Machine learning 1

Linear regression and logistic regression

Piotr Wójcik
pwojcik@wne.uw.edu.pl

academic year 2017/2018

Uniwersytet Warszawski
Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Linear regression

Logistic regression

Linear regression

Linear regression

- Linear regression is one of the simplest supervised learning algorithms
- is used for modeling **quantitative dependent variable**
- although it may seem boring compared to more modern methods of statistical learning, it is still a useful and often used method of modeling
- it can serve as a good **starting point** or **reference point** for newer approaches
- especially that many more fancy methods can be seen as generalizations or extensions of linear regression
- therefore **understanding linear regression is the key to learning more complex methods**

Linear relationship

- the simplest form of the functional dependency between two variables is a **linear function**
- mathematically we can write: $y_i \approx \beta_0 + \beta_1 x_i$
- the linear function has two **parameters** (or **coefficients**): **constant** β_0 and **slope** β_1
- **constant** is the intersection point of the function graph with the vertical axis (says what is the expected value of Y when $X = 0$ – generally has no interpretation)
- **slope** is the most common goal of analysis – it shows how Y will change when X changes by 1

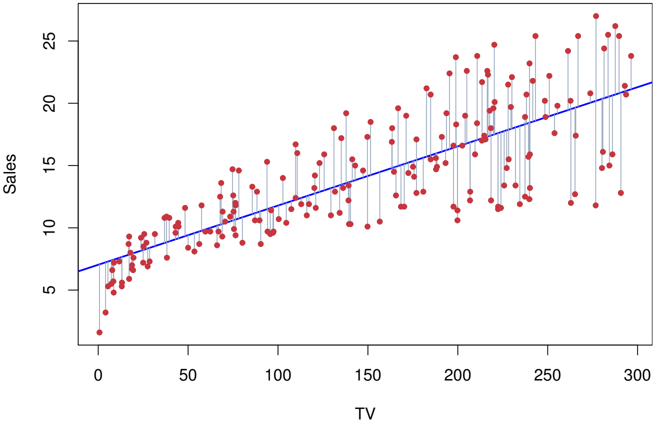
Simple, hyper-plane of regression

- the relationship line is called **the regression line**
- however, the line depicts the relationship only in case of simple regression (with one explanatory variable)
- in the case of two explanatory variables, the result is the **regression surface**
- in the case of many variables we talk about the so-called **regression hypersurface**

Random disturbance

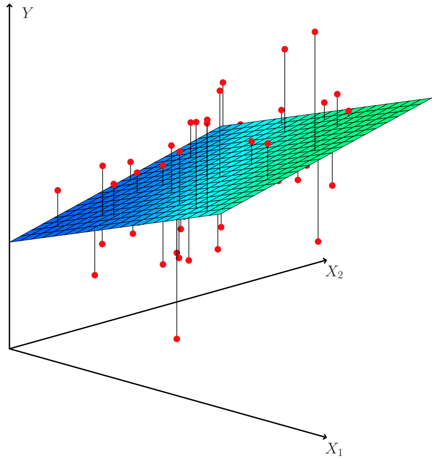
- no line/hypersurface fits the data perfectly
- therefore regression function takes the form: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- the last element (ϵ_i) is called an **error term** or **random disturbance**
- it can be imagined as a distance (vertical) between data point and regression line
- it describes the part of variability of y_i, which cannot be explained by explanatory variables.
- in case of many explanatory variables the formula is written as: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i$

Sample regression line



Source: James i in (2017), s. 62

Sample regression surface



Source: James i in (2017), s. 73

Estimation of linear regression model

- true parameter values $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are not known
- they are estimated based on relationships observed in a sample
- estimators will not reflect true values perfectly
- estimation results on two samples will be usually different
- in linear regression estimates will be **unbiased** – if we estimated the model on many different samples, then the obtained estimates would be on **average** equal to the true (unknown) parameters

Ordinary Least Squares (OLS)

- **estimators** of parameters based on the sample are usually denoted as $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$
- estimates are different from true values: $\hat{\beta} \neq \beta$
- value of variable Y resulting from the estimated model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}$$

  or in matrix notation

$$Y = X\hat{\beta}$$

is called a **fitted value** or **theoretical value**

Residuals

- The difference between empirical (observed) and theoretical value is called model **residual**:

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}) = y_i - \hat{y}_i$$

or in matrix notation

$$e = Y - X\hat{\beta} = Y - \hat{Y}$$

- Model residuals are estimates of the random disturbances, but are different (because $\hat{\beta} \neq \beta$)

Ordinary Least Squares (OLS) – cont'd

- a good model is one for which the residuals are small (the predicted values of the modeled variable are **close** to real values)
- there are many definitions of **closeness**, but the most common is minimizing the sum of squares of all residuals
- if the sum of squared residuals (called also **Residual Sum of Squares**, **RSS**) is small, the model fits the data well
- we are looking for such parameters $\hat{\beta}$, for which **the sum of squared deviations** of empirical values from theoretical **is as small as possible**

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

Assumptions of linear regression model

1. there is a **linear relationship** between the explained variable Y and the explanatory variables X (in practice it is often a sufficiently good approximation)
2. X explanatory variables are **non-random**, do not affect random error values
3. the expected value of the random error **is equal to 0**
4. individual random components **are NOT correlated – lack of autocorrelation** of random components
5. the variance of random disturbances is **the same** for all observations – the random component is **homoscedastic**

Testing the statistical significance of parameters

- after estimating the regression coefficients, one can test their **statistical significance**
- statistical significance means a situation in which **the actual parameter value is different from zero**
- we are testing the **null hypothesis**, assuming that the actual model coefficient β_i **equals zero**
- if the null hypothesis **is rejected** in favor of the **alternative hypothesis**, that the coefficient is **different from zero**, we will say that the variable to which it refers **is significant** in the model
- it can be formulated as: $H_0 : \beta_i = 0$ vs. $H_a : \beta_i \neq 0$

Testing statistical significance of parameters - cont.

- the test statistic has the form:

$$t = \frac{\hat{\beta}_i}{SE(\beta_i)}$$

where $SE(\beta_i)$ means the standard error of the β_i

- test statistic t has a t-Student distribution with $n-2$ degrees of freedom

Testing the significance of the entire model

- apart from testing statistical significance of **individual parameters**, the **joint significance** of the entire model is also examined
- the null hypothesis is set that all **estimated parameters apart from the constant** are simultaneously 0:
  $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against the alternative hypothesis
  H_a: at least one parameter $\beta_i \neq 0$
- if the only statistically significant parameter is a constant, the values of the dependent variable do not depend in any way on the values of the independent variables
- then the estimated model is not able to explain the variability of the analyzed phenomenon

Testing the significance of the entire model – F test

- in this case the F test is used, the test statistic of which is in the form:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

where TSS means the total sum of squares (**TSS**), i.e. the sum of the squared deviations of the Y variable from the average Y:
$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- it may happen that the **model as a whole is significant**, although all variables **individually will be insignificant**

Measure of model fit – R^2

- R^2 measures the proportion of variability of Y, that can be explained by variables X

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

- always takes values from 0 to 1
- does not depend on the scale of Y
- R^2 closer to 1 generally means a better model
- a value close to 0 can mean that:
  - the linear model is incorrect or
  - the estimation error is large (resulting, for example, from not taking into account significant variables)
- however, there is no clear indication of how high the value of R^2 is good enough – it depends on the application

Measure of model fit – adjusted R^2

- **removing variable** from the model usually **will lower** R^2
- **adding a regressor** (even not significant) most often **increases** R^2
- it does not mean, however, that you should add them to maximize R^2
- for a model with a large number of variables, one should use **adjusted** R^2

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

- simplified interpretation: the adjusted R^2 is a measure of fit for statistically significant variables

Qualitative explanatory variables

- often **qualitative variables** (discrete) appear among explanatory variables
- these are variables that take one of a finite number of values
- special case is the **dummy variable**, **binary variable**, taking only two possible values
- if the variable is binary, then for modeling **it should be coded into** 0 and 1
- if the variable has more than 2 levels, for **each level** of the variable variable one should create a **separate dunmy** variable, which takes the value 1 if the original variable is at a given level and 0 otherwise
- and to the model we finally include **all but one** (reference level)

Logistic regression

Logistic regression

- the correct model for a binary dependent variable is the **logistic regression**, which models **the probability** that the qualitative dependent variable will take a specific value, e.g. $Pr(Y = 1|X)$
- often one of the variable levels is called a **success** and the other **default**
- in order for the resulting probability to fall within an range $[0, 1]$, the appropriate transforming function is used

Logistic function

- in logistic regression one uses **logistic function**:

$$p(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

- transforming the above equaion one can show that:

$$ln\left(\frac{p(X)}{1 - p(X)}\right) = X\beta$$

- the quantity $\frac{p(X)}{1-p(X)}$ is called **odds** and takes values between 0, for $p(X)$ close to 0, and ∞, for $p(X)$ close to 1
- in turn **natural logarithm of odds** is called **logit** – in logistic regression this quantity **linearly depends** on explanatory variables

Logistic regression – cont'd

- however, since the relationship between $p(X)$ and X is **not linear**, the β parameters can not be interpreted as a change in the expected value of $p(X)$ caused by the unit change of X
- the effect of the unit change of X on $p(X)$ **will not be constant** – will also depend on the value of the variable X
- however, the sign of the β parameter can be interpreted as **direction of influence** of unit change of the X into $p(X)$
- a positive estimate of β means that an increase of X increases the probability of $p(X)$ and *vice versa*

Estimation of the logistic regression model

- the logistic model is estimated by th **Maximum Likelihood** (ML) method
- **intuitively**: we are looking for such estimates β, for which the probability of success $\hat{p}(x_i)$ for each observation of i reflects the actual value of the target variable as much as possible
- in other words, we try to find the parameters of the model, which application for modeling $p(X)$ will give values close to 1 for all **real successes** and values close to 0 for all **real defaults**

Similarities to linear regression

- we can measure the precision of estimates by calculating their standard errors
- z statistics play the same role as t statistics in a linear regression – they can be used to test the individual significance of particular variables
- one can use qualitative explanatory variables, decoding them into dummy variables, analogously to linear regression

Multinomial logit

- If the dependent variable assumes k levels with the probabilities of $p_1, p_2, p_3, \ldots, p_k$, one can use **generalized logit** model (also known as **multinomial logit**)
- in this case, one of the values of the explained variable becomes the reference level
- estimated $k-1$ equations model the influence of explanatory variables on the ratio of probabilities that the explained variable will take **individual levels** (in relation to the reference)
- if for example level k is assumed as the reference – then **generalized logits** are equal to:

$$ln(\frac{p_1}{p_k}), ln(\frac{p_2}{p_k}), \ldots, ln(\frac{p_{k-1}}{p_k})$$

Multinomial logit – cont'd

- it is difficult to assume that individual explanatory variables have the same effect on each relation of probabilities
- so in this case each generalized logit is used as **explained variable** in a separate equation with **different** values of estimated parameters
- for example for a dependent variable with three levels, two equations can be estimated:

$$ln(\frac{p_1}{p_3}) = \beta_{0,1} + \beta_{1,1}x_1 + \ldots + \beta_{p,1}x_p$$

$$ln(\frac{p_2}{p_3}) = \beta_{0,2} + \beta_{1,2}x_1 + \ldots + \beta_{p,2}x_p$$

Multinomial logit – cont'd

- as the reference level, it is recommended to select the category of the target variable with **highest frequency** – this ensures greater stability of the model
- in the selection of the reference level, the sensibility of interpretation is also important
- multinomial logit is thus *de facto* a multi-equation model – usually requires a much larger sample than models for binary variables
- the model assumes **Independence of Irrelevant Alternatives**, **IIA** – which means that adding or omitting selected levels of the target variable does not affect the relationship between the remaining levels
- models of this type are often used to model consumer preferences – choosing one of a set of available products

Thank you for your attention