QBUS6840 Lecture 4

Time Series Regression

Professor Junbin Gao

The University of Sydney Business School

Outline

- Simple linear regression.
- Issues in time series regression.
- Multiple regression.
- Some useful time series predictors.
- Selecting predictors.
- Residual diagnostics.

Readings: FPP2 Chapter 5 (`https://otexts.org/fpp2/regression.html`) and BOK Chapter 6.

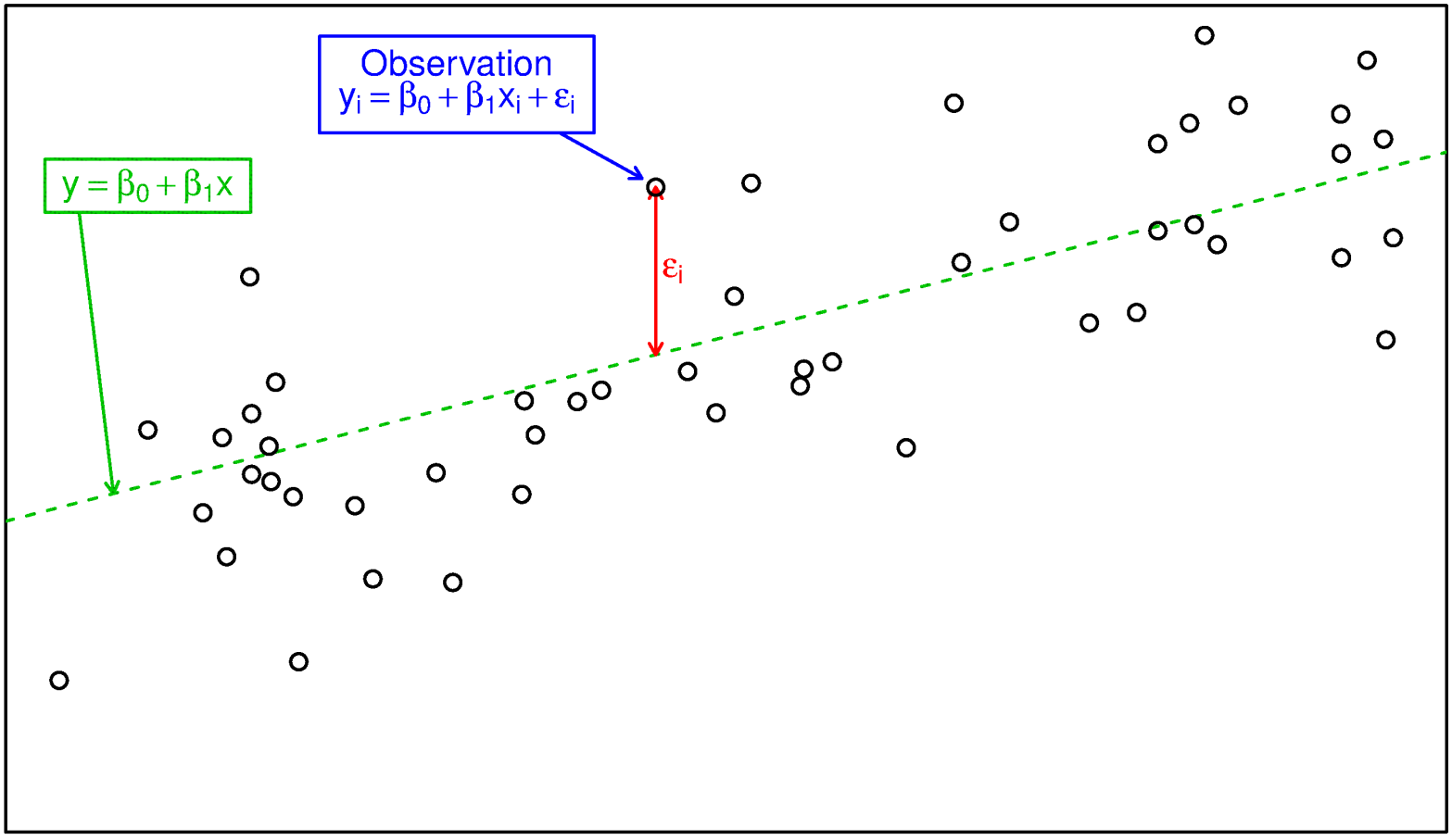
Review: The simple linear model

- Let forecast/dependent and predictor variables are assumed to be related by the simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- The parameters β_0 and β_1 determine the intercept and the slope of the line respectively.

The simple linear model: a visual example



An example of data from a linear regression model.

The simple linear model

- Notice that the observations do not lie on the straight line but are scattered around it.
- The random error ε_i captures anything that may affect y_i other than x_i. We assume that these errors:
  1. Are not serially correlated: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.
  2. Have zero conditional mean: $E(\varepsilon_i | X_i) = 0$.
  3. Homoscedasticity: $\text{Var}(\varepsilon_i) = \sigma^2$.
- It is also useful to have the errors normally distributed with constant variance in order to produce prediction intervals and to perform simplified statistical inference.

Least squares estimation

最小二乘估计

- In practice, of course, we have a collection of observations but we do not know the values of β_0 and β_1. These need to be estimated from the data. We call this *fitting a line through the data*.

- There are many possible choices for β_0 and β_1, each choice giving a different line. The least squares principle provides a way of choosing β_0 and β_1 effectively by minimizing the sum of the squared errors. That is, we choose the values of β_0 and β_1 that minimize

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Least squares estimation

- Using simple algebra, it can be shown that the resulting **least squares estimators** are

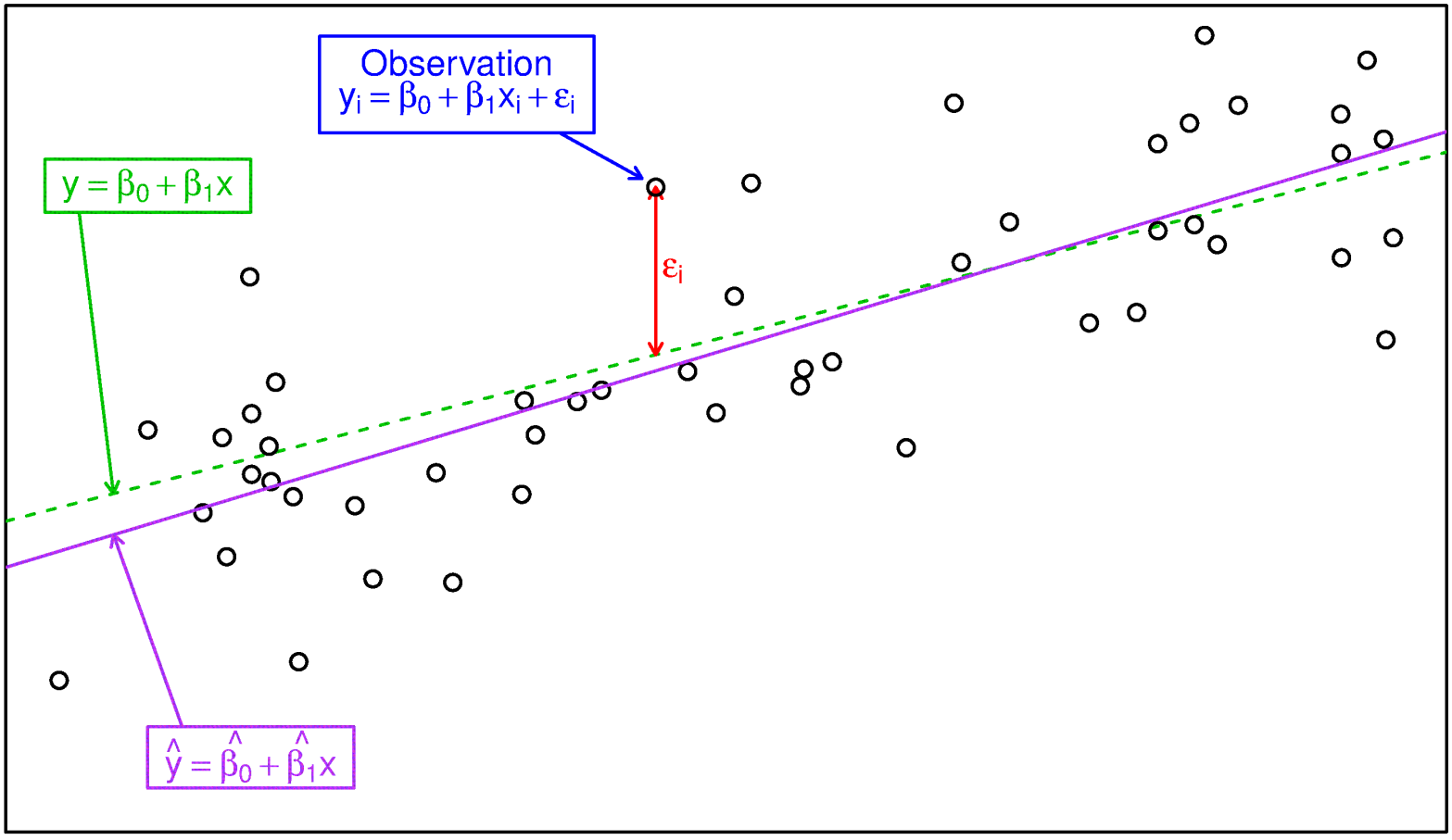
$$\hat{\beta}_1 := \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{x}$ is the average of the x observations and $\bar{y}$ is the average of the y observations.

Least squares estimation: visual example



Estimated regression line for a random sample of size N.

Least squares estimation

- We imagine that there is a true line denoted by $y = \beta_0 + \beta_1 x$, which we do not know. Therefore we obtain estimates β_0 and β_1 from the observed data to give the regression line.

- We use the regression line for forecasting. For each value of x^*, we can forecast a corresponding value of y using $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

Fitted values and residuals

- The forecast values of y obtained from the observed x values are called fitted values. We write these as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, for $i = 1, \ldots, N$. Each $\hat{y}_i$ is the point on the regression line corresponding to observation x_i.

- The difference between the observed y values and the corresponding fitted values are the "residuals":

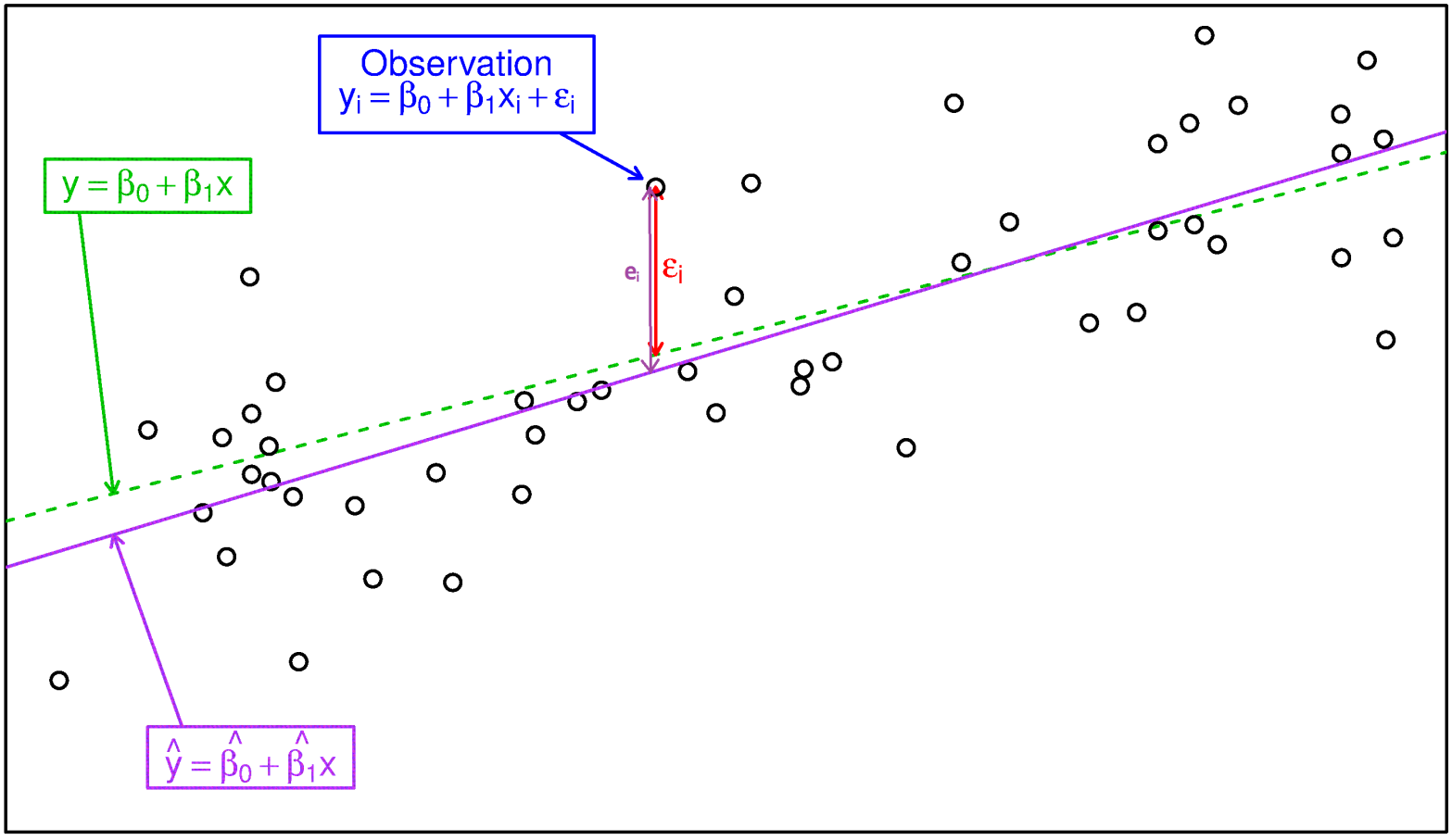
$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- The residuals have some useful properties including the following two:

$$\sum_{i=1}^{N} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} x_i e_i = 0.$$

- The residual e_i is different from the error ϵ_i

The residual e_i is different from the error ϵ_i



The residual e_i is different from the error ϵ_i

Forecasting with regression

- Forecasts from a simple linear model are easily obtained using the equation
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
- When this calculation is done using an observed value of x from the data, we call the resulting value of $\hat{y}$ a "fitted value". When the value of x is new (i.e., not part of the data that were used to estimate the model), the resulting value of $\hat{y}$ is a genuine forecast.

Goodness-of-fit

- A common way to summarize how well a linear regression model fits the data. It is calculated as the square of the correlation between the observed y values and the predicted $\hat{y}$ values

- Alternatively, it is also computed as

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

它也是回归模型考虑（或解释）的预测变量变化的比例。

- It is also the proportion of variation in the forecast variable that is accounted for (or explained) by the regression model.

- If the predictions are close to the actual values, we would expect R^2 to be close to 1. On the other hand, if the predictions are unrelated to the actual values, then $R^2 = 0$. In all cases, R^2 lies between 0 and 1.

- R is not reliable

Forecasting with regression

- Assuming that the regression errors are normally distributed, an approximate 95% **forecast interval** (also called a prediction interval) associated with this forecast is given by

预测间隔

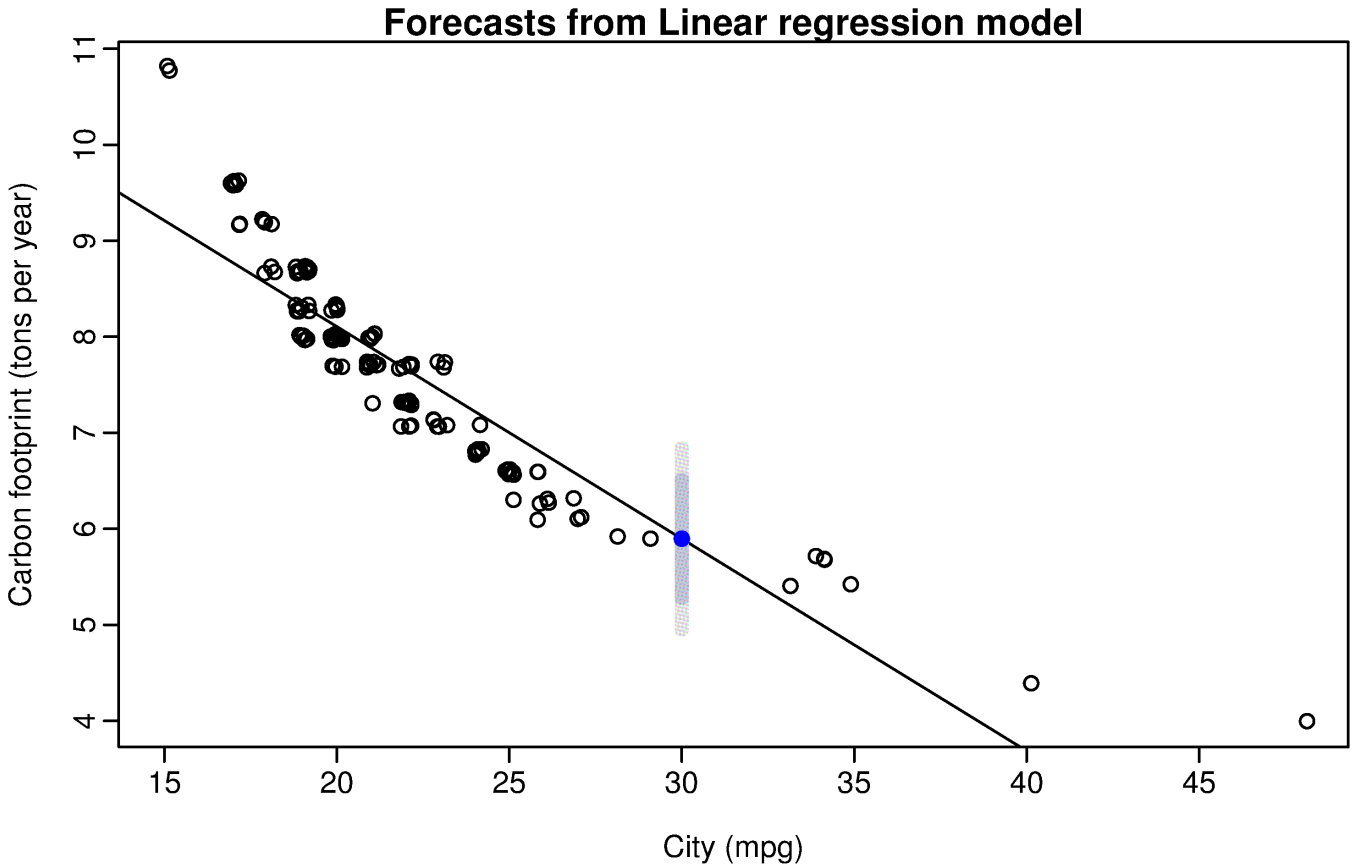
$$\hat{y} \pm 1.96 s_e \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{(N-1)s_x^2}},$$

where N is the total number of observations, $\bar{x}$ is the mean of the observed x values, s_x is the standard deviation of the observed x values and s_e is the standard error of the regression, which are defined as

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}; \quad s_e = \sqrt{\frac{1}{N-k-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

- The equation shows that the forecast interval is wider when x is far from $\bar{x}$. That is, we are more certain about our forecasts when considering values of the predictor variable close to its sample mean.

Forecasting with regression



Forecast with 80% and 95% forecast intervals for a car with $x = 30$ mpg in city driving.

See an example of using Python sklearn package for linear regression in `Lecture04_Example01.py`

Non-linear functional forms

- Although the assumption of linear relationship is often adequate, there are cases for which a non-linear functional form is more suitable.

- The most commonly used transformation is the (natural) logarithmic.

- A **log-log** functional form is specified as

$$\log y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i.$$

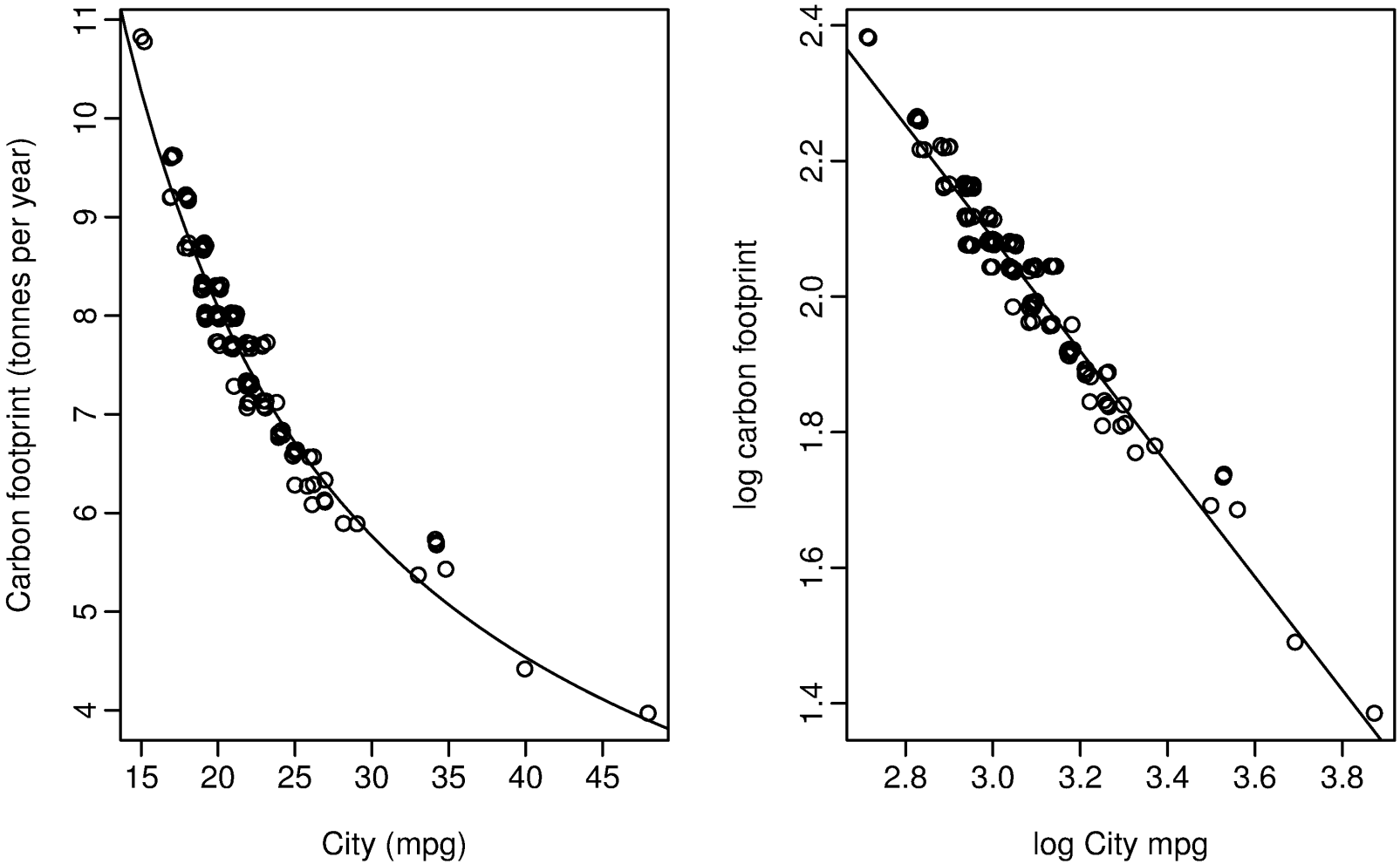
  In this model, the slope β_1 can be interpeted as an elasticity: β_1 is the average percentage change in y resulting from a 1% change in x.

- The model is equivalent to

$$y_i = e^{\beta_0} x_i^{\beta_1} e^{\epsilon_i} = B_0 x_i^{\beta_1} E_i$$

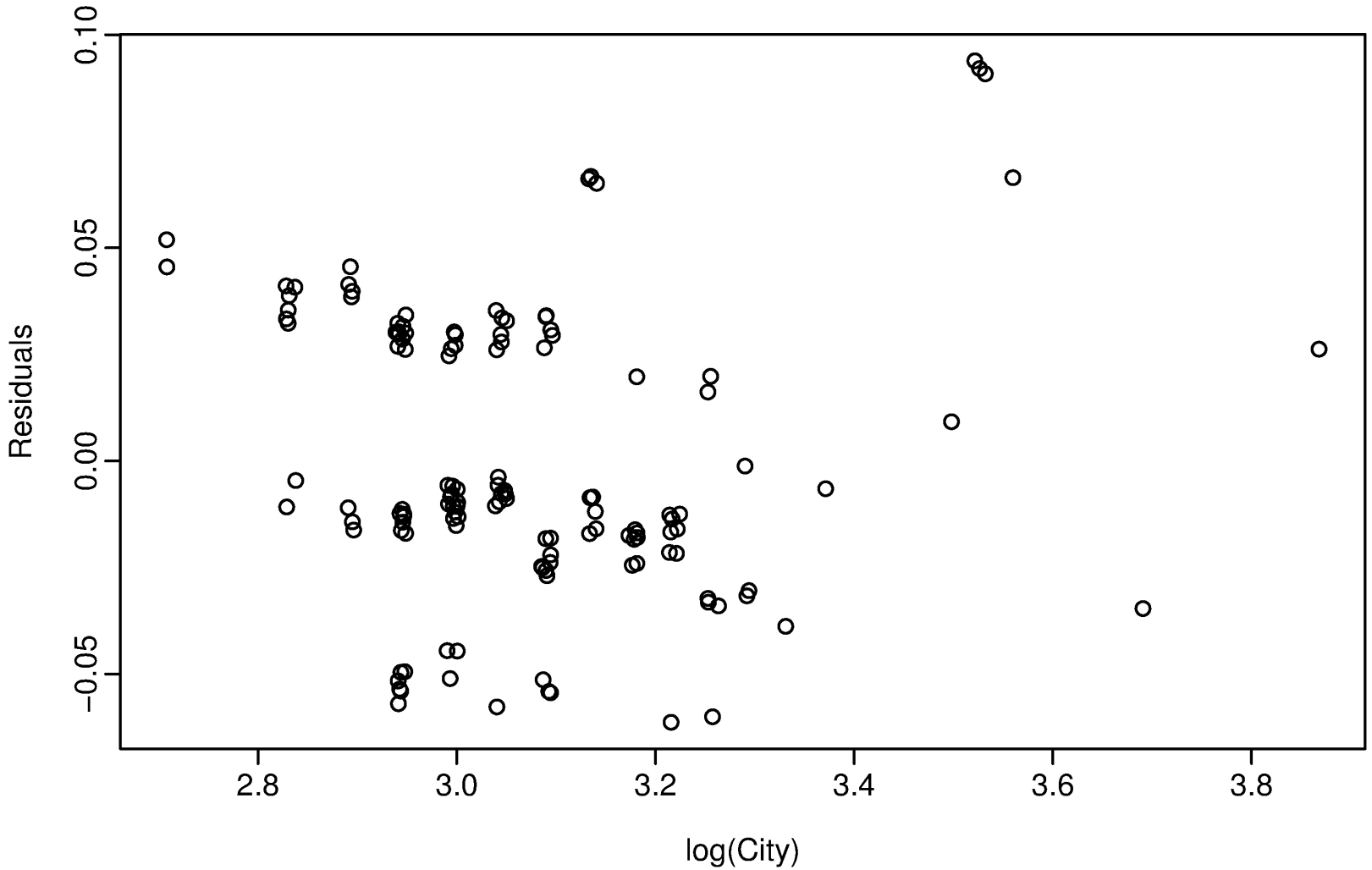
  with multiplicative errors rather than additive errors

Non-linear functional forms



Fitting a log-log functional form to the Car data example. Plots show the estimated relationship both in the original and the logarithmic scales.

Non-linear functional forms



Residual plot from estimating a log-log functional form for the Car data example.

Non-linear functional forms

Other useful forms are the log-linear form and the linear-log form:

Model	Functional form	Slope	Elasticity
linear	$y = \beta_0 + \beta_1 x$	β_1	$\beta_1 x/y$
log-log	$\log y = \beta_0 + \beta_1 \log x$	$\beta_1 y/x$	β_1
linear-log	$y = \beta_0 + \beta_1 \log x$	β_1/x	$\beta_1 y$
log-linear	$\log y = \beta_0 + \beta_1 x$	$\beta_1 y$	$\beta_1 x$

Summary of selected functional forms. Elasticities that depend on the observed values of y and x are commonly calculated for the sample means of these.

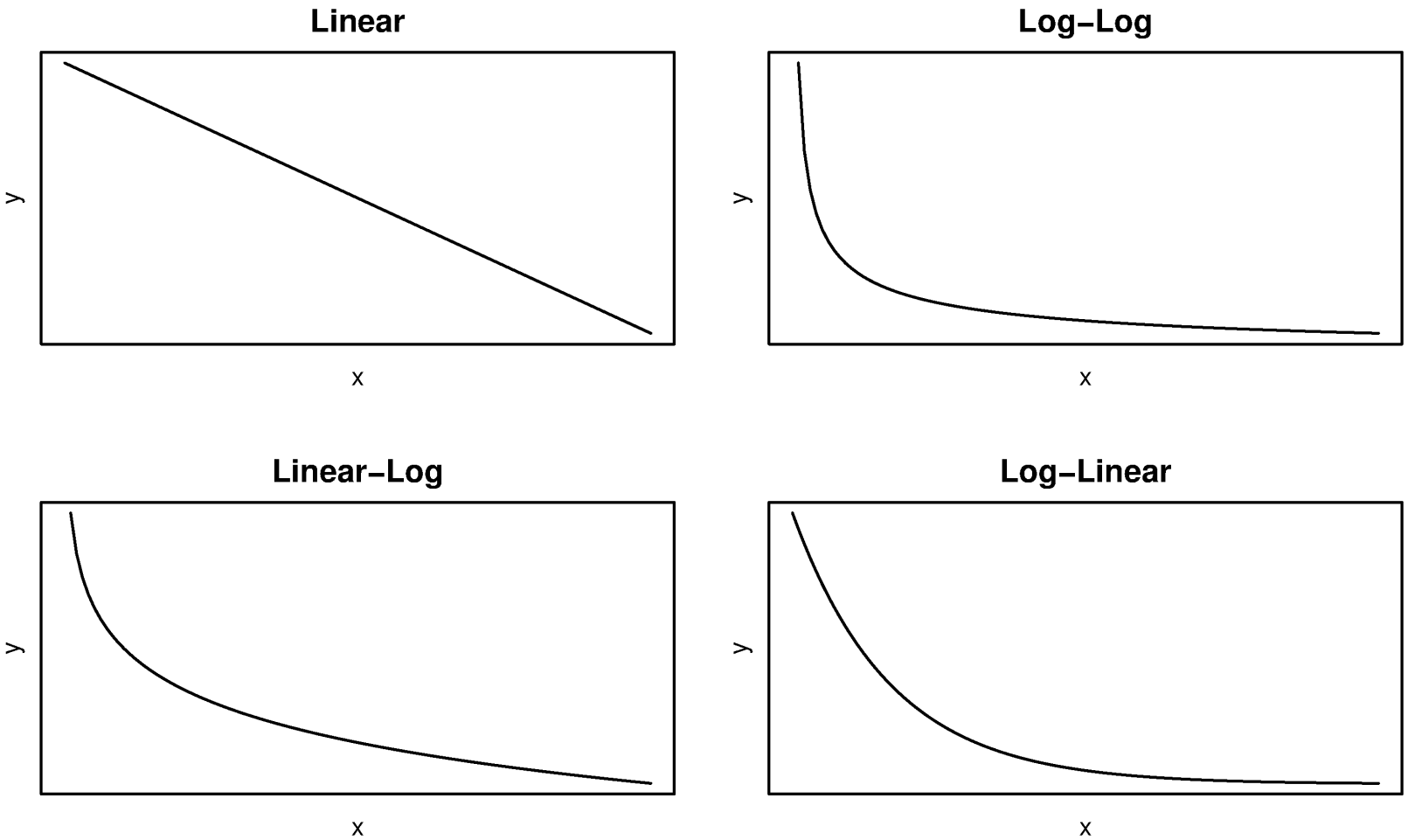
弹性

取决于y和x的观测值的弹性
通常针对这些的样本平均值计算

Note: the elasticity is defined as

$$e := \frac{d \log(y)}{d \log(x)} = \frac{dy}{dx} \frac{x}{y}$$

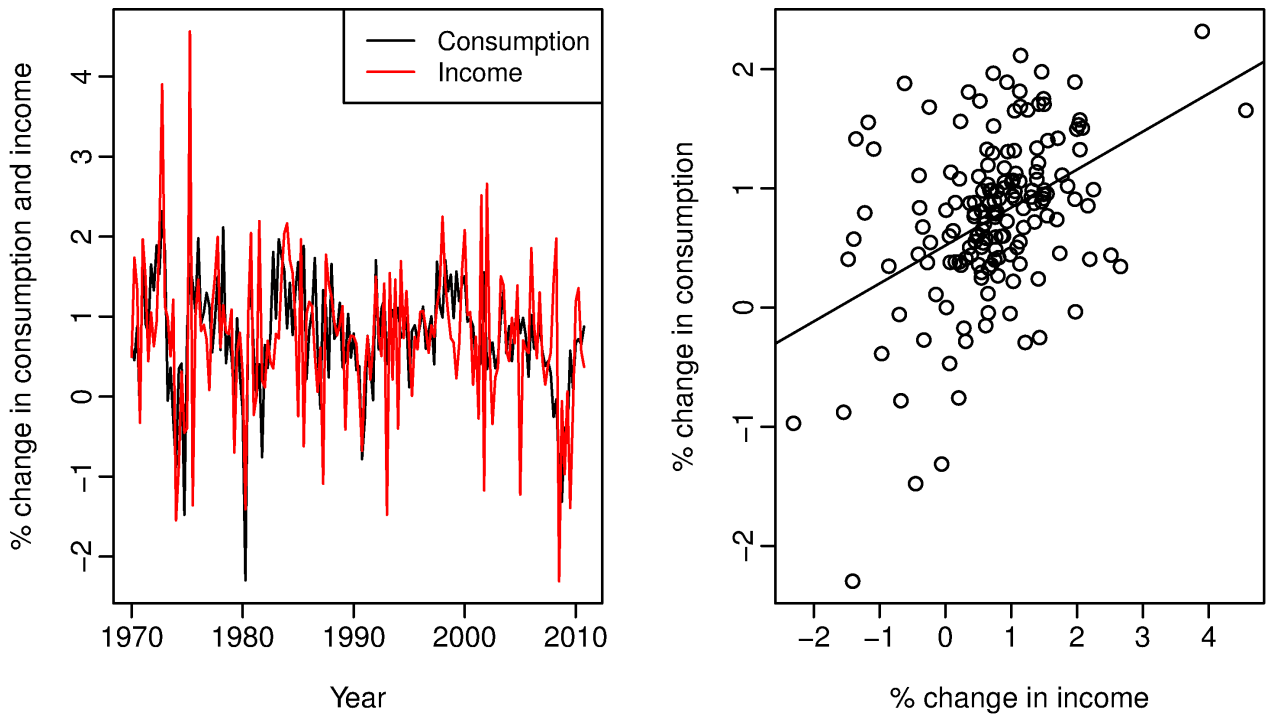
Non-linear functional forms



The four non-linear forms from the previous slide

Regression with time series data

- When using regression for prediction, we are often considering time series data and we are aiming to forecast the future. There are a few issues that arise with time series data but not with cross-sectional data that we will consider in this lecture.



Percentage changes in personal consumption expenditure for the US.

Regression with time series data

- The figure shows time series plots of quarterly percentage changes (growth rates) of real personal consumption expenditure (C_t) and real personal disposable income (I_t) for the US for the period March 1970 to Dec 2010. Also shown is a scatter plot including the estimated regression line

$$\hat{C}_t = 0.52 + 0.32 I_t$$

  We are interested in forecasting consumption for the four quarters of 2011.

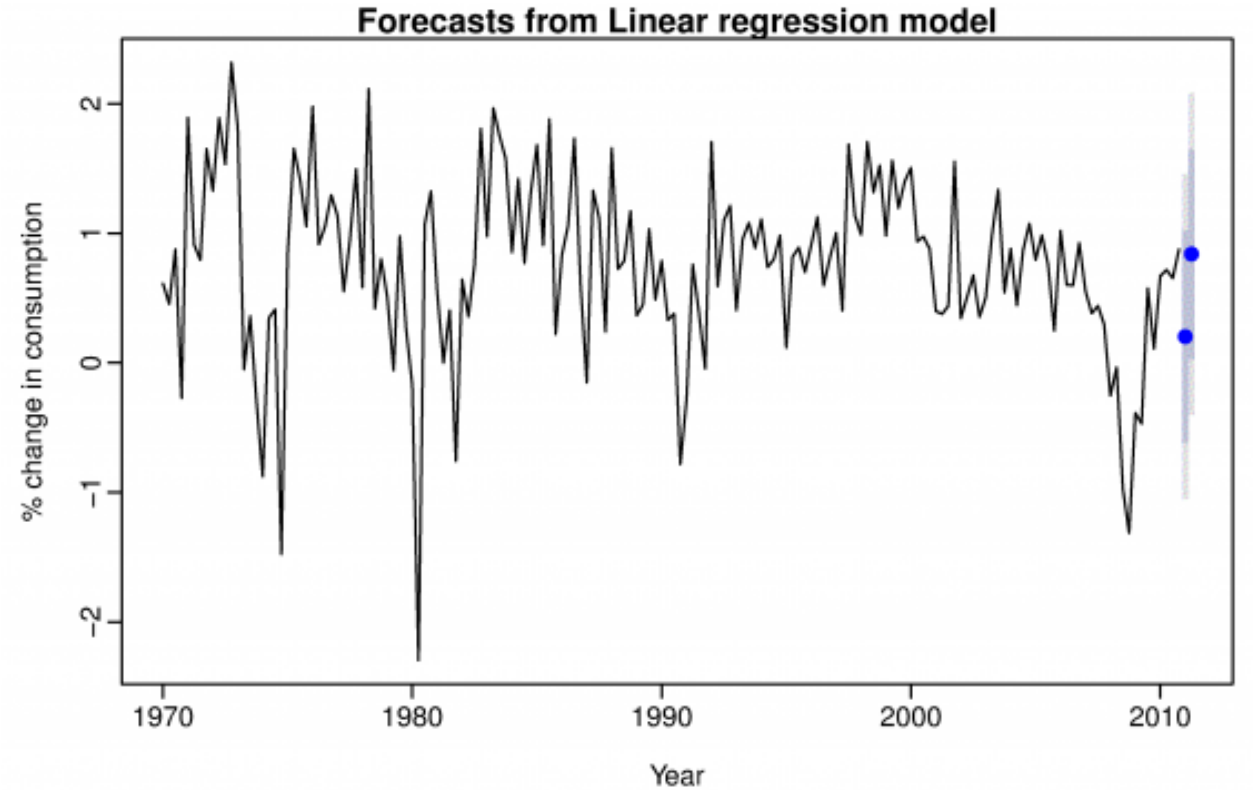
- Using a regression model to forecast time series data poses a challenge in that future values of the predictor variable (income in this case).

Scenario based forecasting

In this setting the forecaster assumes possible scenarios for the predictor variable that are of interest.

For example the US policy maker may want to forecast consumption if there is a 1% growth in income for each of the quarters in 2011. Alternatively a 1% decline in income for each of the quarters may be of interest.

Scenario based forecasting



Forecasting percentage changes in personal consumption expenditure for the US.

Scenario based forecasting

- Forecast intervals for scenario based forecasts do not include the uncertainty associated with the future values of the predictor variables. They assume the value of the predictor is known in advance.

- An alternative approach is to use genuine forecasts for the predictor variable.

Ex-ante versus ex-post forecasts

- When using regression models with time series data, we need to distinguish between two different types of forecasts that can be produced, depending on what is assumed to be known when the forecasts are computed.

- Ex ante forecasts are those that are made using only the information that is available in advance. For example, ex ante forecasts of consumption for the four quarters in 2011 should only use information that was available before 2011. These are the only genuine forecasts, made in advance using whatever information is available at the time.

- Ex post forecasts are those that are made using later information on the predictors. For example, ex post forecasts of consumption for each of the 2011 quarters may use the actual observations of income for each of these quarters, once these have been observed. These are not genuine forecasts, but are useful for studying the behaviour of forecasting models.

事后对2011年每季度的消费预测可以使用这些季度中每一季度的实际收入观察。一旦观察到这些
这些不是真正的预测，但对于研究预测模型的行为很有用

Ex-ante versus ex-post forecasts

不应使用预测期的数据估算产生事后预测的模型。
也就是说，事后预测可以假设预测变量（x变量）的知识，
但不应该假设要预测的数据（y变量）的知识。

- The model from which ex-post forecasts are produced should not be estimated using data from the forecast period. That is, ex-post forecasts can assume knowledge of the predictor variable (the x variable), but should not assume knowledge of the data that are to be forecast (the y variable).

- A comparative evaluation of ex ante forecasts and ex post forecasts can help to separate out the sources of forecast uncertainty. This will show whether forecast errors have arisen due to poor forecasts of the predictor or due to a poor forecasting model.

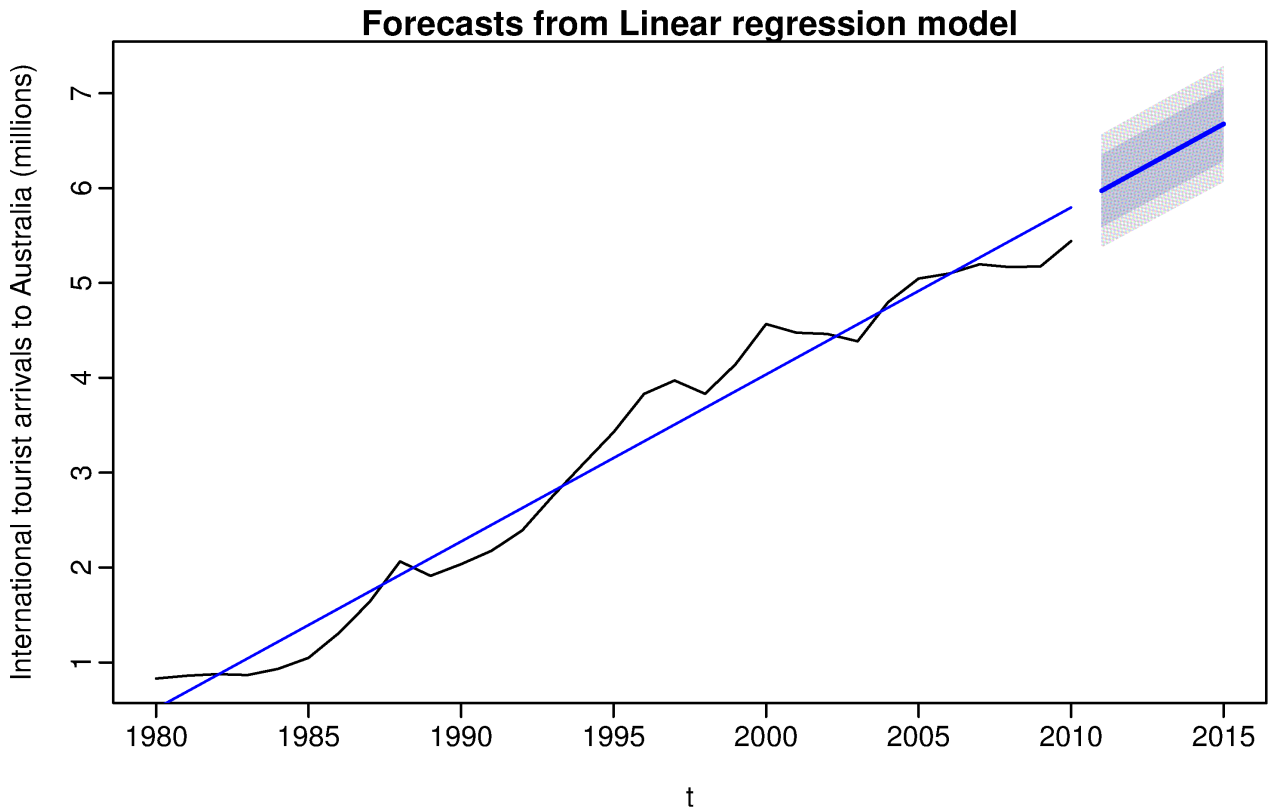
Linear trend

- Using regression we can model and forecast the trend in time series data by including $t = 1, \ldots, T$, as a predictor variable:

$$T_t = \beta_0 + \beta_1 t + \varepsilon_t.$$

- The following figure shows a time series plot of aggregate tourist arrivals to Australia over the period 1980 to 2010 with the fitted linear trend line $\hat{T}_t = 0.3375 + 0.1761t$. Also plotted are the point and forecast intervals for the years 2011 to 2015.

Linear trend



Forecasting international tourist arrivals to Australia for the period 2011-2015 using a linear trend. 80% and 95% forecast intervals are shown.

Residual autocorrelation

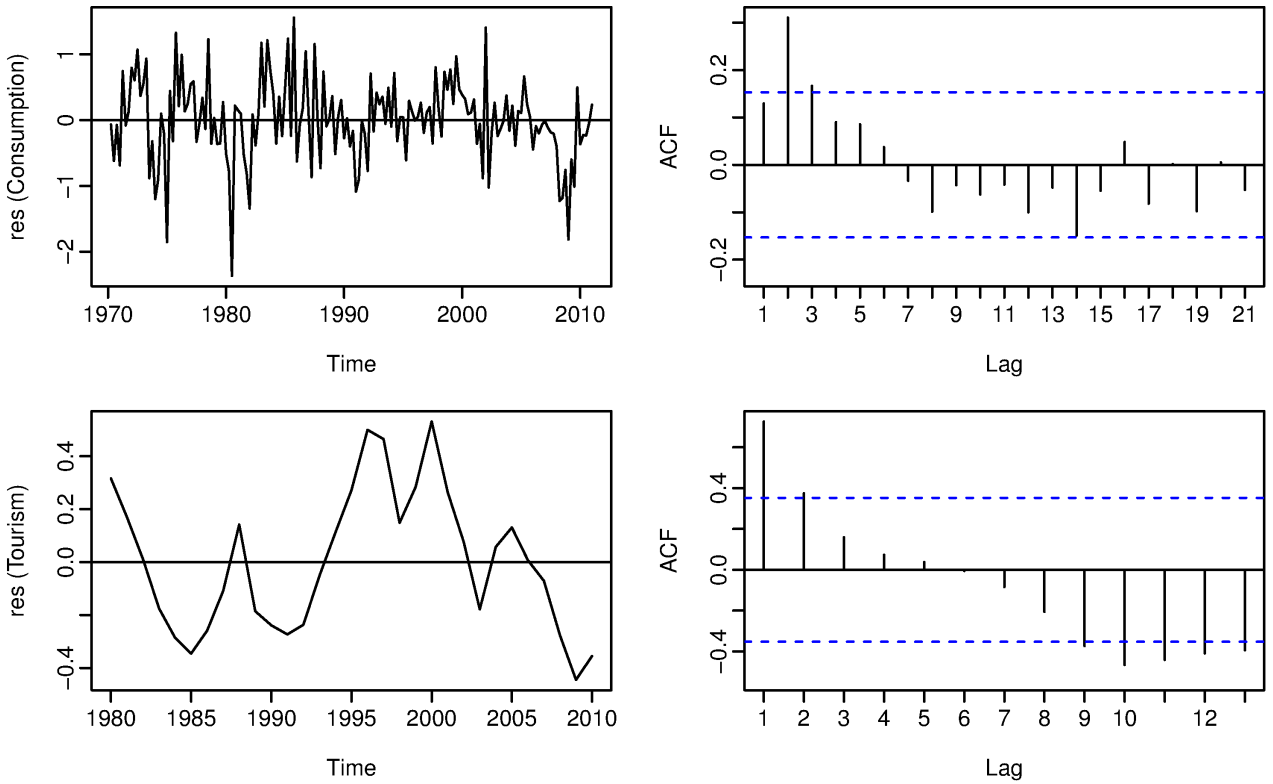
- With time series data it is highly likely that the value of a variable observed in the current time period will be influenced by its value in the previous period, or even the period before that, and so on.

自相关

- When fitting a regression model to time series data, it is very common to find autocorrelation in the residuals, which violates the assumption of no autocorrelation in the errors.

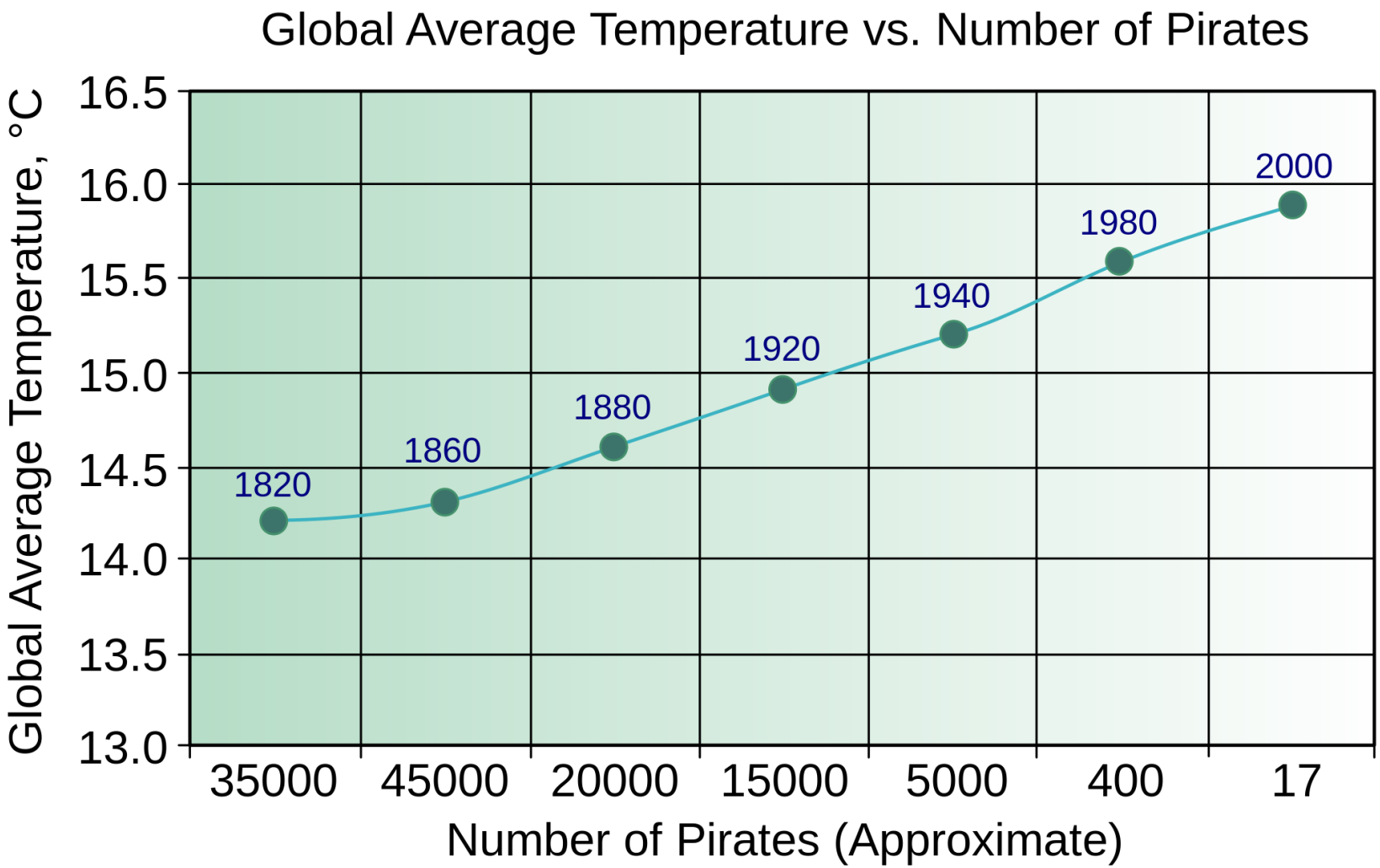
- Some information left over should be utilized in order to obtain better forecasts.

Residual autocorrelation



Residuals from the regression models for Consumption and Tourism. Because these involved time series data, it is important to look at the ACF of the residuals to see if there is any remaining information not accounted for by the model.

The problem of spurious regressions

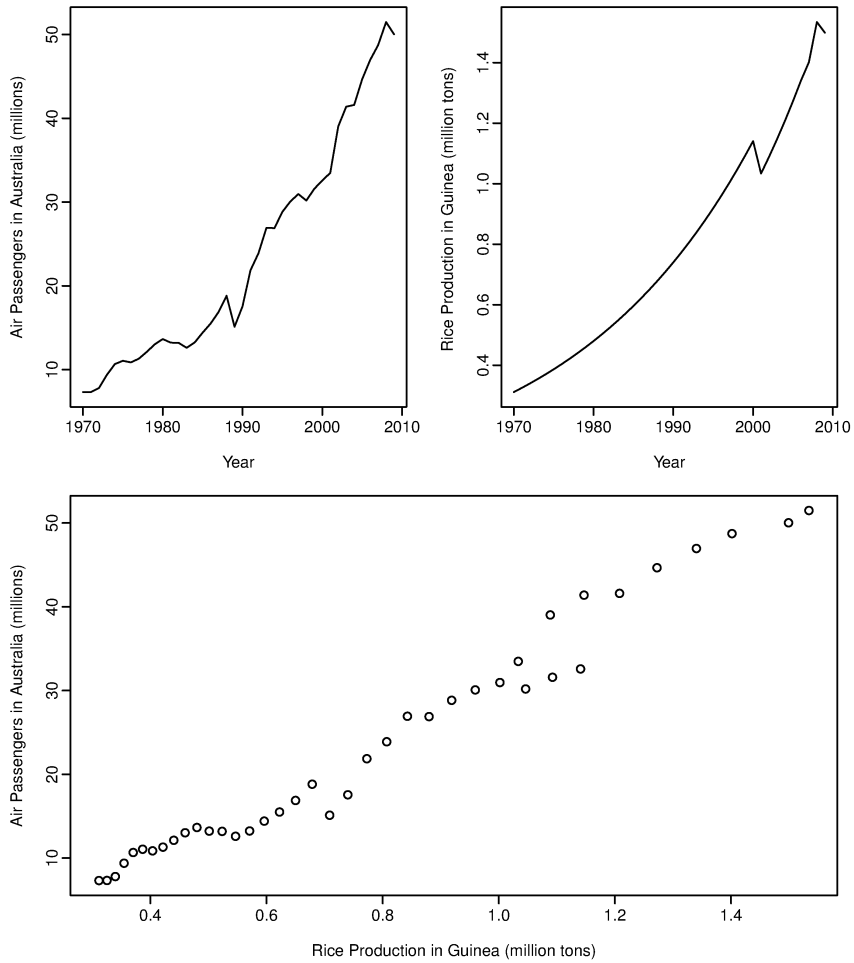


Spurious regression

- More often than not, time series data are non-stationary; that is, the values of the time series do not fluctuate around a constant mean or with a constant variance. We need to address the effect non-stationary data can have on regression models.

- For example consider the two variables plotted in below, which appear to be related simply because they both trend upwards in the same manner. However, air passenger traffic in Australia has nothing to do with rice production in Guinea.

Spurious regression



Trending time series data can appear to be related, as shown in this example in which air passengers in Australia are regressed against rice production in Guinea.

Spurious regression

虚假的

- Regressing non-stationary time series can lead to spurious regressions. High R^2s and high residual autocorrelation can be signs of spurious regression.

- Cases of spurious regression might appear to give reasonable short-term forecasts, but they will generally not continue to work into the future.

Multiple linear regression

- The general form of a multiple regression is

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + e_i,$$

  where y_i is the variable to be forecast and $x_{1,i}, \ldots, x_{k,i}$ are the k predictor variables. Each of the predictor variables must be numerical. The coefficients measure the marginal effects of the predictor variables (that is, holding all others constant).

- The assumptions regarding the error term are the same as before.

Organising Data

Item	Target	Predictor 1	Predictor 2		Predictor k
1	y_1	$x_{1,1}$	$x_{2,1}$	...	$x_{k,1}$
2	y_2	$x_{1,2}$	$x_{2,2}$	...	$x_{k,2}$
3	y_3	$x_{1,3}$	$x_{2,3}$	...	$x_{k,3}$
⋮	⋮	⋮	⋮	...	⋮
i	y_i	$x_{1,i}$	$x_{2,i}$	...	$x_{k,i}$
⋮	⋮	⋮	⋮	...	⋮
N	y_N	$x_{1,N}$	$x_{2,N}$	...	$x_{k,N}$

Estimation of the model

- The values of the coefficients $\beta_0, \ldots, \beta_k$ are obtained by finding the minimum sum of squares of the errors. That is, we find the values of $\beta_0, \ldots, \beta_k$ which minimize

$$\ell(\beta_0, \beta_1, ..., \beta_k) := \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{1,i} - \cdots - \beta_k x_{k,i})^2.$$

- This is called "least squares" estimation because it the least value of the sum of squared errors. In practice, the calculation is always done using a computer package. Finding the best estimates of the coefficients is often called "fitting" the model to the data.

- When we refer to the estimated coefficients, we will use the notation $\hat{\beta}_0, \ldots, \hat{\beta}_k$.

Fitted values, forecast values, residuals and R^2

- Predictions of y can be calculated by ignoring the error in the regression equation. That is

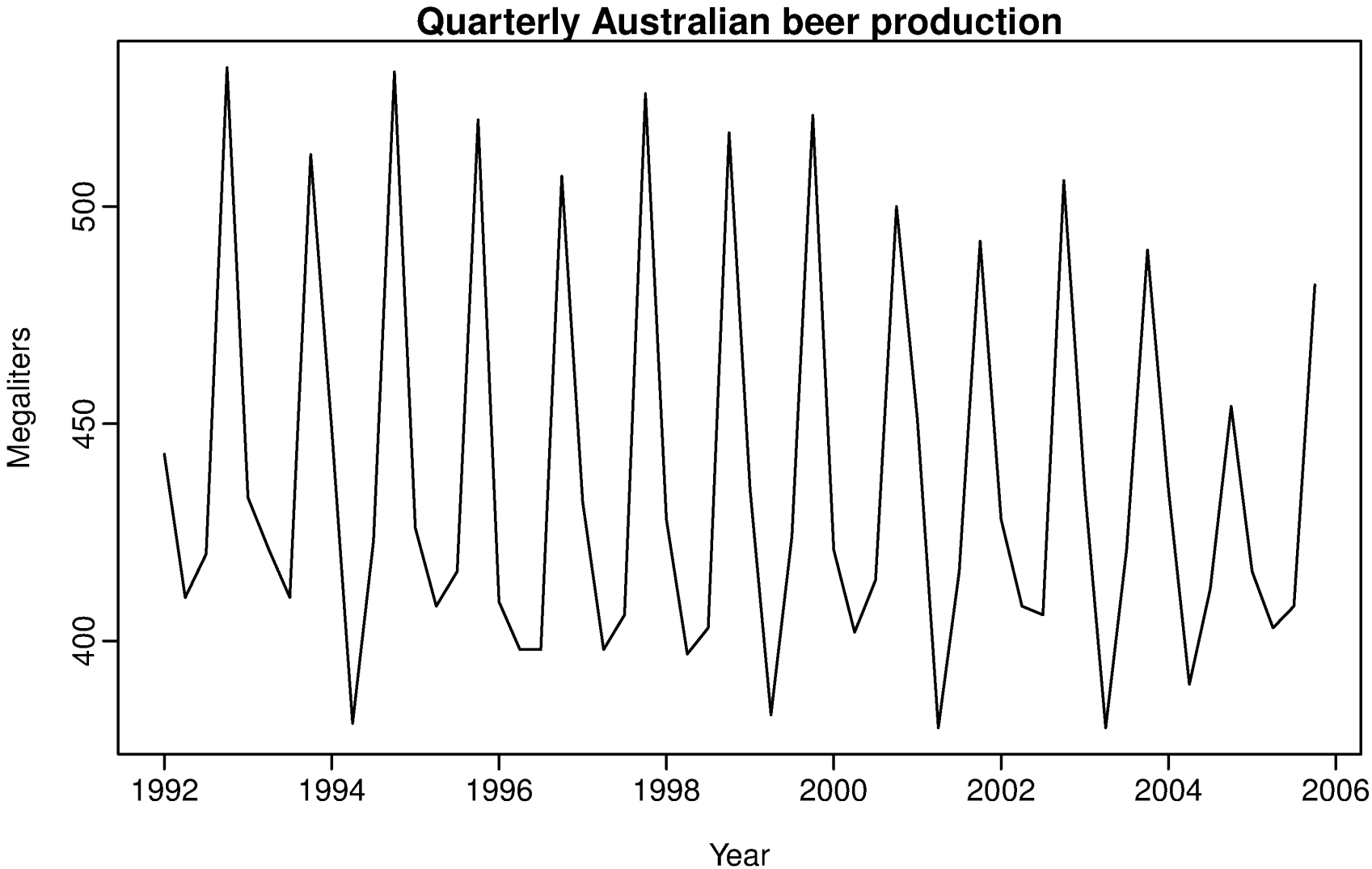
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k.$$

  Plugging in values of $x_1, \ldots, x_k$ into the right hand side of this equation gives a prediction of y for that combination of predictors.

- The value of R^2 can also be calculated as the proportion of variation in the forecast variable that is explained by the regression model:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Australian quarterly beer production



Australian quarterly beer production.

Seasonal Dummy variables

- Suppose we are forecasting daily data with weekly patterns. Then the following dummy variables can be created.

Day (t)	D1	D2	D3	D4	D5	D6
Sunday	0	0	0	0	0	0
Monday	1	0	0	0	0	0
Tuesday	0	1	0	0	0	0
Wednesday	0	0	1	0	0	0
Thursday	0	0	0	1	0	0
Friday	0	0	0	0	1	0
Saturday	0	0	0	0	0	1
Sunday	0	0	0	0	0	0
Monday	1	0	0	0	0	0
Tuesday	0	1	0	0	0	0
Wednesday	0	0	1	0	0	0
Thursday	0	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Seasonal Dummy variables

- Notice that only six dummy variables are needed to code seven categories. That is because the seventh category (in this case Sunday) is specified when the dummy variables are all set to zero.
- Putting a seventh dummy variable for the seventh category is known as the "dummy variable trap" because it will cause the regression to fail.
- The general rule is to use one fewer dummy variables than categories. So for quarterly data, use three dummy variables; for monthly data, use 11 dummy variables; and for daily data, use six dummy variables.
- The interpretation of each of the coefficients associated with the dummy variables is that it is a measure of the effect of that category relative to the omitted category.

Explanation

- See `Lecture04_Example04.py`, consider a synthetic four-seasonal time series

  $2.01,\ -1.99,\ 1.98,\ -2.05,\ 1.89,\ -2.0,\ 1.93,\ -1.95,\ 2.10,\ -1.79,\ 1.87,\ -1.85$

  with a horizontal trend
- Four seasons data with (horizontal) linear trend

$$\widehat{y}_t = \beta_0 + 0 * t + \beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t}$$

where dummy variables $d_{2,t}, d_{3,t}, d_{4,t}$ are defined as
when t is a time of season ONE, then $d_{2,t} = d_{3,t} = d_{4,t} = 0$;
when t is a time of season TWO, then $d_{2,t} = 1, d_{3,t} = d_{4,t} = 0$;
when t is a time of season THREE, then $d_{2,t} = 0, d_{3,t} = 1, d_{4,t} = 0$;
when t is a time of season FOUR, then $d_{2,t} = 0, d_{3,t} = 0, d_{4,t} = 1$;

Explanation: conti...

- Copy the model here again

$$\widehat{y}_t = \beta_0 + \beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t}$$

- Question: What is the model forecast when $t = 1$? (We pretend we know $\beta_0,\ \beta_2,\ \beta_3,\ \beta_4$)
- When $t = 1$, it is season One. So $d_{2,1} = d_{3,1} = d_{4,1} = 0$, hence

$$\widehat{y}_1 = \beta_0$$

- When $t = 2$, what is the model forecast? $\widehat{y}_2 = \beta_0 + \beta_2$
- How about $t = 7$ and 8? $\widehat{y}_7 = \beta_0 + \beta_3$ and $\widehat{y}_8 = \beta_0 + \beta_4$

Explanation: conti...

- All the model forecast along the observed data are

2.01,	-1.99,	1.98,	-2.05,	1.89,	-2.0,
β_0	$\beta_0 + \beta_2$	$\beta_0 + \beta_3$	$\beta_0 + \beta_4$	β_0	$\beta_0 + \beta_2$
1.93,	-1.95,	2.10,	-1.79,	1.87,	-1.85
$\beta_0 + \beta_3$	$\beta_0 + \beta_4$	β_0	$\beta_0 + \beta_2$	$\beta_0 + \beta_3$	$\beta_0 + \beta_4$

- Finally we estimate $\beta_0, \beta_2, \beta_3, \beta_4$ such that the model forecasted series

$$\beta_0, \beta_0 + \beta_2, \beta_0 + \beta_3, \beta_0 + \beta_4, \beta_0, \beta_0 + \beta_2, \beta_0 + \beta_3, \beta_0 + \beta_4, \beta_0, \beta_0 + \beta_2, \beta_0 + \beta_3, \beta_0 + \beta_4$$

is as close as possible to the observed series

$$2.01,\ -1.99,\ 1.98,\ -2.05,\ 1.89,\ -2.0,\ 1.93,\ -1.95,\ 2.10,\ -1.79,\ 1.87,\ -1.85$$

- We can see the coefficients associated with the other seasons are measures of the difference between those seasons and the first season.

Example: Australian quarterly beer production

- We can model the Australian beer production data using a regression model with a linear trend and quarterly dummy variables:

$$y_t = \beta_0 + \beta_1 t + \beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t} + e_t,$$

  here $d_{i,t} = 1$ if t is in quarter i and 0 otherwise. The first quarter variable has been omitted, so the coefficients associated with the other quarters are measures of the "*difference*" between those quarters and the first quarter.
- Why? Consider the time (quarterly) $t = 14$ which is the second quarter of year 3. Hence $d_{2,14} = 1$ and $d_{3,14} = d_{4,14} = 0$. The prediction for this time point is

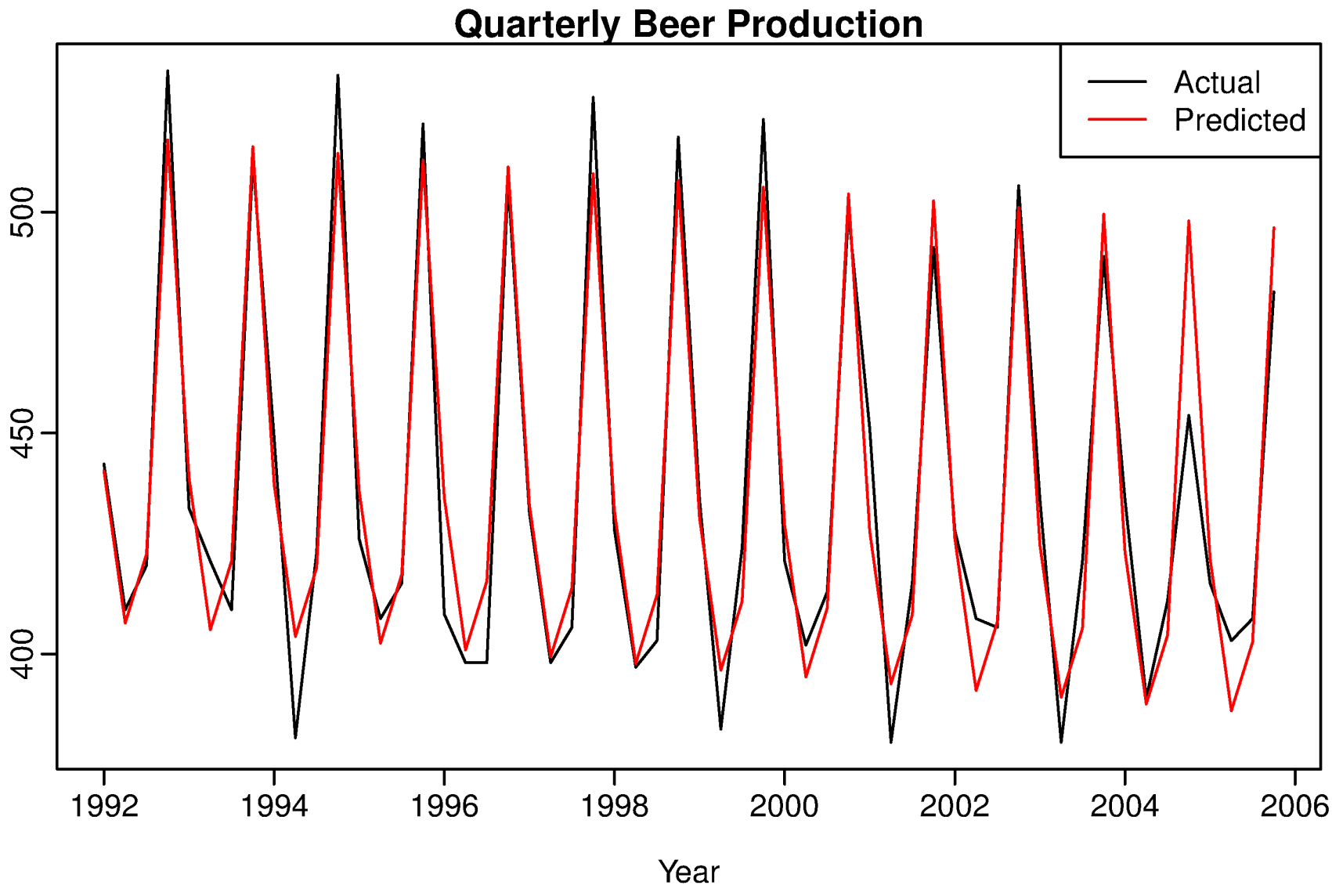
$$\widehat{y}_{\text{2nd quarter of year 3}} = \beta_0 + \beta_1 \times 14 + \beta_2$$

- What is the prediction for the first quarter of year 3, corresponding to $d_{2,13} = d_{3,13} = d_{4,13} = 0$, which give

$$\widehat{y}_{\text{1st quarter of year 3}} = \beta_0 + \beta_1 \times 13$$

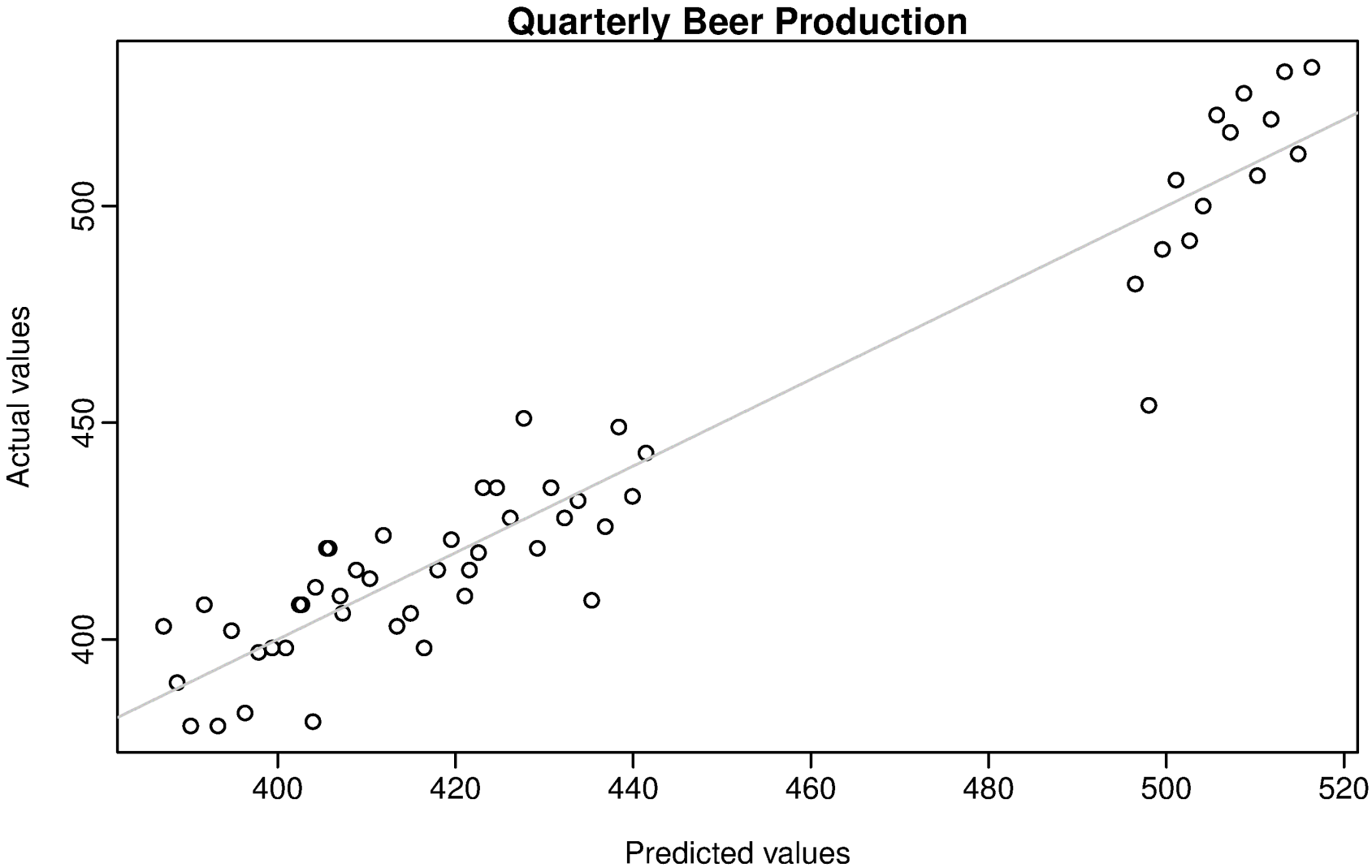
  Hence $\beta_2 = \widehat{y}_{\text{2nd quarter of year 3}} - \widehat{y}_{\text{1st quarter of year 3}} - \beta_1$

Example: Australian quarterly beer production



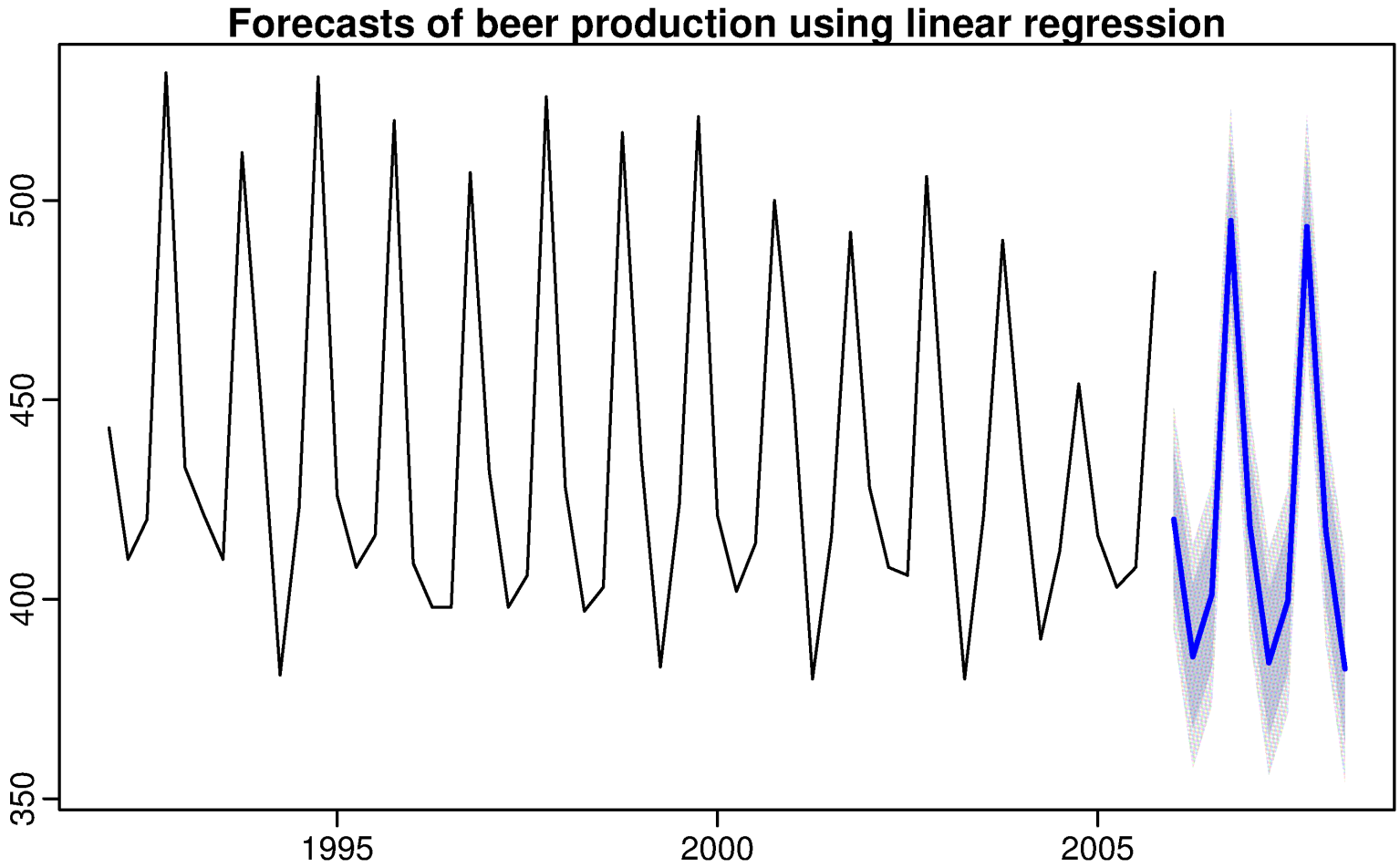
Time plot of beer production and predicted beer production.

Example: Australian quarterly beer production



Actual beer production plotted against predicted beer production.

Example: Australian quarterly beer production



Forecasts from the regression model for beer production. The dark blue region shows 80% prediction intervals and the light blue region shows 95% prediction intervals.

Other common dummy predictors

- See `Lecture04_Example02.py` if we cannot rightly identify the seasonal period, then we may assume a wrong model.

- Outliers: If there is an outlier in the data, rather than omit it, you can use a dummy variable to remove its effect. In this case, the dummy variable takes value one for that observation and zero everywhere else.

- Public holidays: For daily data, the effect of public holidays can be accounted for by including a dummy variable predictor taking value one on public holidays and zero elsewhere.

Trend

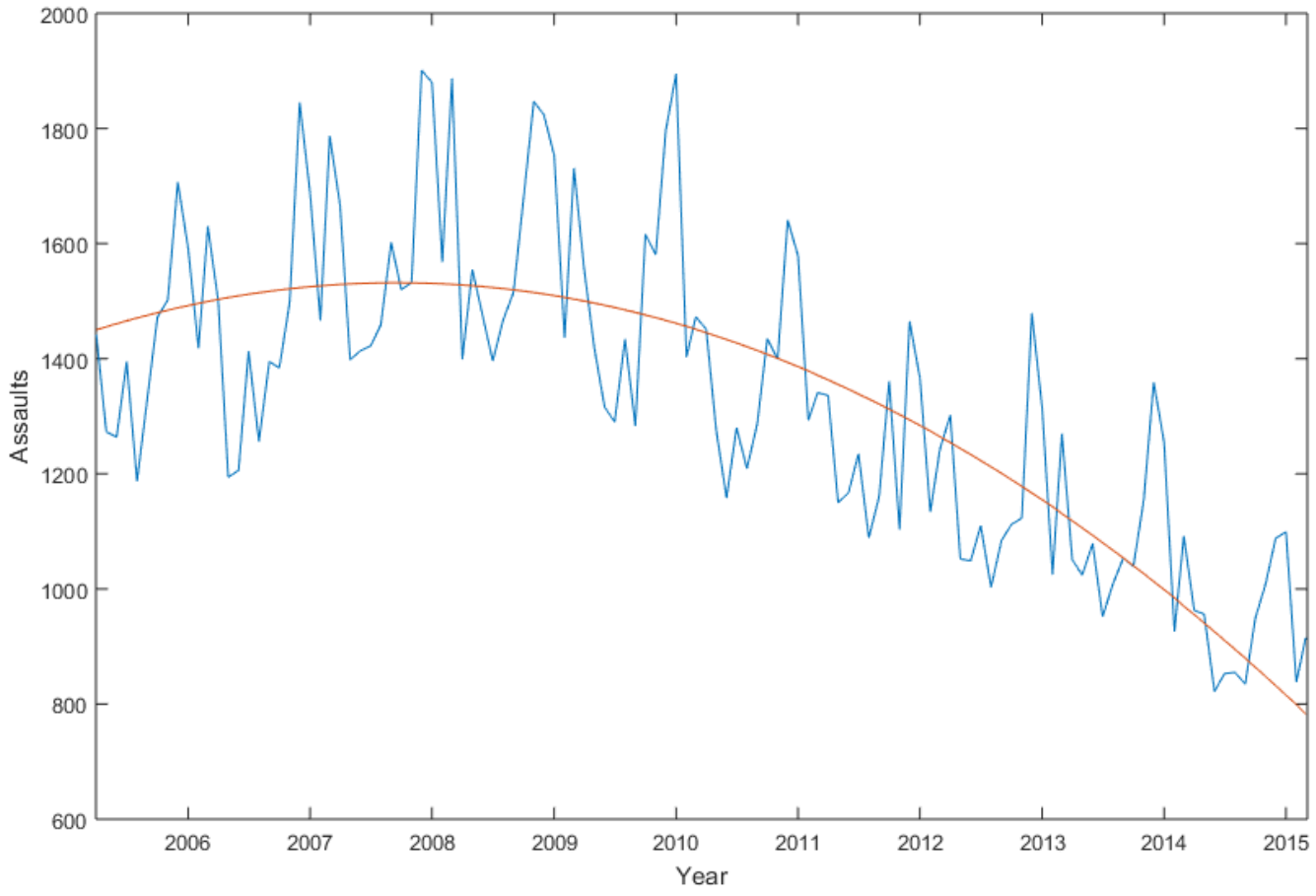
- A linear trend is easily accounted for by including the predictor $x_{1,t} = t$.
- A quadratic or higher order trend is obtained by specifying

$$x_{1,t} = t, \quad x_{2,t} = t^2, \quad \ldots$$

- However, it is not recommended that quadratic or higher order trends are used in forecasting. When they are extrapolated, the resulting forecasts are often very unrealistic. 推断

Example: a quadratic trend

Figure: Alcohol related assaults in NSW



See another example in `Lecture04_Example03.py`

Trend

- A better approach is to use a piecewise linear trend which bends at some time. If the trend bends at time τ, then it can be specified by including the following predictors in the model.

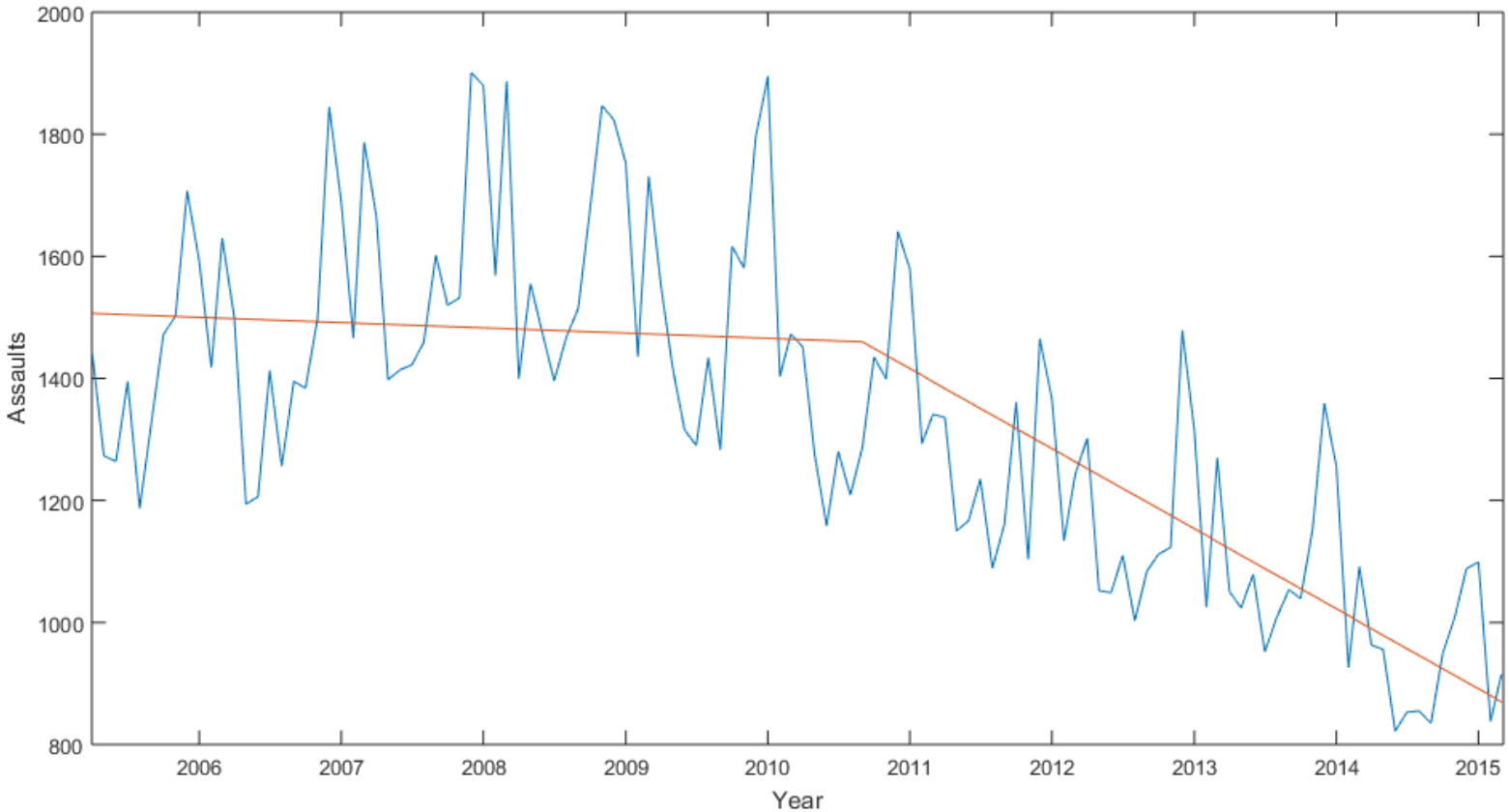
$$x_{1,t} = t$$

$$x_{2,t} = \begin{cases} 0 & t < \tau \\ (t - \tau) & t \geq \tau \end{cases}$$

- If the associated coefficients of $x_{1,t}$ and $x_{2,t}$ are β_1 and β_2, then β_1 gives the slope of the trend before time τ, while the slope of the line after time τ is given by $\beta_1 + \beta_2$.

Example: piecewise trend

Figure: Alcohol related assaults in NSW



Intervention variables

- It is often necessary to model interventions (干涉) that may have affected the variable to be forecast. For example, competitor activity, advertising expenditure, industrial action, and so on, can all have an effect.
- When the effect lasts only for one period, we use a spike variable. This is a dummy variable taking value one in the period of the intervention and zero elsewhere. A spike variable is equivalent to a dummy variable for handling an outlier. For example

$$x_t = \begin{cases} 1 & t = t_0 (\text{intervention time}) \\ 0 & \text{otherwise} \end{cases}$$

- Other interventions have an immediate and permanent effect. If an intervention causes a level shift (i.e., the value of the series changes suddenly and permanently from the time of intervention), then we use a step variable. For example

$$x_t = \begin{cases} 1 & t >= t_0 (\text{intervention time}) \\ 0 & t < t_0 \end{cases}$$

Trading days

- The number of trading days in a month can vary considerably and can have a substantial effect on sales data. To allow for this, the number of trading days in each month can be included as a predictor. An alternative that allows for the effects of different days of the week has the following predictors.

$$
\begin{aligned}
x_1 &= \#\text{ Mondays in month;} \\
x_2 &= \#\text{ Tuesdays in month;} \\
&\vdots \\
x_7 &= \#\text{ Sundays in month.}
\end{aligned}
$$

Distributed lags

- It is often useful to include advertising expenditure as a predictor. However, since the effect of advertising can last beyond the actual campaign, we need to include lagged values of advertising expenditure. So the following predictors may be used.

$$
\begin{aligned}
x_1 &= \text{advertising for previous month;} \\
x_2 &= \text{advertising for two months previously;} \\
&\vdots \\
x_m &= \text{advertising for } m \text{ months previously.}
\end{aligned}
$$

Selecting predictors: adjusted R^2

- Computer output for regression will always give the R^2 value. However, it is not a good measure of the predictive ability of a model.
- In addition, R^2 does not take overfitting into account. Adding any variable tends to increase the value of R^2, even if that variable is irrelevant. For these reasons, forecasters should not use R^2 to determine whether a model will give good predictions.
- An equivalent idea is to select the model which gives the minimum sum of squared errors (SSE), given by $\text{SSE} = \sum_{i=1}^{N} e_i^2$.
- Minimizing the SSE is equivalent to maximizing R^2 and will always choose the model with the most variables, and so is not a valid way of selecting predictors.

Selecting predictors: adjusted R^2

- An alternative, designed to overcome these problems, is the adjusted R^2:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1},$$

where N is the number of observations and k is the number of predictors. This is an improvement on R^2 as it will no longer increase with each added predictor.

- Maximizing $\bar{R}^2$ is equivalent to minimizing the following estimate of the variance of the forecast errors:

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N-k-1}.$$

Maximizing $\bar{R}^2$ works well as a method of selecting predictors, although it does tend to err on the side of selecting too many predictors.

Cross-validation

- Cross-validation is a very useful way of determining the predictive ability of a model (see also Section 5.2). In general, leave-one-out cross-validation for regression can be carried out using the following steps.
  1. Remove observation i from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation (x_i, y_i). (This is not the same as the residual because the i-th observation was not used in estimating the value of $\hat{y}_i$.)
  2. Repeat step 1 for $i = 1, 2, ..., N$.
  3. Compute the MSE from $e_1^*, e_2^*, ..., e_N^*$. We shall call this the CV.
- For many forecasting models, this is a time-consuming procedure, but for regression there are very fast methods of calculating CV so it takes no longer than fitting one model to the full data set. The equation for computing CV is given in Section 5.7 `https://otexts.com/fpp2/regression-matrices.html#regression-matrices`.
- Under this criterion, the best model is the one with the smallest value of CV.

Akaike's Information Criterion

- We define Akaike's Information Criterion as

$$\text{AIC} = N \log\left(\frac{\text{SSE}}{N}\right) + 2(k+2),$$

where N is the number of observations used for estimation and k is the number of predictors in the model. Different computer packages use slightly different definitions for the AIC. The $k+2$ part of the equation occurs because there are $k+2$ parameters in the model — the k coefficients for the predictors, the intercept and the variance of the residuals.

- The model with the minimum value of the AIC is often the best model for forecasting. For large values of N, minimizing the AIC is equivalent to minimizing the CV value.

Schwarz Bayesian Information Criterion

- A related measure is Schwarz's Bayesian Information Criterion (known as SBIC, BIC or SC):

$$\text{BIC} = N \log\left(\frac{\text{SSE}}{N}\right) + (k+2)\log(N).$$

- As with the AIC, minimizing the BIC is intended to give the best model. The model chosen by BIC is either the same as that chosen by AIC, or one with fewer terms. This is because BIC penalizes the SSE more heavily than the AIC.

- Many statisticians like to use BIC because it has the feature that if there is a true underlying model, then with enough data the BIC will select that model. 基础模型

Best subset regression

- Where possible, all potential regression models can be fitted and the best one selected based on one of the measures discussed here. This is known as "best subsets" regression or "all possible subsets" regression.

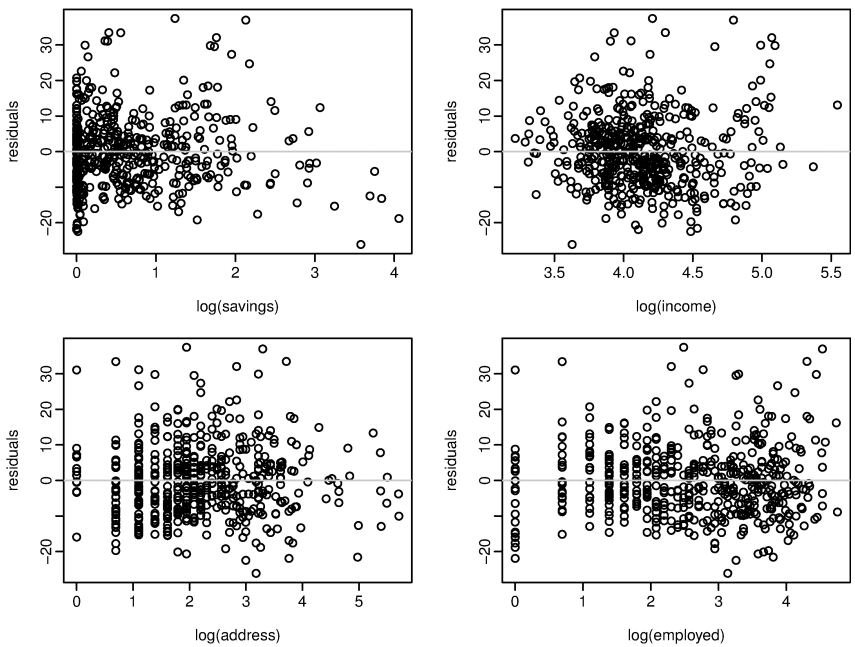
- It is recommended that one of CV, AIC or AICc be used for this purpose. If the value of N is large enough, they will all lead to the same model.

- While $\bar{R}^2$ is very widely used, and has been around longer than the other measures, its tendency to select too many variables makes it less suitable for forecasting than either AIC or AICc.

Residual diagnostics

Scatterplots of residuals against predictors

- Do a scatterplot of the residuals against each predictor in the model. If these scatterplots show a pattern, then the relationship may be nonlinear and the model will need to be modified accordingly.

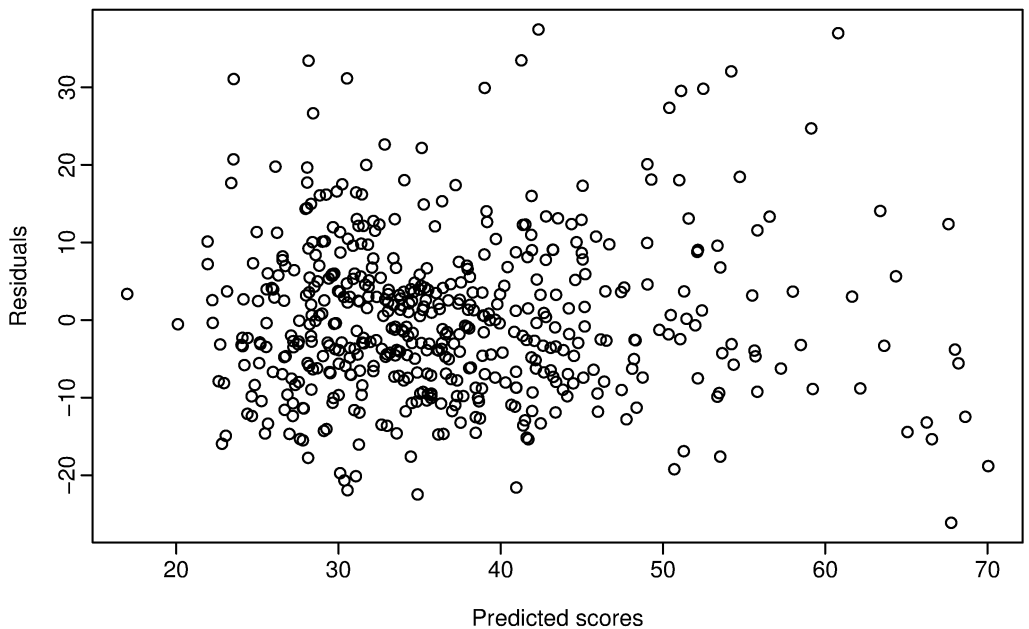


- It is also necessary to plot the residuals against each predictor NOT in the model. If these show a pattern, then the predictor may need to be added to the model.

Residual diagnostics

Scatterplot of residuals against fitted values

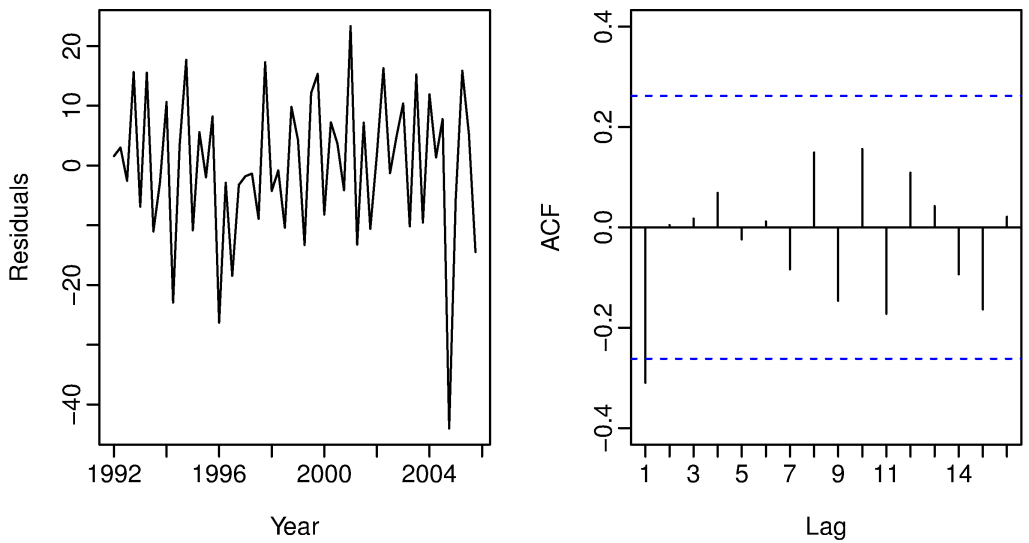
- A plot of the residuals against the fitted values should show no pattern. If a pattern is observed, there may be "heteroscedasticity" in the errors.



Residual diagnostics

Autocorrelation in the residuals

- There is an outlier in the residuals (2004:Q4) which suggests there was something unusual happening in that quarter.
- There is a small amount of autocorrelation left in the residuals (seen in the significant spike in the ACF plot)

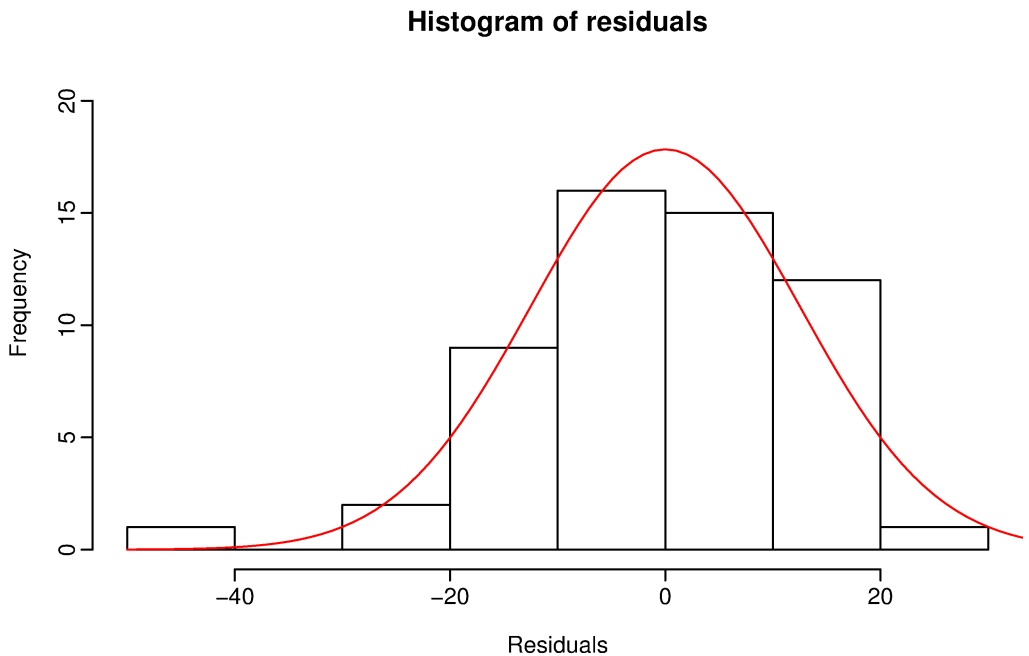


Residuals from the regression model for beer production.

Residual diagnostics

Histogram of residuals

- Finally, it is a good idea to check if the residuals are normally distributed.
- The residuals seem to be slightly negatively skewed, although that is probably due to the outlier.



Histogram of residuals from regression model for beer production.

Key takeaways

- Time series observations are not independent, so that there are particular issues that arise in this context.
- Regressing trending variables is a no-no (spurious regression).
- We can use multiple regression to model seasonality, trend, consider other predictors, and obtain forecasting intervals.
- Predictor selection.
- Run residual diagnostics, especially to check that there is no autocorrelation in the residuals.