate normality**

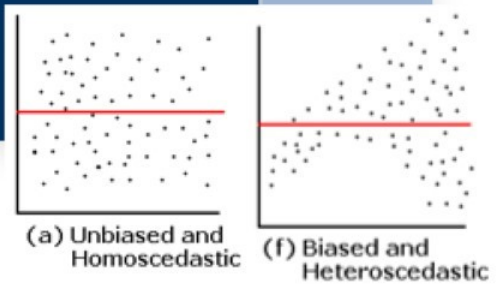
Multiple regression assumes that the residuals are normally distributed.

- Tests for Normality
  - Q-Q-plots or simple histograms.
  - Goodness-of-fit test, e.g., a Kolmogorov-Smirnov test on the residuals. The null hypothesis of K-S test is that the data is normally distributed.
- If there is no multivariate normality, then a transformation of the variables (e.g., a logarithmic transformation) is advised.

**Assumption 4: residuals are independent**

- In other words, there is no **autocorrelation** between the error terms. Independent means that the error term of a certain observation i does not say anything about the error term of observation j.
- Autocorrelation can be tested with a scatterplot or with **an autocorrelation test, e.g. Durbin-Watson test**. The null hypothesis for this test is that there is no linear autocorrelation between the error terms.

# Multiple linear regression: Assumptions




**Assumption 5: homoscedasticity**

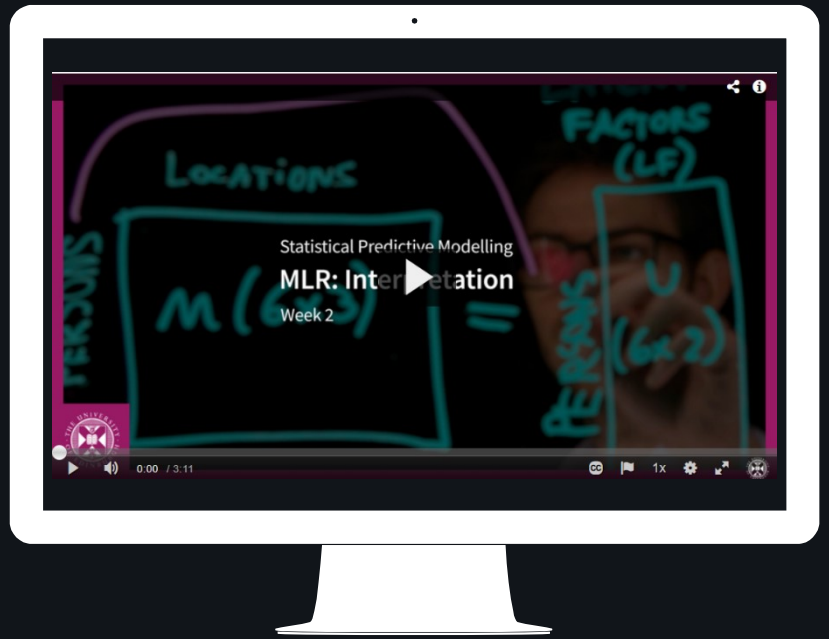
Constant variance of the error terms across all observations.

- Can be identified with a residual plot
- If the error terms are homoscedastic, we see a chaotic scatterplot of the error terms with no real relationship. We don't notice a changing spread of the error terms, instead the spread of the error terms is constant and we have a constant variance.
- If there is no homoscedasticity (or heteroscedasticity), we see a pattern in the scatter plot. Often this pattern has a funnel shape and this represents a changing spread of the error terms.

To summarise, multiple linear regression has 5 main assumptions:

- 2 related to the predictors: linear relationship between predictors and response, no multicollinearity
- 1 related to the predictors and the error terms: multivariate normality
- 2 related to the error terms: no autocorrelation and homoscedasticity

# MLR: Interpretation




Watch the following the video via this [link](https://media.ed.ac.uk/media/MLRA+Interpretation/0_jb4haxp9/122596071)

https://media.ed.ac.uk/media/MLRA+Interpretation/0_jb4haxp9/122596071

- Try the following exercise:

4 - Activity 2 - Build a multiple regression model.ipynb

**Box office revenues.**

Instructions

- Produce a simple linear regression model for both models and have a look at the coefficients and output. You might also want to make a scatterplot for both models.
- Make a multiple linear regression model and interpret the coefficients and the model accuracy.

## Activity: Build a multiple linear regression model

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.756 |
| Model: | OLS | Adj. R-squared: | 0.686 |
| Method: | Least Squares | F-statistic: | 10.83 |
| Date: | Mon, 12 Oct 2020 | Prob (F-statistic): | 0.00720 |
| Time: | 20:24:20 | Log-Likelihood: | -37.029 |
| No. Observations: | 10 | AIC: | 80.06 |
| Df Residuals: | 7 | BIC: | 80.97 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -8.4879 | 13.908 | -0.610 | 0.561 | -41.376 | 24.400 |
| x1 | 0.2066 | 0.064 | 3.217 | 0.015 | 0.055 | 0.358 |
| x2 | 10.6473 | 6.566 | 1.622 | 0.149 | -4.879 | 26.174 |

| | | | |
|---|---|---|---|
| Omnibus: | 2.741 | Durbin-Watson: | 2.477 |
| Prob(Omnibus): | 0.254 | Jarque-Bera (JB): | 1.466 |
| Skew: | -0.920 | Prob(JB): | 0.480 |
| Kurtosis: | 2.631 | Cond. No. | 488. |

**F-statistics:** Whether at least one predictor is related to the response variable?

$$Box\ office\ sales = \hat{\beta}_0 + \hat{\beta}_1 Advertising + \hat{\beta}_2 Theaters + \varepsilon$$

## Activity: Multiple linear regression outputs and interpretation

| | |
|---|---|
| Adj. R-squared: | 0.686 |
| F-statistic: | 10.83 |
| Prob (F-statistic): | 0.00720 |

$$Box\ office\ sales = \hat{\beta}_0 + \hat{\beta}_1 Advertising + \hat{\beta}_2 Theaters + \varepsilon$$

**To test whether there is at least one predictor related to the response, we perform an F-test.**

$H_0$: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs $H_1$: at least one at least one of the $\beta_j$ in $H_0$ is $\neq 0$

- Note that if apply F-test on simple linear regression ⇔ testing significance of $\beta_1$
- The F-test uses an F-statistic ~ F distribution. An F distribution is characterised by two degrees of freedom:
  - p (the number of predictors)
  - n-p-1 (with n the number of observations).

  E.g. 10 observations and 2 predictors: F(2, 7)

| | |
|---|---|
| Df Residuals: | 7 |
| Df Model: | 2 |

- **Reject the Null hypothesis $H_0$:**
  - when the p-value is smaller than 0.05 (value depend on significance level)
  - Or, F-statistics close to 1

## Activity: Multiple linear regression outputs and interpretation

- F-test
  - $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - $H_A: at\ least\ one\ \beta_j\ is\ non - zero$
  - Follows an *F*-distribution with *p* and *n-p-1* degrees of freedom
  - Reject $H_0$ if $F > 1$ or $p - value < 0.05$

**Activity: Multiple linear regression outputs and interpretation**

```
R-squared:                       0.756
Adj. R-squared:                  0.686
```

- The adjusted R-squared
  - $\bar{R}^2 = 1 - \frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}}$ with d = degrees of freedom

The measure was introduced because the regular R-squared always increases when more predictors are added, even if these predictors are weakly related to the response variable. The adjusted R-squared solves this issue by controlling the number of predictors in both the nominator and the denominator of the equation.

$d$ stands for the degrees of freedom and this makes sure that the R-squared will not increase when adding more predictors. The reason is that when all the correct predictors are included in the model, adding more predictors will only slightly decrease the residual sum of squares.

## Activity: Multiple linear regression outputs and interpretation

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -8.4879 | 13.908 | -0.610 | 0.561 | -41.376 | 24.400 |
| x1 | 0.2066 | 0.064 | 3.217 | 0.015 | 0.055 | 0.358 |
| x2 | 10.6473 | 6.566 | 1.622 | 0.149 | -4.879 | 26.174 |

- Finally, let's have a look at the coefficients and their p-value. We see that x1 (or advertising) is significant and that an increase in advertising of £1000 would lead to an increase in box office sales of £207, while keeping the number of theatres constant. The number of theatres does not significantly increase model performance with a p-value of 0.15. Playing the movie in one extra theatre would lead to an increase in box office revenues of £11, while keeping the advertising spending fixed.

## Activity: Multiple linear regression outputs and interpretation

# 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 $607 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

**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.

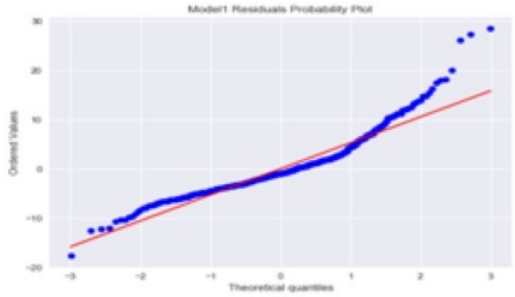
D: R-squared and adjusted R-squared both assume that all predictors explain variation in the data.

# 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.

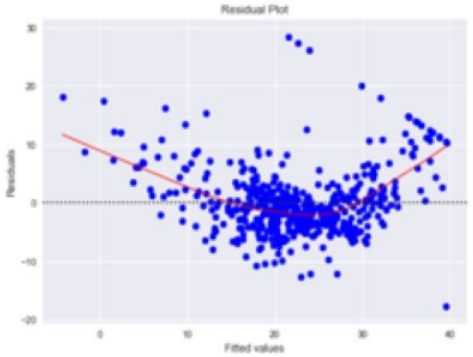
| | ZN | RM | LSTAT |
|---|---|---|---|
| **ZN** | 1.000000 | 0.311991 | -0.412995 |
| **RM** | 0.311991 | 1.000000 | -0.613808 |
| **LSTAT** | -0.412995 | -0.613808 | 1.000000 |



# Quiz

**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