Data Science

Regression and Prediction

Séverine Affeldt

Université Paris Cité
Centre Borelli, UMR 9010
UFR Sciences Fondamentales et Biomédicales

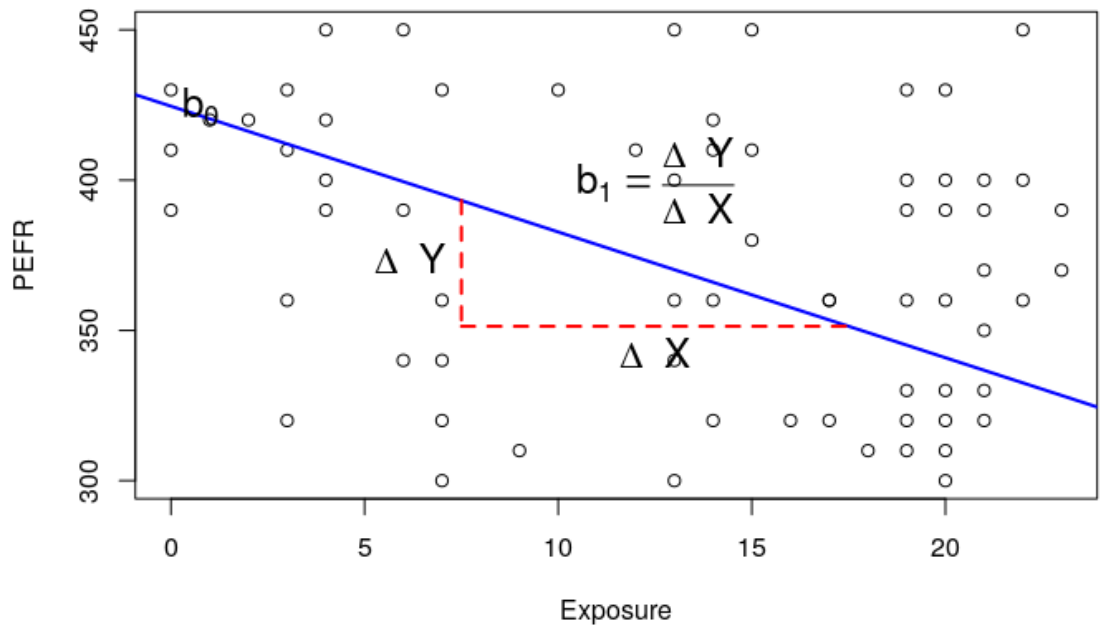
2023-2024

Simple Linear Regression

Objective

Simple Linear Regression estimates *how much* Y will change when X changes by a certain amount. We try to predict the variable Y (*reponse*, *dependent variable*, *target*, *outcome*) from X (*predictor*, *independent variable*, *feature*) using a linear relationship:

$$Y = b_0 + b_1 X$$



Prediction with an error terms

Predictions (*fitted values*) do not fall exactly on the line. There are prediction errors (*residuals*). The regression equation should include an error term e_i.

$$Y_i = b_0 + b_1 X_i + e_i$$

Prediction with an error terms

Predictions (*fitted values*) do not fall exactly on the line. There are prediction errors (*residuals*). The regression equation should include an error term e_i.

$$Y_i = b_0 + b_1 X_i + e_i$$

Residuals

The fitted values are denoted with the *hat* notation,
$\hat{Y}_i = \hat{b_0} + \hat{b_1} X_i$ where $\hat{b_0}$, $\hat{b_1}$ are estimated (vs. known).

The residuales $\hat{e}_i$ are computed as $\hat{e}_i = Y_i - \hat{Y}_i$.

Least Squares

The regression line is the estimate that minimizes the sum of squared residual values, also called the *residual sum of squares* or *RSS*.

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{b_0} - \hat{b_1} X_i)^2$$

$\hat{b_0}$, $\hat{b_1}$ are the values that minimize *RSS*.

Simple_Linear_Regression.Rmd | Simple_Linear_Regression.ipynb

Multiple Linear Regression

Multiple Linear Regression objective

Explaining the values taken by the **outcome** Y based on the p **features** $\{X_j\}_{1 \leq j \leq p}$ selon,

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p + e$$

All the concepts in simple linear regression, such as fitting by least squares and the definition of fitted values and residuals, extend to the multiple linear regression setting. Hence, fitted values are computed as,

$$\hat{Y}_i = \hat{b_0} + \hat{b_1} X_{1,i} + \hat{b_2} X_{2,i} + ... + \hat{b_p} X_{p,i}$$

Multiple Linear Regression objective

Explaining the values taken by the **outcome** Y based on the p **features** $\{X_j\}_{1 \leq j \leq p}$ selon,

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p + e$$

All the concepts in simple linear regression, such as fitting by least squares and the definition of fitted values and residuals, extend to the multiple linear regression setting. Hence, fitted values are computed as,

$$\hat{Y}_i = \hat{b_0} + \hat{b_1} X_{1,i} + \hat{b_2} X_{2,i} + ... + \hat{b_p} X_{p,i}$$

Matrix notation

$$(n, 1) = (n, p+1) \times (p+1, 1) + (n, 1)$$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & \cdots & \cdots & \cdots & X_{1,p} \\ \vdots & \vdots & \cdots & \cdots & \cdots & \vdots \\ 1 & X_{i,1} & \cdots & X_{i,j} & \cdots & X_{i,p} \\ \vdots & \vdots & \cdots & \cdots & \cdots & \vdots \\ 1 & X_{n,1} & \cdots & \cdots & \cdots & X_{n,p} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_i \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix} = \mathbf{Xb} + \mathbf{e}$$

Similarly as for the simple linear regression, we estimate the parameters $\{b_j\}_{0 \leq j \leq p}$ that **minimize** the **residuals sum of squared** (RSS)

Similarly as for the simple linear regression, we estimate the parameters $\{b_j\}_{0 \leq j \leq p}$ that **minimize** the **residuals sum of squared** (RSS)

Similarly, for multiple linear regression

$$RSS = S(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{b_0} - \hat{b_1}X_i)^2 = (\mathbf{Y} - \mathbf{Xb})^2$$

RSS is quadratic in $\mathbf{b}$ with positive-definitive Hessian $\Rightarrow \exists$ a global minimum at $\mathbf{b} = \hat{\mathbf{b}}$

Similarly as for the simple linear regression, we estimate the parameters $\{b_j\}_{0 \leq j \leq p}$ that **minimize** the **residuals sum of squared** (RSS)

Similarly, for multiple linear regression

$$RSS = S(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{b_0} - \hat{b_1}X_i)^2 = (\mathbf{Y} - \mathbf{Xb})^2$$

RSS is quadratic in $\mathbf{b}$ with positive-definitive Hessian $\Rightarrow \exists$ a global minimum at $\mathbf{b} = \hat{\mathbf{b}}$

With matrix notation

$$\begin{aligned} S(\mathbf{b}) &= (\mathbf{Y} - \mathbf{Xb})^T(\mathbf{Y} - \mathbf{Xb}) \\ &= \mathbf{YY}^T - 2\mathbf{b}^T\mathbf{X}^T\mathbf{Y} + \mathbf{b}^T\mathbf{X}^T\mathbf{Xb} \end{aligned}$$

hence we should solve,

$$\frac{\partial S}{\partial \mathbf{b}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T X\mathbf{b} = 0$$

$$\frac{\partial S}{\partial \mathbf{b}} = 0 \Leftrightarrow \boxed{\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})}$$

$$\frac{\partial S}{\partial \mathbf{b}} = 0 \Leftrightarrow \boxed{\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})}$$

$$\frac{\partial S}{\partial \mathbf{b}} = 0 \Leftrightarrow \boxed{\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_i X_{i,1} & \cdots & \sum_i X_{i,p} \\ \sum_i X_{i,1} & \sum_i X_{i,1}^2 & \cdots & \sum_i X_{i,1}X_{i,p} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_i X_{i,p} & \sum_i X_{i,1}x_{i,p} & \cdots & \sum_i X_{i,p}^2 \end{pmatrix}$$

A symmetric matrix that indicates the relationship **between the features** variables $\{X_j\}$

If centered values $\frac{1}{n}(\mathbf{X}^T\mathbf{X}) \Leftrightarrow$ variance-covariance matrix of $(\mathbf{X}_i, \mathbf{X}_j)$

If centered reduced values $\frac{1}{n}(\mathbf{X}^T\mathbf{X}) \Leftrightarrow$ cross-correlation matrix of $(\mathbf{X}_i, \mathbf{X}_j)$

$$\frac{\partial S}{\partial \mathbf{b}} = 0 \Leftrightarrow \boxed{\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_i X_{i,1} & \cdots & \sum_i X_{i,p} \\ \sum_i X_{i,1} & \sum_i X_{i,1}^2 & \cdots & \sum_i X_{i,1}X_{i,p} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_i X_{i,p} & \sum_i X_{i,1}x_{i,p} & \cdots & \sum_i X_{i,p}^2 \end{pmatrix}$$

A symmetric matrix that indicates the relationship **between the features** variables $\{X_j\}$

If centered values $\frac{1}{n}(\mathbf{X}^T\mathbf{X}) \Leftrightarrow$ variance-covariance matrix of $(\mathbf{X}_i, \mathbf{X}_j)$
If centered reduced values $\frac{1}{n}(\mathbf{X}^T\mathbf{X}) \Leftrightarrow$ cross-correlation matrix of $(\mathbf{X}_i, \mathbf{X}_j)$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_i Y_i \\ \sum_i Y_i x_{i,1} \\ \vdots \\ \sum_i Y_i x_{i,p} \end{pmatrix}$$

A matrix that indicates the relationships **between the features $\{\mathbf{X}_j\}$ and the outcome Y**

If centered values $\frac{1}{n}(\mathbf{X}^T\mathbf{Y}) \Leftrightarrow$ covariance matrix of $(\mathbf{X}_i, \mathbf{Y})$
If centered reduced values $\frac{1}{n}(\mathbf{X}^T\mathbf{Y}) \Leftrightarrow$ correlation matrix of $(\mathbf{X}_i, \mathbf{Y})$

$$\frac{\partial S}{\partial \mathbf{b}} = 0 \Leftrightarrow \boxed{\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_i X_{i,1} & \cdots & \sum_i X_{i,p} \\ \sum_i X_{i,1} & \sum_i X_{i,1}^2 & \cdots & \sum_i X_{i,1}X_{i,p} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_i X_{i,p} & \sum_i X_{i,1}x_{i,p} & \cdots & \sum_i X_{i,p}^2 \end{pmatrix}$$

A symmetric matrix that indicates the relationship **between the features** variables $\{X_j\}$

If centered values $\frac{1}{n}(\mathbf{X}^T\mathbf{X}) \Leftrightarrow$ variance-covariance matrix of $(\mathbf{X}_i, \mathbf{X}_j)$
If centered reduced values $\frac{1}{n}(\mathbf{X}^T\mathbf{X}) \Leftrightarrow$ cross-correlation matrix of $(\mathbf{X}_i, \mathbf{X}_j)$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_i Y_i \\ \sum_i Y_i x_{i,1} \\ \vdots \\ \sum_i Y_i x_{i,p} \end{pmatrix}$$

A matrix that indicates the relationships **between the features $\{\mathbf{X}_j\}$ and the outcome $\mathbf{Y}$**

If centered values $\frac{1}{n}(\mathbf{X}^T\mathbf{Y}) \Leftrightarrow$ covariance matrix of $(\mathbf{X}_i, \mathbf{Y})$
If centered reduced values $\frac{1}{n}(\mathbf{X}^T\mathbf{Y}) \Leftrightarrow$ correlation matrix of $(\mathbf{X}_i, \mathbf{Y})$

Hence, b_j is large if $\mathbf{X}_j$ is strongly related to $\mathbf{Y}$ AND weakly related to $\mathbf{X}_i$, $i \neq j$

As for the simple linear regression, we can decompose the error variance as,

$$Total SS = Residual SS + Explain SS$$

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

As for the simple linear regression, we can decompose the error variance as,

$$Total SS = Residual SS + Explain SS$$

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

RMSE: Root Mean Squared Error

- the most important performance metric
- measures the overall accuracy of the model
- basis for comparing to other models

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

As for the simple linear regression, we can decompose the error variance as,

$$Total SS = Residual SS + Explain SS$$

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

RSE: Residual Standard Error

- very similar to RMSE in practice
- replace the number of observations by the degree of freedom

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

As for the simple linear regression, we can decompose the error variance as,

$$Total SS = Residual SS + Explain SS$$

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

R^2: R-squared statistic or Coefficient of determination

- indicates the quality of the model
- measures the proportion of the variation in the data that is accounted for in the model
- assesses how well the model fits the data
  - $R^2 \approx 1$: good
  - $R^2 \approx 0$: bad

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

R^2 (minor) issue (when n is large...)

R^2 increases with the number of features (even if they are not relevant!)

Increasing the number of $\{X_j\}$ implies that the degree of freedom decreases. We can adjust the R^2 by the degree of freedom to solve the issue.

Adjusted R^2

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \\ &= 1 - \frac{n-1}{n-p-1}(1-R^2) \end{aligned}$$

SE and *t-statistic* t_b

The *t-statistic* (mirror image of the p-value) measures the extent to which a coefficient is **statistically significant**, ie. outside the range of what a random chance arrangement of predictor and target variable might produce.

$$t_b = \frac{\hat{b}}{SE(\hat{b})}$$

The **higher the t-statistic** (or lower the *p-value*), the **more significant the predictor**. It can guide choice of variables to include as predictors.

Multiple_Linear_Regression.Rmd | Multiple_Linear_Regression.ipynb

Multiple Linear Regression & Model Selection

Objective

Get a good **trade-off** between the **model precision** (quantified by R^2 or SSR) and the **model complexity** (quantified by the number of number of variables).

Akaike criteria (AIC) metric to be minimized

AIC penalizes adding terms to a model.

$$AIC = n \ln \frac{RSS}{n} + 2(p+1)$$

Schwartz criteria (BIC) to be minimized

$$BIC = n \ln \frac{RSS}{n} + \ln(n)(p+1)$$

When $n > e^2 \approx 7$, the BIC criteria strongly penalizes complex models, *ie.* preferentially proposes a model with few variables.

NB: The dependent variables redundancy is implicitly taken into account, as redundant variables won't improve the RSS but will increase the model complexity.

Stepwise regression through *backward elimination*

Start from a model with all the variables and iteratively remove a variable from this complete model whenever its removal implies the strongest reduction of the AIC criteria,

1. Compute AIC for the complete model
2. Consecutively remove each variable and recompute the AIC. Permanently remove the variable that induces the strongest AIC reduction
3. If none, STOP. Else, go back to (1)

The CONSO dataset contains $n = 27$ observations. We consider the 4 variables `prix`, `cylindrée`, `puissance` and `poids` (and the constant!). Using the `stepAIC` function from the `R MASS` package,

Step	Current model	AIC	Removal
1	$Y \sim b_1 \times prix + b_2 \times cylindree + b_3 \times puissance + b_4 \times poids + b_0$	-18.69	$\underline{puissance} \rightarrow -20.6188$ $prix \rightarrow -20.0081$ $cylindree \rightarrow -17.4625$ $poids \rightarrow -12.1155$
2	$Y \sim b_1 \times prix + b_2 \times cylindree + b_3 \times poids + b_0$	-20.6188	$\underline{prix} \rightarrow -21.9986$ $cylindree \rightarrow -17.6000$ $poids \rightarrow -13.3381$
3	$Y \sim b_1 \times cylindree + b_2 \times poids + b_0$	-21.9986	$cylindree \rightarrow -13.3049$ $poids \rightarrow -0.2785$

Optimal model

At step (3), no variable removal improves the AIC criteria. The optimal model is,

$$y = 1.392276 + 0.01311 \times cylindree + 0.004505 \times poids$$

Multiple_Linear_Regression_StepReg.Rmd | Multiple_Linear_Regression_StepReg.ipynb

Interpreting the Regression Equation

Key terms

Correlated variables When the features are highly correlated, it is difficult to interpret the individual coefficients.

Multicollinearity When the features have perfect, or near-perfect, correlation, the regression can be unstable or impossible to compute.

Confounding variables An important predictor that, when omitted, leads to spurious relationships in a regression equation.

Main effects The relationship between a feature and the outcome variable, independent of other variable

Interactions An interdependent relationship between two or more features and the response

Multiple_Linear_Regression_InterpretReg.Rmd |
Multiple_Linear_Regression_InterpretReg.ipynb → TBD!

Bibliography

- Rakotomalala, Ricco. “Pratique de la Régression Linéaire Multiple – Diagnostic et sélection de variables.”
- Wikipedia!