UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Regression – Part II

Simple Linear Regression (SLR)

Simple linear regression: the basics

Parametric methods

- *A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model. No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs.*
- Parametric methods make assumptions about the functional form of f. We have only a fixed finite number of parameters. In statistics, 'parametric' means data are assumed to follow a specified probability distribution associated with finite number of parameters (μ, σ of Gaussian).
- E.g. linear regression, logistic regression, naive Bayes, simple neural networks

Non parametric methods

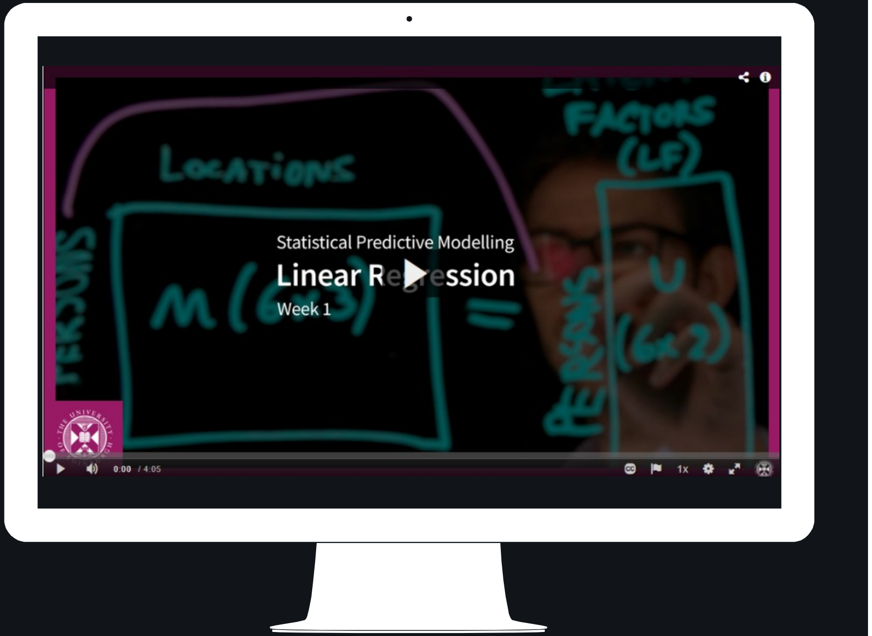
- Non-parametric techniques don't make any assumption about the functional form and the number of parameters is (potentially) infinite.
- Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features.
- E.g. k-Nearest Neighbors, Decision Trees like CART and C4.5

Russell, S. and Norvig, P., 2002. Artificial intelligence: a modern approach.

Simple linear regression: the basics

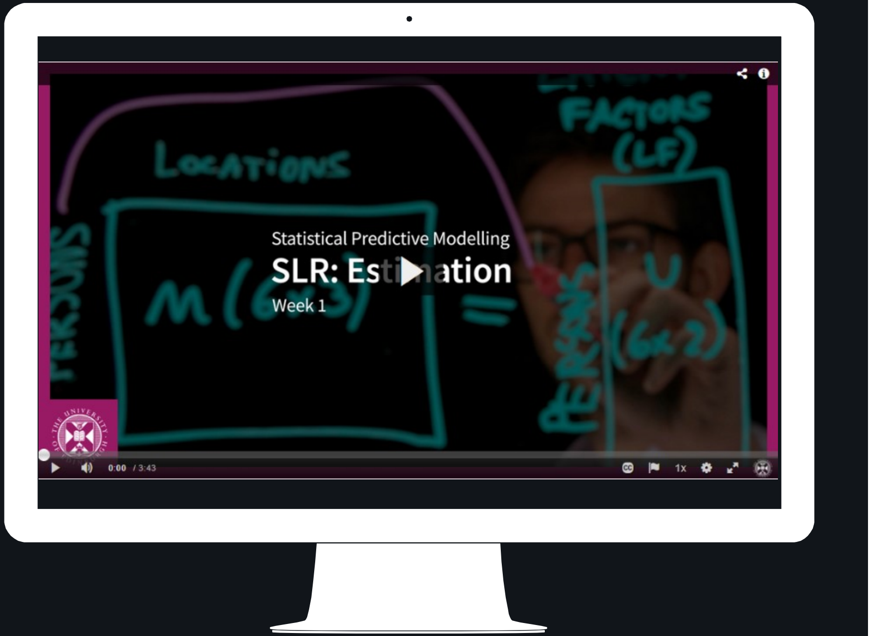
Simple vs multiple linear regression		
	Independent variables	Dependent variable
Simple Linear Regression	One discrete or continuous variable	Continuous
Example	• Visitor count vs advertising spending • Number of friends on Facebook vs number of groups	
Multiple Linear Regression	Two or more discrete or continuous variables	Continuous
Example	• Visitor count using economic growth, advertising spending, and season • Number of friends on Facebook using number of groups and number of photos	

Linear Regression



- Please watch the following video via this [link](https://media.ed.ac.uk/media/Linear+Regression/1_5osz0279/122596071)
- https://media.ed.ac.uk/media/Linear+Regression/1_5osz0279/122596071

SLR: Estimation



- Please watch the following video via this link
- https://media.ed.ac.uk/media/SLRA+Estimation/1_rza6j2wn/122596071

Estimate coefficients: example

- Recall that the formula for the estimate of β_1 is:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The formula for β_0 is:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Please calculate the coefficients for the box office sales example.

1. Determine the mean of x and y
2. Determine the numerator and the denominator
3. Use these values to calculate β_1
4. Use β_1 to calculate β_0

box office sales	
x_i	y_i
29	23
49	12
89	36
110	27
210	45

Estimate coefficients: example

- Recall that the formula for the estimate of β_1 is:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The formula for β_0 is:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Please calculate the coefficients for the box office sales example.

1. Determine the mean of x and y
2. Determine the numerator and the denominator
3. Use these values to calculate β_1
4. Use β_1 to calculate β_0

box office sales	
x_i	y_i
29	23
49	12
89	36
110	27
210	45

A closer look at estimating coefficients

Answer: $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i = 14.179 + 0.1481\, x_i$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

1. Determine the mean of x and y

$$\bar{x} = \frac{29 + 49 + 89 + 110 + 210}{5} = 97.4\,, \bar{y} = \frac{23+12+36+27+46}{5} = 28.6$$

2. Determine the numerator and the denominator

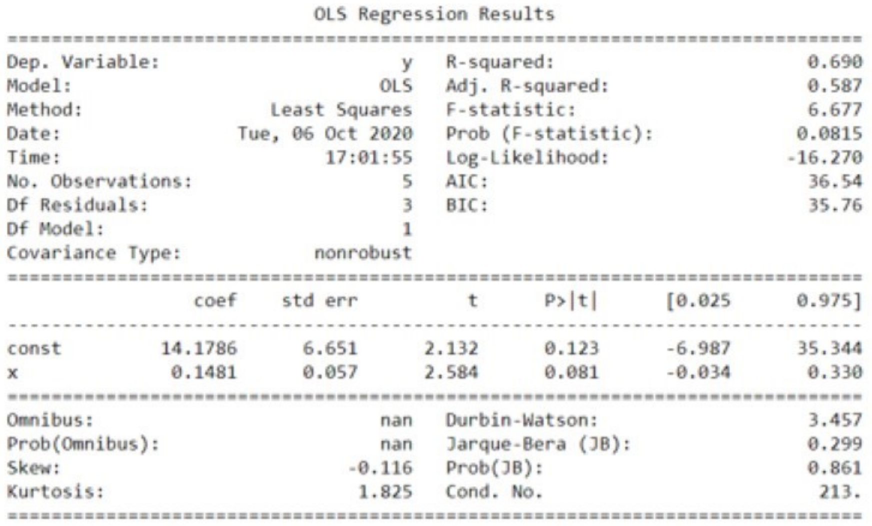
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
29	-68.4	4678.56	23	-5.6	383.04
49	-48.4	2342.56	12	-16.6	803.44
…	…	…	…	…	…

3. Use these values to calculate $\widehat{\beta_1} = \frac{2950.8}{19929.2} = 0.1481$

4. Use $\widehat{\beta_1}$ to calculate $\widehat{\beta_0} = 28.6 - 0.1481 \times 97.4 = 14.179$

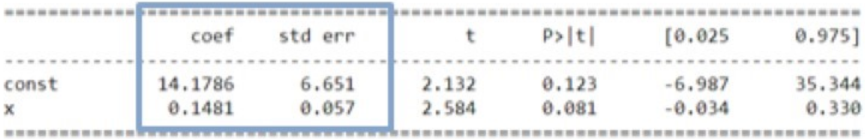
box office sales	
x_i	y_i
29	23
49	12
89	36
110	27
210	45

- Please try the following file about implementing linear regression.
- 1 - How to code simple linear regression.ipynb



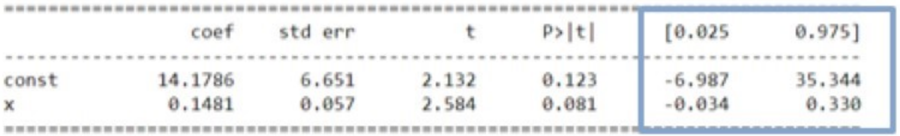
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.690
Model:                            OLS   Adj. R-squared:                  0.587
Method:                 Least Squares   F-statistic:                     6.677
Date:                Tue, 06 Oct 2020   Prob (F-statistic):             0.0815
Time:                        17:01:55   Log-Likelihood:                -16.270
No. Observations:                   5   AIC:                             36.54
Df Residuals:                       3   BIC:                             35.76
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         14.1786      6.651      2.132      0.123      -6.987      35.344
x              0.1481      0.057      2.584      0.081      -0.034       0.330
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   3.457
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.299
Skew:                          -0.116   Prob(JB):                        0.861
Kurtosis:                       1.825   Cond. No.                         213.
==============================================================================
```

Activity: Implementing linear regression in Python

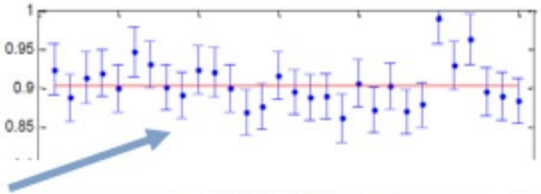


```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         14.1786      6.651      2.132      0.123      -6.987      35.344
x              0.1481      0.057      2.584      0.081      -0.034       0.330
==============================================================================
```

- The **`const`** and **x** represent the intercept and the slope respectively
- The other columns `std err, t, P>|t|` and `[0.025 and 0.975]` tell us something about the accuracy of the coefficient estimates.
- **`std err:`** Remember that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are estimates, so we don't know their true value. If we would calculate our estimates on another set of observations, we would get another regression line. -- How much will the coefficient estimates vary when using different samples?



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         14.1786      6.651      2.132      0.123      -6.987      35.344
x              0.1481      0.057      2.584      0.081      -0.034       0.330
==============================================================================
```



95% Confidence Interval `[0.025 and 0.975]`:

Strictly speaking a 95% confidence interval means that if we were to take 100 different samples (datasets) and compute a 95% confidence interval from each sample, then approximately 95 of the 100 confidence intervals will contain the 'true' value. In practice, however, we select one random sample and generate one confidence interval, which may or may not contain the true value.

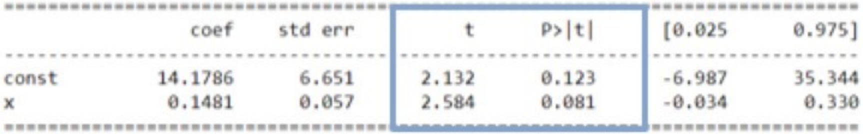
Repeated on numerous different samples of the same size, the fraction of calculated confidence intervals encompass the true value would tend toward 95%.
From a pre-experiment point of view: **there is a 95% probability that the calculated confidence interval in some future experiments will contain the true value.**

BE CAREFUL:

The confidence interval does NOT reflect the variability in the unknown parameter.

This does NOT mean that for one sample there is 95% probability that the true value will lie within this interval.

It also does NOT mean that 95% of the sample data lies within this interval.

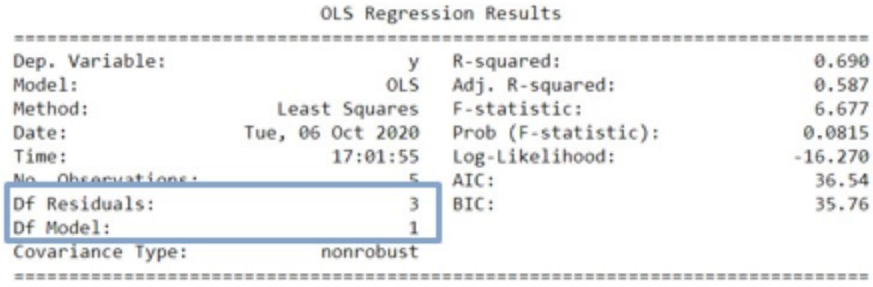


```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         14.1786      6.651      2.132      0.123      -6.987      35.344
x              0.1481      0.057      2.584      0.081      -0.034       0.330
==============================================================================
```

`t and P>|t|:`

They tell us something about whether or not the relationship between the predictor and the response is significant. -- a hypothesis test

- The null hypothesis $\mathrm{H}_0\colon \beta_1 = 0$
- The alternative hypothesis $\mathrm{H}_1\colon \beta_1 \neq 0$
  To test the null hypothesis, we need to determine if our estimate $\widehat{\beta_1}$ is sufficiently far from zero that we can be confident that β_1 is non-zero. To test this hypothesis, linear regression performs a t-test and the outputs of this test are the t-statistic (**`t`**) and the p-value (`P>|t|`).
- **If the t-statistic is very large,** the alternative hypothesis is likely to be true. A basic rule to **reject the null hypothesis** in favour of the alternative hypothesis is when the –**p-value is smaller** than 0.05.
  This would mean that the probability of observing such an extreme value for β_1 is 5%, given that the null hypothesis is true. Hence, it is very unlikely that the null hypothesis is true and we can say that the predictor has a significant influence on the 5% significance level.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.690
Model:                            OLS   Adj. R-squared:                  0.587
Method:                 Least Squares   F-statistic:                     6.677
Date:                Tue, 06 Oct 2020   Prob (F-statistic):             0.0815
Time:                        17:01:55   Log-Likelihood:                -16.270
No. Observations:                   5   AIC:                             36.54
Df Residuals:                       3   BIC:                             35.76
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
```

- The **`Df Model`** represents the degrees of freedom of the model. In this case, we have 1 degree of freedom, since we have 1 predictor.
- The degrees of freedom of the residuals (**`Df Residuals`**) is equal to 3, since we have 5 observations but have to estimate the intercept and the slope (so 5-2=3).

The **`R-squared (`R^2`)`** is a measure of goodness-of-fit and shows us the explanatory power of our model. The R^2 indicates **how much variance in the response variable (y) is explained by the regressed model**. In other words, the R^2 shows how much of the variability in y is explained by the predictor x.

$$\mathrm{R}^2 = \frac{TSS - RSS}{RSS} = 1 - \frac{RSS}{TSS}$$

RSS represents the **residual sum of squares**, which can be defined as the variability that is unexplained by the model.

TSS stands for the **total sum of squares** and is defined as: $TSS = \sum(y_i - \bar{y})^2$. It represents the amount of variability in the response before the model is built.

- In simple linear regression, the R^2 can also be computed as the squared correlation between predictor and response.
- The R^2 is a number between 0 and 1. The closer to 0 (1) means that the model does not explain (explains) a lot of the variability in the response.
- A low value for the R2 might also be an indication that the relationship is not linear.

Quiz

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

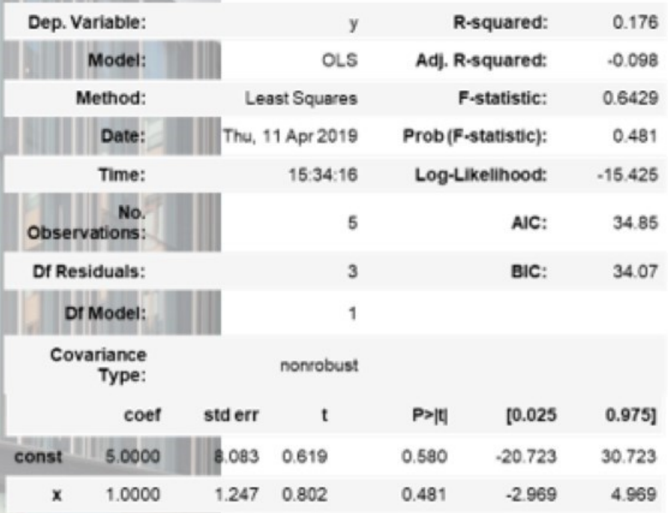
- A. This is a supervised learning problem, since you have an outcome variable (house prices)
- B. If we regressed house prices on the size, we would have a continuous response variable and a discrete predictor.
- C. The relationship between house prices and size will be approximated by a straight line.
- 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



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.

- Please read and study the following file
- 2 - Making predictions with simple linear regression.ipynb

- Then try the following exercise:
- 3 - Activity 1 - Build a simple linear regression model.ipynb + social_media.csv

Activity: Applying Simple Linear Regression Model

UNIVERSITY OF EDINBURGH
Business School