Lecture 5

Regression Analysis

Ashish Dandekar

Lecture Overview

Linear regression

Analysing OLS regression

- Validating the assumptions
- Analysing Multiple Linear Regression

Topics in regression

- Polynomial Regression
- Regularisation
- Difference-in-differences analysis

Linear Regression

Revisiting Regression

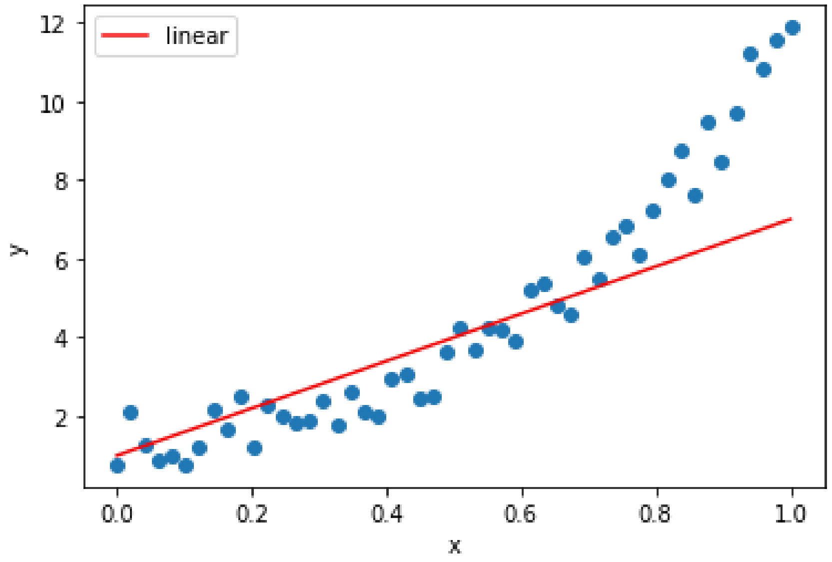
Equation of a line

$$y = a + bx$$

where a is the intercept and b is the slope.

In higher dimensions where $x \in \mathrm{R}^d$,

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_d x_d$$



Note!

- x is called as the **predictor**, explanatory, independent or exogenous variable.
- y is called as the **response**, outcome or dependent or endogenous variable.

Types of Regression

Given a labeled dataset $\mathsf{D} = \{(x_i, y_i)\}$ of n points where $x_i \in \mathsf{R}^d$, $y_i \in \mathsf{R}$.

Ordinary Least Squares (OLS)

$$\hat{b} = \arg\min_{b \in \mathsf{R}^{d+1}} (Y - Xb)^T (Y - Xb)$$

Weighted Least Squares (WLS)

$$\hat{b} = \arg\min_{b \in \mathsf{R}^{d+1}} (Y - Xb)^T W (Y - Xb)$$

Weights can be used to balance outliers in the data!

Types of Regression

Simple Linear Regression

Given a labeled dataset $D = \{(x_i, y_i)\}$ of n points where $x_i \in R$, $y_i \in R$.

$$y = a + bx$$

Multiple Linear Regression

Given a labeled dataset $D = \{(x_i, y_i)\}$ of n points where $x_i \in R^d$, $y_i \in R$.

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_d x_d$$

Multivariate Linear Regression

Given a labeled dataset $D = \{(x_i, y_i)\}$ of n points where $x_i \in R^d$, $y_i \in R^k$.

$$y_{ij} = b_{0j} + \sum_{l=1}^{k} b_{lj} x_{ik} + \epsilon_{ik}$$

Simple Linear Regression

Let's focus on OLS simple linear regression: $y = a + bx$.

Intercept (a)

- It is an estimate of the response when all inputs are zero.
- It equals to average value of the response.

Slope (b)

- It is the estimate of the change in the response per unit change in the predictor.

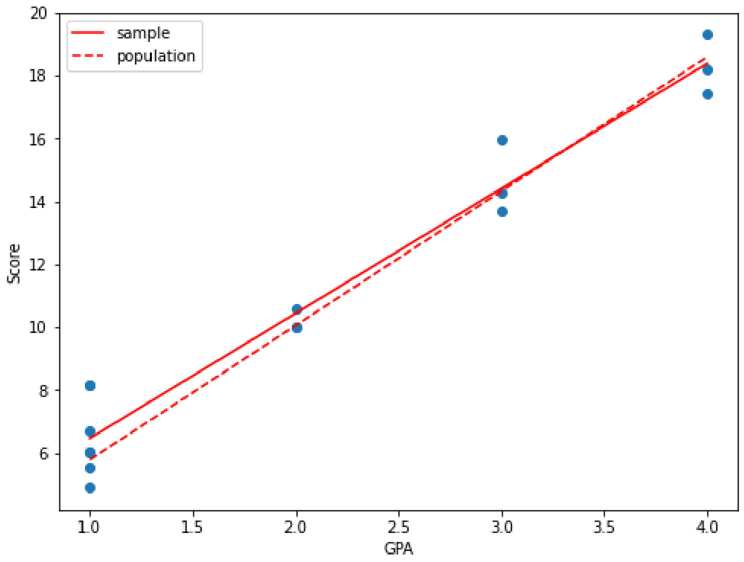
EXAMPLE

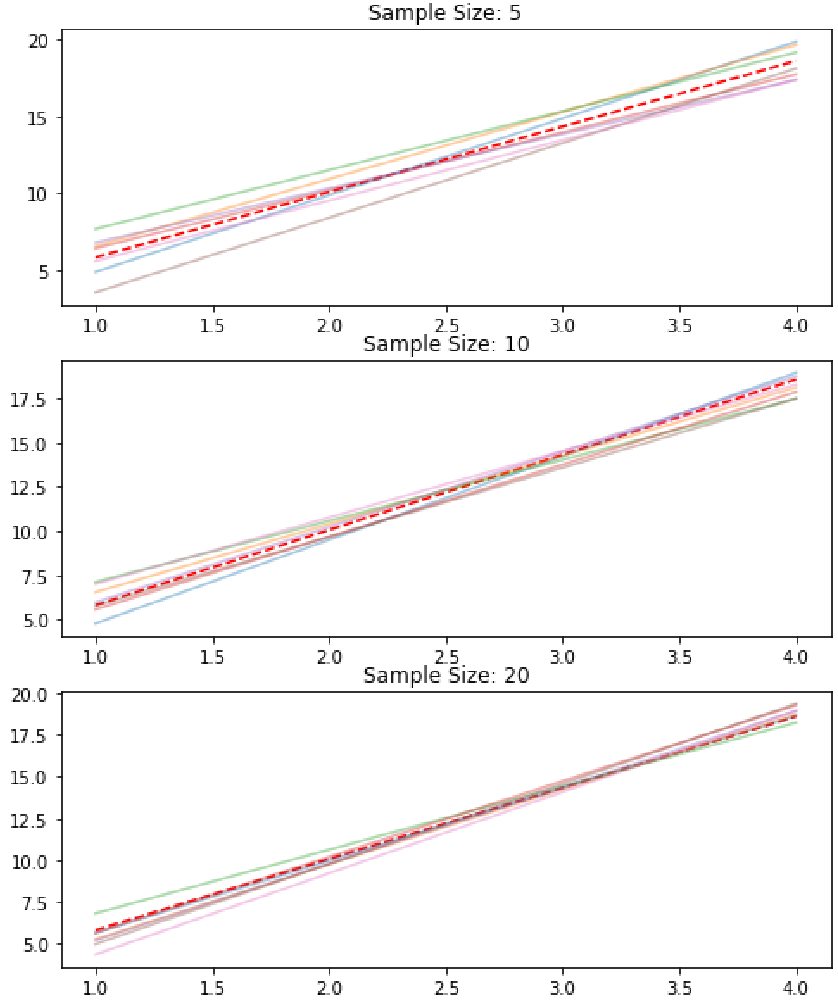
$$GPA = 0.5 + 1.1 \cdot Score$$

$$GPA = 0.5 + 0.85 \cdot Gender$$

$$GPA = 0.5 + 1.1 \cdot Score + 0.85 \cdot Gender$$

Population Versus Sample

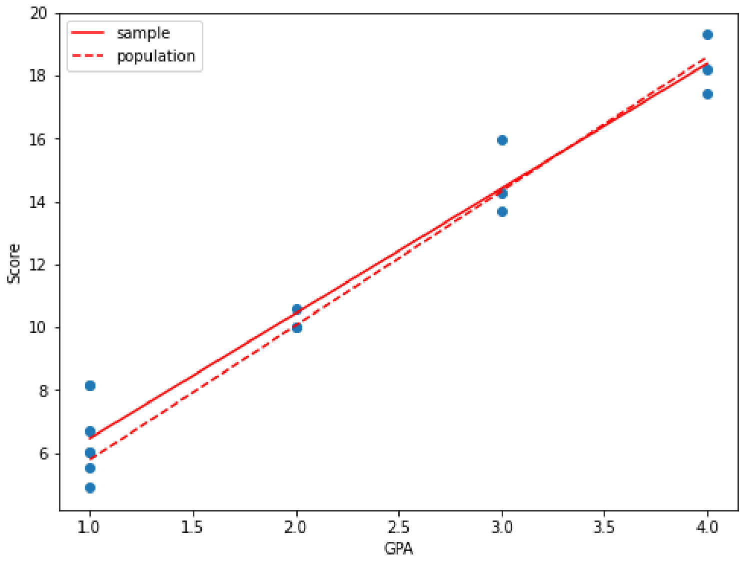




What do we actually estimate?

$$\hat{b_0} + \hat{b_1}x \rightarrow \hat{y}$$

Noisy Data



What do we actually estimate?

$$\hat{b}_0 + \hat{b}_1 x \rightarrow \hat{y}$$

Noisy data model

- The population line can be modelled as:

$$\mu_Y = E[Y] = b_0^* + b_1^* x$$

- The individual response can be modelled as:

$$y_i = b_0^* + b_1^* x + \epsilon_i$$

where, $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ is called as *white noise* in the data.

Analysing OLS Regression

Typical Result of OLS Regression


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.897
Method:                 Least Squares   F-statistic:                     131.6
Date:                Wed, 13 Jan 2021   Prob (F-statistic):           1.66e-08
Time:                        21:53:47   Log-Likelihood:                -29.553
No. Observations:                  16   AIC:                             63.11
Df Residuals:                      14   BIC:                             64.65
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.4891      1.082    -10.616      0.000     -13.810      -9.168
x1             4.7135      0.411     11.472      0.000       3.832       5.595
==============================================================================
Omnibus:                        1.741   Durbin-Watson:                   1.323
Prob(Omnibus):                  0.419   Jarque-Bera (JB):                1.369
Skew:                          -0.567   Prob(JB):                        0.504
Kurtosis:                       2.124   Cond. No.                         7.83
==============================================================================
```

Significant predictors

What do we actually estimate?

We estimate the coefficients of the regression line on a specified sample of the dataset. Thus, the randomness of the sampling procedure makes the coefficient as random variables!

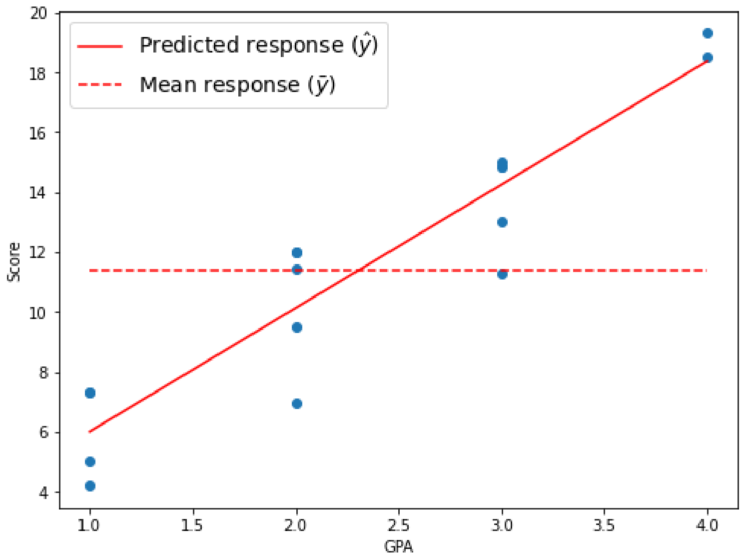
$$\hat{b_0} + \hat{b_1} x \rightarrow \hat{y}$$

t-test for the coefficients

The OLS regression result shows the result of t-test with the null hypothesis $b_i = 0$.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-11.4891	1.082	-10.616	0.000	-13.810	-9.168
x1	4.7135	0.411	11.472	0.000	3.832	5.595

Different kinds of errors



$$e_i = y_i - \hat{y}_i$$

$$e_i \leftarrow \text{Residual}$$

$$y_i \leftarrow \text{Actual response}$$

$$\hat{y}_i \leftarrow \text{Predicted response}$$

Residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Explained sum of squares (ESS)

$$ESS = \sum_{i=1}^{n} (\bar{y}_i - \hat{y}_i)^2$$

Total sum of squares (TSS)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Coefficient of Determination

Result
p-values
Errors
❯ R-squared
Validity
Multiple LR

Coefficient of Determination

模型对因变量的解释能力

- It is also known as R^2 value and it is defined as follows:

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

- It quantifies the fraction of total variability in the response that is explained by the model.
- It typically lies between 0 *(bad)* and 1 *(good)*.

Past Exam MCQ

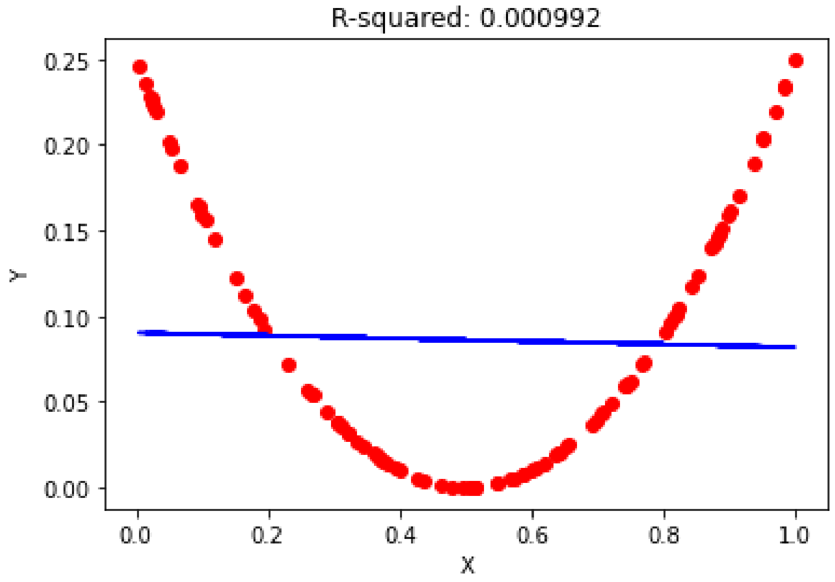
The R-squared on training dataset is 0.92 whereas on the test dataset it is -0.2. You are shocked looking at the negative value. Which of the following is a valid inference based on the observation?

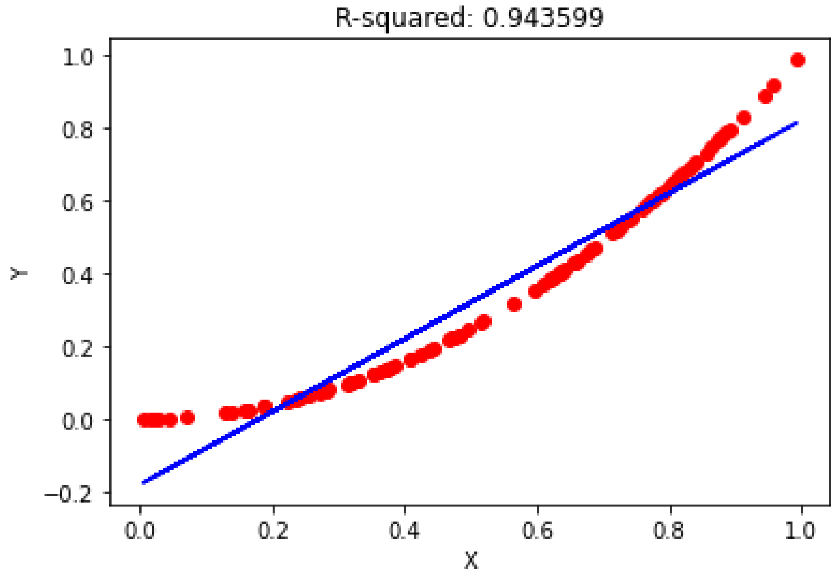
1. The training data does not truly represent the population.
2. There is a bug in the implementation since R^2 can't be negative.
3. The training dataset does not follow the linearity assumption of the regression model.
4. This is an evidence of multi-collinearity.

Coefficient of Determination

Result
p-values
Errors
❯ R-squared
Validity
Multiple LR

R^2 quantifies the strength of linear relationship.





R squared的值很好

但是fit实际不好：
这是线性的线，而R^2衡量的是线性关系

Revisiting the assumptions

Linearity

Predictor and response are linearly related to each other.

Normality

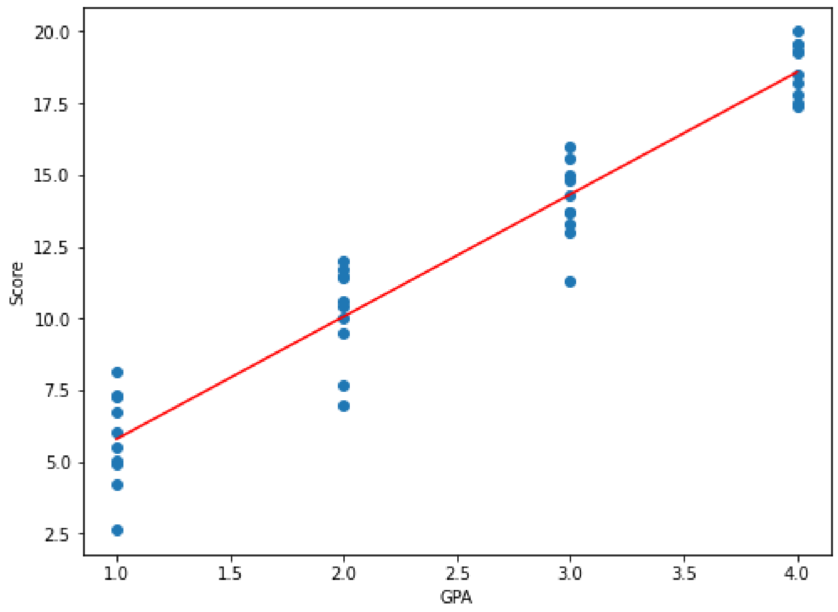
For a given value of predictor, the residuals are normally distributed.

Homoscadasticity

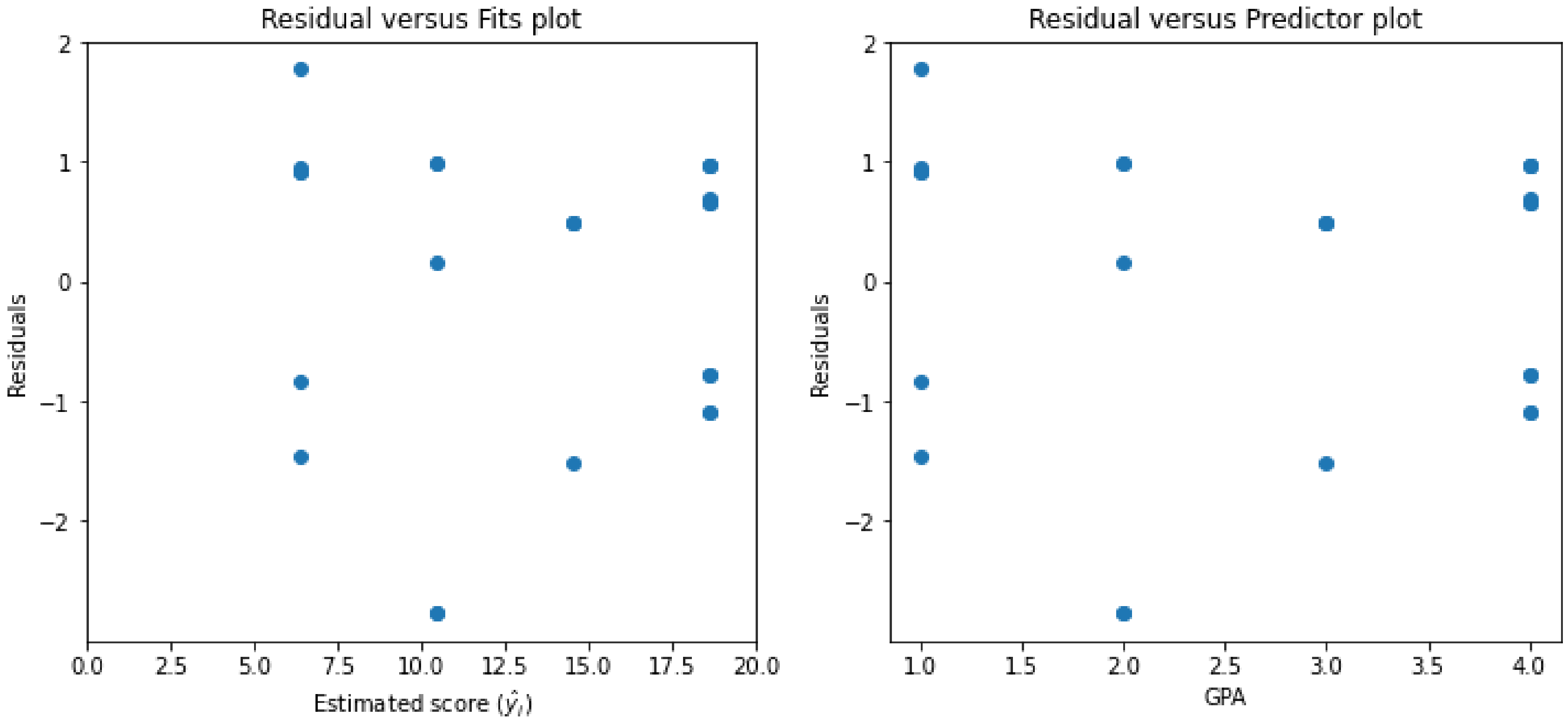
For a given value of predictor, the residuals have a constant and equal variance.

Independence

For a given value of predictor, the residuals are independent of each other.

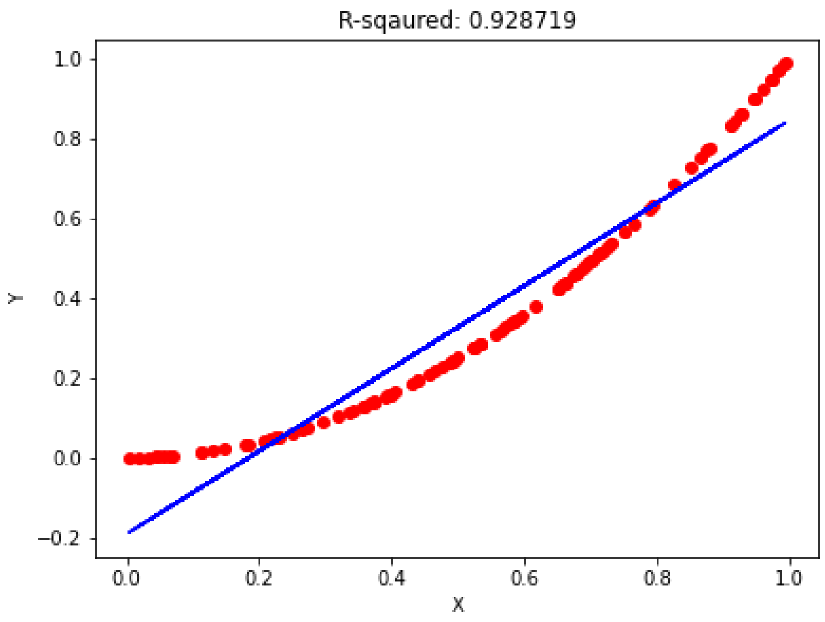


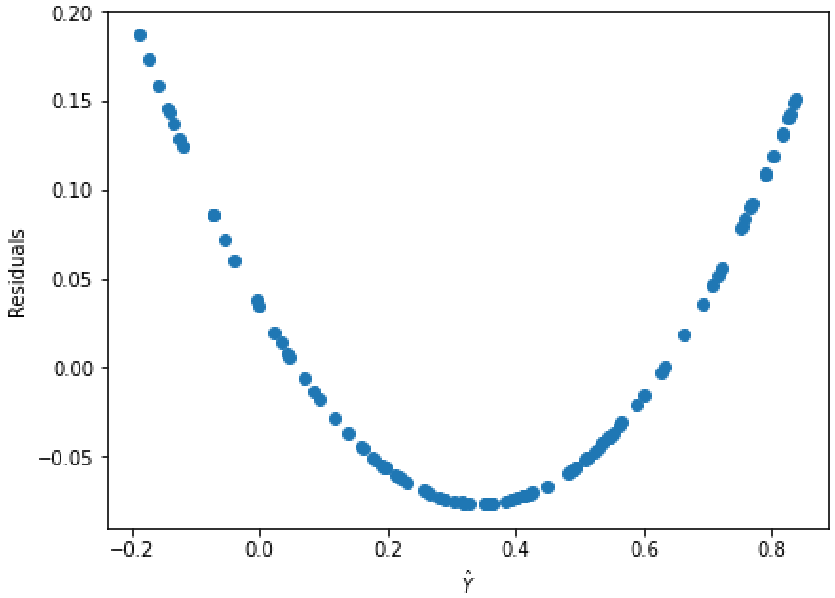
Residual Plot



The plot should comprise of randomly scattered points.

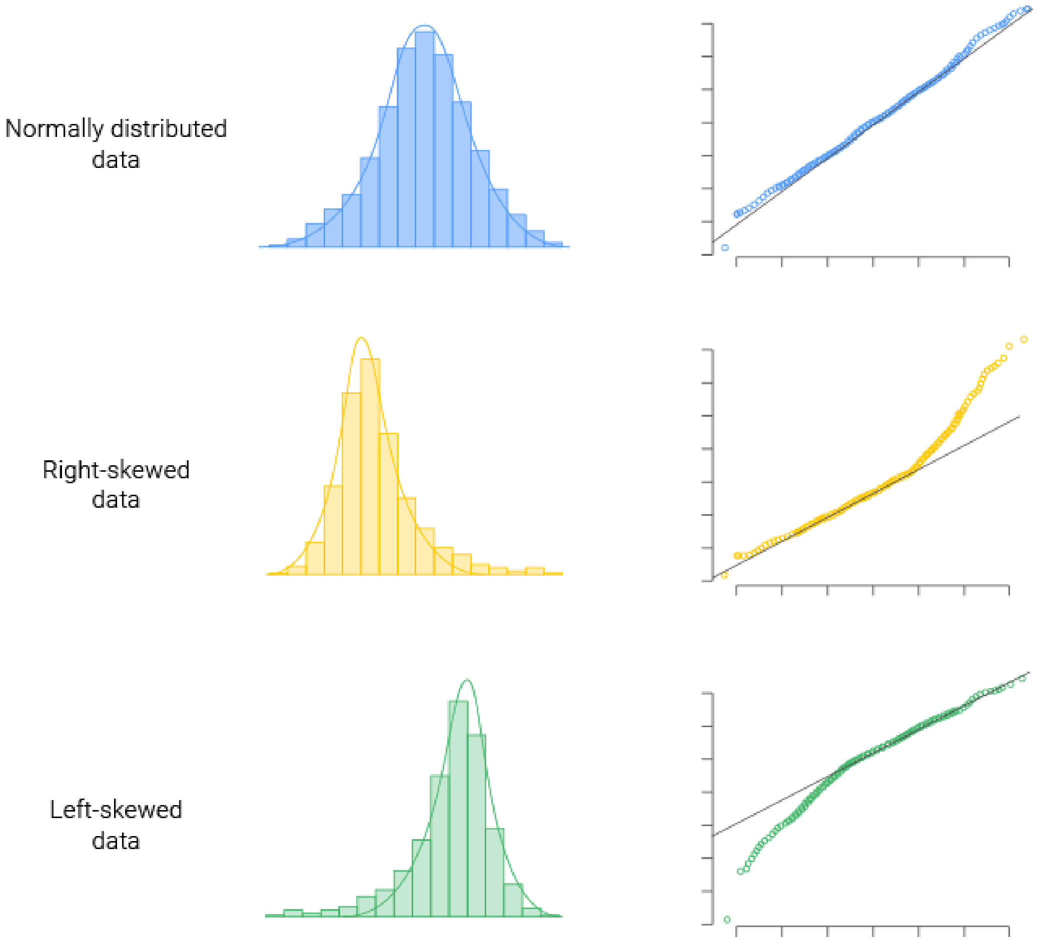
Assessing Linearity





Any pattern in the residual plot is an indicator of non-linear relationship between the predictor and response.

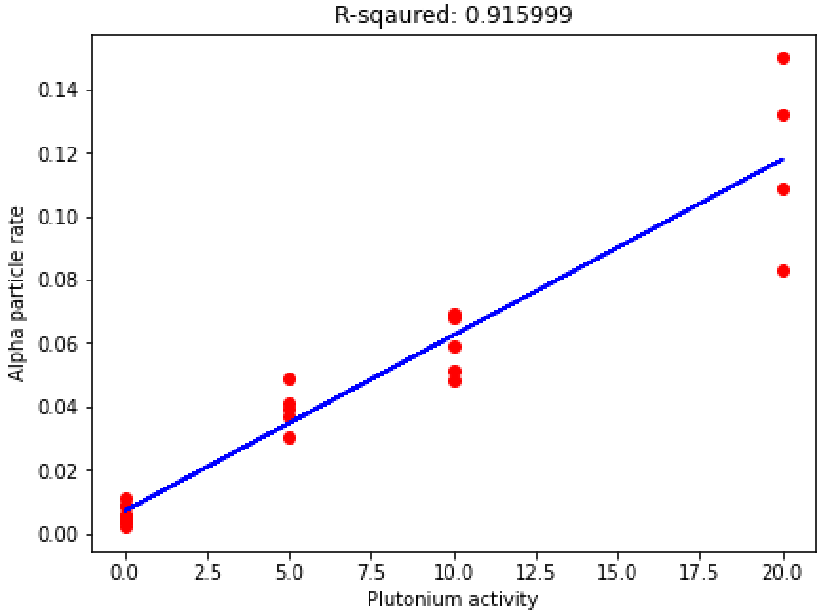
Assessing Normality

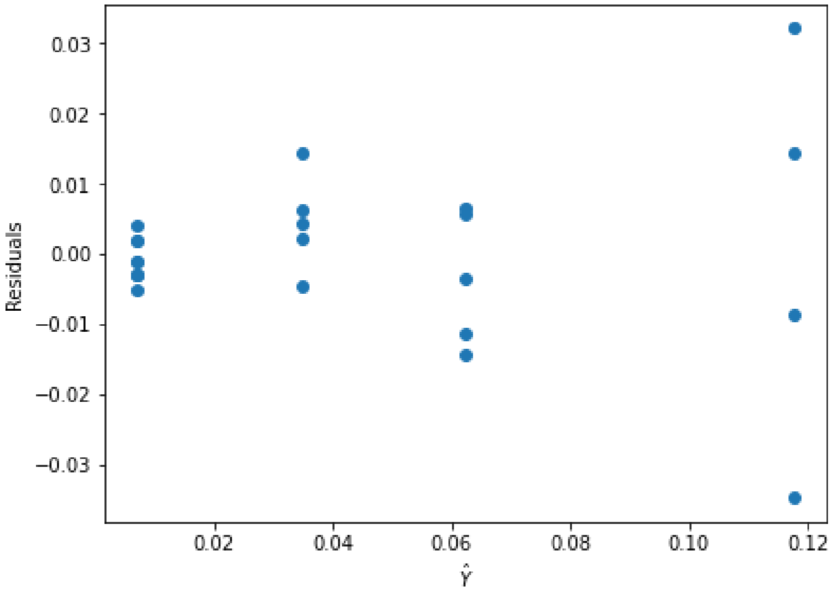


QQ Plots

Theoretical quantiles versus Observed Quantiles

Assessing Homoscadasticity





If the residuals are more spread out for some predicted values and less for others, it suggests violation of homoscadsticity assumption. The data is said to be show the evidence heteroscadasticity.

Adjusted R^2

Do you remember Multiple Linear Regression?

Given a labeled dataset $\mathsf{D} = \{(x_i, y_i)\}$ of n points where $x_i \in \mathsf{R}^d$, $y_i \in \mathsf{R}$.

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_dx_d$$

Adjusted R^2

Addition of predictors to the regression provides more information for training. Since OLS regression minimises residual error, it tends to improves the R^2.

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{n-1}{n-d-1}$$

< R^2

若R^2与adjusted R^2 相差很大，
那么新加的信息（列）不能很好的predict.
也就没有contribute a lot

加更多的predictors --> add more info
--> 可能过拟合 --> R^2 更大

Adjusted R^2

检验模型整体，衡量全部 xi 是否有用

F statistical test is used to assess the overall significance of the multiple linear regression. It works on the following null hypothesis:

$$\mathrm{H}_0 : b_1 = b_2 = b_3 = \ldots = b_d = 0$$

但不知道具体是哪一个

If the p-value is less than 0.05, then we reject H_0, which means the there is at least one predictor which is useful to explain the response.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.897
Method:                 Least Squares   F-statistic:                     131.6
Date:                Wed, 13 Jan 2021   Prob (F-statistic):           1.66e-08
Time:                        21:53:47   Log-Likelihood:                -29.553
No. Observations:                  16   AIC:                             63.11
Df Residuals:                      14   BIC:                             64.65
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.4891      1.082    -10.616      0.000     -13.810      -9.168
x1             4.7135      0.411     11.472      0.000       3.832       5.595
==============================================================================
Omnibus:                        1.741   Durbin-Watson:                   1.323
Prob(Omnibus):                  0.419   Jarque-Bera (JB):                1.369
Skew:                          -0.567   Prob(JB):                        0.504
Kurtosis:                       2.124   Cond. No.                         7.83
==============================================================================
```

F统计量的p值

Is this test even useful?

t-tests for individual predictor only test the efficacy of that predictor in isolation.

It is useful if you are working with very high dimensional data. You can use it to eliminate some of the features.

Multicollinearity

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other.

For examples: including both left and the right shoe size in the analysis!

How to detect?

- Correlation matrix (part of EDA) 若correlation很高，就移除一些项
- Unstable coefficients for correlated features
- t-test for the coefficients is insignificant but F-test remains significant.
- Variance Inflation Factor (VIF)

Multicollinearity

How harmful is multicollinearity?

Multicollinearity may not significantly affect the quantitative predictive performance of the model; but may severely affect the qualitative interpretation fo the model.

For instance: Consider that a certain y is related to x_1 as $2x_1$. Let x_2 be another another features with a perfect correlation with x_1. If we train a multiple regression with both x_1 and x_2, then it may lead to multiple answers such as:

- $y = x_1 + x_2$
- $y = 0.8x_1 + 0.2x_2$
- $y = 2x_2$
- ...

Every model will have a different physical interpretation!

Topics in Linear Regression

Polynomial Regression

If the linearity assumptions is violated by the data, one can transform the predictors to include their higher order terms and approach it as a *multiple linear regression*.

Degree	Forms
2	$y = b_0 + b_1 x_1 + b_2 x_1^2$
2	$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2$

Structural multicollinearity

The introduction of higher-order terms (and feature transformation) introduces multi-collinearity. In order to subdue the effect, it is recommended to **center** the feature by subtracting the mean from the actual values.

How to solve it by Python?

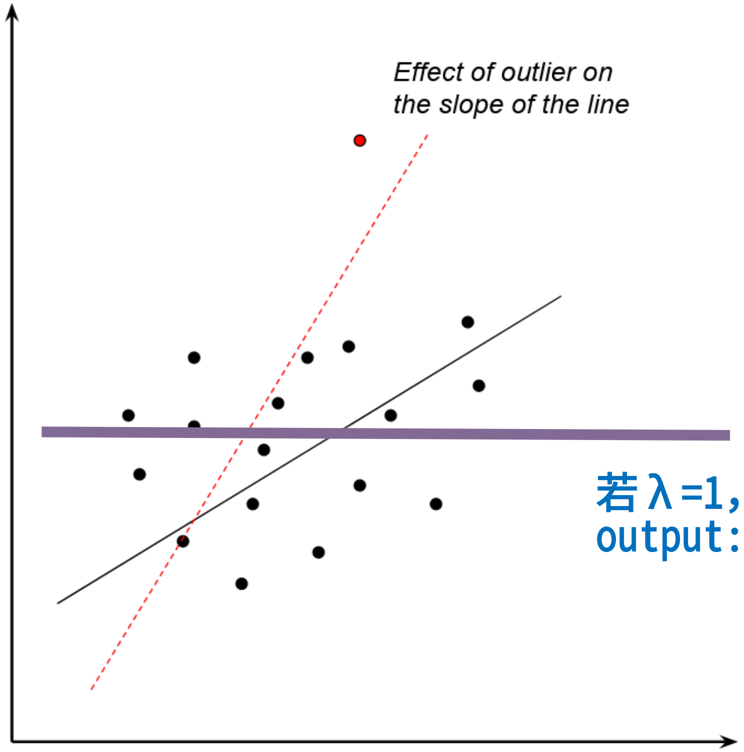
```
from sklearn.preprocessing \\
import PolynomialFeatures

pf = PolynomialFeatures(degree = 2)
pf.fit_transform(df)
```

Regularised Regression

Regularised regression adds **regularisation term** to the loss function of OLS regression.

仅依赖b(斜率)

$$\hat{b} = \underset{b \in \mathrm{R}^{d+1}}{\arg\min}\ \ell_D(b) + \lambda \mathrm{R}(b), \ \ 0 \leq \lambda \leq 1$$



若λ=1，线会变成紫线
output: avg

Ridge regression

L2范数

$$\mathrm{R}(b) = \sum_i b_i^2$$

若有离群点，使用这个

Used to minimise the impact of outliers.

LASSO regression

L1范数

$$\mathrm{R}(b) = \sum_i |\ b_i\ |$$

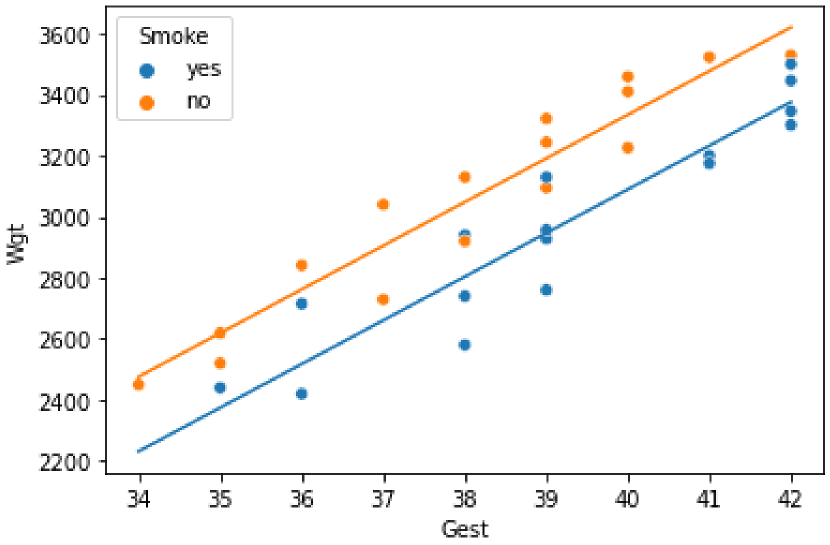
Used to perform feature selection.

一些特征会=0，特征选择

Additive Effect

Two non-interacting variables are said to have an *additive effect* on the response. In such a case, the regression takes the following form:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 \text{ ...}(b_3 \text{ is statistically insignificant})$$



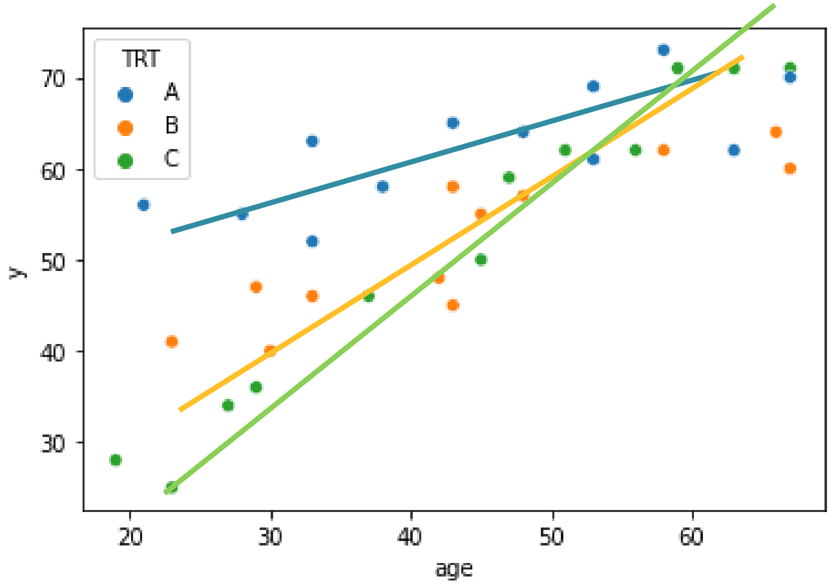
We can uniquely determine the effect of gestation period *(Gest)* on the weight of the baby *(Wgt)* if we know the smaoking habit *(Smoke)* of the mother.

Interaction Effects

An interaction effect occurs when the effect of one predictor on the response is not constant across different values or levels of another predictor. In such a case, the regression takes the following form:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 \text{ ...(} b_3 \text{ is statistically significant)}$$

cross lines



We can independently determine the effect of age *(age)* on the efficacy of the vaccine *(y)* if we know the treatment *(TRT)* received by the subject.

Difference-in-Differences Analysis

Difference-in-differences (DID) analysis is a statistical method used to evaluate the causal impact of an intervention or treatment.

What do we need?

- A group that received the treatement *a.k.a.* Treatment Group
- A group that did not receive the treatement *a.k.a.* Control Group
- Data about these two groups before and after intervention.

Difference-in-Differences Analysis

The linear regression is formulated as follows:

$$y_i = b_0 + b_1 T_i + b_2 C_i + b_3 (T_i C_i) + \epsilon_i$$

- y_i is the response.
- T_i is the dummy variable for the time period before (0) /after (1).
- C_i is the dummy variable for the group control (0) / treatment (1).

	Before $(T = 0)$	After $(T = 1)$
Control $(C = 0)$	$y_i = b_0 + \epsilon_i$	$y_i = b_0 + b_1 + \epsilon_i$
Treatment $(C = 1)$	$y_i = b_0 + b_2 + \epsilon_i$	$y_i = b_0 + b_1 + b_2 + b_3 + \epsilon_i$

Difference-in-Differences Analysis

	Before ($T = 0$)	After ($T = 1$)
Control ($C = 0$)	$y_i = b_0 + \epsilon_i$	$y_i = b_0 + b_1 + \epsilon_i$
Treatment ($C = 1$)	$y_i = b_0 + b_2 + \epsilon_i$	$y_i = b_0 + b_1 + b_2 + b_3 + \epsilon_i$

The difference for the treatment group

$$E[Y \mid T = 1, C = 1] - E[Y \mid T = 0, C = 1] = b_1 + b_3$$

The difference for the control group

$$E[Y \mid T = 1, C = 0] - E[Y \mid T = 0, C = 0] = b_1$$

Difference in differences

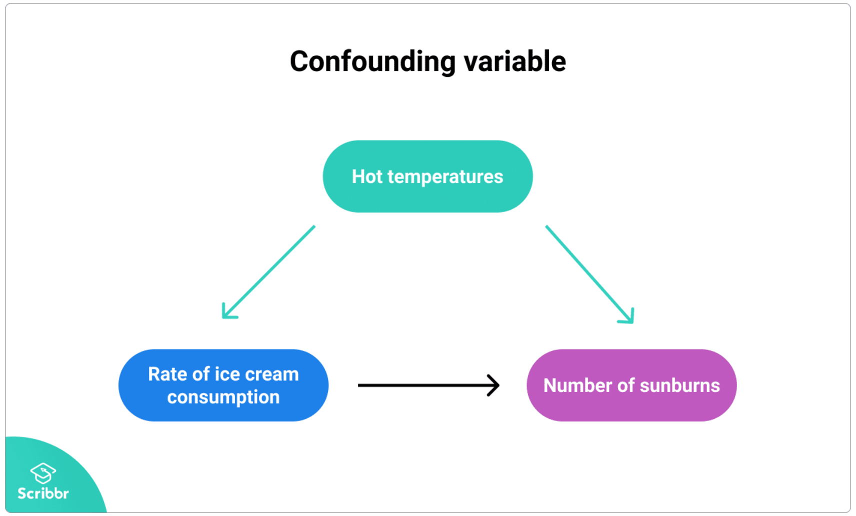
$$E[DID] = b_3$$

If b_3 is statistically significant, the treatment is said to be effective!

Confounder's Bias

Confounding Variable

It is the variable that is not accounted for in the experiment, but affect both the predictor and the response.



Things to remember

- Thorough study of the domain
- Restrict the sample.
- Include confounding variables as controls in the study.

Summary

Recipe of the linear regression

- Keep aside $20 - 30\%$ of data for the model valiation purpose.
- Obtain summary statistis of various predictors.
- Obtain scatterplots and correlation among various features.
- Preprocess categorical variables.
- Perform feature selection.

Recipe of the linear regression

- Run linear regression. Check R^2, adjusted R^2 and F-statistic. If they are unusual
  - Assess the statistical significance of the predictors.
  - Assess the possibility of multicollinearity.
- Obtain the residual plots.
  - Check linearity assumption.
  - Check normality assumption.
  - Check for heteroscadsticity.

Recipe of the linear regression

- If all steps are successful, perform model validation.
- If model validation is unsuccessful, re-start from the second step!

Thank you!

Feel free to reach out to me at

dcsashi (at) nus (dot) edu (dot) sg