An Introduction to Statitical Learning with Applications in Python by James et al.:

Chapter 3 Linear Regression

Summary by Ryoko Ito

London, November 2023

Outline

- 3.1 Simple linear regression
  - 3.1.1 Estimating the coefficients, 3.1.2 Assessing the accuracy of the coefficient estimates, 3.1.3 Assessing the accuracy of the model (Residual standard error, R2)
- 3.2 Multiple linear regression
  - 3.2.1 Estimating the regression coefficients, 3.2.2 Some important questions (Hypothesis testing, Is there a relationship? Which variable is important? Is the model fit good? Can we predict?)
- 3.3 Other considerations in the regression model
  - 3.3.1 Qualitative predictors (binary variables), 3.3.2 Extensions of the linear model (interaction / higher order variables), 3.3.3 Potential problems (assumption violations)

3.1 Simple linear regression

Only 2 variables, X and Y. (Random variables. Their realizations will be denoted by x and y.)
Example: X = TV advertising (in £, say), Y = sales (£)
We ask: *Is there a relationship between X and Y?*

Assume a linear relationship ("regress Y on X"):

$$Y \approx \beta_0 + \beta_1 X. \tag{1}$$

This is usually written as

$$\begin{aligned} & Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim (0, \sigma^2) \\ \Rightarrow \quad & \mathbb{E}[Y|X] = \beta_0 + \beta_1 X. \end{aligned}$$

i.e. the centre of the distribution of Y is given by the RHS of (1). ϵ determines the nature of variability around the centre.

Aside

Assumptions / empirical features of ϵ affects estimation procedure (e.g. independence, Gaussian).

3.1 Simple linear regression

A typical analysis seeks to estimate β_0 and β_1.
X and Y are random variables. Their observations (realizations) are denoted by x and y.

If we denote the estimated quantities using $\hat{\cdot}$, our model predicts y as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The estimation (prediction) error at x is:

$$e \equiv y - \hat{y} = y - \hat{\beta}_0 + \hat{\beta}_1 x.$$

This e is not necessarily related in any way to ϵ. Whether the model is “good” in some sense is determined by whether e satisfies what we assume about ϵ. (Diagnostics.)

3.1.1 Estimating the coefficients

We have pairwise observations (x_i, y_i) for $i = 1, \ldots, n$.
From these, our model gives predictions $\hat{y}_i$ and errors e_i.

One way of quantifying the overall error (*loss*) of the model is the *residual sum of squares* (RSS) given by $\sum_{i=1}^{n} e_i^2$.

We can try to find βs that minimize this loss function (the method of least squares).

Aside

RSS is just one of many loss functions. The loss function defines the analysts' preferences.

- Analysts using RSS must really dislike larger errors. (Penalize squarely.)
- If Y is such that outliers exist but are infrequent, RSS puts extra focus on outlying days, even at the cost of increasing errors on all other (calmer) business-as-usual days.

3.1.1 Estimating the coefficients

It is easy to show (by taking the first order conditions etc) that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

where $\bar{\cdot} \equiv \sum_{i=1}^{n} \cdot_i / n$ (i.e. sample mean).

$\hat{\beta}_1$ numerator: Sample covariance measuring how much x and y co-vary.
$\hat{\beta}_1$ denominator: Sample variance of x.
$\Rightarrow$ $\hat{\beta}_1$ measures by how much y co-varies with x, after taking into account variability in x.

$\hat{\beta}_0$ re-centers $\bar{y}$ (by takind the diff. from the main RHS covariate $\hat{\beta}_1 \bar{x}$) to ensure that the residuals e are on average zero.

3.1.2 Assessing the accuracy of the coefficient estimates

(2) means that $\hat{\beta}_0$ and $\hat{\beta}_1$ are just sample statistics (i.e. they are a realization of random variables).

If they have Gaussian distributions, we can check if the data suggest that β_1 is likely to be of a particular value.

E.g. we can ask "can $\beta_1 = \beta_{H_0}$ be true *according to the data*?"
Hypothesis testing with the null hypothesis H_0: $\beta_1 = \beta_{H_0}$ and the alternative H_1: $\beta_1 \neq \beta_{H_0}$ (two sided test).
If we set $\beta_{H_0} = 0$, we're asking if there is any relationship between x and y.

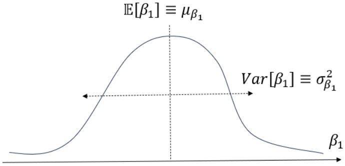

$\mathbb{E}[\beta_1] \equiv \mu_{\beta_1}$
$Var[\beta_1] \equiv \sigma^2_{\beta_1}$
β_1

Figure: Distribution of β_1.

3.1.2 Assessing the accuracy of the coefficient estimates

Method 1: Confidence interval (C.I.)
Idea: If $\beta_i \sim N(\mu_{\beta_i}, \sigma_{\beta_i})$, then $(\beta_i - \mu_{\beta_i})/\sigma_{\beta_i} \sim N(0,1)$. In standard normal, it is highly unlikely (less than 5% chance) to get a draw outside $(-1.96, 1.96)$. Two-sided test. Can take 1% critical values etc.

Infeasible. We don't know population parameters μ_{β_i} and σ_{β_i}.
$\Rightarrow$ Use the sample counterpart:

$$\hat{\beta}_i \pm 1.96\, S.E.(\hat{\beta}_i), \tag{3}$$

where

$$S.E.(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right), \; S.E.(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

If the C.I. (3) excludes β_{H_0}, then we say that "there is not enough evidence in the data to suggest $\beta_i = \beta_{H_0}$ at the 5% (confidence) level". "5% level" is the *size* of the test: a caveat that there's 5% chance this conclusion is wrong.

3.1.2 Assessing the accuracy of the coefficient estimates

Method 2: p-value

The above idea also means that $\hat{t} \equiv (\hat{\beta}_i - \beta_{H_0})/S.E.(\hat{\beta}_i)$ should be roughly standard normal under the hypothesis $\beta_i = \beta_{H_0}$. ("Roughly" because t-distribution with n-2 degrees of freedom is to used especially in small sample.)

$\Rightarrow$ P-value $=$ the probability of a sandard normal draw being more extreme (larger in magnitude) than $|\hat{t}|$.

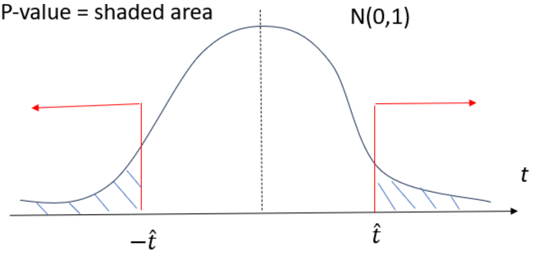


Figure: Distribution of t-statistics.

3.1.2 Assessing the accuracy of the coefficient estimates

Aside: Diagnostic statistics in big data analysis

Notice that in (3), the standard error becomes smaller when n is large. C.I. becomes tight.
$\Rightarrow$ The test becomes very good at rejecting false null. The *power* of statistics $\uparrow$.

Is it a good thing?

Suppose the truth is β_1 =1e-5 i.e. negligible. When the power of the stat becomes so great, we get very good at rejecting even the smallest deviation from zero. Then rejecting $H_0 : \beta_1 = 0$ in favor of $H_1 : \beta_1 \neq 0$ is correct (statistical significance), but it's probably not meaningful (economically insignificant).

Always consider economic v.s. statistical significance.
Problem discarding economically meaningless coefficients when the variable universe is large and model selection is automated.

3.1.2 Assessing the accuracy of the coefficient estimates

Aside: Diagnostic statistics in big data analysis

For these reasons, it is typically recommended to use Akaike information criterion (AIC) or Schwarz (Bayesian) information criterion (SIC, BIC) alongside other diagnostic stats in model selection.

They are not influenced by the sample size in the above sense.

Look for a model with smaller AIC/SIC.

SIC is often preferred over AIC as SIC penalizes the inclusion of additional covariates more than AIC. (We prefer a parsimonious model. Avoid overfitting.)

3.1.2 Assessing the accuracy of the coefficient estimates

Aside: Residual standard error and test statistics

Notice also that in (3), the population variance parameter σ^2 of $\epsilon \sim (0, \sigma^2)$ from Slide 3 appears.

ϵ is formally called the *irregular* term (compared with "error" which can also mean prediction error etc).

The derivation (validity in some sense) of (3) crucially depends on the assumption that ϵ is independent and identically distributed (i.i.d.).

But in practice, data would deviate from this assumption. What matters is by how much.

Lack of "independence": typically in the sense of serially/mutually correlated ϵ (heteroscedastic as opposed to homoscedastic errors). E.g. GARCH models.

3.1.2 Assessing the accuracy of the coefficient estimates

Aside: Residual standard error and test statistics

Mutually correlated ϵ affects the *efficiency* of the statistics (i.e. how fast the sample stats approach the limiting distribution $N(\cdot, \cdot)$ as $n \to \infty$).

We don't need to worry about it that much if the sample size is large (which is usually the case in finance).

Non-identically distributed ϵ: More complex / advanced issue. Different model may be needed (e.g. higher dimensional model).

3.1.3 Assessing the accuracy of the model

In (3), we cannot observe σ^2. Infeasible.

Estimate it using *residual standard error* (RSE):

$$RSE = \sqrt{RSS/(n-2)},$$

where RSS was given in Slide 5.

If RSE is “large”, the model can be seen to be not fitting the data well. Often in terms of too much noise compared to signal (signal-to-noise ratio).

3.1.3 Assessing the accuracy of the model

R-squared statistics: Measures the proporion of variability in y explained by the RHS covariates.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{4}$$

where TSS is the *total sum of squares* $\sum_i (y_i - \bar{y})^2$.

Useful and easy to use as it is bounded between 0 and 1. More commonly used than RSE.

But it increases with the number of covariates.

Adjusted R-squared is better as it has a penalty for adding more covariates.
(Also use AIC / SIC alongside as discussed above.)

3.2 Multiple linear regression

We can have more covariates on the RHS, X_i for $i = 1, ..., p$, say. Then we have a multiple linear regression model:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon = \mathbf{X}^\top \boldsymbol{\beta} + \epsilon \tag{5}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^\top$ and $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)^\top$.
Then the least squares estimator is:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i. \tag{6}$$

Compares with (2) (after rearrangement).

3.2 Multiple linear regression

Remark on stability of β_i

Note that we have a matrix inverse quantity on the RHS of (6), which is $\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1}$. This is valid only if the matrix is invertible (*nonsinglar*), which is the case when the matrix is full-rank (i.e. $rank(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top) = p$).

Then the RHS covariates $X_1, X_2, \ldots, X_p$ cannot be mutually correlated (i.e. no *multicollinerarity*). Otherwise the inverse computation and estimated coefficients become unstable / unreliable.

But in practice, we always find some degree of correlation among the covariates (using sample correlation statistics). What matters is by how much we depart from nonsingularity.

$\Rightarrow$ Check by how much coefficients change by removing some covariates. If some coefficients swing a lot, there is a problem. (More later.)

3.2 Multiple linear regression

Since we have multiple covariates, we can perform hypothesis tests on some of them jointly.

F-statistics:
Suppose we want to test the null hypothesis: $H_0 : \beta_j = 0$ for $j = 3, 6, 9$, say. (Can be any or all of them.)
Model with restrictions. The number of restrictions are $q = 3$.

Alternative hypothesis: $H_1 : \beta_j \neq 0$ for $j = 3, 6, 9$ i.e. they are significant.

1. Estimate the model with and without the restrictions.
2. Get the RSS of the two specifications and compute:

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p-1)} = \frac{explained\ variation}{unexplained\ variation}$$

   F follows the F-distribution with degrees of freedom q and n-p-1 under the null.

3.2 Multiple linear regression

Remark: F-test and t-test

The restriction doesn't have to be at zero (i.e. can be any number). Just need to compute the RSS under the restricted and unrestricted models, and then look at the F-stat. If it is statistically significantly different from zero at a given confidence level, we reject the null like we would in a t-test.

The t-test from earlier is a special case of F-test for when $q = 1$.

As was the case with t-test, the confidence interval in F-tests become narrower as $n \to \infty$. So then everything becomes significant i.e. it becomes difficult to reject only a very negligible deviation from the null. Can't drop covariates.

3.2 Multiple linear regression

When selecting a model, one can do a forward / backward / mixed selection.

- Forward = specific to general. Start simple and keep adding covariates.
- Backward = general to specific. Start from all covariates and keep dropping insignificant covariates.
- Mixed = back and forth.

In practice, we usually end up going back and forth.

It is a good practice to look at all diagnostic statistics to select a good model: t-stat, adjusted R-squared, AIC / SIC, etc.

Remark on forcasting, uncertainty, and purpose

The estimated model can be used to make predictions. Denoting forecast variables by $\tilde{\cdot}$, we compute:

$$\tilde{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \tilde{x}_j$$

This is a result of taking the first moment (expectation) of the model given $\tilde{x}_j$.

$\tilde{y}$ may be wrong due to uncertainty in βs or variability in x_j, ϵ, etc.

So some analysts would also present scenario analysis, confidence interval around $\tilde{y}$, etc to check variability in $\tilde{y}$ instead of making a point-prediction.

Remark on forcasting, uncertainty, and purpose

If Y is a continuous random variable (e.g. height) instead of a discrete one (e.g. age), then by definition, the probability of a random draw hitting any specific number from the support (domain, along the real line) is zero.

Particularly in financial / economic applications, the act of forming and assessing forecasts/expectations is often as important as getting $\tilde{y}$ correct.

This is when expectations themselves are known to affect future outcomes (feedback / endogenous effects). E.g. Expected inflation affects consumer spending patterns, which affects real economy.

The objective of the analysis may not be about getting it spot-on correct. (Can be theoretically impossible as mentioned above.)

3.3 Other considerations in the regression model

Dummy variable:
Qualitative covariates / dummy variables are very useful as they can reflect "it is or isn't" outcomes.

E.g. Seasons. $x_1 = 1$ if spring, 0 if not. $x_2 = 1$ if summer, 0 if not. $x_3 = 1$ if autumn, 0 if not. $x_4 = 1$ if winter, 0 if not.

$$\hat{y} = \sum_{j=1}^{4} \hat{\beta}_j x_j = \hat{\beta}_j \textit{ if season } j. \tag{7}$$

This is a level shift.

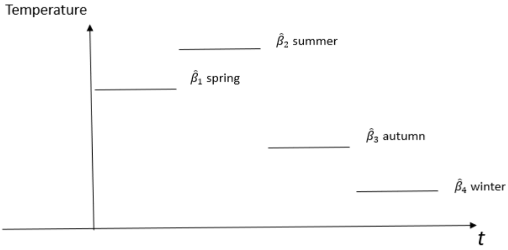


Figure: Seasonal dummies.

3.3 Other considerations in the regression model

Remarks on overfitting and multicolinearity

Excessive use of dummy variables can lead to overfitting in the sense of fitting the data "too well".

Extreme example is one dummy variable per data point. Perfect fit.

Notice that β_0 is excluded from (7). This is because $\sum_j x_j = 1$ so we avoid its perfect correlation with the intercept "covariate" $\mathbf{1}$.

If you prefer to include β_0 in (7), drop just one dummy x_j for any j. The model is still the same.

3.3.2 Extensions of the linear model

Interaction terms involve two or more covariates multiplying the effect of each other.

Example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

This can be rewritten as:

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$

X_1's overall effect on Y changes with X_2 by $\beta_3 X_2$. Likewise for X_2.

Example: Y is sales, X_1 is advertisement cost by mode (tv, radio, posters), and X_2 is location.

3.3.2 Extensions of the linear model

Quadratic terms can be used to capture an effect of a variable stronger than linear.

Example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

X_1's overall effect on Y increases squarely with X_1.

3.3.3 Potential problems

Other non-linear transformations include $\sqrt{\cdot}$ and $\log(\cdot)$ (for LHS and/or RHS).

Coefficient on log-transformed data can be interpreted as the effect after 1 % change in the variable.

Usually good to check by visually inspecting the data (as in scatter plot etc) to see if higher order effects exist before fitting.

Always check residuals to see if the model is reasonable.

3.3.3 Potential problems

Heteroscedasticity is when the variance σ^2 of $\epsilon_i \sim (0, \sigma^2)$ (i.e. the error term) is not constant.

Always inspect residual^2. Can use *heteroscedasticity consistent standard error* in diagnostics. But always think why there's heteroscedasticity.

Good examples in finance.

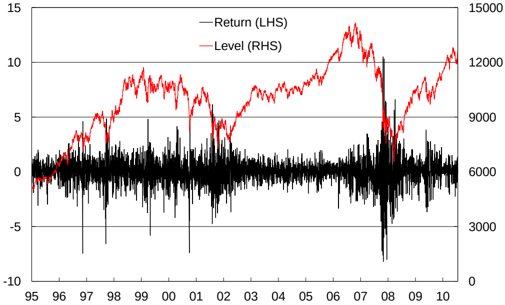


Figure: Dow Jones Industrial Average. Daily returns. Black: returns. Red: price (or index) level.

3.3.3 Potential problems

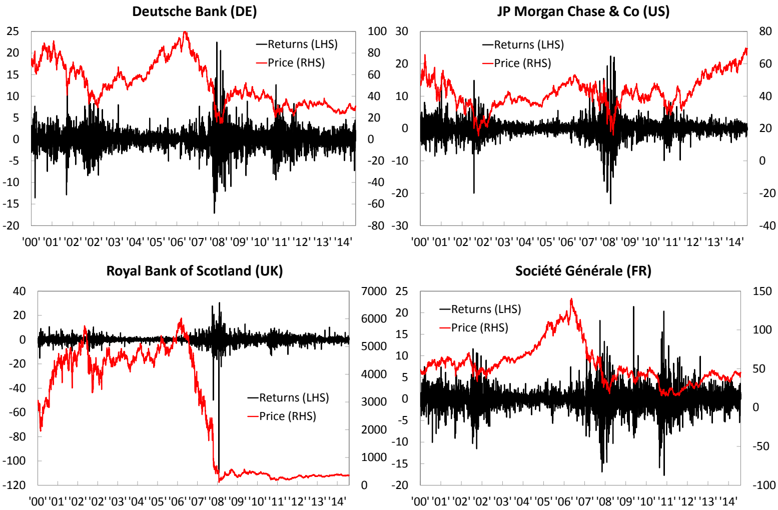


Figure: Stock price dynamics of selected banks. Daily returns.

3.3.3 Potential problems

Volatility of equity returns go up when prices are falling (the *leverage effect*). Not consistent with constant σ^2 assumption.

When heteroscedasticity is observed, need to think why. Might need (or want) to model heteroscedasticity.

E.g. Gaussian GARCH(1,1):

$$\begin{aligned} r_t &= \mu_t + \sigma_t z_t, \quad z_t \sim \text{i.i.d. } N(0,1) \\ \sigma_t^2 &= \delta + \beta \sigma_{t-1}^2 + \alpha r_{t-1}^2. \end{aligned}$$

We can write $\varepsilon_t = \sigma_t z_t$, then $\mathbb{E}[\varepsilon_t] = 0$ but σ_t^2 is serially dependent.

3.3.3 Potential problems

Modelling the degree of dispersion (i.e. risk / uncertainty) around the centre.

σ_t scales the support of the distribution of z_t.

When σ_t is large, it is more likely to observe large $|r_t|$.

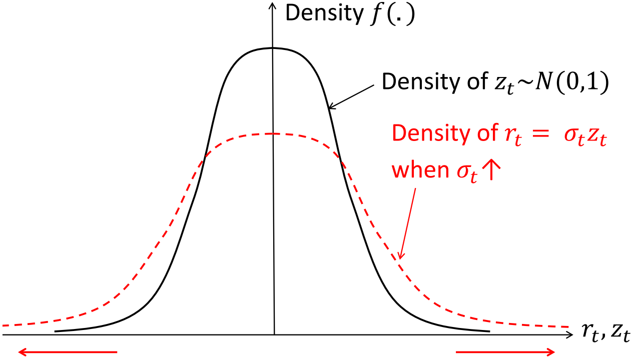
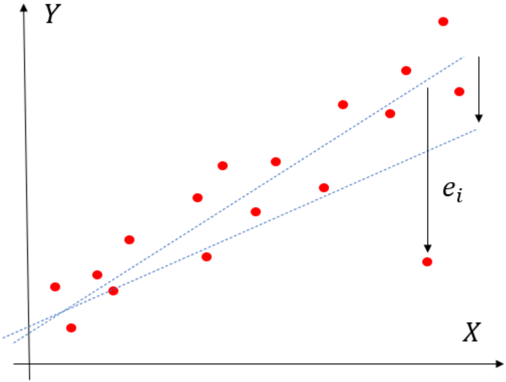


Figure: Picture illustration of the idea behind GARCH.

3.3.3 Potential problems

Outliers are "unusual" values of Y given X (unusual from the viewpoint of a Gaussian yardstick).

OLS can be sensitive to outlying data points as it tries to minimize MSE (i.e. particularly dislikes large e_i^2). Pulls the fitted line towards outliers.



3.3.3 Potential problems

The leverage statistic can be used to assess if a data point looks extreme.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

We have $\sum_i h_i = 1$. Leverage stat larger than 0.2 may be worth investigating further.

If a high leverage point is a corrupt data point, it is possible to simply remove that point. But usually we should never remove any datapoints unless we have a very good reason to do so.

Alternatively, one can aim to minimize mean absolute deviation (MAE) $\sum_i |e_i|/n$, which is more robust than MSE.

More generally, outliers (in the sense of Gaussian distributions) may suggest that we have non-Gaussian data.