Class notes 3

Introduction to regression analysis

The two-variable regression model (simple regression)

The two-variable regression model examines the relation between a *dependent variable* (also called *response variable*) Y and an *independent variable* (also called *predictor variable*) X. Our interest lies in modeling the mean response of Y for different values of X. Assume that the (conditional) mean of Y can be adequately described by a linear equation of the form

$$\mu_{Y|X} = B_0 + B_1 X$$

This equation contains the unknown *population parameters* B_0 and B_1, which quantitatively link X to the mean of Y. Instead of writing $\mu_{Y|X}$, we may write the model in terms of Y. To do so, we must include a random *error term* into the formulation

$$Y = B_0 + B_1 X + \varepsilon$$

In this equation, the *intercept* (also called *constant term*) B_0 represents the mean value of Y when $X = 0$. In many cases, it is difficult to attach a meaningful interpretation to the constant term. The *slope* parameter B_1 relates the two variables in the model. The value of B_1 represents the change in the mean of Y resulting from a one-unit increase in X. Obviously, if Y is influenced by X, B_1 must be different from zero.

Objectives of regression analysis

The main objectives of regression analysis are:

- *Estimation*; estimate the unknown parameters B_0 and B_1.
- *Testing hypotheses*; evaluate hypotheses concerning the values of B_0 and B_1.
- *Prediction*; predict Y for given values of X.

Estimation

Suppose that estimates of B_0 and B_1 can be obtained from a realized sample of size n (later we will see how these estimates are computed using the information in the data). Then an expression for the fitted (or estimated) regression is given by

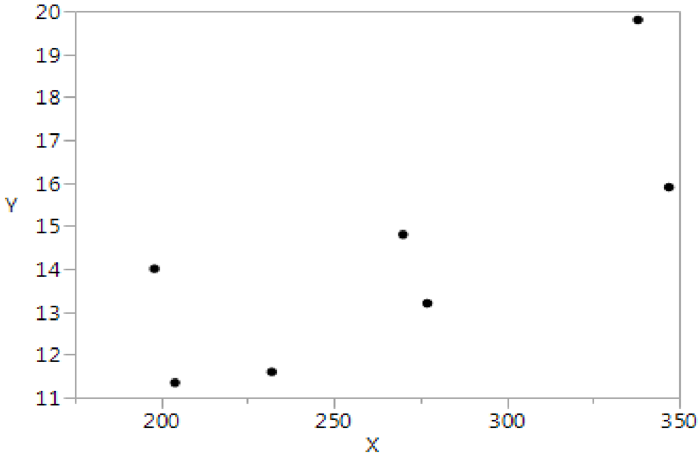
$$\hat{Y} = b_0 + b_1 X$$

Using b_0 and b_1 allows us to estimate ε using the expression

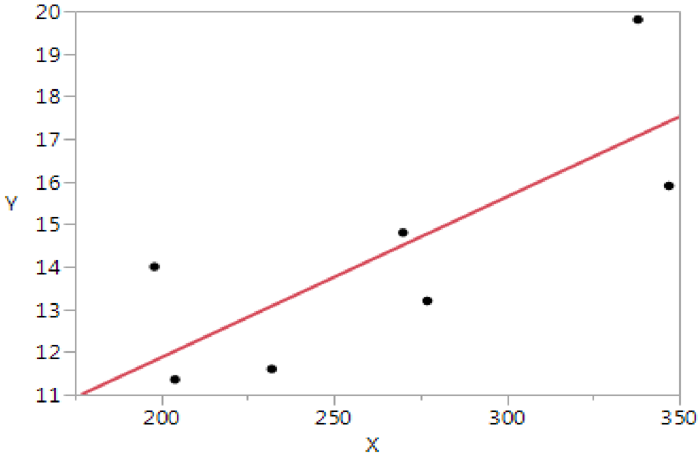
$$\begin{aligned}\hat{\varepsilon} &= Y - (b_0 + b_1 X)\\ &= Y - \hat{Y}\end{aligned}$$

The estimated errors are known as the *residuals*.

Before we get into the mathematical details on how to compute b_0 and b_1, it is useful to visually examine a plot of the data. As an example, let $\{y_i, x_i\}_{i=1}^{n}$ denote a set of n pairs of measurements (or observations) on Y and X (in this case $n = 7$). A scatter plot of the data is given by:



From the plot, we immediately see what appears to be a positive relationship between Y and X. We may ask the following questions; how can we fit a straight line of the form $\hat{Y} = b_0 + b_1X$ to the data? That is, how do we find the parameters b_0 and b_1 so that the line fits the data in the best possible way? Again, look at the plot, but now with the regression line fitted



The regression line predicts the mean of Y given values of X. The best line will minimize the deviation between the actual points y_i and the fitted points $\hat{y}_i$. The deviations are the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$. To avoid the problem that positive and negative deviations cancel out when using summation, it is useful to consider the squared deviations

$$\hat{\varepsilon}_i^2 = (y_i - \hat{y}_i)^2$$

The estimation process then becomes a problem of finding the (estimated) parameters that jointly minimize the sum of the squared residuals. Mathematically, we formulate this as

$$\min_{b_0, b_1} L = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

where the function L can be written in the form

$$
\begin{aligned}
L &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2
\end{aligned}
$$

To obtain the estimated parameters, first differentiate L with respect to b_0, b_1, and then set the resulting equations equal to zero

$$
\begin{aligned}
\frac{\partial L}{\partial b_0} &= \frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \\
\frac{\partial L}{\partial b_1} &= \frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0
\end{aligned}
$$

Fortunately, the solutions to these equations are simple

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

and

$$
b_0 = \bar{y} - b_1 \bar{x}
$$

The estimated error variance is obtained by

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\
&= \frac{RSS}{df}
\end{aligned}
$$

where k is the number of slope parameters in the model and $RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$ is the *Residual Sum of Squares*. For the two-variable regression model, $k = 1$. The estimators of B_0 and B_1 are known as *Ordinary Least Squares* or just OLS. Regression analysis and the use of OLS is by far the most implemented statistical procedure in empirical research in the social sciences.

Note that a set of regression estimates are *specific for the sample used in their estimation.* Thus, different samples, taken from the same population, give different values of b_0 and b_1. The uncertainty in b_0 and b_1 can be estimated. The standard errors associated with b_0 and b_1 are denoted $SE(b_0)$ and $SE(b_1)$, and are given by the expressions

$$
SE(b_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
$$

$$SE(b_1) = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Performing the computations by hand is quite cumbersome. Thus, we typically use computers to obtain the parameter estimates and their associated standard errors.

Multiple regression

In the more general case when there are k independent variables, the regression model takes the form

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

In this equation, B_0 represents the mean of Y when $X_1, \dots, X_k$ are simultaneously zero. The value of the slope parameter B_1 represents the change in the expected value of Y resulting from a one-unit increase in X_1, *holding* $X_2, \dots, X_k$ fixed. Similar interpretation applies for $B_2, \dots, B_k$.

The fitted regression "line" (in this case it is actually a plane in a multidimensional space) takes the form

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

The estimated errors follows from the expression

$$\begin{aligned}\hat{\varepsilon} &= Y - (b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k)\\ &= Y - \hat{Y}\end{aligned}$$

Estimation

Estimation works the same as for the two-variable regression model with the slight modification that the loss function has to be minimized with respect to $b_0, \dots, b_k$ (instead of just b_0 and b_1). Thus, based on the data, we obtain the parameter estimates by

$$\min_{b_0,\dots,b_k} L = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

and

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\\ &= \frac{RSS}{df}\end{aligned}$$

The estimated uncertainty associated with the $b_0, \dots, b_k$ are given by $SE(b_0), \dots, SE(b_k)$.

Assumptions

For completeness, we briefly state the underlying assumptions of the linear regression model:

- The data $\{y_i, x_{1i}, x_{2i}, \dots, x_{ki}\}_{i=1}^{n}$ is obtained in a way that satisfy the requirements for a random sample.
- The regression model is linear, which implies that the model can be written in the form

  $$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k + \varepsilon$$

- The mean value of ε (given the Xs) is zero (this assumption is somewhat related to the previous one).
- The variance of ε (given the Xs) is constant (or homoscedastic).
- There is no exact linear relationship among the Xs.
- Distributional assumption (matters for hypothesis testing):

  a. *Small sample size*: if n is small, it must be assumed that the error term is normally distributed. That is, $\varepsilon \sim N(0, \sigma^2)$ (given the Xs).

  b. *Large sample size*: if n is large, the CLT applies and no distributional assumption is needed.

A measure of model fit

It is desirable to have some measure of how well the regression model fits the data. In other words, how much of the variation in the dependent variable Y is accounted for by the variation in the independent variables $X_1, \dots, X_k$?

Based on the data, the fitted is

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

and

$$\begin{aligned} y_i &= b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + \hat{\varepsilon}_i \\ &= \hat{y}_i + \hat{\varepsilon}_i \end{aligned}$$

From these equations (and the assumption that ε and the Xs are unrelated), we can write

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

or simply

$$TSS = ESS + RSS$$

where TSS is short for *Total Sum of Squares*, ESS is short for *Estimated Sum of Squares* and RSS is short for *Residual Sum of Squares*. The decomposition of TSS (into ESS and RSS) allows us to obtain the fraction of variation accounted for by the Xs.

This fraction is referred to as R^2, and is obtained by

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

The R^2 is a scale free measure of fit that is bounded between 0 and 1, and hence, its interpretation is straightforward. When RSS is small compared to TSS, R^2 is close to 1, which indicates a strong model fit. On the other hand, when RSS is large compared to TSS, R^2 is close to 0, which indicates a weak model fit.

For the two-variable regression model, R^2 can be derived directly from the estimated correlation between X and Y in the following way

$$R^2 = r_{XY}^2$$

We immediately see that if the measured correlation (positive or negative) is strong, R^2 will be close to 1. The corresponding expression for the multiple regression model is

$$R^2 = r_{Y\hat{Y}}^2$$

An important property of R^2 is that it increases, or at least stays the same, as more independent variables are added to the model. For instance, consider the following two models

$M1$: $\quad Y = B_0 + B_1X_1 + B_2X_2 + \varepsilon$

$M2$: $\quad Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \varepsilon$

Even if X_3 in the second model is an *irrelevant* variable, R_{M2}^2 will in practice be larger than R_{M1}^2. Thus, it becomes impossible to rank the two models in terms of statistical adequacy using R^2. To avoid this problem, we may use the an alternative measure called the adjusted R^2. This measure is obtained by

$$\bar{R}^2 = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right]$$

Where, as previously, k is the number of slope parameters in the model (such that $k + 1$ is the total number of parameters). The $\bar{R}^2$ takes into account the number of independent variables in the model. Thus, if X_3 is an irrelevant variable, estimating the two models from above, one would expect R_{M1}^2 to be larger than R_{M2}^2.

Inference

Testing the individual parameters

As previously stated, one aim of regression analysis is to test model parameters. More specifically, we would like to evaluate claims about the effect of the independent variables on the dependent variable. There are three ways to formulate the alternative hypothesis. The appropriate formulation depends on the research question.

A test of significance may be one-tailed:

$$\begin{aligned} H_0 &: B = B^* \\ H_A &: B > B^* \end{aligned}, \qquad \begin{aligned} H_0 &: B = B^* \\ H_A &: B < B^* \end{aligned}$$

or two-tailed

$$\begin{aligned} H_0 &: B = B^* \\ H_A &: B \neq B^* \end{aligned}$$

It can be shown that under the null-hypothesis, the standardized OLS estimators follow a t-distribution with $n - k - 1$ degrees of freedom.

The test statistic (or t-statistic) is given by

$$t = \frac{b - B^*}{SE(b)}$$

Typically, we are interested in the effect of the independent variables on the dependent variable, which means that we put less emphasis on testing the significance of the constant term B_0. Note that there are exceptions.

There are several approaches as how to carry out the test:

- The use of critical values and rejection regions
- The use of p-values
- The use of confidence intervals

For any these three approaches, one must first choose a significance level α. Typical values of α are 0.01, 0.05, or 0.1.

The critical value approach

When an appropriate α has been chosen, we need to determine the critical value of the test. Note that the critical value for a one-tailed significance test is different from the critical value of a two-tailed test. The critical value of a one-tailed test is either t_α or $-t_\alpha$, depending on the formulation of the alternative hypothesis. The critical value of a two-tailed test is $t_{\alpha/2}$.

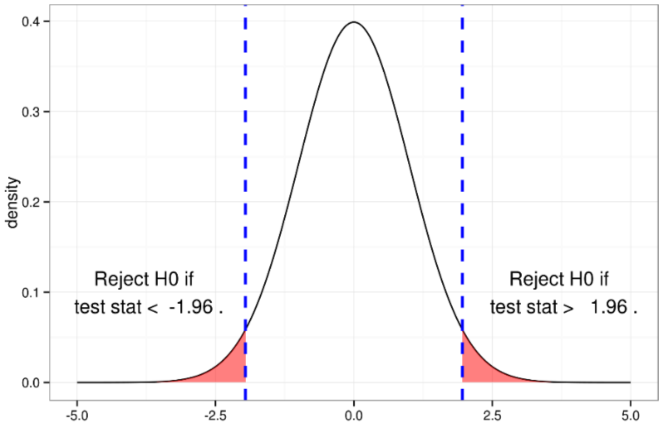
Rules for when to reject the hull-hypothesis are summarized in the following table:

H_0	H_A	Reject H_0 when
$B = B^*$	$B > B^*$	$t > t_\alpha$
$B = B^*$	$B < B^*$	$t < -t_\alpha$
$B = B^*$	$B \neq B^*$	$\lvert t \rvert > t_{\alpha/2}$

where t to denotes the realized value of the test statistic.

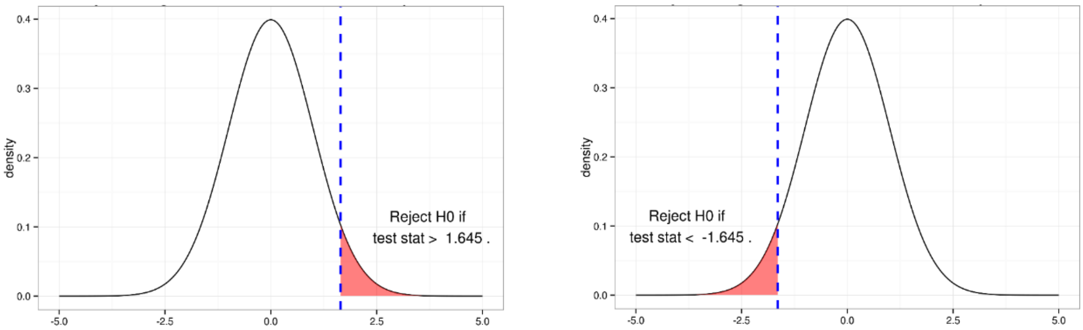
Graphically, the decision rules (or rejection regions) are depicted in the following way:

Regions of rejection for a two-tailed 5 % hypothesis test (t-distribution $n > 100$)



For a one-tailed tests, the 5 % rejection region is located solely in one tail of the distribution.

Regions of rejection for a one-tailed 5 % hypothesis test (t-distribution $n > 100$)



The p-value approach

Instead of using the critical values to determine when to reject the null hypothesis, we may perform the test using p-values.

The p-value is defined the same way as previously stated. That is:

> *The p-value of a test is the probability of obtaining a test statistic as extreme, or more extreme, as the one realized from the sample, given a true null hypothesis.*

So, how to obtain the p-value? The following table summarize how to obtain the p-value for the three formulations of the alternative hypothesis:

H_0	H_A	p-value =
$B = B^*$	$B > B^*$	$P(T \geq t \mid H_0)$
$B = B^*$	$B < B^*$	$P(T \leq t \mid H_0)$
$B = B^*$	$B \neq B^*$	$2 \cdot P(T \geq \|t\| \mid H_0)$

The rejection rule is this case is

Reject the null hypothesis when the *p*-value $< \alpha$

One problem of using the *p*-value approach is that it generally requires computational power to obtain the *p*-value. In some special cases, we may be able to approximate the *p*-value using the t-table. Also, recall that when df tend to infinity (when df becomes a large number), the t-distribution tend to a standard normal distribution. Thus, we can we use the normal table to find the *p*-value when n is large.

Confidence intervals for the parameters

A $100 \cdot (1 - \alpha)\%$ confidence interval for the true parameter B is given by

$$b \pm t_{\alpha/2} \cdot SE(b)$$

When computing a confidence interval, one is applying a procedure for which the true population parameter (in this case B) is contained in the interval in $100 \cdot (1 - \alpha)\%$ of the cases. In other words, if we were to compute a large number of confidence intervals based on a large number of samples (drawn from the same population), we would expect the true population parameter B to be located in the intervals in $100 \cdot (1 - \alpha)\%$ of the cases. *It would be wrong to claim that there is a* $100 \cdot (1 - \alpha)\%$ *chance that the true parameter falls within an already computed interval.*

The confidence interval approach for testing B (only two-tailed hypothesis)

The test of significance (the ordinary *t*-test) and the confidence interval approach always give the same conclusions. The confidence interval approach is an intuitive and an appealing way of testing the individual parameters.

Consider the following (tow-tailed) hypothesis

$$H_0 : B = B^*$$

$$H_A : B \neq B^*$$

The rejection rule is that if B^* is not contained in the interval, the null hypothesis is rejected. Note that our presentation is limited to only consider two-tailed hypothesis tests. One could expand the framework to also consider one-tailed hypothesis testing.

Testing multiple restrictions

The F-test is used to evaluate hypotheses involving more than one parameter. Such hypotheses are typically referred to as *joint hypotheses*. The versatility of the F-test allows us to handle a variety of testing problems.

The testing procedure is slightly more involved than applying the simple t-test. Broadly speaking, the test involves comparing two models, one that is *restricted* and one that is *unrestricted*. An example will illustrate how to perform a test for multiple restrictions.

Example (testing two restrictions)

Consider the following two regression models

M1 (the restricted model): $Y = B_0 + B_1X_1 + B_2X_2 + \varepsilon$

M2 (the unrestricted model): $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \varepsilon$

As can be seen, M1 is nested within M2. The difference between the two models is that M1 is restricted in the sense that $B_3 = 0$, $B_4 = 0$. The question becomes; which model is statistically better? Stated differently, does X_3 and X_4 *jointly* add any predictive power to the model? We may use the R^2 a measure of fit to determine which model is better. However, since $R^2_{M2} \geq R^2_{M1}$, R^2 is not a suitable measure to rank the two models. Alternatively, we may use $\bar{R}^2$. An even better alternative is to perform formal test, which in this case will be the F-test.

The hypothesis to be evaluated is

$$H_0 : B_3 = 0, B_4 = 0$$

$$H_A : \text{At least one } B_j \neq 0 \text{ for } j = 3,4$$

The number of restrictions in this case is 2 (corresponding to the number of equality signs under the null hypothesis).

To perform the F-test, we need to obtain the test statistic, also known as the F-statistic. The F-statistic is based on the *Residual Sum of Squares* of both the restricted model (RSS_r) and the unrestricted model (RSS_{ur}), where $RSS_r \geq RSS_{ur}$. If the unrestricted model is a significantly better predictor of Y, as compared to the restricted model, RSS_{ur} will be considerably smaller than RSS_r (recall that RSS is related to model fit).

The F-statistic is given by

$$F = \frac{\frac{RSS_r - RSS_{ur}}{m}}{\frac{RSS_{ur}}{n-k-1}}$$

where m is the number of restrictions.

The decision rule is to reject the null-hypothesis when $F > F_{\alpha,(m,n-k-1)}$, where $F_{\alpha,(m,n-k-1)}$ is the critical value at the α-level. We find the critical value in the F-table. In the table, m is the numerator df (denoted df_1) and $n-k-1$ is the denominator df (denoted df_2). Alternatively, we may use the *p*-value approach. As for the t-test, the null-hypothesis is rejected when the *p*-value is less than α.

Recall that R^2 is derived directly from RSS. Thus, it is no surprise that the F-statistic can be obtained from the R^2 of the two models, the restricted and the unrestricted.