Linear
&
non-linear
modeling

Objectives

- Simple linear regression: ordinary least squares, coefficient estimate.
- Residual modelisation and analysis.
- Transform target variable.
- Multiple Regression and normal equation.
- Dummy variables, variables interaction and polynomial regression.
- Inference in Regression: statistical test and confidence interval for coefs and predictions.
- Metrics used to evaluate the models performance: R2, R2 adjusted, AIC, BIC.
- Methods used for variable selection.
- How to deal with collinearity and dataset with big number of variables.
- Logistic regression and GLM.

Review notations

$\overline{X} = \frac{1}{n} \sum_{1}^{n} X_i$ observed means for random variables ***X***

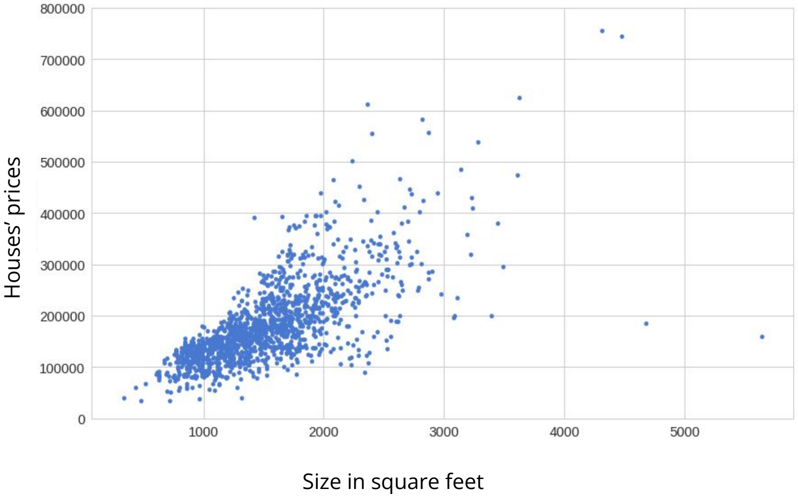
$S^2 = \frac{1}{n-1} \sum_{1}^{n} \left(X_i - \overline{X} \right)^2$ is the empirical variance of ***X***

Empirical standard deviation is defined as $S = \sqrt{S^2}$

Normalizing the data/random variable is defined as $Z_i = \frac{X_i - \overline{X}}{S}$

Problem

Let’s say you have a data set that contains the prices of houses sold in your city the last year. For each house of your that set, you also have the size of the house in square feet.



Your objective is to find a relationship between house price and it’s size. Or more precisely, predict house price based on house size in square feet.

What is linear regression model

Linear regression is a prediction model that establish a linear relationship between a target variable and a set of explanatory variables.

The case of one explanatory variable is called simple linear regression.

$$\widehat{Y} = aX + b$$

The **Gauss–Markov theorem** states that ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the **errors** in the linear regression model are **uncorrelated**, have **equal variances** and **expectation value of zero.**

$$Y = aX + b + \varepsilon$$

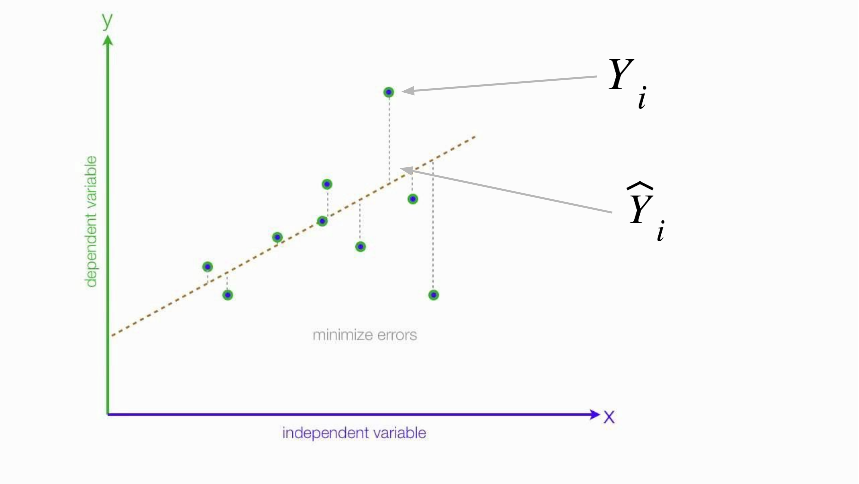
Simple linear regression

Objective function

We need to minimize the quantity :

$$\sum_{1}^{n} (Y - \widehat{Y})^2 = \sum_{1}^{n} (Y - (aX + b))^2$$

we find that :



$$\widehat{a} = \frac{Cov(X, Y)}{Var(X)}$$

$$\widehat{b} = \bar{y} - \widehat{a}.\bar{x}$$

$$\widehat{a} = \frac{\sum_{1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i}^{n} (X_i - \overline{X})^2}$$

Covariance and Correlation

Covariance signifies the direction of the linear relationship between the two variables

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

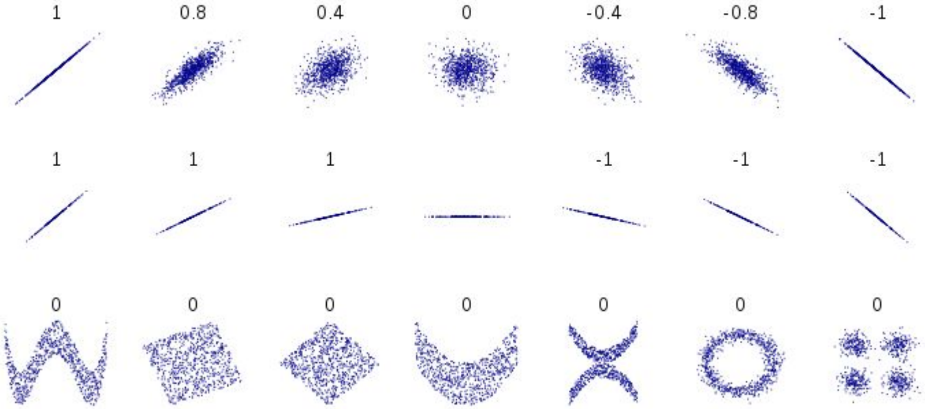
$$Cov(X,X) = VAR(X)$$

Empirical covariance is defined as

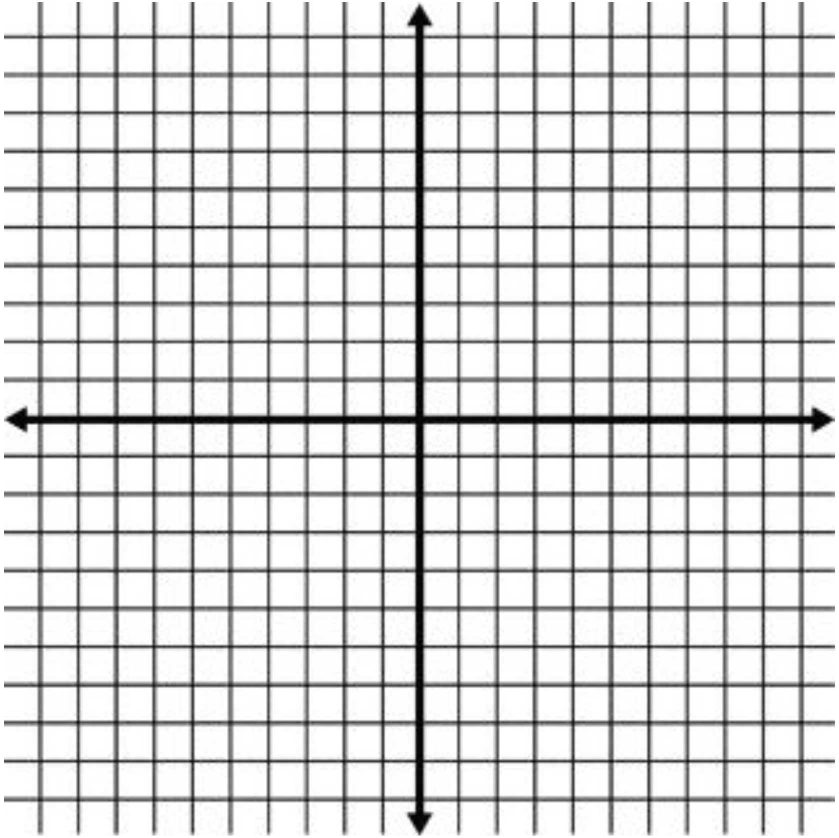
$$Cov(X,Y) = \frac{1}{n-1}\sum_{1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

Correlation refers to the scaled form of covariance.

$$Cor(X,Y) = \frac{Cov(X,Y)}{S_x S_y}$$



Example of OLS



Errors and Residuals

A statistical error (or disturbance) is the amount by which an observation differs from its expected value:

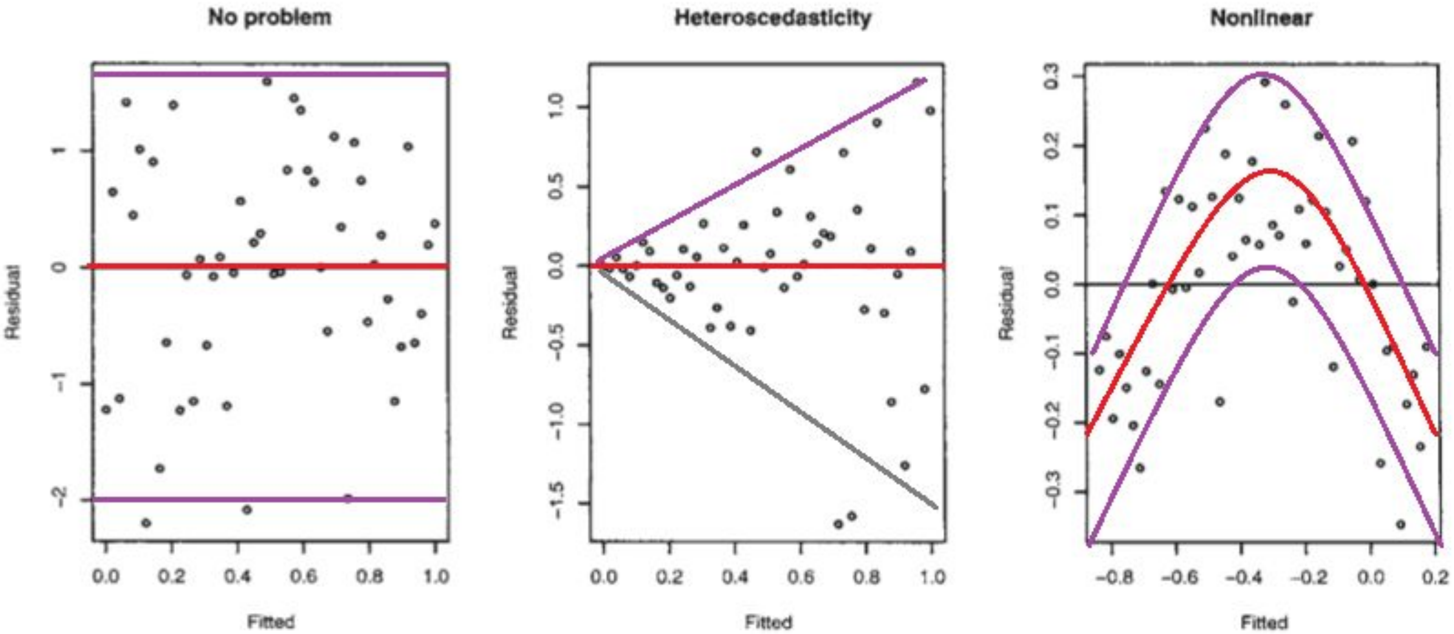
$$Y_i = aX_i + b + \varepsilon_i \quad \longrightarrow \quad \varepsilon_i = Y_i - \left(aX_i + b \right)$$

Residual is observable estimate of the unobservable statistical error :

$$r_i = \widehat{\varepsilon}_i = Y_i - \left(\widehat{a}X_i + \widehat{b} \right)$$

Residuals analysis

Scatterplot



Residuals analysis

Scatterplot

- A plot of the residuals against the fitted values should show no pattern. If a pattern is observed, there may be “heteroscedasticity” in the errors.

- That is, the variance of the residuals may not be constant. To overcome this problem, a transformation of the target variable (such as a logarithm or square root) may be required.

- Do a scatterplot of the residuals against predictor. If these scatterplots show a pattern, then the relationship may be nonlinear and the model will need to be modified accordingly.

Studentized residual

Under the assumption of **normality of the errors**, we can state that:

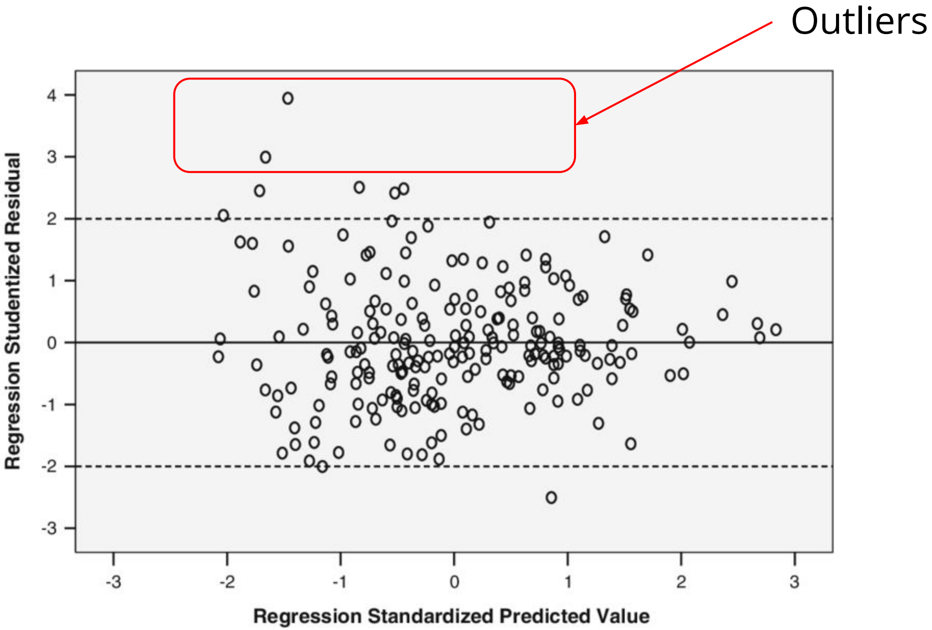
$$t_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

is a **Student's t-distribution with *n* − 2**

$$\widehat{\sigma} = \sqrt{\frac{1}{n-2}\sum_1^n \widehat{\varepsilon}_j^{\,2}}$$

$$h_{ii} = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_1^N \left(X_j - \overline{X}\right)^2}$$

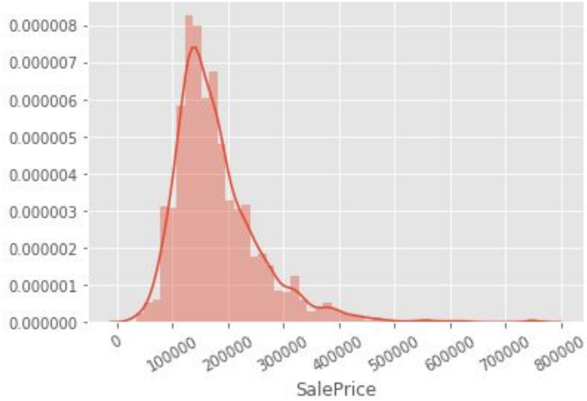
Example of outliers detection

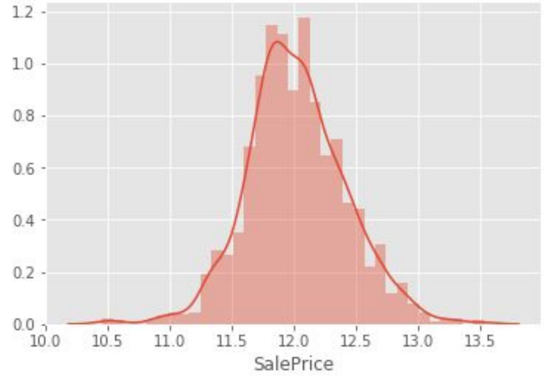


Target transformation

Target transformation is necessary to cope with **heteroscedasticity**. Usually heteroscedasticity is linked to the **skewness** of the target distribution.

To overcome this problem, you should apply a transformation of the target variable, such as a logarithm or square root.





Multiple Regression

The **Multiple Regression** model relates more than one predictor and one response.

Here is the linear regression equation for ***p+1*** regressors plus the intercept :

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p + \varepsilon$$

Where ***y*** and $x = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix}$ are element of one observation taken from the sample.

We consider first element of ***x*** equal to 1 to reflect the intercept in our model.

$$y = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} + \varepsilon$$

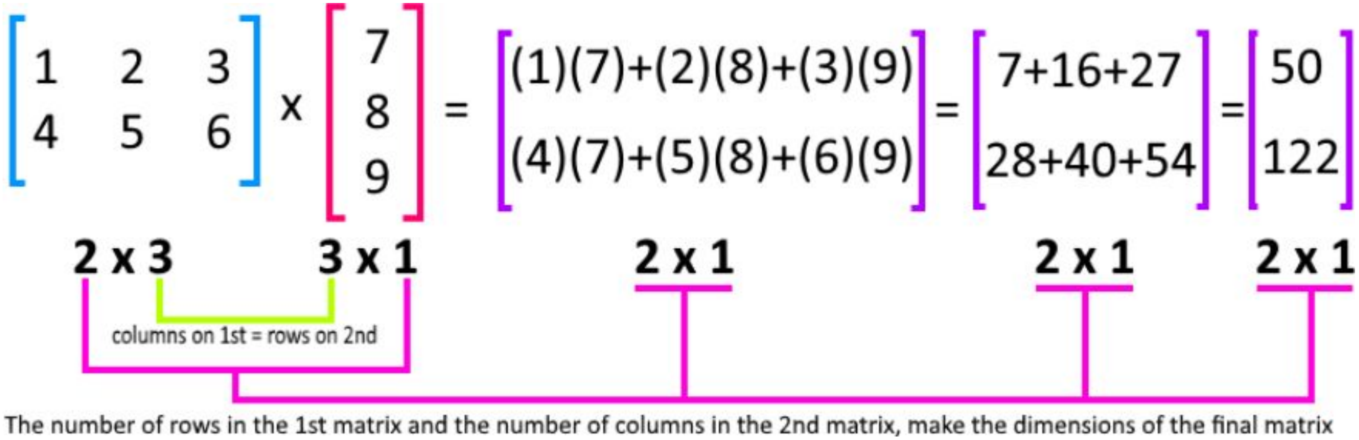
Matrix notation

Linear regression in matrix notation of sample of ***n*** :

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix} \longrightarrow Y = X.a + E$$

For regression with intercept, we have: $x_{i1} = 1 \ \ \forall i$

Matrix multiplication - Example



$$\underbrace{\begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\end{bmatrix}}_{2 \text{ x } 3} \text{ x } \underbrace{\begin{bmatrix}7\\8\\9\end{bmatrix}}_{3 \text{ x } 1} = \underbrace{\begin{bmatrix}(1)(7)+(2)(8)+(3)(9)\\(4)(7)+(5)(8)+(6)(9)\end{bmatrix}}_{2 \text{ x } 1} = \underbrace{\begin{bmatrix}7+16+27\\28+40+54\end{bmatrix}}_{2 \text{ x } 1} = \underbrace{\begin{bmatrix}50\\122\end{bmatrix}}_{2 \text{ x } 1}$$

columns on 1st = rows on 2nd

The number of rows in the 1st matrix and the number of columns in the 2nd matrix, make the dimensions of the final matrix

Norm of a vector

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{pmatrix}$$

$$\|v\|^2 = v^T v = v_1^2 + v_2^2 + \cdots + v_p^2$$

Example:

$$\sqrt{\sum_i \mid v_i \mid^2} = \sqrt{\mid -1 \mid^2 + \mid -2 \mid^2 + \mid 3 \mid^2 + \mid 4 \mid^2} = \sqrt{30}$$

Matrix transpose

A^T The transpose of a matrix is an operator which flips a matrix over its diagonal. That is, it switches the row and column indices of the matrix.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{bmatrix}^{\mathrm{T}}$$

Matrix inverse

B is the inverse of ***A*** if it verifies the following equality:

$$B.A = A.B = I_n \qquad B = A^{-1}$$

$$AX = Y \iff X = A^{-1}Y$$

- Non-square matrices (*m*-by-*n* matrices for which *m* ≠ *n*) do not have an inverse
- A square matrix that is *not* invertible is called **singular** or **degenerate**.
- A square matrix that is invertible if it has a **full column rank *n***; **not perfect multicollinearity exists in the predictor variables**, meaning no linear relationship exists between two or more predictor variables.

Ordinary least squares (OLS)

We need to find the vector $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$ that minimizes the OLS:

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{1}^{n} \left(y_i - \sum_{j=1}^{p} a_j x_{ij} \right)^2$$

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{1}^{n} \left(y_i - X_i . a \right)^2 = \| Y - X . a \|^2$$

Normal equation

For **OLS** methods, the analytical solution is given by:

$$\hat{a} = (X^T X)^{-1} X^T Y$$

Under the hypotheses that ***X*** has a **full column rank**: $rank(X) = p$

$$rank(X^T X) = rank(X X^T) = rank(X)$$

Normal equation is not to consider if :

- In the case of perfect multicollinearity, the parameter will be non-identifiable (no unique solution).
- There is too little data available compared to the number of parameters to be estimated (e.g., fewer data points than regression coefficients).

$$rank(X) \leq min(n, p)$$

If predictors are highly but not perfectly correlated, it can reduce the precision of parameter estimates.

Binary and categorical variables

Dummy variables are **binary variables** that take on value of **1** when the measurement is in a particular group, and **0** when the measurement is not. Example:

- House has a garden: garden = 1
- House doesn’t have a garden: garden = 0

Dummy variables also work for **categorical variables.** Exemple:

Type of foundation: Brick, Cinder Block, Slab, Stone or Wood.

Reference category is the group whose binary variable has been eliminated.

Predictors transformation and interaction

We call **features engineering**, the creation of new predictors by applying nonlinear operations to the initial predictors/variables.

Examples:

$$f_1(X) = \sqrt{x_1} \qquad f_2(X) = \log(x_2) \qquad f_3(X) = \sin(x_3)$$

Interactions between variables are reflected by the multiplication of two predictors values.

$$f(X) = x_1 x_2$$

Polynomial regression

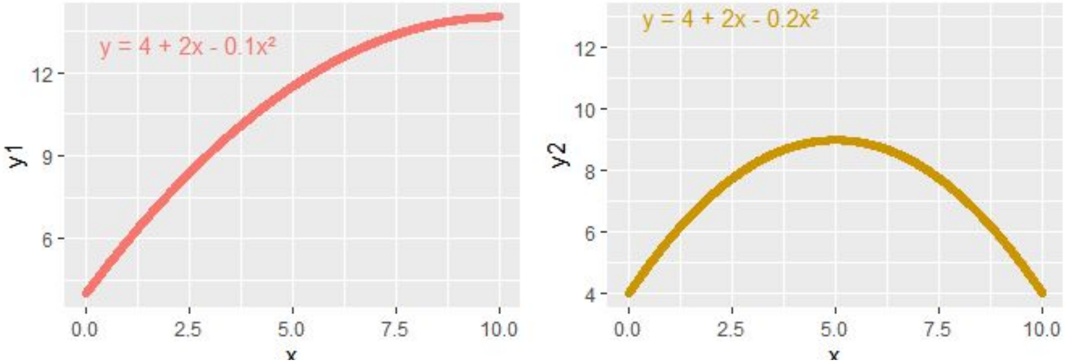
The case of polynomial regression is a special case of multiple linear regression with feature engineering based on power functions:

$$f_k(x) = x^k$$

If we apply this set of function with ***k*** going from ***0*** to ***p*** :

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_p x^p + \varepsilon$$

Example of polynomial regression



Inference in Regression

Under the assumption of **normality of the errors**, we can state that:

$$\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

From the normal equation we have:

$$E(\hat{a}) = a$$

$$VAR(\hat{a}) = \sigma^2 (X^T X)^{-1}$$

$$\hat{a} \sim \mathcal{N}\left(a, \sigma^2 (X^T X)\right)$$

The variance of the estimator can be approximated using square least errors:

$$\widehat{\sigma}^2 = \frac{\sum_1^n (\widehat{\varepsilon}_i)^2}{n - p} = \frac{\|\varepsilon\|}{n - p}$$

significance of coefficient - statistical test

In order to test if a coefficient ***a*** is significant, we build a statistical test based on the distribution of the estimator ***â***.

H0 - null hypothesis: $a = 0$

H1 - alternative hypothesis: $a \neq 0,\ a < 0 \ or \ a > 0$

Under **H0** $\dfrac{\hat{a}}{s_{\widehat{a}}}$ has a **Student's t-distribution with *n* − *p*** degrees of freedom.

To conduct this test, we should also choose a **significance level.** Usually we choose 5% or 1%.

significance of coefficient - simple regression

Formulas for the simple linear regression:

$$Y = a_1 + a_2 X + \varepsilon$$

Coefficients estimators:

$$\hat{a}_2 = \frac{Cov(X, Y)}{Var(X)}$$

$$\hat{a}_2 = \frac{\sum_1^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_i^n \left(X_i - \overline{X}\right)^2}$$

$$\hat{a}_1 = \bar{y} - \hat{a}_2 . \bar{x}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_1^n \hat{\varepsilon}_j^{\,2}}$$

$$s_{\hat{a}_1} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_1^N \left(X_j - \overline{X}\right)^2}}$$

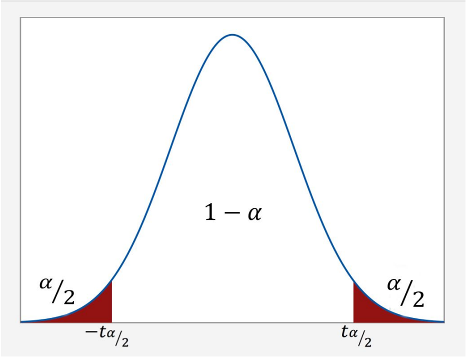
$$s_{\hat{a}_2} = \frac{\hat{\sigma}}{\sqrt{\sum_1^N \left(X_j - \overline{X}\right)^2}}$$

Confidence interval - coefficients

Based on the distribution of the estimators, we can build a confidence interval.

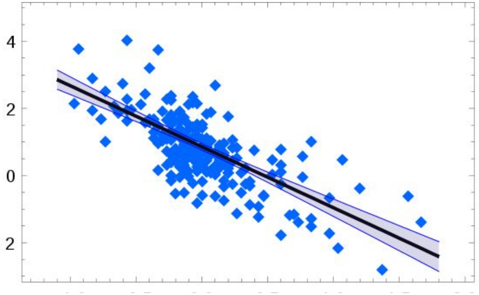
$$a \in \left[\widehat{a} - S_{\widehat{a}}\, t_{\frac{\alpha}{2}, n-p}, \widehat{a} + S_{\widehat{a}}\, t_{\frac{\alpha}{2}, n-p}\right]$$

$t_{\frac{\alpha}{2}, n-p}$ value of Student's t-distribution with ***n − p*** degrees of freedom with level of significance α



<u>Example of simple linear regression</u>

In order to represent this information graphically, in the form of the confidence bands around the regression line. It can be shown that the confidence band has an hyperbolic form.



Confidence interval -prediction

Under the assumption of normality of the errors, we can state that $\widehat{Y}$ follows a normal distribution with variance:

$$VAR(\widehat{Y}) = \sigma^2 X(X^T X)^{-1} X^T$$

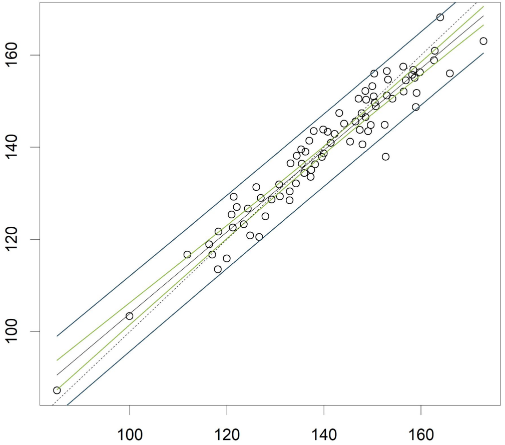
Simple linear regression case :

$$VAR(\widehat{Y}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right)$$

Based on the distribution of the estimators, we can build a confidence interval.

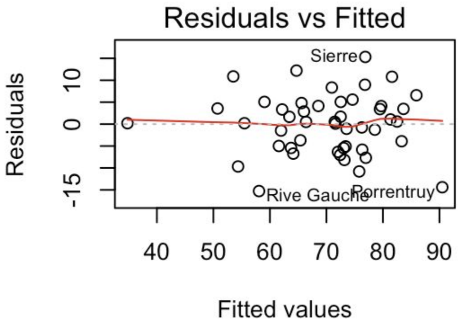
$$\left[\widehat{Y} - S_{\widehat{Y}} t_{\frac{\alpha}{2}, n-p}, \widehat{Y} + S_{\widehat{Y}} t_{\frac{\alpha}{2}, n-p} \right]$$

$t_{\frac{\alpha}{2}, n-p}$ value of Student's t-distribution with ***n − p*** degrees of freedom with level of significance.

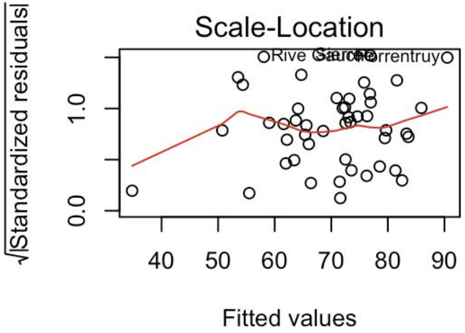


Diagnostic plots

Residuals vs Fitted (most important): plots ordinary residuals vs fitted values. Used to detect patterns for missing variables, heteroskedasticity, …

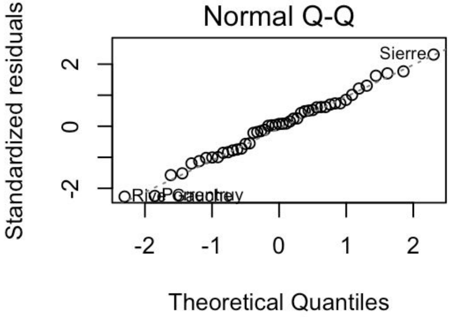


Scale-Location: plots standardized residuals vs fitted values. Used to detect patterns in residuals.

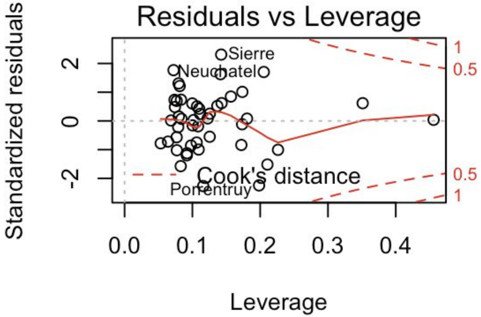


Diagnostic plots

Normal Q-Q (*quantile-quantile)*: plots theoretical quantiles for standard normal vs actual quantiles of standardized residuals. It’s used to evaluate normality of the errors.



Residuals vs Leverage: plots Cooks distances comparison of fit at that point vs potential for influence of that point. It’s used to detect any points that have substantial influence on the regression model.



Cook's distance of observation ***i*** is defined as the sum of all the changes in the regression model when observation ***i*** is removed from it.

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)} \right)^2}{p\widehat{\sigma}^2}$$

Residual analysis - DFFITS (Kuh, and Welsch)

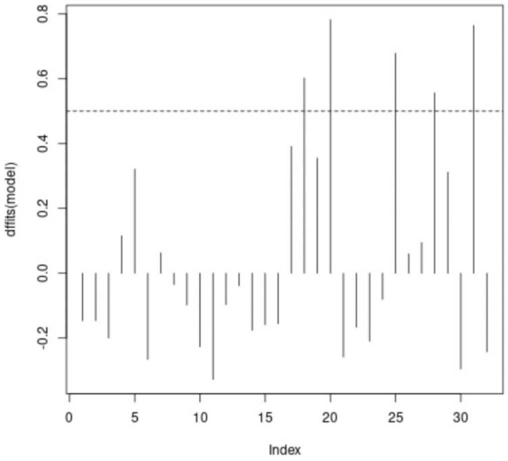
The DFFITS statistic is a scaled measure of the change in the predicted value for the ***i*th** observation, and is calculated by deleting the *i*th observation.

A large value indicates that the observation is very influential in its neighborhood of the $\mathbf{X}$ space.

$$\text{DFFIT} = \widehat{y_i} - \widehat{y_{i(i)}}$$

$$\text{DFFITS} = \frac{\text{DFFIT}}{s_{(i)}\sqrt{h_{ii}}}$$

$$\text{DFFITS} = t_{i(i)}\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$



Large values of DFFITS indicate influential observations.

A size-adjusted cutoff recommended by Belsley is Kuh, and Welsch: $2\sqrt{\frac{p}{n}}$

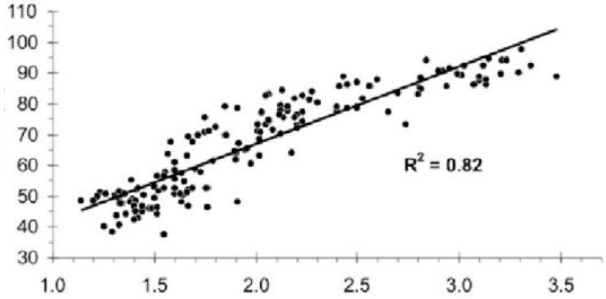
Coefficient of determination

***R*²** is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{SSE}{SST}$$

***R*²** is used as a measure of the goodness of fit of a model. In regression, the ***R*²** coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.

***R*²** of 1 indicates that the regression predictions perfectly fit the data.



Can also be defined as: $R^2 = Cor(\,Y, \widehat{Y})^{\,2}$

Adjusted R squared

Inflation of R^2

R^2 is at least weakly increasing with increases in the number of regressors in the model. Therefore, R^2 alone cannot be used as a meaningful comparison of models with very different numbers of independent variables.

Adjusted R^2

The use of an adjusted R^2 is an attempt to account for the phenomenon of inflation of R^2.
The most used adjustment of R^2 is defined as

$$R^2 = 1 - \frac{SSE}{SST}$$

$$\text{adj } R^2 = 1 - \frac{SSE*df_t}{SST*df_e}$$

$$\longrightarrow$$

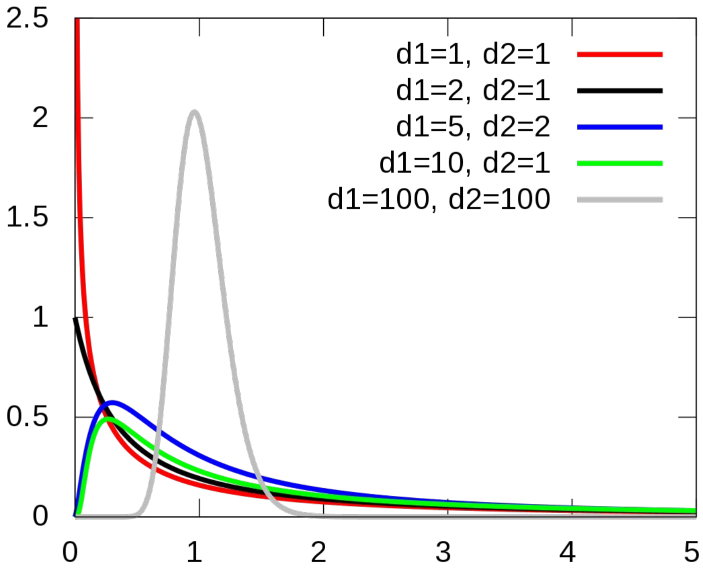
$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p}$$

F-Distribution

The F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled independent chi-squared variables:

$$S_1 \sim \chi^2_{d_1} \qquad S_2 \sim \chi^2_{d_2} \qquad S_1 \perp S_2$$

$$F(d_1, d_2) = \frac{S_1/d_1}{S_2/d_2}$$



The F-test for Simple Linear Regression

The F-test of the overall significance is a specific form of the F-test. It compares a model with no predictors to the model that you specify. A regression model that contains no predictors is also known as an intercept-only model.

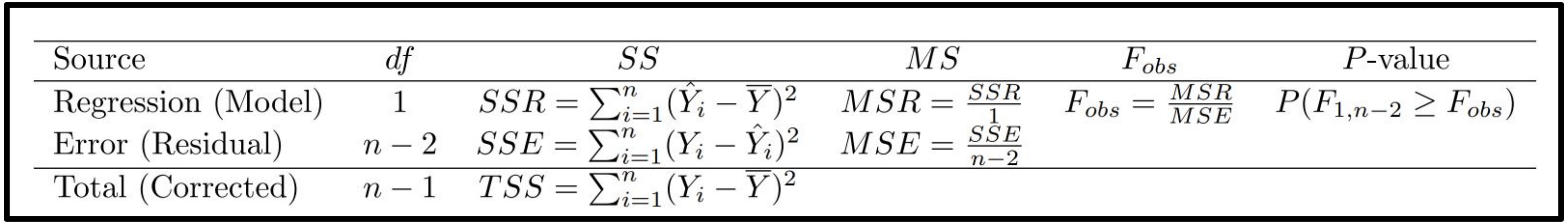
H0 - null hypothesis: The fit of the intercept-only model and your model are equal.

$$a = 0$$

H1 - alternative hypothesis: The fit of the intercept-only model is significantly reduced compared to your model.

$$a \neq 0$$

The F-test - Building the statistique



Source	df	SS	MS	F_{obs}	P-value
Regression (Model)	1	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{1,n-2} \geq F_{obs})$
Error (Residual)	$n-2$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$		
Total (Corrected)	$n-1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

SSR: Sum of Squares Regression

SSE: Sum of Squares Error

TSS: Total Sum of Squares

MSR: Mean Squares Regression

MSE: Mean Squares Error

$$MSR = \frac{SSR}{1} = \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\widehat{a}X_i + \widehat{b} - \overline{Y}\right)^2 = \widehat{a}^2\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$MSE = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2}{n-2} = \widehat{\sigma}^2$$

$$E(MSR) = \sigma^2 + a^2\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$E(MSE) = E\left(\widehat{\sigma}^2\right) = \sigma^2$$

Under **H0** the whe have $a = 0$

$$F_{Obs} = \frac{MSR}{MSE} \sim F(1, n-2)$$

The F-test for Multiple Linear Regression

Consider two models, 1 and 2, where model 1 is 'nested' within model 2. Model 1 is the restricted model, and model 2 is the unrestricted one. For any choice of parameters in model 1, the same regression curve can be achieved by some choice of the parameters of model 2.

The model with more parameters will always be able to fit the data at least as well as the model with fewer parameters.

H0 - null hypothesis: model 2 does not provide a significantly better fit than model 1.

H1 - alternative hypothesis: model 2 provides a significantly better fit than model 1.

Under the null hypothesis **H0** that, *F* will have an *F* distribution, with ($p_2 - p_1$, $n - p_2$) degrees of freedom.

$$F = \frac{\left(\frac{SSE_1 - SSE_2}{df_1 - df_2}\right)}{\frac{SSE_2}{df_2}} = \frac{\left(\frac{SSE_1 - SSE_2}{p_2 - p_1}\right)}{\frac{SSE_2}{n - p_2}} \sim F\left(p_2 - p_1, n - p_2\right)$$

Model selection-variables selection

Differents approchs are possible to perform variables/model selection:

- Backward Elimination, Forward Selection and Stepwise Regression
- Regularization: Lasso, Ridge…
- Cross validation.

Remark: These methods can also fix collinearity problem and help when dataset has big number of variables.

Backward Elimination & Forward Selection

Backward Elimination:

This is a “top-down” method, which begins with a ”Complete” Model, with all potential predictors. The analyst then chooses a significance level to stay in the model (SLS). The model is fit, and the predictor with the lowest t-statistic in absolute value (largest P-value) is identified. If the P-value is larger than SLS, the variable is dropped from the model. Then the model is re-fit with all other predictors (this will change all regression coefficients, standard errors, and P-values). The process continues until all variables have P-values below SLS.

Forward Selection:

This is a ”bottom-up” method, which begins with all ”Simple” Models, each with one predictor. The analyst then chooses a significance level to enter into the model (SLE). Each model is fit, and the predictor with the highest t-statistic in absolute value (smallest P -value) is identified. If the P-value is smaller than SLE, the variable is entered into the model. Then all two variable models including the best predictor in the first round, with each of the other predictors. The best second variable is identified, and its P-value is compared with SLE. If its P-value is below SLE, the variable is added to the model. The process continues until no potential added variables have P-values below SLE.

Stepwise regression

This approach is a hybrid of forward selection and backward elimination. It begins like forward selection, but then applies backward elimination at each step.

In forward selection, once a variable is entered, it stays in the model. In stepwise regression, once a new variable is entered, all previously entered variables are tested, to confirm they should stay in the model, after controlling for the new entrant, as well as the other previous entrant.

GLM-Generalized Linear Models

The **generalized linear model** (**GLM**) is a generalization of ordinary linear regression defined by given formula:

$$E(Y) = g^{-1}(Xa)$$

The GLM consists of 3 elements:

1. An exponential family of probability distributions. ex: Normal distribution, Exponential, Poisson, Bernoulli…
2. A linear predictor : $\eta = Xa$
3. A link function ***g*** such that : $E(Y) = \mu = g^{-1}(\eta)$

The unknown parameters, ***a***, are typically estimated with maximum likelihood,

For ordinary linear regression the distribution is **normal** and ***g(x)=x.***

Logistic regression (logit models)

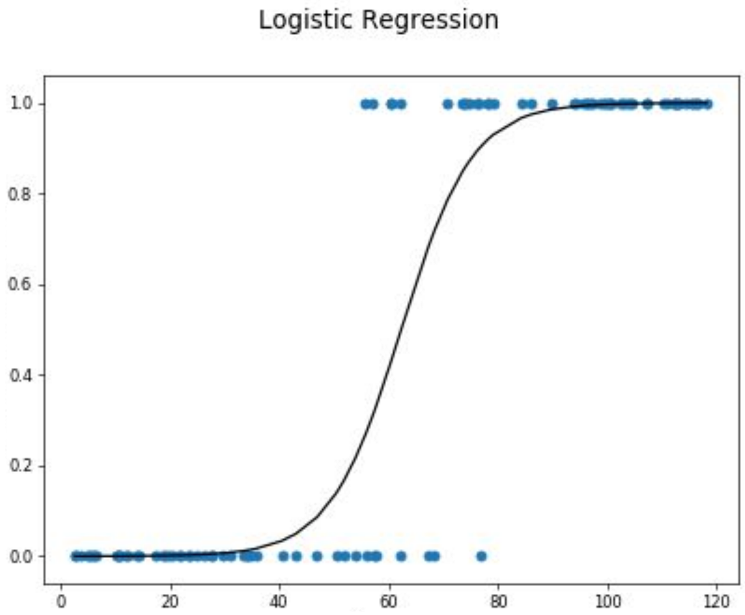
When the response data, *Y*, are binary (taking on only values 0 and 1), the distribution function is generally chosen to be the **Bernoulli** distribution and the interpretation of μ_i is then the probability *p* of Y_i taking the value 1. $E(Y) = \mu = p$

Logit link function

The most typical link function is the canonical logit link:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

$$E(Y) = \mu = g^{-1}(\eta) = \frac{exp(\eta)}{1 + exp(\eta)}$$



Annexes

Normal equation - proof

OLS

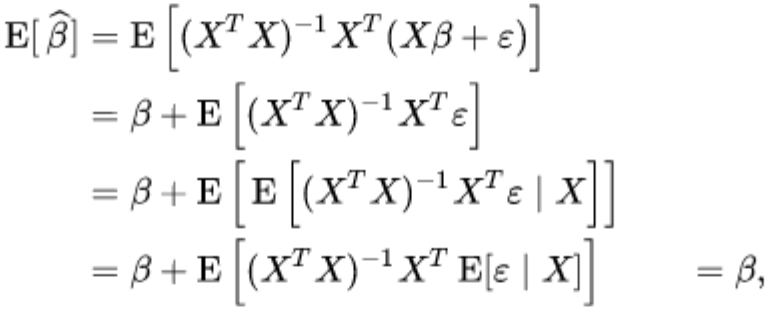
$$\begin{aligned}
L\left(D,\vec{\beta}\right) &= \|X\vec{\beta}-Y\|^2 \\
&= \left(X\vec{\beta}-Y\right)^{\mathsf{T}}\left(X\vec{\beta}-Y\right) \\
&= Y^{\mathsf{T}}Y - Y^{\mathsf{T}}X\vec{\beta} - \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y + \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\vec{\beta}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L\left(D,\vec{\beta}\right)}{\partial\vec{\beta}} &= \frac{\partial\left(Y^{\mathsf{T}}Y - Y^{\mathsf{T}}X\vec{\beta} - \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y + \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\vec{\beta}\right)}{\partial\vec{\beta}} \\
&= -2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\vec{\beta}
\end{aligned}$$

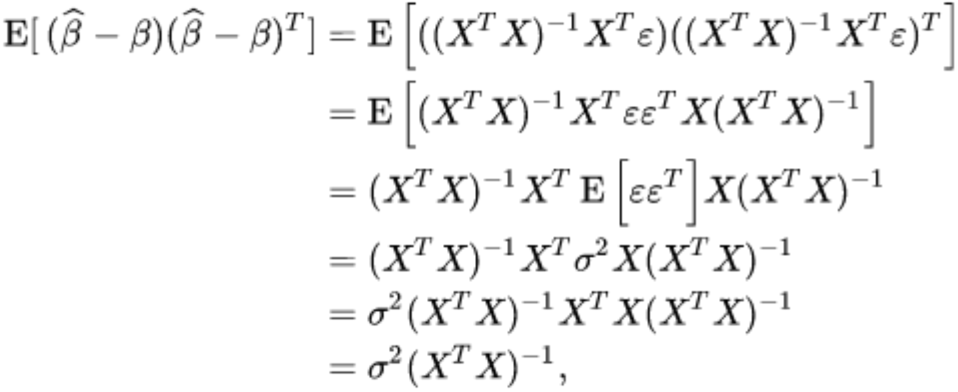
To minimize we have to solve the equation:

$$\begin{aligned}
-2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\vec{\beta} &= 0 \\
\Rightarrow X^{\mathsf{T}}Y &= X^{\mathsf{T}}X\vec{\beta} \\
\Rightarrow \hat{\vec{\beta}} &= \left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}Y
\end{aligned}$$

Coefficient estimator - proof



$$\begin{aligned}
\mathrm{E}[\,\widehat{\beta}\,] &= \mathrm{E}\left[(X^T X)^{-1} X^T (X\beta + \varepsilon)\right] \\
&= \beta + \mathrm{E}\left[(X^T X)^{-1} X^T \varepsilon\right] \\
&= \beta + \mathrm{E}\left[\mathrm{E}\left[(X^T X)^{-1} X^T \varepsilon \mid X\right]\right] \\
&= \beta + \mathrm{E}\left[(X^T X)^{-1} X^T \,\mathrm{E}[\varepsilon \mid X]\right] \qquad = \beta,
\end{aligned}$$



$$\begin{aligned}
\mathrm{E}[\,(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)^T] &= \mathrm{E}\left[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T\right] \\
&= \mathrm{E}\left[(X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}\right] \\
&= (X^T X)^{-1} X^T \,\mathrm{E}\left[\varepsilon\varepsilon^T\right] X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1},
\end{aligned}$$