Data Analysis: Intro to Regression

Mark Hendricks

August Review

UChicago Financial Mathematics

Outline

Regression

OLS Mathematics

Regression analysis in finance

Regression applications in finance include...

- Risk-management. Find how a portfolio return is impacted by some factor/instrument.
- Forecasting. Build forecasts of financial and macroeconomic variables. (inflation, yields, etc.)
- Pricing. The fundamental asset pricing equation is a linear relation between risk and return.

Beyond regression

Nonlinear analysis is also important.

- Options Pricing. Differential equations requiring martingale methods, simulation, finite differenc, etc.
- Value at Risk. Model the tail of the distribution of profits and losses.
- Volatility Models Need non-linear timeseries models such as GARCH.

Linear regression model

Consider a **linear regression model** involving two variables, y and x.

$$y = \alpha + \beta x + \epsilon$$

- y is referred to as the **regressand**, or explained variable.
- x is referred to as the **regressor**, covariate, or explanatory variable.
- α and β are the (constant) parameters of the model.

Example: Portfolio factor sensitivity

Decompose the hedge fund return into a market-driven and market-neutral return.

$$r_p = \alpha + \beta r_{\mathrm{mkt}} + \epsilon$$

- random total portfolio return denoted by r_p
- random return on the S&P 500, denoted by r_{mkt}.

Interpret...

- $\beta = 0, 1, 2$
- $\alpha = -.01, 0, .01$.

Example: Portfolio decomposition

Continuing the example from above,

$$r_p = \alpha + \beta r_{\text{mkt}} + \epsilon$$

We may want to know "how much" of r_p is explained by r_{mkt}.

- R-squared (R^2) is a metric of the variation explained in the regression model.
- Is the hedge-fund driven by market returns if $\beta = 1$, $R^2 = .10$? How about $\beta = .5$, $R^2 = .50$?

(Notation: R^2 is standard notation in regression analysis—nothing to do with my choice of variable name r_p, r_{mkt}.)

Univariate regression

When there is only one regressor, x, we will see that the OLS estimator is simply:

$$\beta = \frac{\operatorname{cov}(y, x)}{\operatorname{var}(x)}$$

And that the R-squared statistic is simply

$$R^2 = [\operatorname{corr}(y, x)]^2$$

So why bother with regression if we just need covariances and variances?

Multiple regression

In the case of multiple regressors, the OLS statistics are not so easily formed.

- Augment our hedge-fund regression with a second regressor: a US dollar index, $r_{\$}$.

$$r_p = \alpha + \beta_1 r_{\text{mkt}} + \beta_2 r_{\$} + \epsilon$$

- The formulas for β_1 and β_2 do not follow as easily:

$$\beta_1 \neq \frac{\text{cov}\left(r_p, r_{\text{mkt}}\right)}{\text{var}\left(r_{\text{mkt}}\right)}$$

- The R-squared stat captures the correlation between r_p and the combined space spanned by both r_{mkt} and $r_{\$}$.

Caution!

Remember that the multi-variable beta is not the same as the univariate beta!

$$r_p = \alpha + \beta_1 r_{\text{mkt}} + \beta_2 r_{\$} + \epsilon$$

- Perhaps r_p is positively correlated with $r_{\$}$, and thus would have a positive beta if regressed on only $r_{\$}$.
- But β_2 is not a measure of this pairwise comovement!
- β_2 gives the impact on r_p if we hold r_{mkt} constant!
- Thus, when the regressors are correlated, multi-variable betas can be quite different from their univariate counterpart.

Units

When interpreting the regression coefficients, be careful to remember the underlying units.

$$r_p = \alpha + \beta_1 r_{\text{mkt}} + \beta_2 r_{\$} + \epsilon$$

- The volatility of r_{mkt} is three times larger than the volatility of $r_{\$}$.
- Thus, even if β_2 is larger than β_1, we need to remember that one-unit changes in $r_{\$}$ happen less frequently.
- In this situation it may be more helpful to report $\beta_1\sigma_1$ and $\beta_2\sigma_2$ to help convey the one-standard deviation impact from each factor.

Outline

Regression

OLS Mathematics

Multivariate linear regression

In a multivariate regression model with k regressors,

$$\begin{aligned} y =& \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \epsilon \\ =& \alpha + \sum_{j=1}^{k} \beta_j x_j + \epsilon \\ =& \mathbf{x}' \boldsymbol{\beta} + \epsilon \end{aligned}$$

- The last line defines $\mathbf{x}$ such that the first element is the constant 1, and the first element of $\boldsymbol{\beta}$ is α.
- Including the regression constant in the vector notation will simplify the algebra, as we will always consider the case where the first regressor is a constant.

Data from the regression model

A sample of n observations is denoted as $(y_i, \mathbf{x}_i)$ for $i = 1, 2, \ldots n$.

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

where

$$\mathbf{x}_i \equiv \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,k} \end{bmatrix} \qquad \boldsymbol{\beta} \equiv \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Regression estimate

Consider a sample estimate of $\boldsymbol{\beta}$, denoted by $\boldsymbol{b}$.

Then

$$y_i = \mathbf{x}_i' \boldsymbol{b} + e_i$$

where e_i denotes a sample residual,

$$e_i = y_i - \mathbf{x}_i' \boldsymbol{b}$$

This is estimated regression, as opposed to the population regression equation above.

Ordinary least squares

The **ordinary least squares estimator** of β minimizes the sum of squared sample errors:

$$\boldsymbol{b} \equiv \arg\min_{\boldsymbol{b}_o} \sum_{i=1}^{n} (e_i)^2$$

$$= \arg\min_{\boldsymbol{b}_o} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i' \boldsymbol{b}_o\right)^2$$

OLS problem

Rewrite the OLS problem in matrix notation,

$$\boldsymbol{b} \equiv \arg\min_{\boldsymbol{b}_o} \; \boldsymbol{e}'\boldsymbol{e}$$

$$= \arg\min_{\boldsymbol{b}_o} (\mathbf{Y} - \mathbf{X}\boldsymbol{b}_o)'(\mathbf{Y} - \mathbf{X}\boldsymbol{b}_o)$$

where

$$\mathbf{X} \equiv \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}, \quad \mathbf{Y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{e} \equiv \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

Assumption: Full-rank

Assumption 1: $\mathbf{X}'\mathbf{X}$ is full rank.

Equivalently, assume that there is no exact linear relationship among any of the regressors.

- Clearly, the existence of OLS estimator requires that this assumption be satisfied.
- Multicollinearity refers to the case where this assumption fails.

OLS estimate

Solving the minimization problem above gives the **OLS estimate**:

$$\boxed{\boldsymbol{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}}$$

- This estimate yields sample residuals of

$$\begin{aligned}\boldsymbol{e} =& \mathbf{Y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\ =& \left(\mathcal{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{Y}\end{aligned}$$

- Thus $\boldsymbol{e}$ is orthogonal to $\mathbf{X}$.
- Equivalently, the in-sample correlation between x_i and e_i is zero.

Alternative OLS derivation

Suppose the population correlation between $\mathbf{x}$ and ϵ is zero.

$$0 = \mathbb{E}\left[\mathbf{x}\epsilon\right]$$
$$0 = \mathbb{E}\left[\mathbf{x}\left(y - \mathbf{x}'\boldsymbol{\beta}\right)\right]$$

Thus,

$$\boldsymbol{\beta} = \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1}\mathbb{E}\left[\mathbf{x}y\right]$$

If regression includes a constant, then these terms are covariance matrices, and we can use sample estimators in place of the population moments to get the OLS estimator:

$$\boldsymbol{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$
$$= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i\right)$$

Regression with an intercept

The assumption that X includes a column of $1's$ is important.

- Including a constant in the regression is equivalent to running a regression with demeaned data.
- Running a regression on just a constant regressor and nothing else, would simply pick up the mean in the data.
- Including a constant in the regression means the regressors try to match the variation in the y data, not the overall level.

Example: Risk premia

A fundamental theorem of asset pricing says that there is a linear relation between the risk premium of asset i, π_i, and a certain risk measure, x_i:

$$\pi_i = \alpha + \beta x_i + \epsilon_i$$

The Portfolio Theory class covers this theory in detail, but for now take it as given.

- Test this theory with a linear regression.
- Try both including a constant, α, and without.
- Risk and return data is collected on various industry portfolios.

Example: Regression with and without an intercept

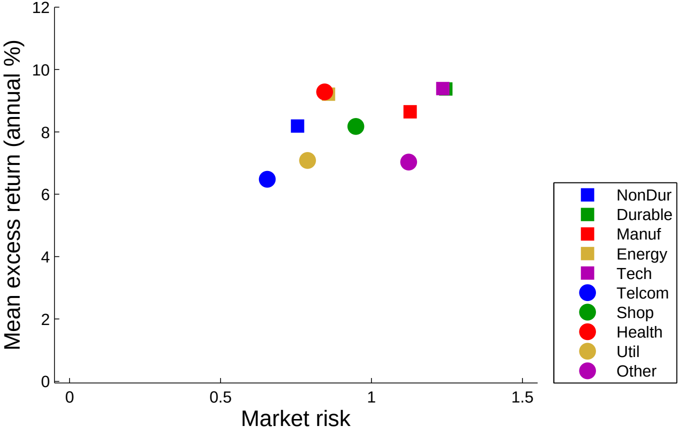

Figure: Data Source: Ken French. Monthly 1926-2011.

Example: Regression with and without an intercept

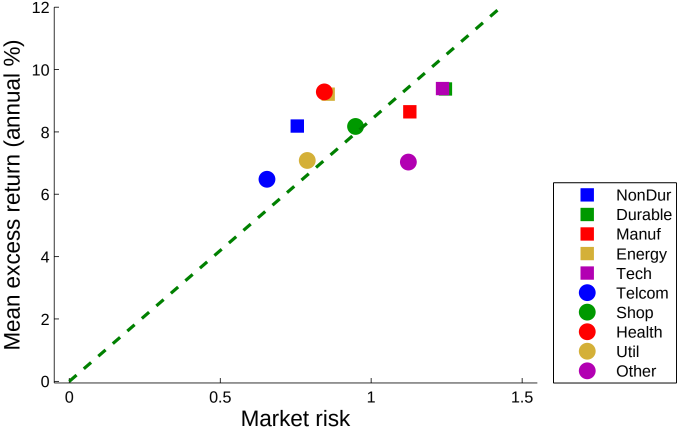

Figure: Data Source: Ken French. Monthly 1926-2011.

Example: Regression with and without an intercept

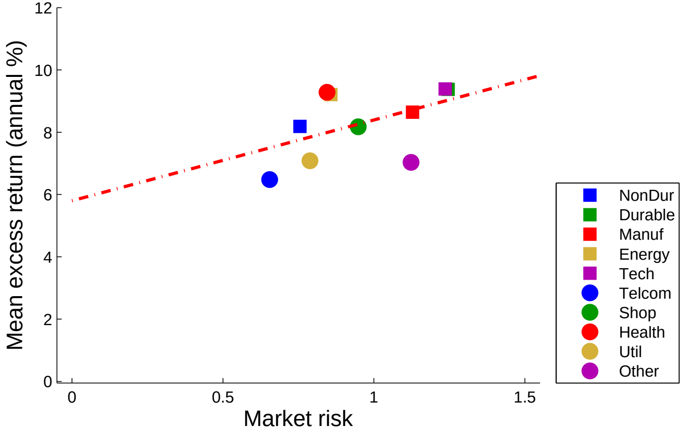

Figure: Data Source: Ken French. Monthly 1926-2011.

Example: Regression with and without an intercept

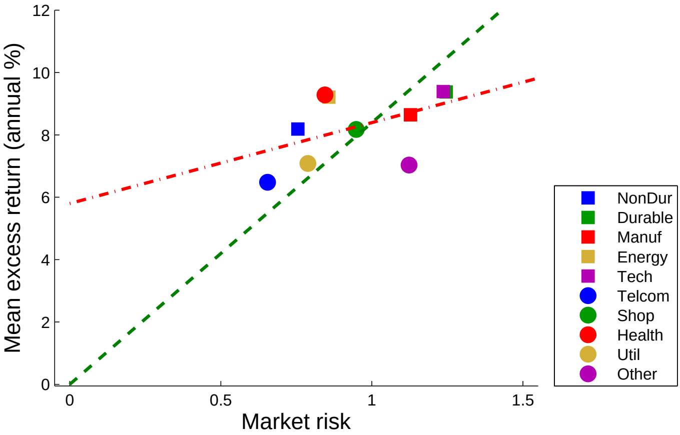

Figure: Data Source: Ken French. Monthly 1926-2011.

Residuals with zero mean

By assuming the model includes a constant,

$$\mathbb{E}[\mathbf{x}\epsilon] = \mathbf{0} \implies \mathbb{E}[\epsilon] = 0$$

By including a constant in the sample estimation,

$$\frac{1}{n}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}' \boldsymbol{e} = 0 \implies \frac{1}{n}\sum_{i=1}^{n} e_i = \bar{\boldsymbol{e}} = 0$$

R-squared

The **R-squared**, or coefficient of determination, in a regression is defined as

$$R^2_{y,x} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

$$= 1 - \frac{\text{error sum of squares}}{\text{total sum of squares}}$$

Algebraically, this is

$$R^2_{y,x} = \frac{\boldsymbol{b} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

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

R-squared versus correlation

Intuitively, the R-squared is the square of the correlation between y and the projection of y onto $\mathbf{x}$.

$$R^2_{y,\mathbf{x}} = [\text{corr}\,(\mathbf{Y}, \mathbf{PY})]^2$$

In a univariate regression of y on x,

$$R^2_{y,x} = [\text{corr}(y, x)]^2$$

Caveat: Regressing on a constant

The interpretation and formula for R-squared does not hold if there is no constant regressor.

- Without a constant, the R-squared will not necessarily be between 0 and 1.
- Without a constant, the R-squared will not necessarily be the square of the correlation between the sample $\mathbf{Y}$ and the projected Y values.
- Without a regressor, the fit can be improved simply by shifting the sample $\mathbf{Y}$ data by a constant.