Regressió Models Mètode

Albert Mata

04/10/2018

Contents

En aquesta assignatura treballem amb la situació més senzilla possible entre els models estocàstics, en la qual se suposa que f és una funció lineal $f = b_0 + b_1 X_1 + b_2 X_2 + ...$ Veurem com és possible ajustar un model que ens permeti descriure la relació entre les variables amb finalitats explicatives o predictives a partir de suposar una relació lineal entre una variable resposta quantitativa i unes variables explicatives que poden ser:

- Contínues (Regressió)
- Categòriques (Anàlisi de la variància -ANOVA-)
- Mixtes (Anàlisi de la covariància)

En concret, tractarem aquestes qüestions:

- Ser capaç d'identificar les variables del problema: quina és la variable resposta i quines són les variables explicatives.
- Saber estimar els paràmetres dels models de regressió i saber determinar la precisió de l'estimació.
- Saber plantejar les qüestions d'interès en termes de contrastos d'hipòtesis i saber resoldre-les.
- Saber utilitzar correctament els mecanismes de diagnosi del model i saber com actuar quan es presentin problemes en algun dels requisits de la metodologia.
- Saber resoldre els contrastos d'hipòtesis plantejats.
- Saber quan s'ha d'utilitzar un disseny o un altre per capturar adequadament la informació d'un experiment planejat.

Unitats:

1. Introducció al model lineal
2. Estimació del model lineal
3. Inferència
4. Regressió lineal simple i múltiple
5. Diagnòstics: comprovant les suposicions
6. Mètodes alternatius
7. Selecció de variables i regularització
8. Variables predictores categòriques

Regression analysis is another term used for linear modeling although regressions can also be nonlinear. Remember "regression to mediocrity".

A regression of a categorical variable as Y on quantitative variables as predictors would involve a qualitative response. A logistic regression could be used, but this will not be covered in this class.

1 Unit 1: Introduction to Linear Regression

1.1 Simple Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x, is regarded as the **predictor**, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the **response**, outcome, or dependent variable.

Simple because it has one predictor variable. **Multiple** when it has two or more predictor variables.

Linear because the parameters enter linearly, but the predictors themselves do not have to be linear. For example, this is a linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 log X_2 + \beta_3 X_1 X_2 + \varepsilon$$

But this is **not** a linear model:

$$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \varepsilon$$

We are only interested in statistical relationships (relationship between variables being not perfect). So we won't consider deterministic relationships where an equation exactly describes the relationship between two variables. Linear models may seem rather restrictive, but because the predictors can be transformed and combined in any way, they are actually very flexible. Linear is also used to refer to straight lines, but linear models can be curved.

The equation for the best fitting line is:

$$\hat{y_i} = b_0 + b_1 x_i$$

We call b_0 **intercept** and b_1 **slope**. But our predictions ($\hat{y_i}$) won't be perfectly correct, as they include some prediction error (or **residual error**):

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

A line that fits the data best will be one for which the *n* prediction errors (one for each observed data point) are as small as possible in some overall sense. One way to achieve this goal is to invoke the **least squares criterion**, which says to minimize the sum of the squared prediction errors. Using some calculus we know that values for b_0 and b_1 can be found with:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

The equation for the best fitting line $\hat{y}_i = b_0 + b_1 x_i$ is also called:

- least squares regression line
- least squares line
- estimated regression equation

Regarding b_1 (the slope):

- $b_1 > 0 \rightarrow$ as *x* increases, *y* tends to increase.
- $b_1 < 0 \rightarrow$ as *x* increases, *y* tends to decrease.

1.2 Simple Linear Regression Model

So, what do b_0 and b_1 actually estimate?

When looking to summarize the relationship between a predictor *x* and a response *y*, we are interested in knowing the **population regression line** $E(Y) = \beta_0 + \beta_1 x$. But we can't collect data on everybody in the population, so we rely on using a sample of data from the population to estimate the population regression line.

The least squares regression line $\hat{y} = b_0 + b_1 x$ estimates the population regression line $E(Y) = \beta_0 + \beta_1 x$. That is, the sample intercept b_0 estimates the population intercept β_0 and the sample slope b_1 estimates the population slope β_1. So we use the sample parameters to learn about the population ones.

Note also that observations usually will not exactly match expected values, as **observation = model + error**:

$$y_i = E(Y_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

1.3 Gauss-Markov Conditions

But in order to draw any conclusions about the population parameters β_0 and β_1 we have to make a few more assumptions about the behavior of the data in a regression setting:

- Does it seem reasonable to assume that the errors for each subpopulation are **normally distributed**?
- Does it seem reasonable to assume that the errors for each subpopulation have **equal variance**?
- Does it seem reasonable to assume that the error for one observation is **independent** of the error for another?

These questions are the basis of the four conditions (LINE) that comprise the simple linear regression model:

- The mean of the response, $E(Y_i)$, at each value of the predictor, x_i, is a **L**inear function of the x_i. This is, the mean of the error, $E(\epsilon_i)$, at each value of the predictor, x_i, is **zero**.
- The errors, ε_i, are **I**ndependent.
- The errors, ε_i, at each value of the predictor, x_i, are **N**ormally distributed.
- The errors, ε_i, at each value of the predictor, x_i, have **E**qual variances (denoted σ^2).

What all these conditions really mean is that the errors, ϵ_i, are independent normal random variables with mean zero and constant variance σ^2.

1.4 Common Error Variance

To get an idea of how precise future predictions would be, we need to know how much the responses (y) vary around the (unknown) mean population regression line $\mu_Y = E(Y) = \beta_0 + \beta_1 x$. As stated earlier, σ^2 quantifies this variance in the responses. The smaller this σ^2 value, the more accurate our predictions will be. But we will rarely know its true value as it is a population parameter. The best we can do is estimate it.

Bàsicament, es tractarà de veure si totes les observacions estan més enganxadetes a la línia o més disperses.

Let's recall the formula for the estimate of the variance of the responses, σ^2, when there is only one population. We estimate this variance using the **sample variance**:

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

- $n-1$ because we would really like to use the unknown population mean μ, but we don't know its value and so we estimate it with $\bar{y}$, losing one degree of freedom.

Analogously, the **mean square error** estimates σ^2, the common variance of the many subpopulations:

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

- $n-2$ because in using $\hat{y}_i$ to estimate μ_Y, we effectively estimate two parameters: the population intercept β_0 and the population slope β_1. That is, we lose two degrees of freedom.

In general, $S = \sqrt{MSE}$ estimates σ and is known as the **regression standard error** or the **residual standard error**.

1.5 Coefficient of Determination (r-squared)

We work with different sums of squares values:

- SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line ($\hat{y}_i$) is from the horizontal "no relationship line" (the sample mean or $\bar{y}$).

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

- SSE is the "error sum of squares" and quantifies how much the data points (y_i) vary around the estimated regression line ($\hat{y}_i$).

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

- SSTO is the "total sum of squares" and quantifies how much the data points (y_i) vary around their mean ($\bar{y}$).

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

Note that SSTO = SSR + SSE. The sums of squares appear to tell the story pretty well: they tell us what part of the variation in the response y (SSTO) is just due to random variation (SSE), not due to the regression of y on x (SSR).

SSR divided by SSTO is what we call **r-squared** (r^2) or **coefficient of determination**:

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Considerations:

- Since r^2 is a proportion, it is always a number between 0 and 1.
- If $r^2 = 1$, all of the data points fall perfectly on the regression line. The predictor x accounts for all of the variation in y.
- If $r^2 = 0$, the estimated regression line is perfectly horizontal. The predictor x accounts for none of the variation in y.
- $r^2 \times 100$ percent of the variation in y is explained by the variation in predictor x.
- We calculate r^2 in a different way when $\beta_0 = 0$, so we can't use r^2 to compare models with and without β_0 parameter.

Cal tenir en compte que el valor depèn tant de la inclinació de la línia com de què les observacions hi estiguin més o menys enganxadetes.

An r-squared value is considered large or not depending on what we're studying. In social sciences a 30% can be considered significant, while in engineering that may be unacceptable.

1.6 (Pearson) Correlation Coefficient r

$$r = \pm\sqrt{r^2}$$

Characteristics:

- Since r^2 is between 0 and 1, the correlation coefficient r is always a number between -1 and 1.
- If the estimated slope coefficient b_1 is negative, then r takes a negative sign.
- If the estimated slope coefficient b_1 is positive, then r takes a positive sign.
- If the estimated slope coefficient b_1 is 0, then r must also be 0.
- One advantage of r is that it is unitless, allowing researchers to make sense of correlation coefficients calculated on different data sets with different units.

Interpretation:

- If $r = 1$, then there is a perfect positive linear relationship between x and y.
- If $r = -1$, then there is a perfect negative linear relationship between x and y.
- If $r = 0$, then there is no linear relationship between x and y.

- All other values of r tell us that the relationship between x and y is not perfect. The closer r is to 0, the weaker the linear relationship. The closer r is to -1, the stronger the negative linear relationship. And, the closer r is to 1, the stronger the positive linear relationship. As is true for the r^2 value, what is considered a large correlation coefficient r value depends greatly on the research area.

So, the sign of r tells us whether the relationship is negative or positive. How fairly close r is to 1 or -1 tells us how fairly strong the linear relationship is (maybe even perfect). Finally, the r^2 value tells us the percentage of the variation in y that is reduced (or explained) by taking into account x.

1.7 r-squared Cautions

1. The coefficient of determination r^2 and the correlation coefficient r quantify the strength of a **linear** relationship. It is possible that $r^2 = 0\%$ and $r = 0$, suggesting there is no linear relation between x and y, and yet a perfect curved or curvilinear relationship exists.

2. A large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function (i.e. a curve) might better describe the trend in the data. So, it is important to plot the data and check by ourselves. Good graphics are vital in data analysis. They help us avoid mistakes and suggest the form of the modeling to come. They are also important in communicating our analysis to others.

3. The coefficient of determination r^2 and the correlation coefficient r can both be greatly affected by just one data point (or a few data points). It's good to check how the situation would change without those points.

4. Correlation (or association) does not imply causation.

- An experiment is a study in which, when collecting the data, the researcher controls the values of the predictor variables.
- An observational study is a study in which, when collecting the data, the researcher merely observes and records the values of the predictor variables as they happen.
- In experiments, one can typically conclude that differences in the predictor values is what caused the changes in the response values. This is not the case for observational studies. Unfortunately, most data used in regression analyses arise from observational studies.

5. Ecological correlations (when we first create groups and calculate rates or averages and then study correlations based on those) tend to overstate the strength of an association.

6. A statistically significant r^2 value does not imply that the slope β_1 is meaningfully different from 0. Statistical significance does not imply practical significance. Just because we get a small P-value and therefore a statistically significant result when testing $H_0 : \beta_1 = 0$, it does not imply that β_1 will be meaningfully different from 0.

7. A large r^2 value does not necessarily mean that a useful prediction of the response y_{new}, or estimation of the mean response μ_Y, can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.

1.8 Hypothesis Test for the Population Correlation Coefficient

As always, we want to draw conclusions about populations, not just samples. To do so, we either have to conduct a hypothesis test or calculate a confidence interval. Let's see how to conduct a hypothesis test for the population correlation coefficient ρ. A researcher willing to learn of a linear association between two variables, when it isn't obvious which variable should be regarded as the response should use either of these tests:

- *t*-test for testing $H_0 : \beta_1 = 0$
- ANOVA F-test for testing: $H_0 : \beta_1 = 0$

So, let's use the *t*-test for testing the population correlation coefficient $H_0 : \rho = 0$. Following standard hypothesis test procedures, we first specify the null and alternative hypotheses:

- **Null hypothesis** $\rightarrow H_0 : \rho = 0$
- **Alternative hypothesis** $\rightarrow H_A : \rho \neq 0$

We then calculate the value of the test statistic:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

And finally we use the resulting test statistic to calculate the *P*-value. As always, the *P*-value is the answer to the question "how likely is it that we'd get a test statistic t^* as extreme as we did if the null hypothesis were true?" The *P*-value is determined by referring to a *t*-distribution with *n - 2* degrees of freedom:

- If the *P*-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α level to conclude that there is a linear relationship in the population between the predictor *x* and response *y*".
- If the *P*-value is larger than the significance level α, we fail to reject the null hypothesis. We conclude "there is **not** enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor *x* and response *y*".

As a summary, it is okay to use the *t*-test for testing $H_0 : \rho = 0$ when:

- It is not obvious which variable is the response.
- The (*x*, *y*) pairs are an independent random sample from a bivariate normal population. So, for each *x*, the *y*'s are normal with equal variances. And viceversa. And either *y* can be considered a linear function of *x* or *x* can be considered a linear function of *y*.

1.9 Final comments

- We should never use a regression model to make a prediction for a point that is outside the range of our data because the relationship between the variables might change.
- Overfitting a regression model results from trying to estimate too many parameters from too small a sample. In regression, a single sample is used to estimate the coefficients for all of the terms in the model. That includes every predictor, interaction, and polynomial term. As a result, the number of terms we can safely accommodate depends on the size of our sample. Larger samples permit more complex models, so

if the question or process we're investigating is very complicated, we'll need a sample size large enough to support that complexity. With an inadequate sample size, our model won't be trustworthy. In multiple linear regression, 10-15 observations per term is a good rule of thumb.

- Don't misunderstand r^2 values as suggesting that the predictor x causes the change in the response y. **Association is not causation**. That is, just because a data set is characterized by having a large r-squared value, it does not imply that *x* causes the changes in *y*.
- We usually use r^2 for the coefficient of determination associated with the simple linear regression model for one predictor and R^2 for the multiple coefficient of determination associated with the multiple linear regression model with more than one predictor.

1.10 Carmona 2004

Con el método de los mínimos cuadrados podemos obtener un modelo de regresión simple (con una sola variable independiente):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

O un modelo de regresión múltiple (con más de una variable independiente):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

El término ϵ representa el error: la parte del modelo no controlable por el experimentador debido a múltiples causas aleatorias. Su cálculo explícito nos permite la evaluación del modelo.

Sobre estos modelos se pueden realizar algunas consideraciones:

- Cuando existe un modelo físico teórico y lineal, podemos utilizar la regresión para estimar los parámetros.
- Si el modelo teórico no es lineal, se puede, en muchos casos, transformar en lineal p.ej. mediante la aplicación de logaritmos. O también se puede estudiar un modelo de regresión polinómico.
- En el modelo múltiple intervienen varias variables predictoras. Conviene analizar si son todas necesarias y si son linealmente independientes.
- Conviene verificar que se cumplan las condiciones de Gauss-Markov.
- Conviene tener en cuenta qué ocurre si las variables predictoras son discretas o lo es la variable dependiente.
- Conviene tener en cuenta que pueden faltar datos, así como la existencia de puntos atípicos y puntos influyentes.