



UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data


CMSE11428 (2020-2021)

Dr Xuefei Lu

The University of Edinburgh Business School

A photograph of a modern, multi-story building with a glass and metal facade, identified as the University of Cambridge Business School. The building has a distinctive corner design with large windows. A sign on the ground floor reads "UNIVERSITY OF CAMBRIDGE Business School".

More about Regressions


- 
- A photograph of a modern building with large glass windows, identified as the University of Cambridge Business School.
- Please study the following file:
 - 5 - Adding predictors to your model.ipynb

Activity: Adding predictors to your model

MLR: Model Selection



- Please watch the following video via this [link](#)
- https://media.ed.ac.uk/media/M/LRA+Model+Selection/1_yc0xwudy/122596071

- 
- A photograph of a modern building with large glass windows, identified as the University of Cambridge Business School.
- Please study the following file:

6 - How to select the best models.ipynb

Then try the following exercise:

7 - Exercise - Build multiple models.ipynb + social_media_2.csv

Activity: Approximation measures and cross-validation

Problems with linearity assumptions

Remember that linear regression had the following major assumptions:

- Linearity between predictors and response
- No multicollinearity
- Multivariate normality
- No autocorrelation
- Homoscedasticity

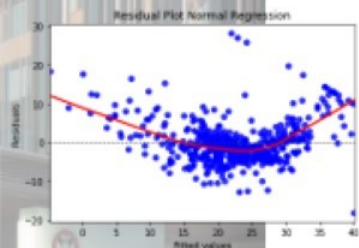
In most cases, a violation of the linearity assumption will cause the most issues. Actually, a violation of the linearity assumptions:

- ➔ often leads to the error being non-normal distributions and heteroscedasticity.
- ➔ your model will have a high bias problem: your model is too simple to capture the true relationship and hence your predictions on both the training and the test set will be far off.

Linear or non-linear?

There are several ways to diagnose whether the relationship is linear or not.

1. To make a scatter plot between the predictors and the response and see what kind of relationship you can spot (linear or non-linear).
2. to make a residual plot for each predictor x_i that depicts the residuals $e_i = y_i - \hat{y}_i$ against observed values of x_i . Disadvantage: need to make a residual plot for each x_i
3. A better way is to make a residual plot in the case of multiple linear regression is to plot the residuals e_i against the fitted values \hat{y}_i .
 - If there is a linear relationship, the points are symmetrically distributed across the horizontal axis with a constant spread.
 - If you observed a curved pattern, \rightarrow systematic errors.
 - A locally estimated scatterplot smoothing (LOESS) curve, if a specific pattern (in general anything other than a quasi-horizontal line) \rightarrow the relationship is probably non-linear
4. A model that is misspecified will also lead to bad predictions: high error rate on both the training and the test sets.



How can we solve problems of non-linearity?

1. To perform non-linear transformations to the dependent and/or independent variables. The transformation most often used is **a logarithmic transformation** to the dependent variable, when:
 - an exponential growth,
 - or strictly positive and clearly skewed.
 - Independent variables that have a skewed pattern can also be transformed using the natural logarithm. when the effect of the independent on the dependent is **multiplicative**.
2. To **add polynomial or interaction terms** of the predictors to the regression equation. For example, if the residual plot shows a parabolic pattern, this often indicates that you should square one of the predictors. However, you should always be careful with adding too many polynomial or interaction terms, since this can cause your model to overfit.

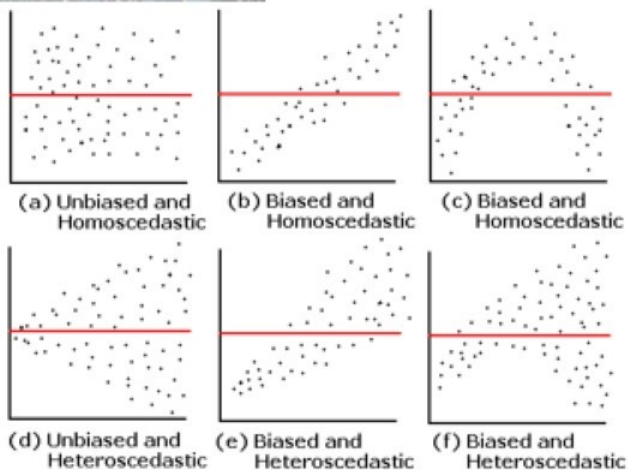
Heteroscedasticity & Normality

- **Heteroscedasticity**
that the **variance of residuals increasing (decreasing) over time**

To spot heteroscedasticity, we can again rely on a residual plot of the residuals against the fitted values. If we notice a funnel shape, this is evidence of heteroscedastic error terms. Often, a logarithmic transformation of the dependent variable can solve this problem.

- **Normality**

However, this assumption may be violated if you are only interested in prediction and not in interpreting the coefficients of your model. In general, you could say that normality is desired, but not required. The normality assumption is mostly violated because the predictors and response are not normality distributed and/or the relationship is non-linear. Solving the non-linear relationship between predictor and response often also solves the normality violation. The solution would be to transform the predictors and/or response using a logarithmic transformation.



Outliers


A classic way to **detect** outliers is to make a boxplot: observations outside the whiskers of the boxplot -- the range of $[Q_1 - 1.5IQR, Q_3 + 1.5 IQR]$, where IQR stands for the interquartile range ($IQR = Q_3 - Q_1$) and Q_1 is the first quartile (or the 25% percentile) and Q_3 the third quartile (or the 75% percentile).

An outlier can impact the regression model in several ways:

1. it can make sure that the predicted value for an observation is far off. This will make sure that your test error increases dramatically and your R-squared decreases. This can be spotted by making a **residual plot** or regression plot and looking at observations that are far off.
2. An outlier can result in violation of the assumptions and hence the functional form of our regression equation is not correct, leading to faulty predictions.

In both cases, you are advised to remove the outliers.

Finally, it could also be that outliers are present, but that they don't impact the performance of your model and don't violate the assumptions. In that case, outliers can remain in the model.



Investigate the possible problems when building a linear regression model on the Boston house prices dataset. You should not focus on all predictor variables, instead regress median house prices (MEDV) on the proportion of lower status of the population (LSTAT) and the average rooms per dwelling (RM).

Consider whether you think the data has the following problems:

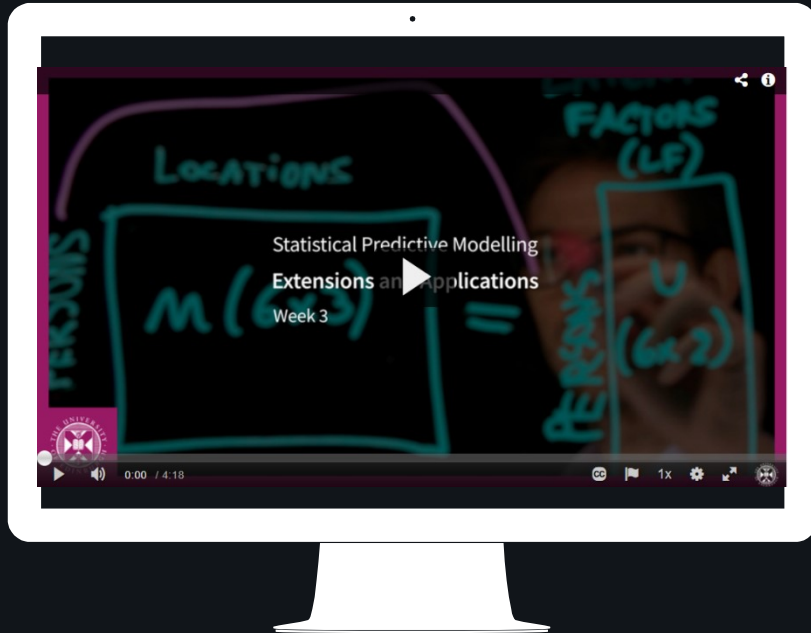
- Non-linearity
- Heteroscedasticity
- Non-normal distributions
- Outliers

If one or more of these problems occurs, how would you solve it?

Activity 3 - Detect linear regression problems.ipynb

Activity: Detect linear regression problems

Extensions and Applications



- Please watch the following video via this [link](#)
- https://media.ed.ac.uk/media/Extensions+and+Applications/1_dzrai648/122596071

A photograph of a modern building with large glass windows, identified as the University of Cambridge Business School.

Try to solve the following exercise:

9 - Activity 4 - Solve problems in Python.ipynb

- You should complete the following three tasks:
 1. Make scatter plots, a residual plot, and a Q-Q plot before building the regression model to detect the problems in the data.
 2. Come up with a strategy to solve these problems. Make sure that you test whether the assumptions are not violated afterwards.
 3. Select the best model and compare the performance of this model with the original one.



UNIVERSITY OF EDINBURGH
Business School