$$\hat{y}_{t+k} = f(X_t) + \varepsilon_t$$

- Linear vs Nonlinear
- Deterministic vs Stochastic
- Data-based vs First principles
- Parametric vs Nonparametric

Linear vs Nonlinear

- **Linear models**: the functional form can be written as a linear function of the model parameters

$$\hat{y}_{t+k} = \alpha_0 + \alpha_1 x_t + \varepsilon_t$$

- **Nonlinear models**: deviations from the classical Gauss Markov assumptions (linearity, homogeneity and independence)

$$\hat{y}_{t+k} = \alpha_0 + \alpha_1 x_t^3 + \alpha_2 / x_t^2 + \varepsilon_t$$

Deterministic vs Stochastic

- **Deterministic models**: model output is completely determined by the model parameters, input variables and initial conditions, i.e., for a given set of inputs, we get one fixed output. For example:

$$y_{n+1} = r y_n (1 - y_n), \text{ Logistic Map.}$$

- **Stochastic models**: there is an inherent randomness in the model, i.e. for a given set of inputs and parameters we will get an ensemble of different outputs. For example:

$$y_{n+1} = \alpha_0 + \alpha_1 y_n + \varepsilon_{n+1}$$

Data based vs First Principles

- **Statistical data-based modelling**
  - Time series modelling (Box and Jenkins models, Regime switching models, Exponential smoothing, Gaussian processes, Machine learning algorithms)

- **Mathematical modelling**
  - Incorporation of prior knowledge
  - Newton's Law (mechanics), Navier-Stokes (fluids), Maxwell's Laws (electromagnetism)
  - Conservation laws of nature (mass, momentum, force)
  - System constraints (non-negative distributions)

- Data based principled (DBP) modelling aims to combine knowledge obtained from first principles, such as conservation laws of nature and system constraints, with information extracted from existing databases and real-time observations

Parametric vs Nonparametric

- **Parametric models**: rely on strong assumptions regarding the true underlying data generating process

$$y_{t+k} = \alpha_0 + \alpha_1 x_t + \varepsilon_t$$

- **Nonparametric models**: attempt to learn the true functional form from the data itself. The data would decide what f would look like.

$$y_{t+k} = f(X_t) + \varepsilon_t$$

- **Semiparametric models**: have both parametric and nonparametric components.

Major steps involved in modelling

- **Time Series Analysis** (plot, scatter plots, ACF, PACF etc.)
- **Pre-processing** (handling missing observations, outliers, transformations)
- **Model Identification** (AR, ARMA, SETAR, ES, ANNs etc.)
- **Model Selection** (AIC, BIC etc.)
- **Parameter Estimation** (OLS, MLE etc.)
- **Forecast Evaluation** (*point* – MAE; *quantile* – coverage; *density* – CRPS etc.)
- **Model Combination** (weighted average)
- **Error Diagnostics** (residual analysis)

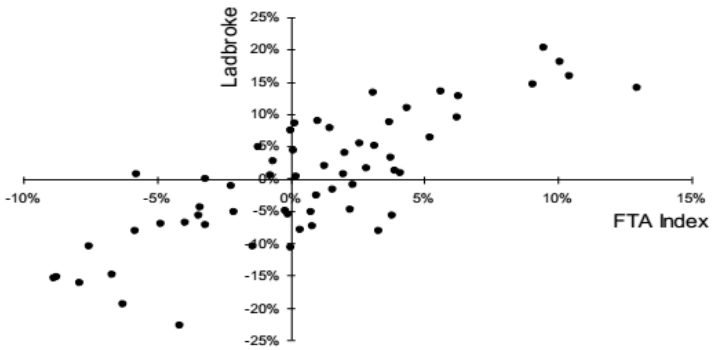
Stationarity

- A time series ($\boldsymbol{y}$) is said to be *strictly stationary* if its joint distribution does not change with shift in time, i.e., the joint distribution of ($y_{t_1}, y_{t_2}, \ldots, y_{t_n}$) and ($y_{t_1+\tau}, y_{t_2+\tau}, \ldots, y_{t_n+\tau}$) are the same for any τ and positive integer n (time invariant)
- For *weak stationarity* assumptions, it is required that the first and second order moments do not depend on time. This means that the time series has a constant mean, constant variance, and constant autocorrelations at each lag
- Tests for stationarity: Dickey-Fuller test, augmented Dickey-Fuller test

Regression analysis

- Regression estimates relationship between a dependent variable (y, also referred to as output/target variable) and a number of independent variables (X, also referred to as input/explanatory/predictor variables). It is used for:
  - Description
  - Prediction

- Ex: how is an investment's return related to market return?



Regression analysis stages

Broadly speaking, there are three main stages involved in the model building process using regression:

1. **Time series plots, scatterplots and correlation analysis**
   - Plot time series to get a feel for underlying structure (trend, seasonality, volatility), relationships between variables, and spot problems (missing data, outliers, irregular sampling)
   - Correlations quantify strength of relationship (you may also consider mutual information)

2. **Model estimation**
   - Ordinary least squares

3. **Diagnostic evaluation**
   - Evaluate model validity
     - Testing for significance of relationship
     - Has anything been left out of model
   - Evaluate usefulness
     - How close the model fits the data (goodness of fit)

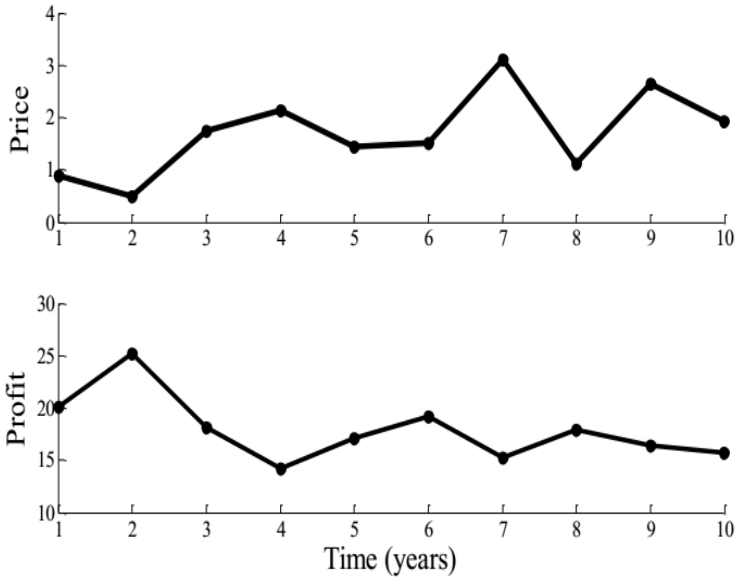
Regression analysis

- We have two variables: price of crude oil ($/litre) and average profit ($/unit sold) for a tyre manufacturing company. The company replies on importing raw materials, so higher oil prices results in lower profits overall
- Aim is to identify relationship between the two factors, and to forecast *profit* (dependent variable) for a future plant
- $Profit = f(Price)$

Time index	Price of Crude Oil	Profit
1	0.9	20.15
2	0.5	25.18
3	1.75	18.10
4	2.14	14.26
5	1.45	17.13
6	1.52	19.14
7	3.11	15.22
8	1.12	17.88
9	2.65	16.40
10	1.92	15.69

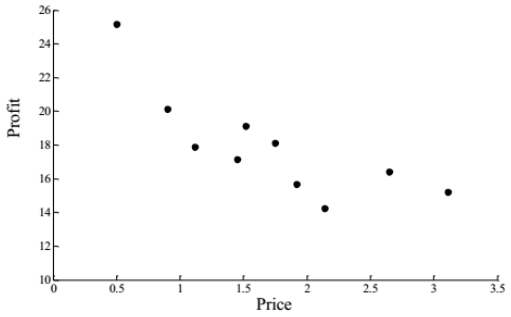
Time series plots, scatterplots and correlation

- Always start the analysis by first plotting the time series to get a feel of the underlying structure



Time series plots, scatterplots and correlation

- Use scatterplots and correlation analysis to get feel of data prior to regression



- Correlation coefficient, r is a measure of the strength of linear relationship between variables $corr(Profit, Price) = -0.80$

Simple linear regression

- Simple linear regression estimates and tests relationship between two variables in a simple linear model

$$\underset{\text{Direction of causality}}{\overset{\longleftarrow}{y = a + bx}}$$

- Basically, we are aiming to find a straight line defining the relationship, and regression model is simply the equation of line of best fit

$$Profit = 23.3 - 3.16Price$$

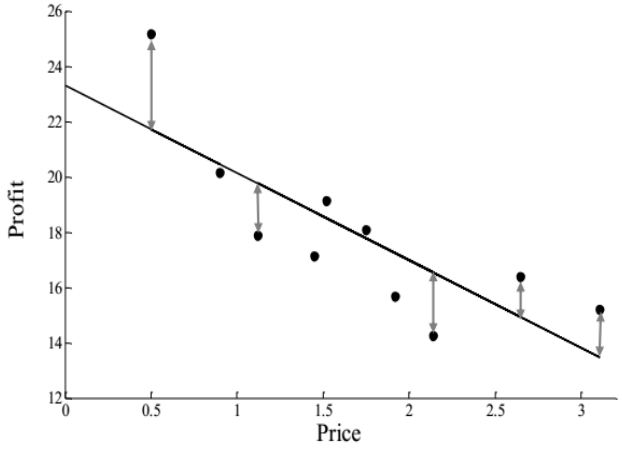
Dependent variable (target variable)	Constant (intercept)	Coefficient (slope)	Independent variable (explanatory)

Simple linear regression

- Model: $Profit = a + bPrice + e$

  $(y = a + bx + e)$

- The residual e_i is the vertical distance of point from the line, i.e., the difference between observed and predicted values of the dependent variable
- Choose line so that residual scatter is minimized. Least squares is traditionally used to estimate a and b. This involves finding the line that minimizes sum of squared residuals (minimizes residual variance): $min \sum_{i=1}^{n} (e_i)^2$



Fit of regression

- **Model**: $Profit = a + bPrice + e$

Time index	Price of Crude Oil	Profit	Fitted	Residuals
1	0.9	20.15	20.46	-0.32
2	0.5	25.18	21.72	3.46
3	1.75	18.10	17.77	0.33
4	2.14	14.26	16.54	-2.28
5	1.45	17.13	18.72	-1.59
6	1.52	19.14	18.50	0.64
7	3.11	15.22	13.47	1.75
8	1.12	17.88	19.76	-1.88
9	2.65	16.40	14.93	1.47
10	1.92	15.69	17.23	-1.54

- Residuals have zero mean and standard deviation, $s = 1.98$
- s is called 'standard error of regression'
- $R^2 = 64.19$
- Adjusted $R^2 = 59.71$

Model evaluation – R^2

- The coefficient of determination, R^2, measures the proportion of variation in dependent variable, y, that has been explained by a statistical model. It is one of the most widely used measure of fit
- R^2 is computed as:

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

- R^2 measures the amount of variance explained by the model given by the ratio of the explained variance (variance of the model's prediction) with the total variance (of the data)

Model evaluation – adjusted R^2

- Advisable to use 'adjusted' R^2 as it penalizes for low number of observations and high number of explanatory variables by using correct degrees of freedom in R^2 formula:

$$R^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where n denotes the total number of observations while k is the number of explanatory variables used in the model

- Note that: $ESS = (\bar{y} - \hat{y}_i)^2$, where *ESS* is the *explained sum of squares* (how well a model represent the data), $TSS = (y_i - \bar{y})^2$, is the *total sum of squares* (variation in observed data), while $RSS = TSS - ESS$, is the *residual sum of squares* (variation in model errors)

Sampling error in regression

- Standard deviation of estimate of population parameter is known as standard error
- In regression, population parameters are a and b

$$y = a + bx$$

- Sample is used to produce estimates for a and b which will have sampling error

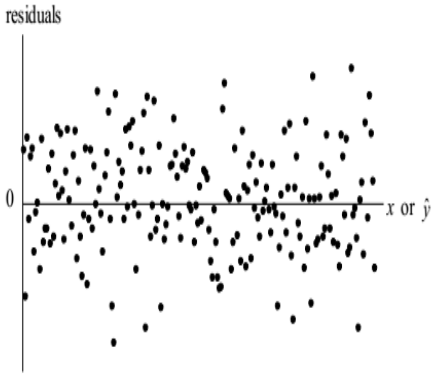
$$Profit = 23.3 \ - \ 3.16Price$$

(19.7, 26.9) (-5.1, -1.2)

- Fundamental question is always whether explanatory variable is useful and should be kept in model
- This question can be answered by testing if b is significantly different from zero. Can compute t-stats, or simply look if the CI spans zero
- Usual not to test significance of the intercept

Residual assumptions

- Model: $y = a + bx$
- Quality of parameter estimates and validity of sig. tests rely upon residuals being $N(0, s)$
- Residuals must be
  - Normally distributed
  - Independent (no autocorrelation)
  - Same variance (no heteroscedasticity)

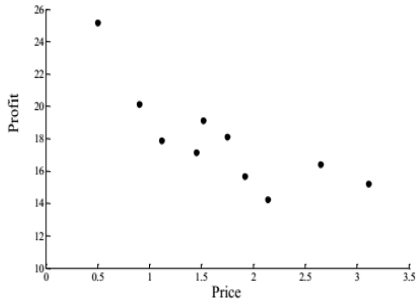


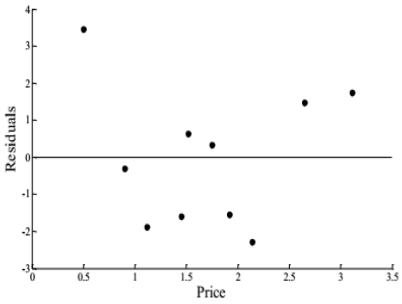
- Intuitively, residuals should be simple randomness remaining after deterministic part of variation in y has been modelled
- Any systematic component in the errors should really be in the model
- It is thus important to check histogram and plots of residuals

Checking residuals

- If there is a pattern in residuals, we have missed something. If each residual is related to it's predecessor, they are described as auto-correlated

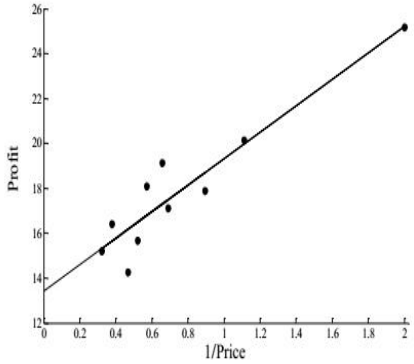
- Plot indicates we are fitting a linear model to a relationship that is non-linear





Data transformations

- Regression can only handle linear relationships so we transform to give a new variable: 1/price (remember the definition of linearity and non-linearity)



$$Profit = 13.4 + 5.90(1/Price)$$

(11.8, 15.0) (4.1, 7.1)

- $corr(Profit, 1/Price) = 0.9365$
- $R^2 = 87.70$
- Adjusted $R^2 = 86.16$
- $DW = 2.65$

Durbin-Watson statistic

- DW statistic evaluates autocorrelation for residuals placed in same order as the observations. This is generally only interesting if there is time order

- $DW = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N}(e_t)^2}$

- $0 < DW < 4$
- $DW = 2$ no autocorrelation
- $DW > 2$ negative autocorrelation
- $DW < 2$ positive autocorrelation

Forecasting and confidence intervals

- Model:

$$Profit = 13.4 + 5.90(1/Price)$$

- If Price of crude oil is say, 2.5, what is forecast for Profit?

$$Profit = 13.4 + 5.90(1/2.5) = 15.76$$

- What is the appx. 95% confidence interval for the above forecast for Profit?

$$15.76 \pm 1.96 \times s$$

where $s = 1.16$ is the standard deviation of the residuals

Formulae for estimation of a and b

- LS regression involves the following minimization:

$$min_{a,b} \sum_{t=1}^{n} e_t^2 = min_{a,b} \sum_{t=1}^{n} (y_t - a - bx_t)^2$$

- Partial differentiation with respect to a:

$$\frac{\partial}{\partial a} \sum_{t=1}^{n} (y_t - a - bx_t)^2 = -2 \sum_{t=1}^{n} (y_t - a - bx_t)$$

Equating this to zero gives: $a = \bar{y} - b\bar{x}$

- Partial differentiation with respect to b:

$$\frac{\partial}{\partial b} \sum_{t=1}^{n} (y_t - a - bx_t)^2 = -2 \sum_{t=1}^{n} x_t (y_t - a - bx_t) = -2 \sum_{t=1}^{n} x_t (y_t - \bar{y} + b\bar{x} - bx_t)$$

Equating this to zero gives:

$$b = \frac{\sum_{t=1}^{n} x_t (y_t - \bar{y})}{\sum_{t=1}^{n} x_t (x_t - \bar{x})} = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})(x_t - \bar{x})} = \frac{cov(x,y)}{var(x)}$$

Regression checklist

- Always start the analysis by first plotting the time series, look at scatterplots to get feel for data and compute correlations
- Fit model to data
- Check coefficients are worth keeping in the model (i.e., are they significantly different from zero)
  - use *t*-stat, confidence intervals, *p*-values
- Check fit of model
  - adjusted R^2, standard error of residuals
- Check residuals for randomness
  - plot residuals, look at standard deviation, normality, use DW statistic
- Judgementally assess the model
- Use the model for description and prediction
  - Generate confidence intervals along with point forecasts to quantify uncertainty