Quantitative Analysis and Empirical Methods

Regression

Jan Rovny

Sciences Po, Paris, CEE / LIEPP

Overview

- Logic of regression
- Interpreting regression results
- Linearity of OLS and curvilinear relationships
- Regression assumptions
- Decomposition of sample variance
- Goodness of fit

Regression

- How do we go about addressing change and response in variables?
- Correlation only tells us the extent to which pairs of variables are a linear function of each other. It does not tell us how change in one translates into change in another.
- Correlation also treats both variables as identical. Correlation between 'smile' and 'flowers' is the same as the correlation between 'flowers' and 'smile'.
- Our answer is Regression:
- Regression models one variable as a **dependent** variable, which is predicted by an **independent** variable (also known as the predictor).
- We write that $y_i = \beta_0 + \beta_1 x_i + \epsilon$

The Regression Model

- $y_i = \beta_0 + \beta_1 x_i + \epsilon$
- This models a relationship between Y - the dependent variable and X - the predictor.
- $\boldsymbol{\beta_0}$ is the intercept – the expected value of y when $x = 0$
- $\boldsymbol{\beta_1}$ is the slope coefficient. It describes the direction and steepness of the regression line. It is the expected change in y for a unit change in x, holding all else constant. This is the most important piece of information for us, because it describes the relationship between x and y.
- $\boldsymbol{x_i}$ is the predictor, treated as fixed (that is non-random or 'error-less') variable.
- $\boldsymbol{\epsilon_i}$ is the stochastic (random) component. It expresses the disturbance or error term. It includes measurement error on y, omitted predictors and idiosyncratic sources of behavior. Error is a very interesting animal (to be discussed later)...

Example 1

- A real example from *Morg05.dta* dataset on wages in the U.S.
- I am interested in seeing how 'gender' affects 'wage.' I thus regress: $wage = \beta_0 + \beta_1 sex + \epsilon_i$
- In R: `model<-lm(wage~sex)`
- My results are the following: $wage = 19.350 + (-3.629)sex$
- What does this mean?
  - $\beta_0 = 19.350$ This is telling us the average value of y when $x = 0$. When does $x = 0$?

Example 1

- A real example from *Morg05.dta* dataset on wages in the U.S.
- I am interested in seeing how 'gender' affects 'wage.' I thus regress: $wage = \beta_0 + \beta_1 sex + \epsilon_i$
- In R: `model<-lm(wage~sex)`
- My results are the following: $wage = 19.350 + (-3.629)sex$
- What does this mean?
  - $\beta_0 = 19.350$ This is telling us the average value of y when $x = 0$. When does $x = 0$?
  - $x = 0$ means that `sex`=0, that is `sex`=male. Therefore, 19.35 is the average wage of a male.

Example 1

- A real example from *Morg05.dta* dataset on wages in the U.S.
- I am interested in seeing how ‘gender’ affects ‘wage.’ I thus regress: $wage = \beta_0 + \beta_1 sex + \epsilon_i$
- In R: `model<-lm(wage~sex)`
- My results are the following: $wage = 19.350 + (-3.629) sex$
- What does this mean?
  - $\beta_0 = 19.350$ This is telling us the average value of y when $x = 0$. When does $x = 0$?
  - $x = 0$ means that `sex`=0, that is `sex`=male. Therefore, 19.35 is the average wage of a male.
  - $\beta_1 = -3.629$ This is telling us the expected change in y when x changes by 1. What does that mean?

Example 1

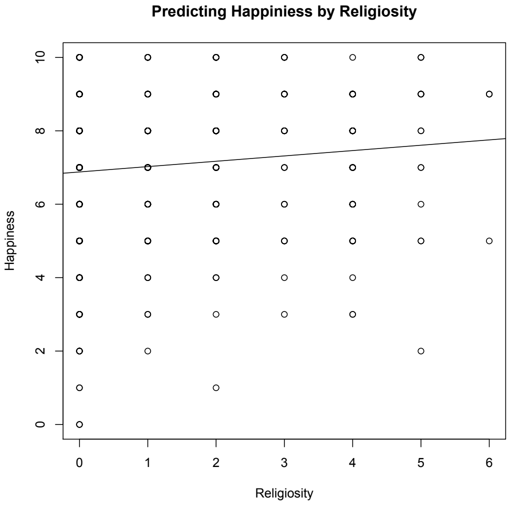
- A real example from *Morg05.dta* dataset on wages in the U.S.
- I am interested in seeing how 'gender' affects 'wage.' I thus regress: $wage = \beta_0 + \beta_1 sex + \epsilon_i$
- In R: `model<-lm(wage~sex)`
- My results are the following: $wage = 19.350 + (-3.629)sex$
- What does this mean?
  - $\beta_0 = 19.350$ This is telling us the average value of y when $x = 0$. When does $x = 0$?
  - $x = 0$ means that `sex`=0, that is `sex`=male. Therefore, 19.35 is the average wage of a male.
  - $\beta_1 = -3.629$ This is telling us the expected change in y when x changes by 1. What does that mean?
  - When x shifts by 1, that is shifts from 0=male to 1=female. Hence -3.690 is the average effect of being a woman on wage. It decreases by $3.69 per hour. An average female wage is thus $19.35 - 3.62 = 15.721$.

Example 2

- Does being more religious lead to greater perceived happines?
- $happy = \beta_0 + \beta_1 religiosity + \epsilon_i$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.8792	0.0813	84.65	0.0000
Religiosity	0.1455	0.0463	3.15	0.0017

Regression Graph



How does it work?

- $\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 * \bar{X}$
- $\hat{\beta}_1$ the covariance of XY divided by the variance of X. It minimizes the sum of squares of the residuals
- This is the so-called **Ordinary Least Squares Estimator**:

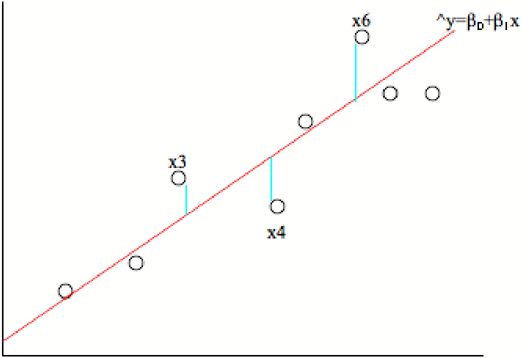


Figure: default

Why an Estimator?

- $\hat{\beta}$s are Estimators, because they estimate the true relationship between X and Y, which is β. (We know samples, but we care about populations, which we do NOT know.)
- Since $\hat{\beta}$s are derived from samples, it is clear that they are likely to vary from sample to sample. The $\hat{\beta}$s are *estimates*, and they thus have a certain variance.
- We can think of estimator variance as the **uncertainty** about the point estimate (our best guess at the true value of β).

Estimator Variance

- From our sample, we know the standard error of the regression $\hat{\sigma} = \sqrt{\frac{\Sigma e_i^2}{N-2}}$ (note that we burn 2 d.f. estimating β_0 and β_1)
- This is the standard deviation of the Y values around the estimated regression line.
- We can derive the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$, and consequently their **standard error**: $s_{\hat{\beta}_0} = \sqrt{\frac{\Sigma x_i^2}{N*\Sigma(x_i-\bar{X})^2}}\sigma$, and $s_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\Sigma(x_i-\bar{X})^2}}$.
- What will be the distribution of our $\hat{\beta}$s?

Estimator Variance

- From our sample, we know the standard error of the regression $\hat{\sigma} = \sqrt{\frac{\Sigma e_i^2}{N-2}}$ (note that we burn 2 d.f. estimating β_0 and β_1)
- This is the standard deviation of the Y values around the estimated regression line.
- We can derive the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$, and consequently their **standard error**: $s_{\hat{\beta}_0} = \sqrt{\frac{\Sigma x_i^2}{N*\Sigma(x_i-\bar{X})^2}}\sigma$, and $s_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\Sigma(x_i-\bar{X})^2}}$.
- What will be the distribution of our $\hat{\beta}$s?
- Remember, the Central Limit Theorem??? Yes, it will be NORMAL!

Estimator Variance

- From our sample, we know the standard error of the regression $\hat{\sigma} = \sqrt{\frac{\Sigma e_i^2}{N-2}}$ (note that we burn 2 d.f. estimating β_0 and β_1)
- This is the standard deviation of the Y values around the estimated regression line.
- We can derive the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$, and consequently their **standard error**: $s_{\hat{\beta}_0} = \sqrt{\frac{\Sigma x_i^2}{N * \Sigma(x_i - \bar{X})^2}}\sigma$, and $s_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{X})^2}}$.
- What will be the distribution of our $\hat{\beta}$s?
- Remember, the Central Limit Theorem??? Yes, it will be NORMAL!
- It follows that $\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}} \sim N(0,1)$ and $\frac{\hat{\beta}-\beta}{s_{\hat{\beta}}} \sim t_{n-2}$
- This is the **t-test** we can see in our statistical output.

The t-test

- The t-test in our statistical output asks the most fundamental question: Is $\hat{\beta} = 0$?
- This is effectively asking, **is my estimate of $\hat{\beta}$ sufficiently different from 0?** Does my variable have any effect?
- Or **What is the chance that the true value of β could be 0?**
- Easy, we did this before with our z- and t-tests.
- We generally take the 95% confidence interval and ask ourselves whether 0 lies outside this interval.
- This tells us the **statistical significance** of a variable

	Estimate	Std. Error	t value	Pr(>|t|)	[95% Conf.	Int.]
(Intercept)	6.8792	0.0813	84.65	0.0000	6.719	7.038
Religiosity	0.1455	0.0463	3.15	0.0017	0.054	0.236

Review

- Regression equation: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- The logic is that we minimize the squared residuals by fitting the 'best line' through the data.
- From our sample data, we obtain **point estimates** of $\hat{\beta}_0$, the intercept, and $\hat{\beta}_1$, the slope coefficient.
- The point estimates give us the 'best guess' of the values
- Then, from the errors we are able to establish the **standard error** of our estimators $(\hat{\beta}_0, \hat{\beta}_1)$
- The standard error tells us the dispersion (or spread) of our estimators, effectively telling us how certain we are about our point estimates.

Interpreting a Regression

- Substantive Significance
  - How strong is the effect of X on Y? Does a change in X lead to a substantial change in Y.
  - This is a matter of argument, but you should report for example that 'having a BA, as opposed to a highschool diploma increases your expected income by so many dollars.'
- Statistical Significance
  - How sure are we about our result? Is it significantly different from 0?
  - This has to do with the size of the standard error of our estimator. We must choose a **level of significance**, which is usually 95%. Then we perform a t-test, on whether our point estimate is significantly different from 0. If yes, we can say that our estimator 'is statistically significant at the .05 level.'
  - An easy way to check what level our estimator is significant at, we look at the **p-value** reported by R.

Regression Output

Predicting Happiness (0-10) with Religiosity (0-6), ESS CZ

	Estimate	Std. Error	t value	Pr(>|t|)	[95% Conf.	Int.]
(Intercept)	6.8792	0.0813	84.65	0.0000	6.719	7.038
Religiosity	0.1455	0.0463	3.15	0.0017	0.054	0.236

- The results of this model suggest that Religiosity is a substantively and statistically significant predictor of happiness.
- Substantively, attending religious services every day as opposed to never increases the expected happiness by about 9% ($6 * 0.1455 = 0.873$, happy is a 10 point scale, thus roughly 0.9 points out of 10)
- Statistically, our t-value of 3.15 is significant at the .05 level (as well as at the .01 level).
- Shortcuts:
  - 1) t-value$>$ 2; 2) confidence interval does not pass through 0

Linearity of Linear Regression

- 'Linear' Regression means that that the β coefficients of the regression are linear, that is they are raised to the first power only.
- Linear Regression, however, can model non-linear relationships between X and Y. That is, linear regression need not be linear in the variables.
- We can thus fit a quadratic model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, which models a curvilinear relationship between X and Y.
- R will fit the βs in such a way as to minimize the square residuals, that is it will draw the 'best fitting' regression curve.
- Example: modeling curvilinear relationships (Functions Calculator)

Regression Assumption #1

- **I. The most important assumption**
- *1. Model is correctly specified*
- Formally: *Mean Independence:* $E(\epsilon_i) = 0$, which means that the mean value of ϵ does not depend on any of the predictors.
- Model includes all relevant predictors in the correct functional form (squares, interactions etc.).
- If this does not hold, there is omitted variable bias, the OLS estimator is biased and inconsistent = WRONG
- Specification error is a central problem for which there is no statistical solution.
- **We must turn to theory!**

Assumptions about Errors

- *2. Linearity*: y is a linear function of the xs.
  - Violation of 1. and 2. causes point estimate bias!
- *3. Normality*: $\epsilon_i \sim N(0, \sigma^2)$ We assume that the error is normally distributed (around the regression line).
  - 3. is important for inference, allows us to use t-tests.
- *4. Homoscedasticity*: $Var(\epsilon_i) = \sigma^2$: variance of errors is constant.
- *5. Nonautocorrelation*: $Cov(\epsilon_i, \epsilon_j) = 0 \ (i \neq j)$, errors are independent. (Problem in time-series data.)
  - 4. and 5. do not effect point estimates, only determine the standard errors.

Decomposition of Sample Variance

- Our main quest is to explain the variance in the dependent variable Y
- The values of Y differ because of the relationship between Y and X, and because of random error.
- The question is, how much of the observed variation on Y is caused by X and how much of it is due to error.
- This effectively tells us how much of the variance of Y is explained by our model (X) and how much of it is due to (unexplainable) error.
- It is thus important to 'decompose' the variance of Y:

Decomposition of Sample Variance 2

- **Total Sum of Squares (TSS)** $= \sum(Y_i - \bar{Y})^2$
  - Is a summary measure of the distances of observations on Y from the mean. It is the total variation of the actual Y values about their sample mean.
- **Regression Sum of Squares (RSS)** $= \sum(\hat{Y}_i - \bar{Y})^2$
  - The vertical distance of the regression line from $\bar{Y}$ is the variation of Y ascribed to X
- **Error Sum of Squares (ESS)** $= \sum e_i^2$
  - The vertical distance of the observed point Y_i from the regression line (or the residual) is the variation in Y ascribed to error

Therefore:

$$\underbrace{\sum(Y_i - \bar{Y})^2}_{\text{TSS}} = \underbrace{\sum(\hat{Y}_i - \bar{Y})^2}_{\text{RSS}} + \underbrace{\sum e_i^2}_{\text{ESS}}$$

Decomposition of Sample Variance 3

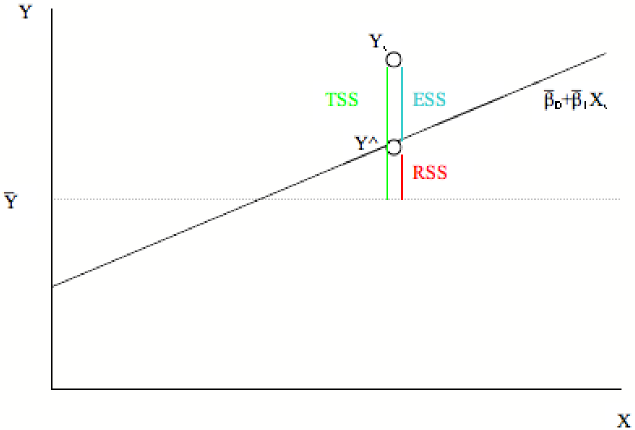


Figure: Variance Decomposition

Goodness of Fit

- This leads to the measure of 'goodness of fit' R^2, which is fundamental for telling us how well our model does in explaining our dependent variable Y
- R^2 is the ratio of variance explained by X and the total variance:

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

- R^2 is bounded between 0 and 1, where 0 means no variance of Y is explained by X and 1 means all variance of Y is explained by X (there is no error).
- R^2 effectively tells us how 'tightly' our observations lie around the regression line.