JACEK WALLUSCH

FUNDAMENTALS OF QUANTITATIVE ANALYSIS

VERLAG DES KAISER WILHELM INSTITUTS ZU POSEN

PUBLISHED BY VERLAG DES KAISER WILHELM INSTITUTS ZU POSEN

http://www.ikbt.org

First printing, October 2017

Contents

List of Figures

List of Tables

Dedicated to my students.

1 Linear Regression

1.1 Linear Regression - Introductory Remarks

We saw in the previous chapter that the value of correlation coefficient increased in a non-linear manner with the coefficient relating variables X and Y. But what if the dependent variable was a function of more than just one variable? Total discount might be primarily driven by the sales volume, but customer's purchasing potential affects the discount as well. And what if a unit-change in price was met with a more than proportionate decrease of the willingness-to-buy? Correlation coefficient is defined on the $(-1, 1)$ interval, and therefore could not fully capture this response. Another what-if scenario assumes the explanatory variable to be equal to 0. Many B2B producers grant their clients standard discounts depending on channel and customer classification. Discount is granted before the customer starts the bargain, which in other words means that even a zero-valued demand is rewarded. Some producers define only an upper limit for standard discount. The decision on exactly what percentage should be granted is left to the sales representatives. The standard discount varies, and therefore needs to be recovered from the sales data. Again, correlation coefficient would be of very limited use. The procedures helping to recover the coefficients describing the discounting function is called regression. There are many algorithms used to estimate the coefficients, there are many types of explaining variables used to approximate the objective function, and also there are many probability distributions employed in order to estimate the coefficients. In this chapter we introduce the linear regression based on the ordinary least squares (OLS).

1.2 Estimated Coefficients

Perhaps the most important reason for running a regression is the estimation of the marginal effects. What is a marginal effect? A marginal effect is a magnitude of response in dependent variable to a change in explanatory one. Business problems refer to quantified relationships between variables. Total discount is a function of ordered quantities, but how much would discount increase if a customer decided to buy 1 000 pieces of equipment more than previously. A price of used car is a function of millage, but wow much will the price decrease if the mileage increased by 1 000 miles? To solve problems like these an analyst performs regression and calculates the marginal effects.

Marginal Effect: magnitude of response of the explaining variable to changes in explanatory variable.

OLS regression refers to a rare case in which the relationship be-

tween variables is linear. For the linear regression the calculations simplify considerably. Marginal effect is calculated as the first derivative of the estimated function with respect to specified explanatory variables. For instance, if total discount is a function of quantities ordered and customer classification, the marginal effect for quantities ordered will be identical to differentiating the discount function with respect to quantities sold. This operation reduces to estimating the quantities coefficient, because we assumed a linear form of the discount function. Technically speaking, the discount function is:

$$D = \alpha_0 + \alpha_1 Q + \alpha_2 CC + \epsilon_D, \tag{1.1}$$

and the result of differentiating is:

$$\frac{\partial}{\partial Q} D = \alpha_1. \tag{1.2}$$

Eq. (1.2) shows that if the quantities increase by one unit, the discount changes by α_1.

Notice, however, that units may be defined in various manners. If a product is sold in tens of thousands units, one does not simply expect a change by one unit to have a large impact on the discount granted. In the B2B world producers sell in both units and in boxes (of, say, 100 units) to reward the stockists and to reduce the logistic cost. A box, although not belonging to standard measures, may be used as one unit.

Example 1.2.1. — *OLS Estimates and Factor Influencing Prices*. *Dr. Plama, an evil genius to whom 'money is the only Esperanto', does not mix up emotions with business. His favourite car, however, reminds him of his early days at the Heidelberg University, hence making him emotional. As the years go by, the car loses its value, and Dr. Plama realises that further hesitation can be very costly. How much would he lose if he postponed the decision of selling the car by another year? How much would Dr. Plama lose if he drove another 10 000 kilometers?*

Being an evil genius, Dr. Plama scrapes the leading Polish on-line auction websites to gather the data and analyses the data. The OLS estimates revealed that the age coefficient was equal to -10 011, and the millage parameter was equal to -1 357. Therefore, Dr. Plama will lose about PLN 1 350 if he drives another 10 000 kilometers, or slightly more than PLN 10 000 if he hesitates another year.

Age has been calculated subtracting the year of build from 2017. The millage was re-calculated by dividing the actual millage by 10 000.

ANOTHER IMPORTANT INSIGHT gained from regression offers elasticity analysis. Recall that price elasticity of demand is calculated as price divided by quantities sold and multiplied by the first derivative

of the demand function, or more formally:

$$\pi = \frac{p}{q} \times \frac{\partial F(d)}{\partial p}. \tag{1.3}$$

Again, for the linear regression the calculations simplify considerably. No matter how many arguments the demand function has, its first derivative with respect to price will be equal to the estimated price coefficient.

Example 1.2.2. — *OLS Estimates and Price Elasticity*. *Wernham Hogg Co. decided to gradually decrease the price of opti-bright laser copy paper from £298 to £240. After reaching the £240 level, the board of Wernham Hogg Co. requested a report describing the consequences of the cut in price, with special emphasis drawn to response of demand.*

The Data Support Deprtment at Wernham Hogg performed OLS estimations summarised in the table.

Coefficient	OLS Estimate	Std. Error	p-value
intercept	203.732	18.085	0.000
slope	-0.497	0.067	0.000

To elucidate the relationship between changes in demand and prices, the Data Support Department calculated the price elasticity of demand. Since the relationship between quantities and price was linear, the first derivative was equal to the estimated price coefficient in the demand function. Substituting the averages for price and quantities, the Data Support Department obtained the average price elasticity equal to $268/70.67 \times (-0.497) = -1.886$. It meant that, on average, a reduction of price by one £1 would result in an increase of quantities sold by 1.9.

Additionally, an elasticity curve has been drawn to depict the inverse relationship between price elasticity and the prices; we expect the elasticity for the initial price to be much larger (in absolute terms) than for the final, reduced priced. Figure 1.1 summarises the price elasticity analysis.

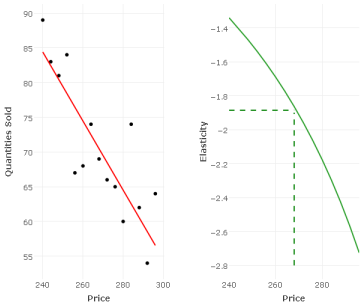

Figure 1.1: Wernham Hogg Co and Price elasticity. The black dots show transactions (combinations of prices and quantities), red line is the OLS line, and the green curve presents the elasticity estimated for price range (£298 to £240). The dashed lines show the elasticity at average price (−1.886).

A LOG-LINEARASATION REQUIRES A SPECIAL COMMENT. In previous examples we assumed that the relationship between prices and quantities was linear. As tempting as it was, sometimes the linear models fitted to the data do not perform particularly well. One way to overcome this obstacle is to take (natural) logarithms from both sides of the estimated equation and to run OLS for the transformed variables. A general form of the price-quants relationship is

Log-linearisation: a transformation of an initially non-linear function to a linear one by taking the (natural) logarithms from both sides of the equation

non-linear and takes the following form:

$$Q = a_0 P^{a_1}. \tag{1.4}$$

Taking logarithms from both sides of (1.4), one obtains:

$$q = a_0 + a_1 p, \tag{1.5}$$

where $q = \ln Q$ and $p = \ln P$. Log-linearisation usually helps to improve the model's fit when the variables take large values (e.g. if we deal with thousands of sold items), but more importantly, it simplifies the calculations of elasticity. The coefficients in (1.4) and (1.5) are identical, therefore once the coefficients of (1.5) are estimated, we can use them to re-write both equations. Recall again equation (1.3), but this time use the first derivative of the demand function (1.4):

$$\frac{\partial \left(a_0 p^{a_1}\right)}{\partial p} = a_1 a_0 p^{a_1 - 1}, \tag{1.6}$$

which leads to:

$$\pi = \frac{p \times a_1 a_0 p^{a_1 - 1}}{a_0 p^{a_1}} = \frac{a_1 a_0 p^{a_1}}{a_0 p^{a_1}} = a_1. \tag{1.7}$$

Therefore, if a log-liearised function is employed, the OLS coefficients become automatically the elasticities. Moreover, the elasticity is constant, which means it does not change along with the changes in prices.

REGRESSION ANALYSIS ALSO HELPS TO OPTIMISE. Imagine a cloud of points depicting all observed combinations of two variables. The OLS line would be drawn in a fashion minimising the distance to the these points. It might be therefore employed to find the optimum value of the explaining variable given a specific value of the explanatory variable.

OLS = oridnary LEAST squares.

Example 1.2.3. — *OLS Estimates and the Optimum Price*. *After concluding business with the Maharajah of Kabur, Dr. Plama has finally decided to sell his car. With the millage and age equal to 66 600 km and 3 years, respectively, what price should Dr. Plama set for his car?*

Dr. Plama employed the OLS estimates to simulate the optimum price:

$$P = 173872.8 - 10011.75 \times A - 1357.895 \times M. \tag{1.8}$$

Substituting 3 and 6.66 for A and M, Dr. Plama priced his car at PLN 134 794.

Dr. Plama used the *point* estimates to obtain the expected value of price under two conditions: millage and age. More specifically,

Recall the definition of conditional expected value from Chapter X.XX.

he used equation (1.8) to approximate the conditional distribution of price, and with the given values of $M = 6.66$ and $A = 3$, estimated the expected, or most probable, price.

What Dr. Plama has simulated for specific values of A and M could be done for the entire range of explanatory variables. Figure 1.2 presents athe actual and predicted prices. Some prices are very close to the fitted line, others are not. Take for instance the cars priced above the PLN 180 000 level. The vehicles are clearly overpriced, but we cannot judge the decision simply by reviewing the distance between the ideal price line and the actual price. Many factors might have influenced the seller's decision, and we cannot evaluate the pricing decision without acquiring the missing information. Bear in mind that data science, econometrics, or statistics can only bring you as far as the data allow to. Bear that in mind especially if you do not plan (or wish) to be directly involved in any sort of data analysis - it will certainly restrain you from asking questions for which there would be no answers.

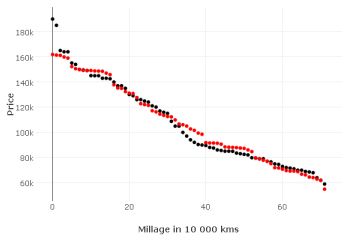

Figure 1.2: Actual Prices and the OLS Line.
The red dotes represent the OLS-predicted price. Black dots depict the actual combination of price and millage.

1.3 *Residuals*

Let us take a closer look at the case of the vastly overpriced vehicles. According to the model, the sellers should have asked for PLN 161 063 and 161 742 . Instead, the prices were set at PLN 189 900 and 184 900. The distance between the actual price and the expected price is called a residual. The residuals capture the behaviour of dependent variable not explained by the model:

Optimum prices calculated for age = 1 and millage = 1.5601 and 2.06

$$\epsilon_i = P_i - \hat{P}_i, \tag{1.9}$$

where $\hat{P}$ is the expected price conditional to age and millage, as it is in (1.8). The term subtracted from the actual price is the forecast based on the OLS model, which produced the expected prices PLN 161 063 and PLN 161 742.

If the model is properly specified, the residuals should have normal distribution with a zero-mean value and a finite variance. Why a zero-mean value? Recall first that the mean value is also the expected value. Now assume for a moment that an estimated model is characterised by a potive-valued residuals. It means that an error term (i.e. the residuals) is on average greater than 0, which in turns means that the modeller has repeatedly committed an error. Moreover, the modeller expected that on average the expected values will deviate from the actual values by some positive distance. Going back to Dr. Plama's dilemma, it would mean that Dr. Plama knew he was going to obtain an 'optimum' price set above the actual, unbiased optimum price. If he did so, he would only be evil, with no genius whatsoever.

1.3.1 *Estimation*

The linear algebra required for estimating the OLS coefficients is not complicated. Three basic operations are needed: transposition, inversion, and multiplication. Let's re-write the price equation (1.8):

$$P = \alpha_0 + \alpha_1 A + \alpha_2 M + \epsilon.$$

1.4 *Significance and Standard Errors*

When a business phenomenon is illustrated by descriptive statistics, an average is usually reported alongside the standard deviation. We want to know what is the average net price, discount granted, or cost, but we also want to know their spread. In the OLS terms, estimated coefficients are used to approximate the average. OLS equivalent of standard deviation is the standard error. A standard routine in statistical inference is to relate the average to standard deviation. If an average price per transaction is equal to £25, and it deviates by £2.5, pricing conditions are considered stable. If the standard deviation approaches average, stability can no longer be claimed. Similar inference can be performed for the estimated coefficients. Moreover, it is much more precise and has well established statistical background. It does not necessarily mean, however, that no controversy surrounds not only the procedure, but also the entire concept.

OLS coefficients and the expected value

The concept mentioned above is called testing for significance. The test statistic is obtained by dividing the estimated coefficient by its standard error. Statistical softwares report all elements required to determine whether or not a coefficient is statistically significant. More technically, to determine whether or not the null hypothesis of insignificance is rejected. Bear in mind that every time significance is mentioned, the null hypothesis actually states the opposite. If the p-value is small, the null of insignificance is rejected, and so the coefficient is statistically significant.

The test statistic is Student-t distributed with the number of degrees of freedom equal to sample size - number of estimated coefficients

Example 1.4.1. — *Individual Significance*. *Sir George Head, OBE, applied for a financial grant to build a bridge between two picks of Mt. Kilimanjaro. His personal assistant, Mr. James Blenkinsop, estimated the impact of various factors on the length of time necessary to finish the works. As pointed out by Mr. Arthur Wilson, the necessity of hiring fully qualified mountaineers might be the leading factors. Based on the experience gathered from the last year's expedition, Sir George questioned the importance of qualified staff. According to Mr. Blenkinsop's estimations covering a sample of 56 similar expeditions and 4 explanatory variables, the qualified-staff-coefficient was 3.22 with a standard error equal to 1.276. Were Mr. Wilson's remarks in order?*

Mr. Blenkinson reported the estimation results. For 52 degrees of freedom and the test value of

56 expeditions and 4 estimated coefficients = 52 degrees of freedom.

$$\frac{3.22}{1.276} = 2.524 \tag{1.10}$$

the p-value was 0.015. The test carried out by Mr. Blenkinson had the null hypothesis specified as *qualified-staff-coefficient is insignificant*. After reviewing the p-value, Mr. Blenkinson rejected the null and proved Sir George wrong (at 1.5% signficance level). Even if right, Mr. Wilson shan't be coming on the expedition, because he's absolutely no confidence in anyone involved in it.

Most textbooks present clear-cut cases with coefficients either highly significant or highly insignificant. The celebrated rule of thumb for significance level, however, cannot be automatically applied to business or economic problems. Boarder-line (in)significance is very likely to be discovered once the results are inspected. Needless to say, the term *boarder-line* is not stricktly defined. As the researchers and analysts are in charge of deciding where the rejection region starts, they also carry the responsibility for their decisions. You should always recall the latter remark every time somebody is refering to statistics in the words of Mark Twain (1907).

Recall again 1%, 5%, and 10% significance level.

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies, and statistics'.

Boarder-line insignificance of individual coefficient may yet not jeopardise the entire model. It is possible to test a hypothesis the estimated are jointly significant. The procedure, often called an overall F-test, has a null hypothesis of the following form:

$$H_0 : \alpha_0 = \alpha_1 = \ldots = \alpha_k = 0, \tag{1.11}$$

which simply reads: all coefficients are jointly equal to 0. Again, checking coefficients' significance we actually test for their insignificance, which means that a small p-value would be in favour of joint significance.

Example 1.4.2. — *Individual vs. Joint Significance*. *After discussing the quality of the rooms with Mrs Alice Richards, Basil Fawlty, the owner of the Fawlty Towers Hotel in Torquay, considers a temporary reduction of prices by 60%. To simulate the possible response of demand, Mr. Fawlty regresses the demand variable against the monthly ratio of satisfied-to-dissatisfied guests, number of guests from Germany (per month), the frequency of gourmet nights hoosted at the hotel(per month), and obviously the price. The first explanatory variable turns out to be insignificant at 15% significance level, whilst the other are found significant. Should Mr. Fawlty exclude the variable of questionable significance from the model?*

Further estimations revealed that the p-value for the overall F-test was equal to 0.024. As the null of joint insignificance has been

rejected, Mr. Fawlty decided to use the original model to evaluate the response of demand to a reduction in price.

Regardless the controversy on systematic mistake of economic and statistical significance (Hoover-Siegler vs. Ziliak-McCloskey), even statistically insignificant coefficients might be important for business analysis. Previous remarks on hypotheses testing, significance level, and p-value are obviously also valid for significance testing.

[With more remarks and text yet to come]

Statistical significance at the 5% or other arbitrary level is neither necessary nor sufficient for proving discovery of a scientific or commercially relevant result, Stephen T. Ziliak and Deirdre N. McCloskey (2009). *Economic and statistical significance are different, but we do not believe that there is any convincing evidence that economists* systematically *mistake the two*, Kevin D. Hoover and Mark V. Siegler (2008).

Point-estimates vs. most probable intervals

STANDARD ERRORS CAN ALSO BE UTILISED in order to replace point-estimate-based simulations and forecasts with most probable intervals. Previously we saw that the optimum price for Dr. Plama's car was equal to PLN 134 794. Using the estimated standard errors we can now present the interval in which the bid price should fall. The standard errors are presented in Table 1.4.

Coefficient	OLS Estimate	Std. Error	p-value
intercept	173 872.8	3 010.9	0.000
age	-10 011.7	840.4	0.000
millage	-1 357.9	272.0	0.000

An interval is obtained by increasing and then decreasing the coefficient values by the values of estimated standard errors. Equation (1.8) takes now the following form:

$$P = (173872.8 \pm 3010.9) - (10011.75 \pm 840.4) \times A - (1357.895 \pm 272) \times M. \tag{1.12}$$

The interval in which the price of Dr. Plama's car should fall is PLN 127 450.4 - 142 137.6.

I am convinced that a successful pricing manager would be equally interested in point-estimates as well as in most probable intervals for the simulated price. The former is very precise, but the latter leaves plenty of rooms for maneuvers and necessary adjustments.

Example 1.4.3. — *Standard Errors and Clustering*. *The Data Department at Cartwright Soap revises its customers classification and regresses the quantities per order against the customer's purchasing potential, . An important business partner, Union Jack Rubber Co.,*

1.5 *Goodness-of-Fit*

Once the coefficients are estimated and they have turned out to be statistically significant, or at least they have met the assumed

criteria for significance, a new set of problems arises. The problems are best summarised by the question: How well does your model fit the data? Precision, as it is in sport, experimental physics, and many other aspects of life, is a virtue no modeller would sacrifice. A forecast based on a model of questionable precision may result in questionable results. We have already seen that precision can be measured. Recall the distance between the forecasted and actual prices in Figure 1.2. Intuition prompts that the smaller the distance, the better the fit. Indeed, the distance is a very important indicative of a goodness-of-fit.

THE MOST POPULAR MEASURE of the goodness-of-fit is the multiple coefficient of determination, more commonly known as R^2 . Coefficient of determination measures the variability of the dependent variable explained by the model. More technically, to estimate the coefficient we need to calculate two sums of squared deviations: from actual average $\hat{y}$, and from the expected value $\bar{y}$:

R^2 pronounced 'r squared'. Please note that in the bi-variate case, the R^2 is simply equal to the squared correlation coefficient (hence its name).

$$R^2 = 1 - \frac{\sum_{i=1}^{T} (y - \hat{y})}{\sum_{i=1}^{T} (y - \bar{y})}. \tag{1.13}$$

You have probably noticed that the numerator in (1.13) is the sum of squared residuals, similar to the term appearing in (1.9). In other words, R^2 relates the variability of the dependent variable to the variability of residuals. Consider again equation (1.8). The denominator is obtained by subtracting the actual prices from the average price, squaring the differences, and summing them up. Similarly, the nominator is the sum of squared distances between observed price and the OLS line. Had the model emulated perfectly the behaviour of prices, the expected values would have become equal the actual values. If that was true, there would be only zero-valued residuals, and the fraction in (1.8) would be 0. Therefore, the closer the fitted (expected, estimated) prices to the actual prices, the better the fit, and the closer to 1 is the R^2.

The R^2 for equation (1.8) is equal to 0.897. Is it large? Yes, it is. The model fits the data reasonably well. What else do we know after inspecting the R^2 of Dr. Plama's model? Cars' age and millage explains 89.7% of price variability. The remaning 10.3% is accounted for unknown factors, not included to the model. But is a large value of R^2 a necessary condition for a model to be helpful?

Example 1.5.1. — *Discount Policy Evaluation*. *According to the official Wernham Hogg Co. pricing guidelines, the discounting decision should be driven by two factors: purchasing potential of customer and size of current purchase. An estimated model revealed that both coefficients were statistically significant, and the R^2 was equal to 0.224.*

A poor fit offers a valuable insight into the discounting habits of Wernham Hogg's sales representatives. The factors mentioned in the official guidelines accounted only for 22.4% of the variability of discounts. Clearly, it suggests that both purchasing potential and size of purchase have significantly influenced the sales reps decisions, there must have been other factors influencing their decisions.

Adjusted R-squared

IF TOO MANY REGRESSORS ARE ADDED TO THE MODEL the fit is usually artificially improved. In order to penalise for the excessive number of explanatory variables, an adjusted R squared, denoted as $\bar{R}^2$, is calculated. The penalising factor is a combination of the number of regressors, k, and the sample size T:

$$\bar{R}^2 = R^2 - \frac{k-1}{T-k}\left(1 - R^2\right). \tag{1.14}$$

Along with the increasing number of explanatory variables, the penalty factor increases. Not surprisingly, the penalty for the price equation (1.8) is small, as the $\bar{R}^2$ is equal to 0.895. After all, there were only three coefficients to estimate, whilst the sample size was 72. In some more fragile cases, however, adjusted R-squared becomes crucial, and the value of $\bar{R}^2$ might even be negative.

Logarithm of likelihood

RESIDUALS ARE ALSO USED to estimate the logarithm of likelihood. This measure too puts a great emphasis on the distance between predicted and actual value of the dependent variable:

$$L = -\frac{T}{2}\left(1 + \ln 2\pi\right) - \frac{T}{2}\ln\frac{\sum_{i=1}^{T}\epsilon_i^2}{T}. \tag{1.15}$$

where the ϵs are the residuals.

Unlike the R^2, the logarithm of likelihood, or log-likelihood for short, is not defined over a specified interval. On a more general notion, a large and positive value of log-likelihood is an indicative of a good fit. Neither *large* nor *small*, however, has not been precisely defined.

When focusing on how good the model fits the data, one should inspect both log-likelihood and the R^2, as the results might sharply contrast. The log-likelihood for the model estimated by Dr. Plama was large, but negative: -769.852, a value that is indeed in a sharp contrast to R^2 = 0.897. The discrepancy depicts the fact that log-likelihood puts even larger emphasis on the (squared) size of the residuals than the R^2 does. The R^2 relates the variability of residuals to the variability of dependent variable; if both residuals and the dependent variable are due to large fluctuations, as it was in the case of Dr. Plama's model, then the R^2 would not capture it. Log-likelihood, however, would and it did. A simple remedy in case of a

linear model is to take logarithms from both sides of the estimated equation. The size of the residuals will be reduced, and so will be their variability. A re-estimated model for the log-linearised variables produced R^2 = 0.839 and log-likelihood = 48.787.

1.6 *Diagnostics*

Bibliography

Index