QBUS6840 Lecture 4

Time Series Regression

Discipline of Business Analytics

The University of Sydney Business School

Recap and an example

In W3, you have

- looked at multivariate and additive decomposition models for time series
- learned how to use moving average for smoothing time series, and estimating trend
- learned methods for seasonal adjustment.

An interesting example (for self-reading): Forecasting at Scale, Facebook Research,
`https://research.fb.com/publications/forecasting-at-scale/`

Table of contents

Readings

Online textbook, chapter 5, or BOK chapter 6.

Outline

Outline

The simple linear regression model

- Assume that the dependent variable Y and the single predictor variable X are related by the simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, ..., N.$$

  - This is cross-sectional data: no time order between $\{(y_i, x_i), i = 1, 2, ..N\}$
  - When the purpose is to forecast Y, Y is often referred to as the forecast variable

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

The simple linear regression model

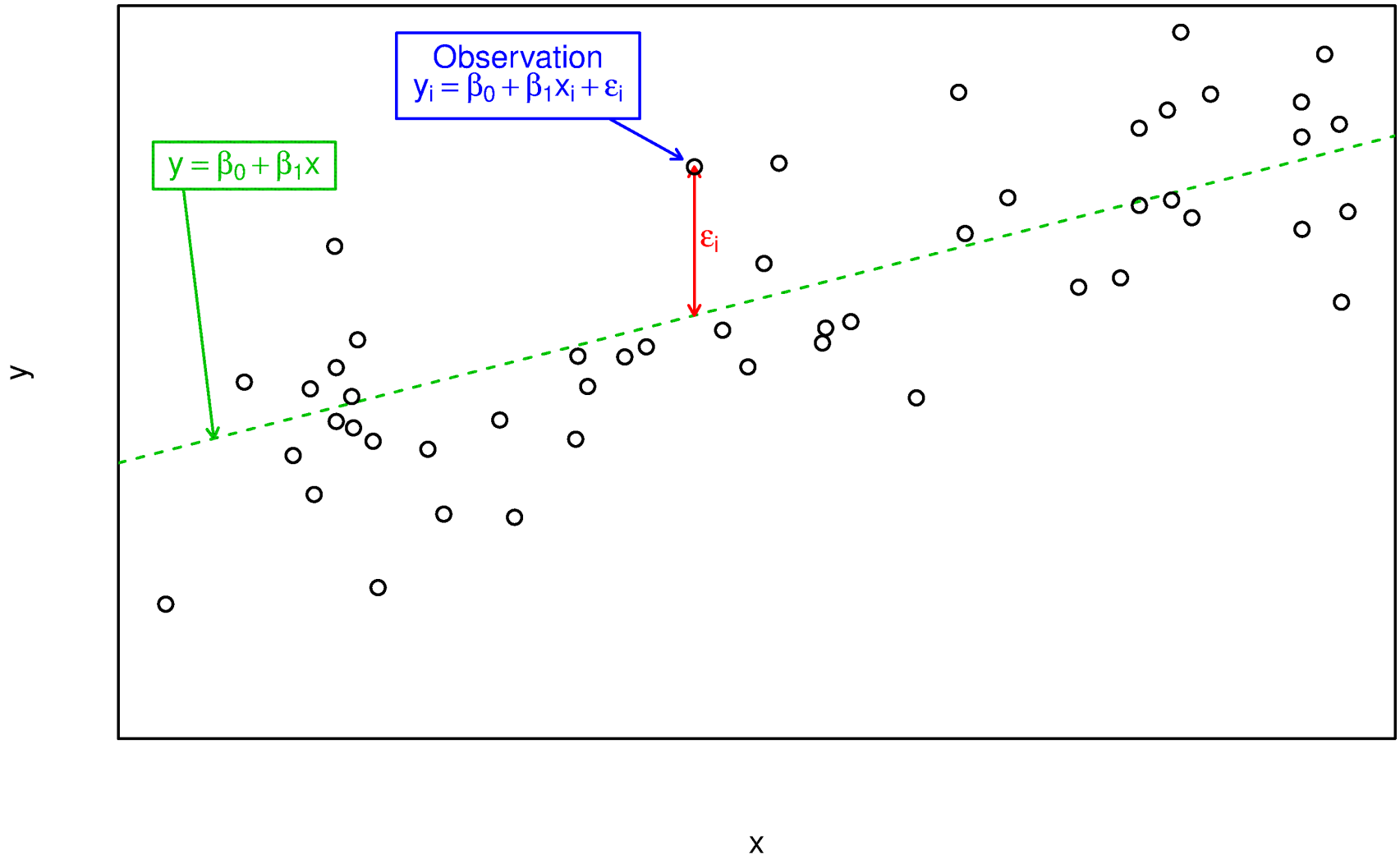


Figure: An example of data from a linear regression model.

The simple linear regression 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 which cannot be explained by x_i. We assume that these errors:
  1. Are not 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

- 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.

Exercise: derive the estimators above, discuss on Ed if you have any questions

Least squares estimation: visualization

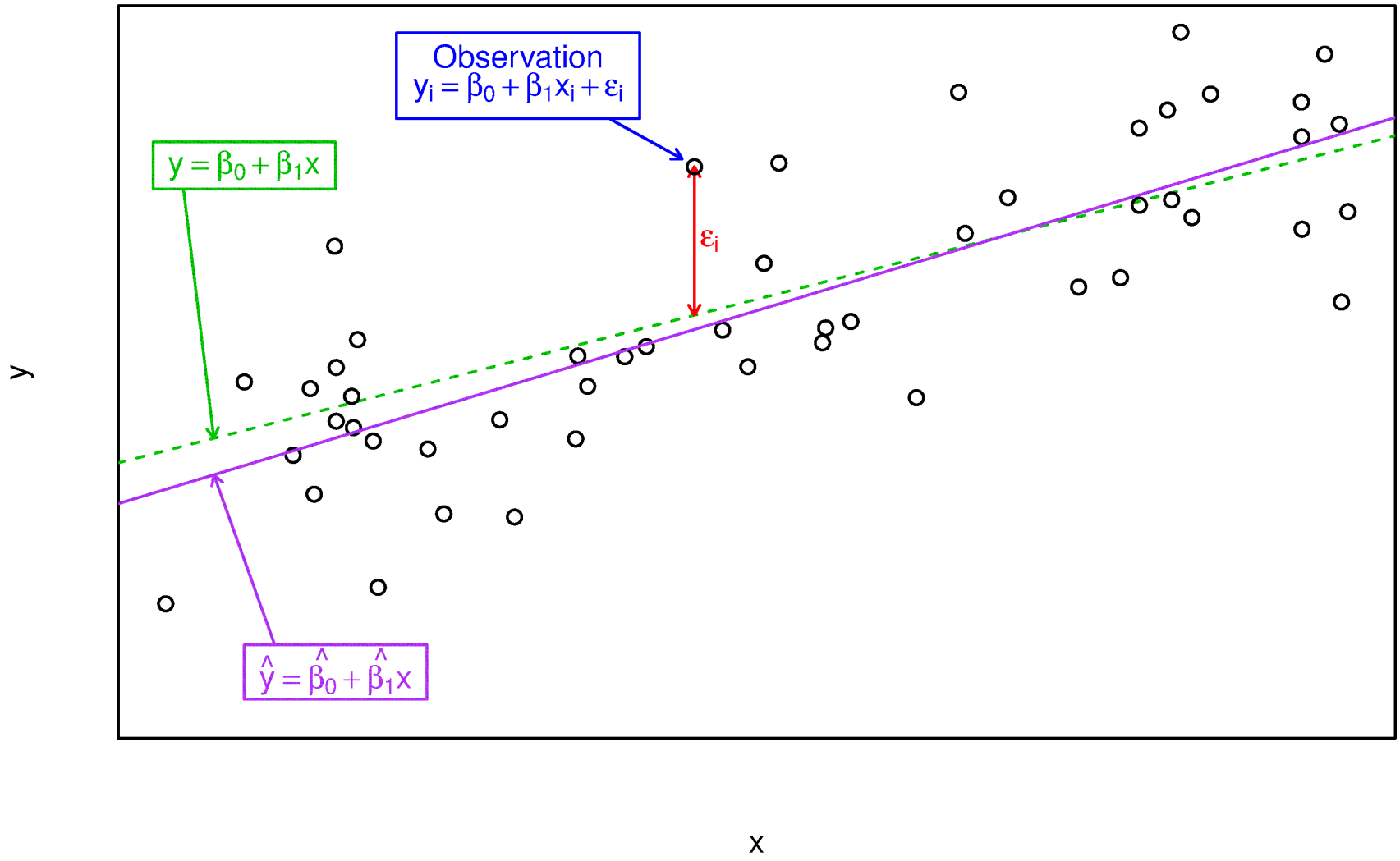


Figure: Estimated regression line for a random sample of size N.

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 $\hat{\beta}_0$ and $\hat{\beta}_1$ 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_i obtained from the observed x_i 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.
  - Note: fitted values are not truly forecasts as the observed data (y_i, x_i) have been used to compute coefficients $\hat{\beta}_0, \hat{\beta}_1$.

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

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

Note: residuals and forecast errors are different!

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

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

Exercise: derive the properties above, discuss on Ed if you have any questions

The residual e_i is different from the error ϵ_i

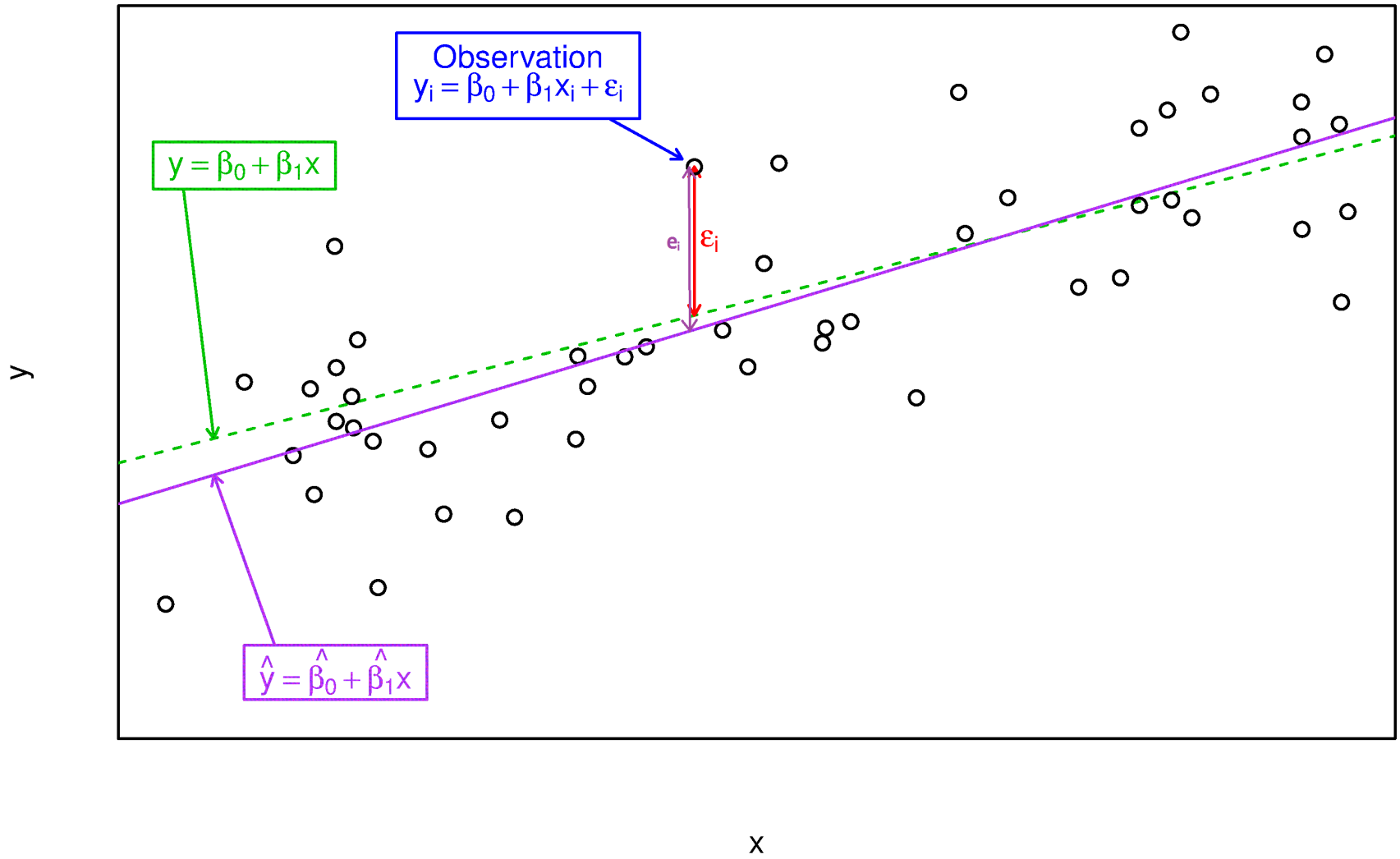


Figure: The residual e_i is an estimate of the random error ϵ_i

Forecasting with regression

- Forecasts from a simple linear model are easily obtained using the equation
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
- Recall that 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

- is a common way to summarize how well a linear regression model fits the data.
- A measure of Goodness-of-fit is the coefficient of determination

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

- It can be explained as 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.
- Note: R^2 is not very reliable, adjusted R^2 is.

Forecasting with regression

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

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

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

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

- 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: car emissions example

Model	Engine (litres)	City (mpg)	Highway (mpg)	Carbon (tons CO2 per year)
Chevrolet Aveo	1.6	25	34	6.6
Chevrolet Aveo 5	1.6	25	34	6.6
Honda Civic	1.8	25	36	6.3
Honda Civic Hybrid	1.3	40	45	4.4
Honda Fit	1.5	27	33	6.1
Honda Fit	1.5	28	35	5.9
Hyundai Accent	1.6	26	35	6.3
Kia Rio	1.6	26	35	6.1
Nissan Versa	1.8	27	33	6.3
Nissan Versa	1.8	24	32	6.8
Pontiac G3 Wave	1.6	25	34	6.6
Pontiac G3 Wave 5	1.6	25	34	6.6
Pontiac Vibe	1.8	26	31	6.6
Saturn Astra 2DR Hatchback	1.8	24	30	6.8
Saturn Astra 4DR Hatchback	1.8	24	30	6.8
Scion xD	1.8	26	32	6.6
Toyota Corolla	1.8	27	35	6.1
Toyota Matrix	1.8	25	31	6.6
Toyota Prius	1.5	48	45	4.0
Toyota Yaris	1.5	29	35	5.9

Forecasting with regression: car emissions example

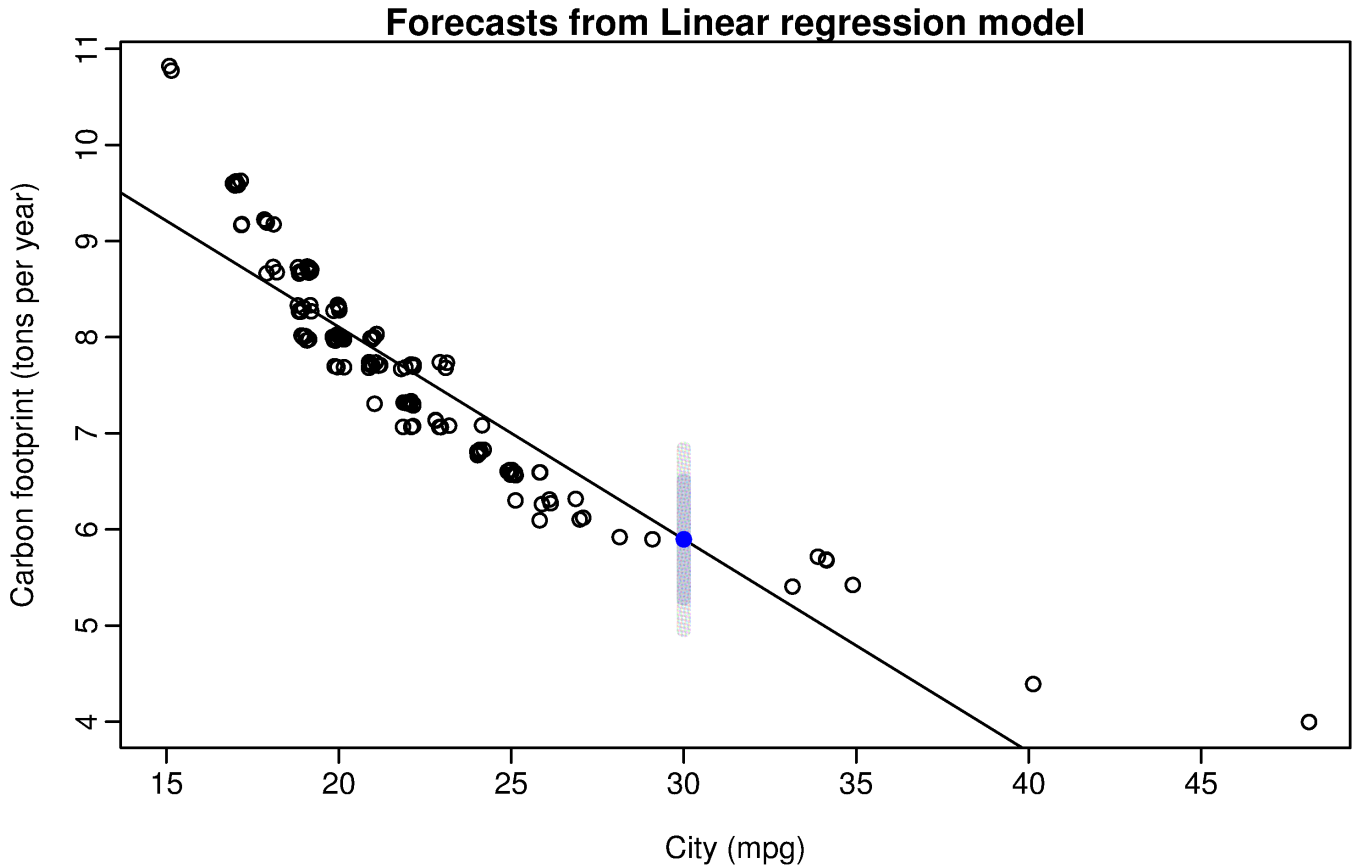


Figure: Forecast with 80% and 95% forecast intervals for a car with $x = 30$ mpg in city driving.

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 might be adequate in some cases, there are situations for which a non-linear functional form is more suitable.
- The most commonly used transformation is the (natural) logarithm.
- 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

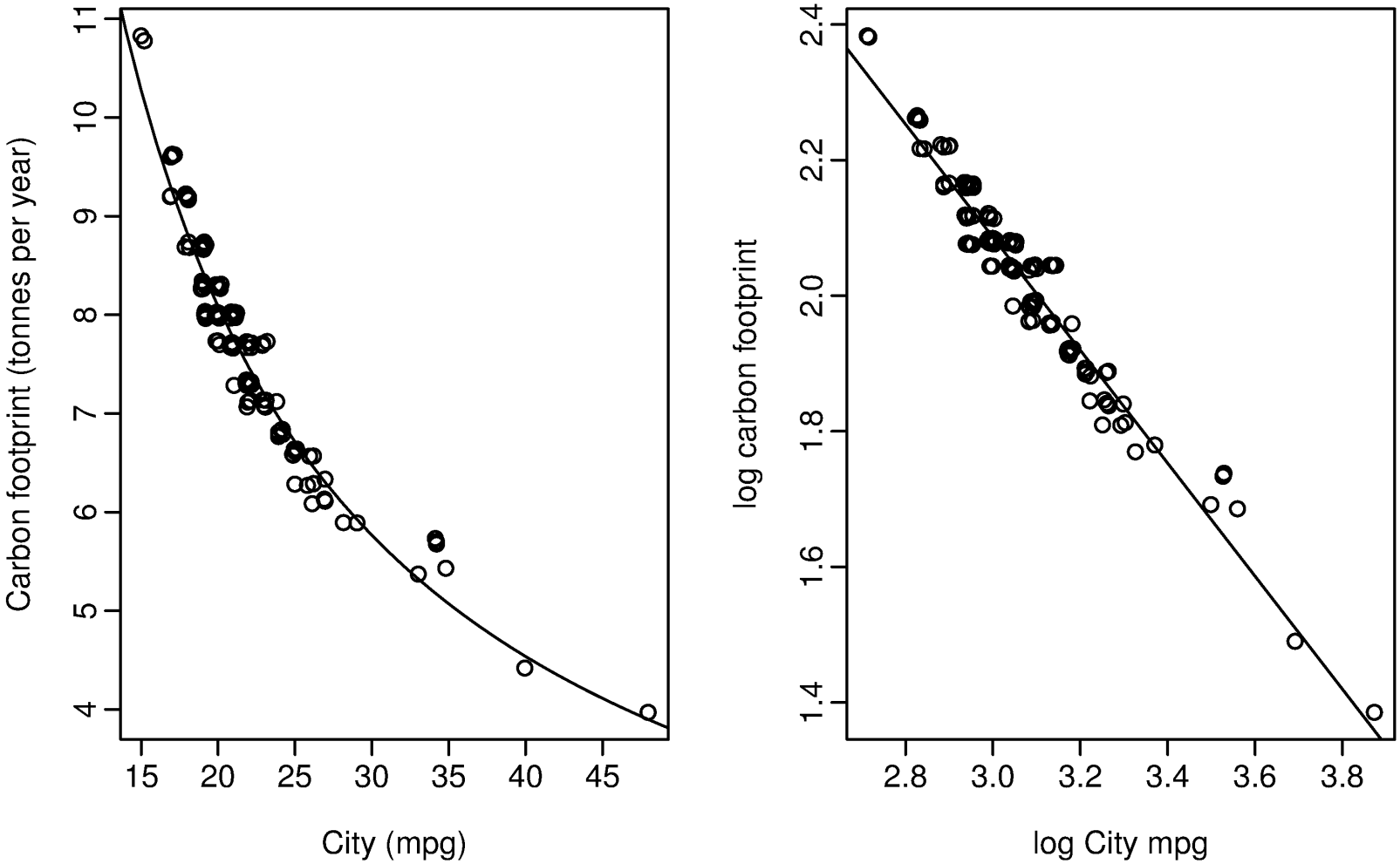


Figure: Fitting a log-log functional form to the Car data example. Plots show the estimated relationship both in the original and the logarithmic scales.

Non-linear functional forms

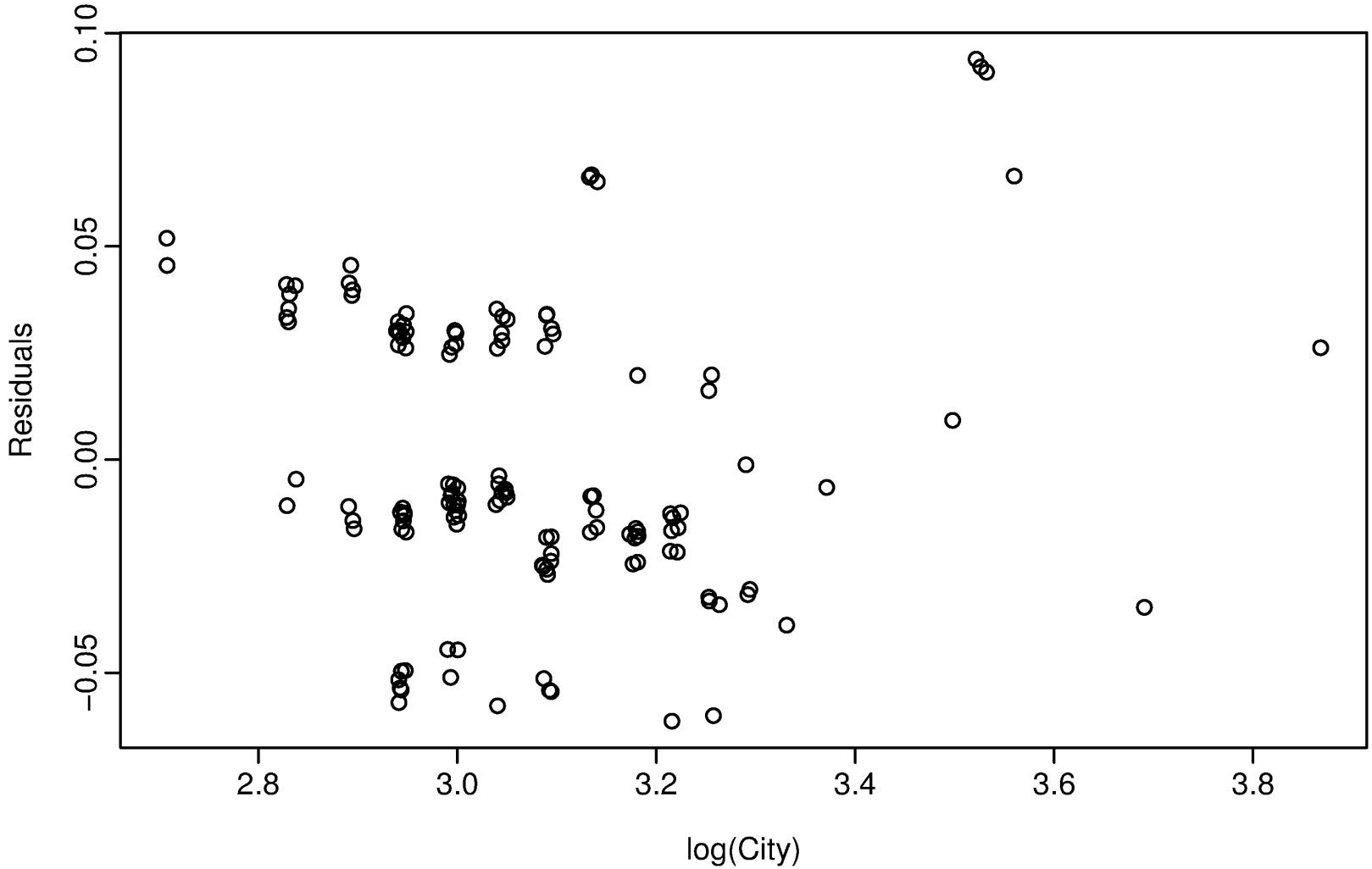


Figure: Residual plot from estimating a log-log functional form for the Car data example.

Non-linear functional forms

Some useful non-linear functional forms

- log-log: $\log y = \beta_0 + \beta_1 \log x$
- linear-log: $y = \beta_0 + \beta_1 \log x$
- log-linear: $\log y = \beta_0 + \beta_1 x$

Note: careful with non-positive values!

Non-linear functional forms

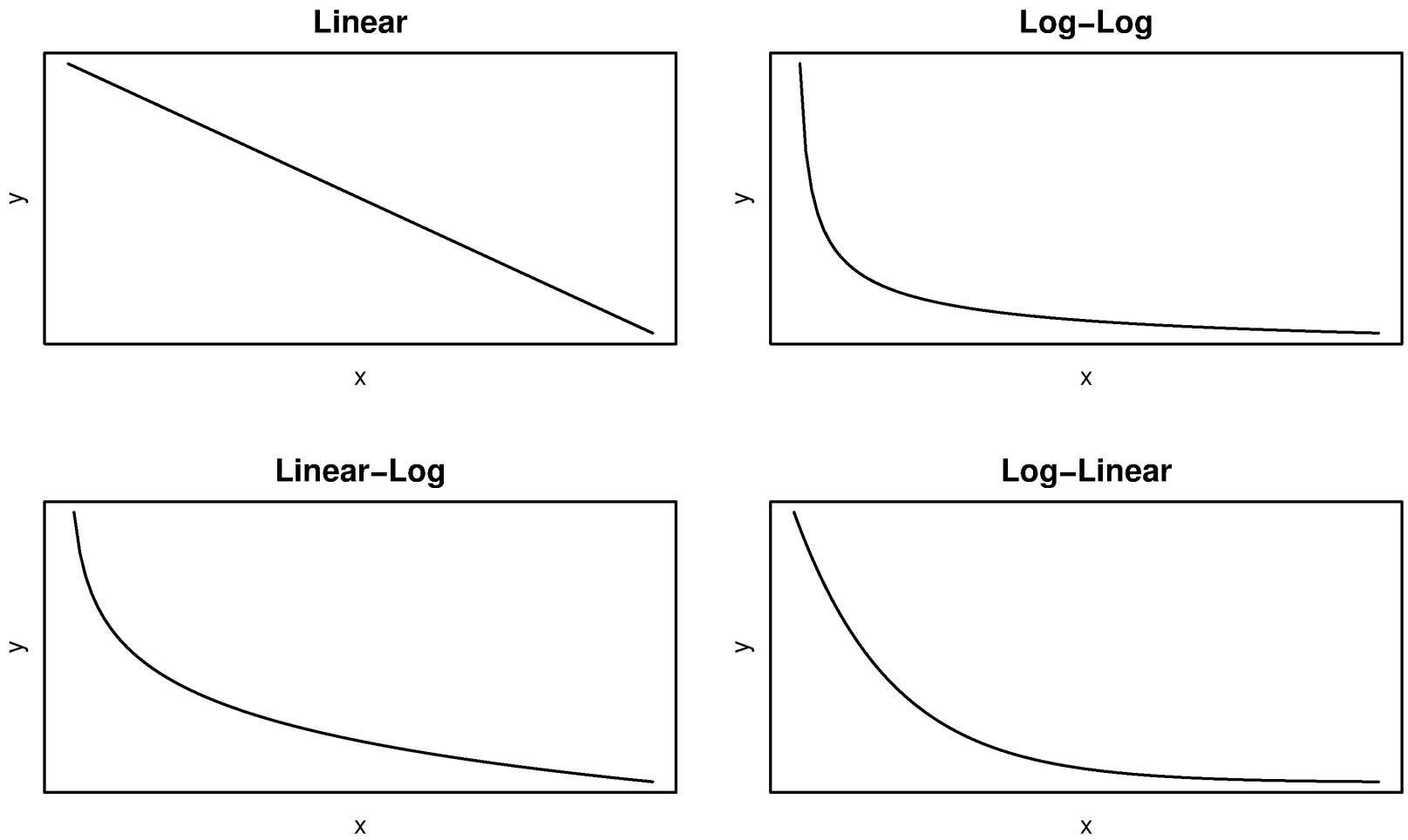


Figure: The four non-linear forms from the previous slide

Outline

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	Forecast variable	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's 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:

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

Outline

Regression with time series data

- Let's look at a simple linear regression model for time series data: the forecast variable Y and the single predictor variable X are related by the simple linear model:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad t = 1, ..., T.$$

$\{(y_t, x_t), t = 1, 2, ..., T\}$ is time series data.

- When using regression with time series data, we often aim to forecast the future. There are a few issues that arise with time series data but not with cross-sectional data.

- A challenge in that future values of the predictor variable x might be unknown.

Regression with time series data

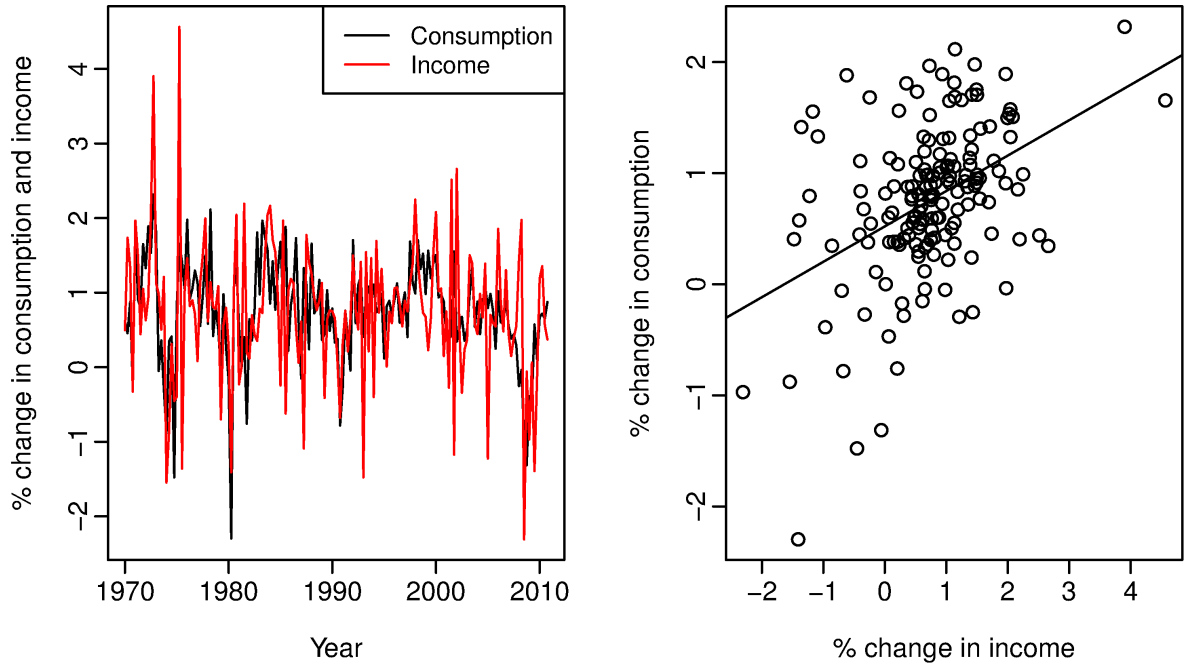


Figure: Time series plots of quarterly percentage changes (growth rates) of real personal consumption expenditure (C_t) and real personal disposable income (I_t) for the US for the period March 1970 to Dec 2010. Also shown is a scatter plot including the estimated regression line $\hat{C}_t = 0.52 + 0.32 I_t$

We are interested in forecasting consumption for the four quarters of 2011. But we don't know the values of the predictor I_t for these four quarters!

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

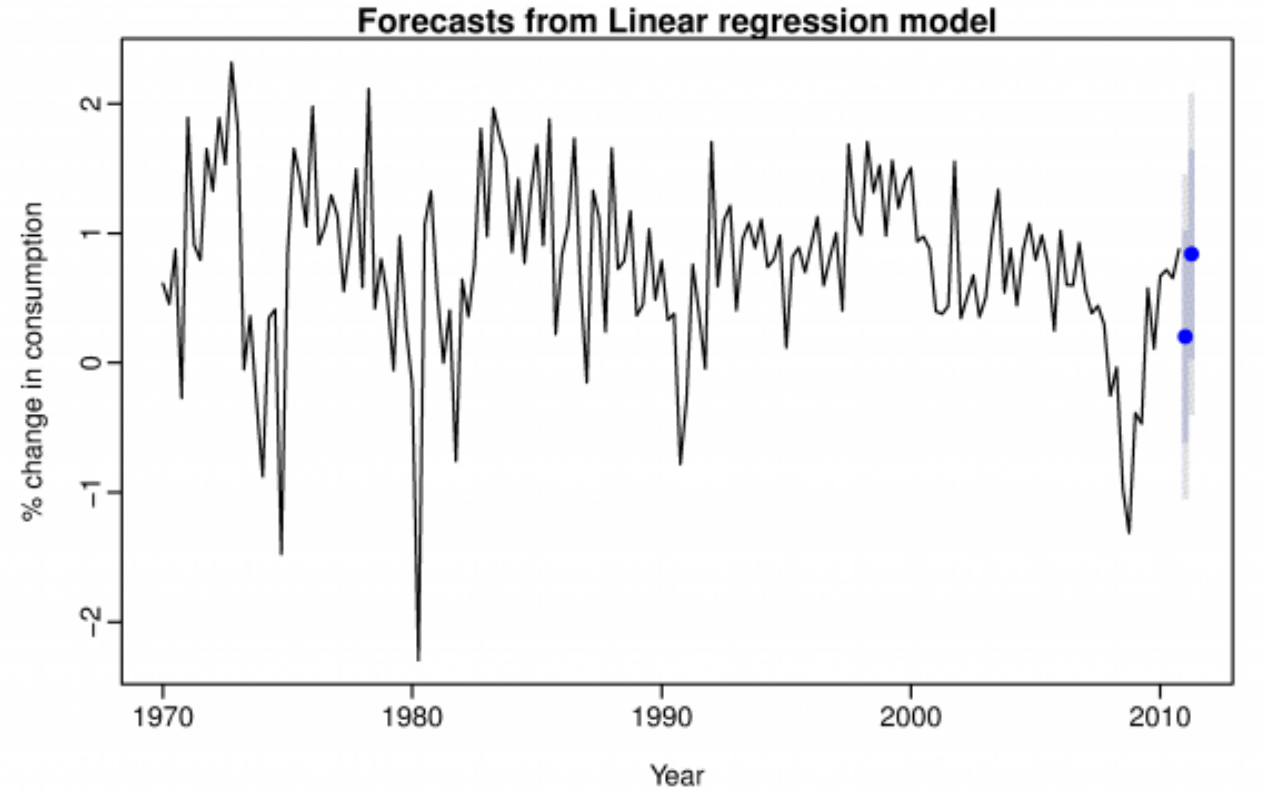


Figure: Forecasting percentage changes in personal consumption expenditure for the US.

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 forecast is a forecast that only uses information available at the time of the forecast.
  - E.g., ex ante forecasts of consumption for the four quarters in 2011 should only use information that was available before 2011.
  - These are genuine forecasts, made in advance using whatever information is available at the time of forecast.

Ex-ante versus ex-post forecasts

- Ex post forecast is a forecast that uses information beyond the time at which the forecast is made.
  - E.g., 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.
  - The better the fit (ex post forecasts) of the forecasting model, the more accurate ex ante forecasts should be.

Outline

Useful predictors

A great advantage of using regression models for time series data is that we can use useful predictors to capture or model many features of the forecast variable. By using suitable predictors, we can capture

- trend component
- seasonal component
- outliers
- etc.

Modelling the trend

- Using regression we can model and forecast the *linear trend* in time series data by including $x_t = t$ as a predictor variable:

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

- The following figure shows a time series plot of aggregate (yearly) 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

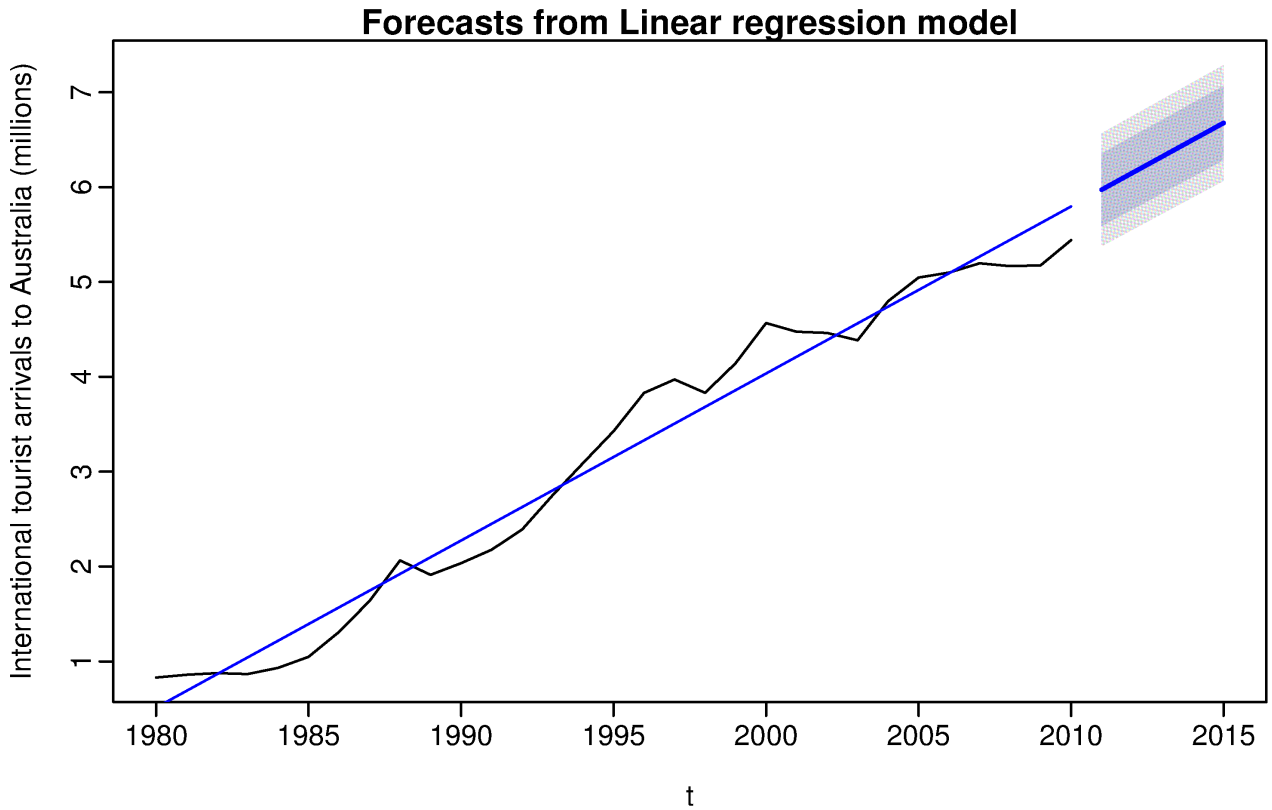


Figure: Forecasting international tourist arrivals to Australia for the period 2011-2015 using a linear trend. 80% and 95% forecast intervals are shown.

Modelling the trend

A quadratic trend is obtained by using two predictors $x_{1,t} = t$, $x_{2,t} = t^2$

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

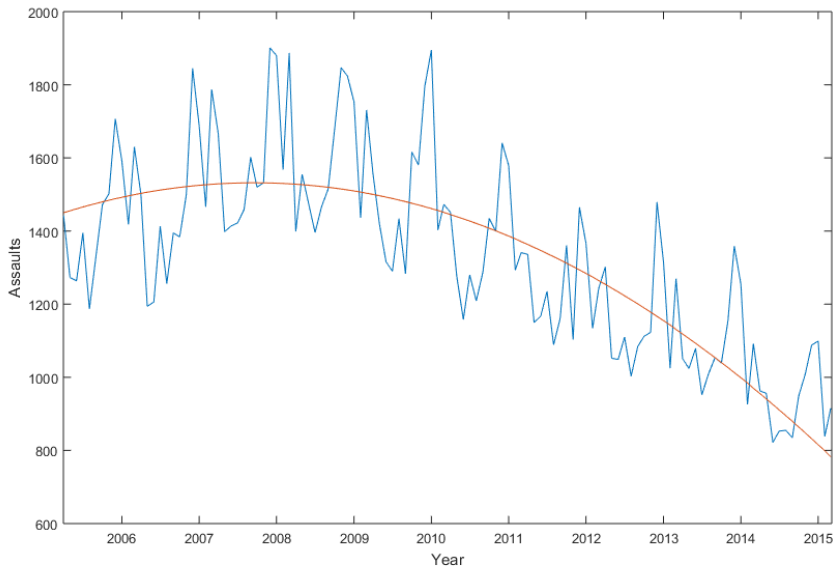


Figure: Alcohol related assaults in NSW

See another example in `Lecture04_Example03.py`

Modelling the trend

- Another 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.

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

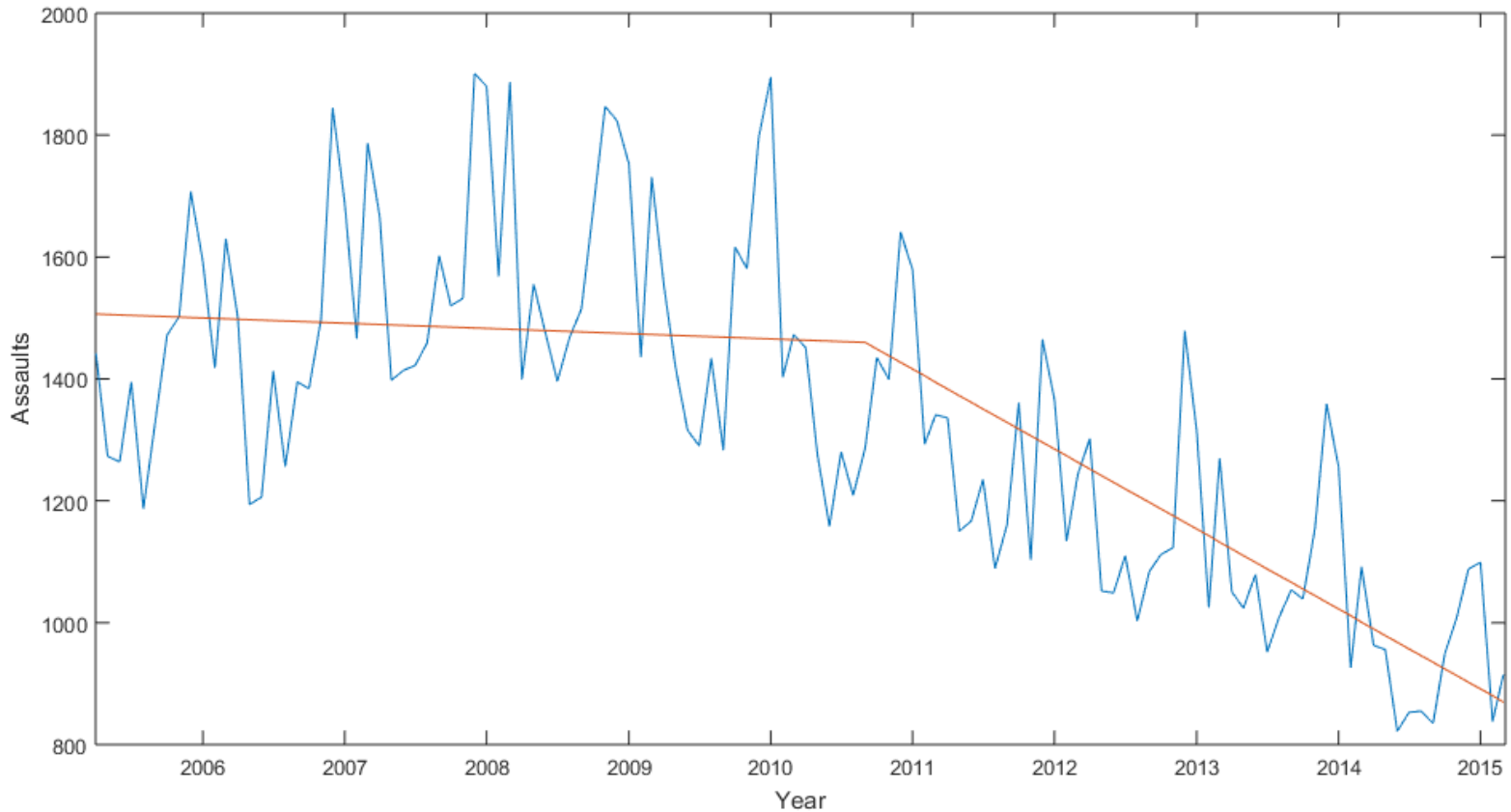
$$x_{2,t} = \begin{cases} 0 & t < \tau \\ (t - \tau) & t \geq \tau \end{cases}$$

- 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 $\beta_1 + \beta_2$.

- Piecewise linear regression is a special case of spline regression. Want to learn more about spline regression? QBUS6810 or QBUS6850

Example: piecewise trend

Figure: Alcohol related assaults in NSW



Seasonal Dummy variables

- We can capture the seasonal component using linear regression as well.
- 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
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

- Consider a four-seasonal time series with a horizontal trend
- Can be forecast by

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

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: cont.

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

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

- When $t = 2$, what is the forecast? $\widehat{y}_2 = \beta_0 + \beta_2$
- How about $t = 7$ and 8? $\widehat{y}_7 = \beta_0 + \beta_3$ and $\widehat{y}_8 = \beta_0 + \beta_4$
- We can see that the coefficients associated with the other seasons are measures of the difference between those seasons and the first season.

Example: Australian quarterly beer production

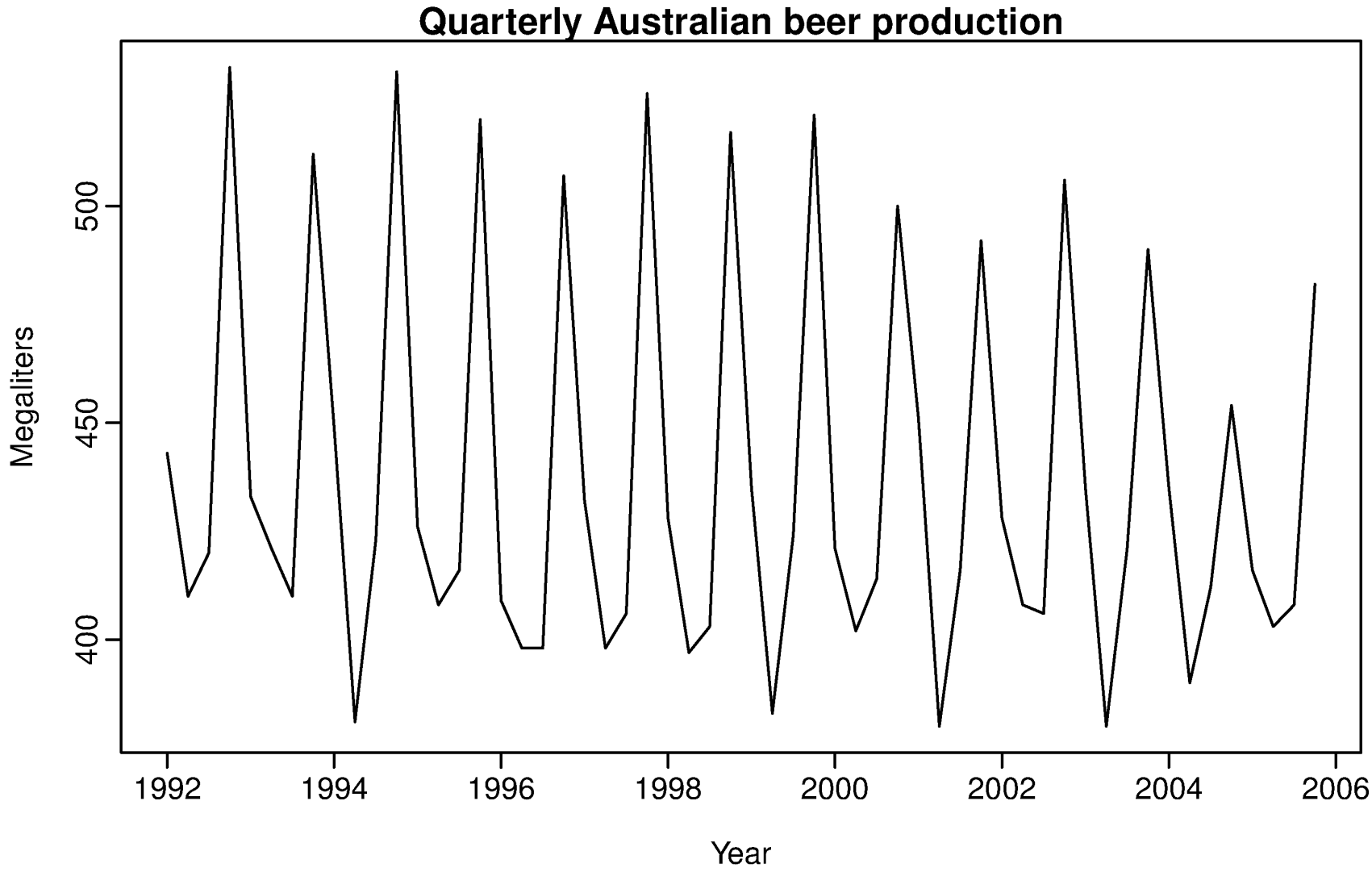


Figure: Australian quarterly beer production: exhibit a seasonal pattern with $M = 4$

Example: Australian quarterly beer production

- It seems that the Australian beer production data has a linear trend component and a seasonal component with $M = 4$. Can we model these components using a linear regression model?

- We can model these 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 caused by seasons* between those quarters and the first quarter

Example: Australian quarterly beer production

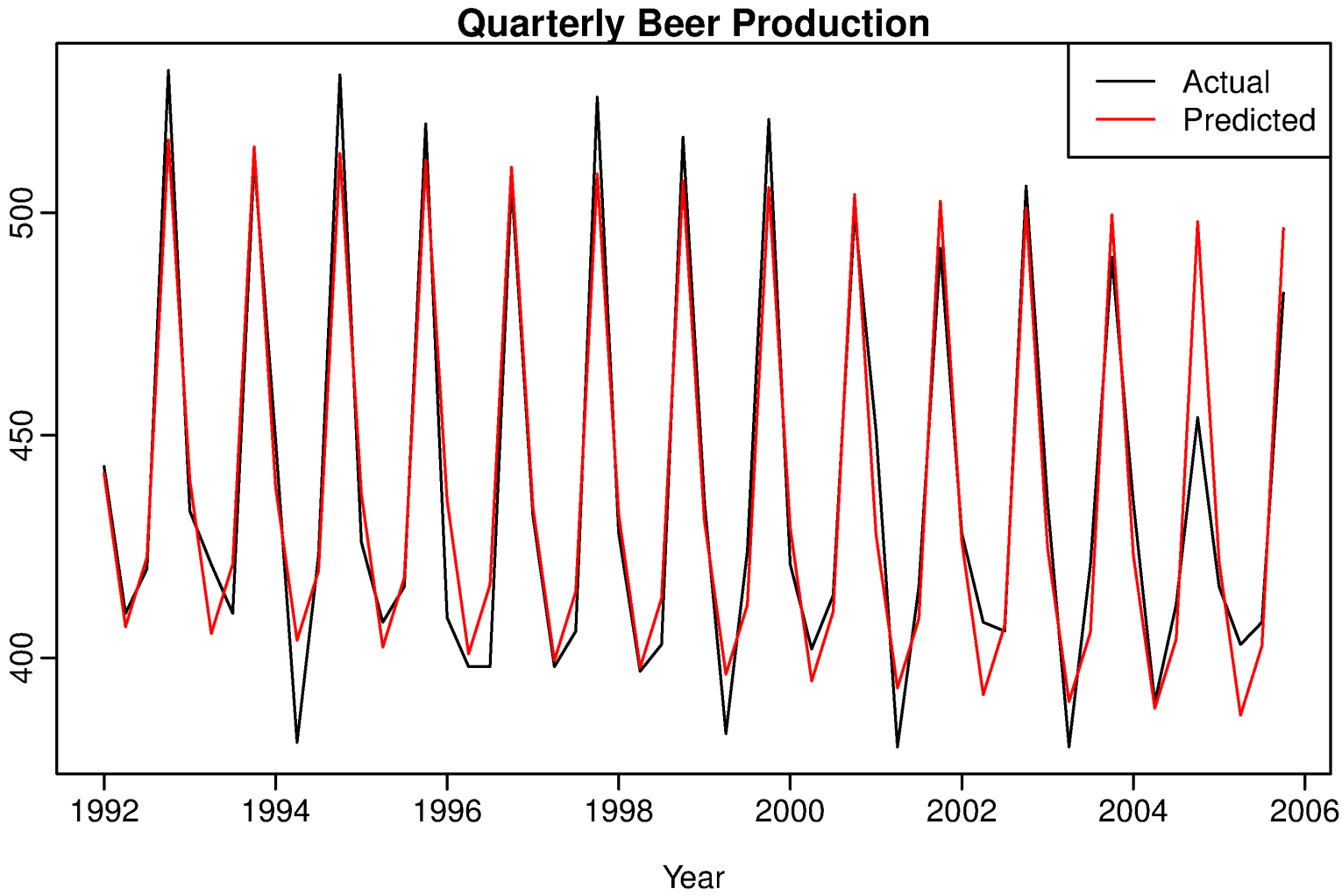


Figure: Time plot of beer production and predicted beer production.

Example: Australian quarterly beer production

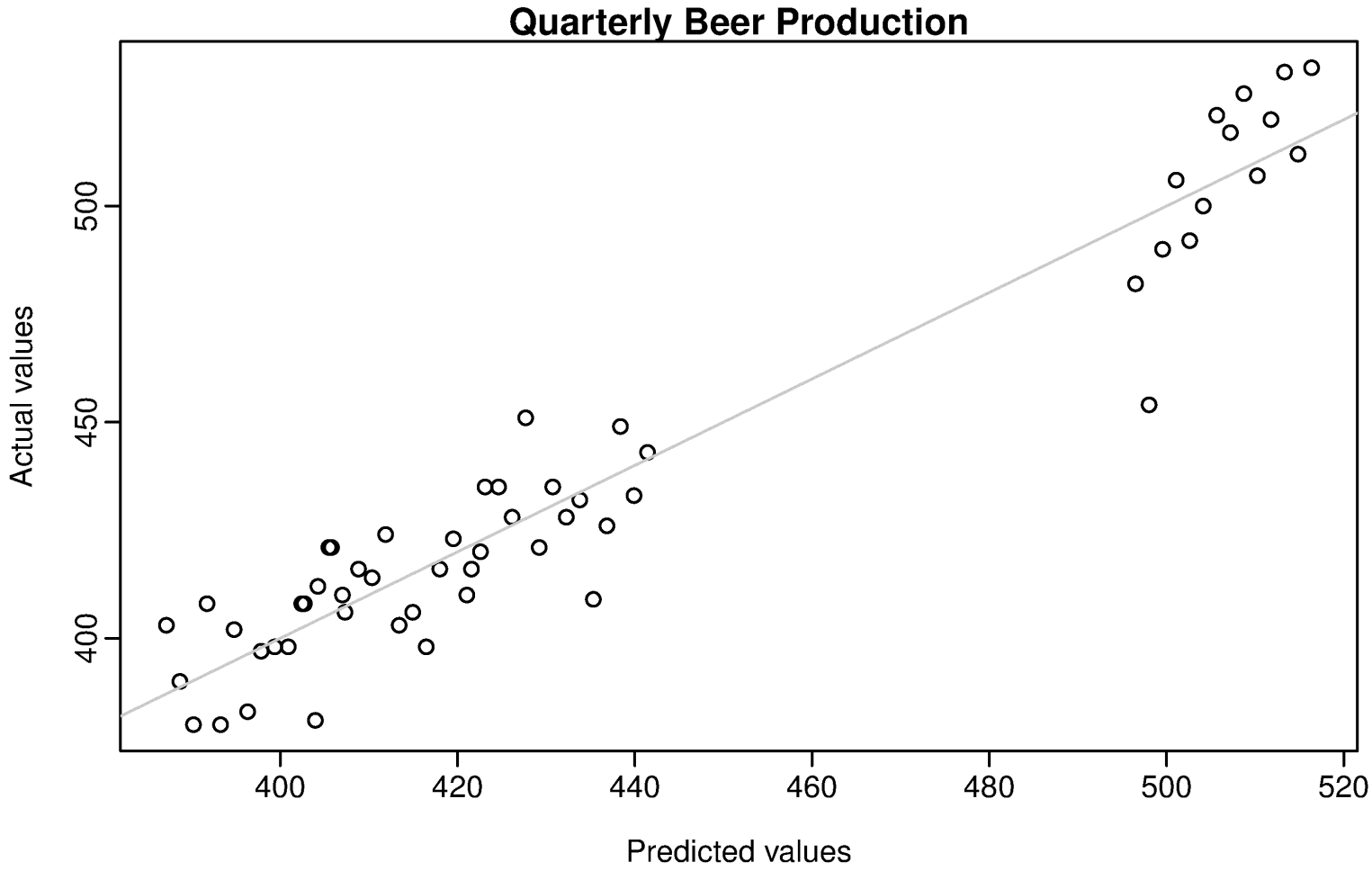


Figure: Actual beer production plotted against predicted beer production.

Example: Australian quarterly beer production

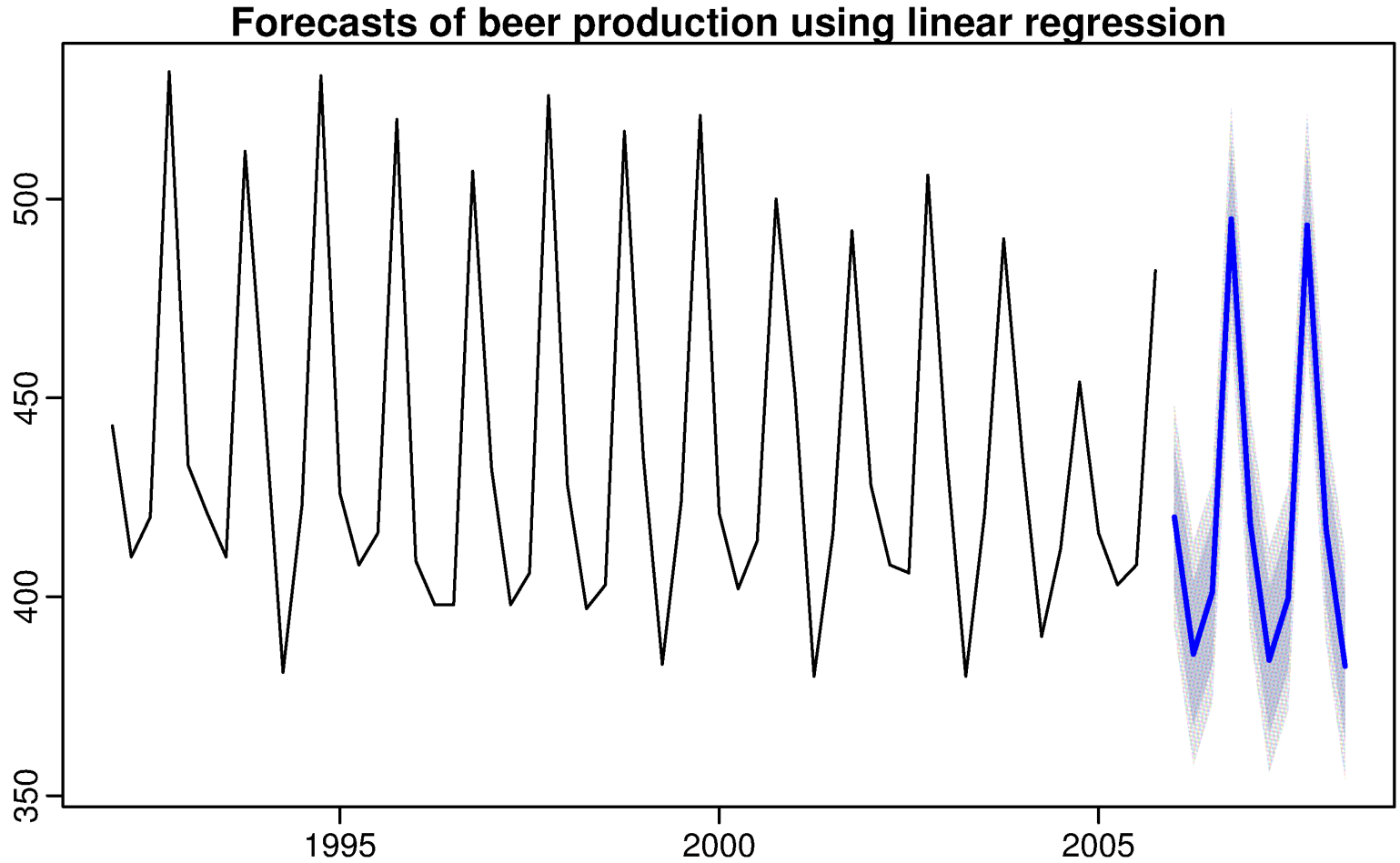


Figure: Forecasts from the regression model for beer production. The dark blue region shows 80% prediction intervals and the light blue region shows 95% prediction intervals.

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.

Intervention variables

- It is often necessary to model interventions 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.

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

Outline

Selecting important predictors

- You can add as many *potential* predictors into the linear regression model as you want
- Then, you need to let the data to decide: which predictors are significant/important? This is known as variable selection problem
- There are many techniques for selecting predictors: adjusted R^2, cross-validation, AIC, BIC, etc.

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 select too many predictors.

Cross-validation

- Cross-validation is a very useful and more objective way of determining the predictive ability of a model (see Week 2's lecture).
- The best model is the one with the smallest value of cross-validated error.

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.

Bayesian Information Criterion

- A related measure is Bayesian Information Criterion (often known as BIC):

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

- We select the model with smallest BIC.
- The model chosen by BIC often has fewer predictors than the model chosen by AIC. This is because BIC penalizes the model complexity more heavily than the AIC.
- If there is a true underlying model, then with enough data the BIC will select that model. But do you believe that the true model exists?

Best subset regression

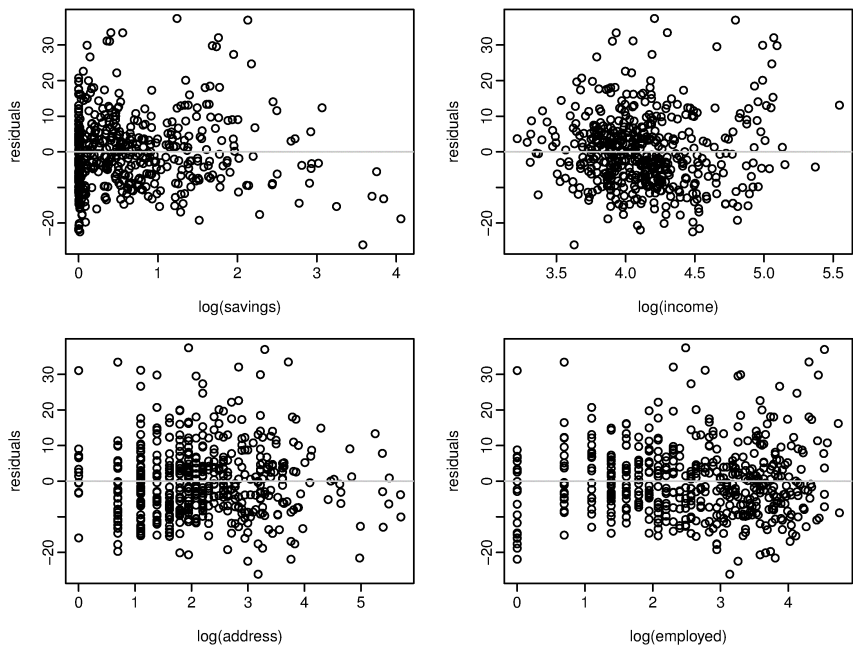
- Where possible, all potential regression models can be fitted and the best one selected based on one of the measures discussed here. This is known as "best subsets" regression or "all possible subsets" regression.
- It is recommended that one of CV or AIC is used for this purpose
- While $\bar{R}^2$ 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

Outline

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.

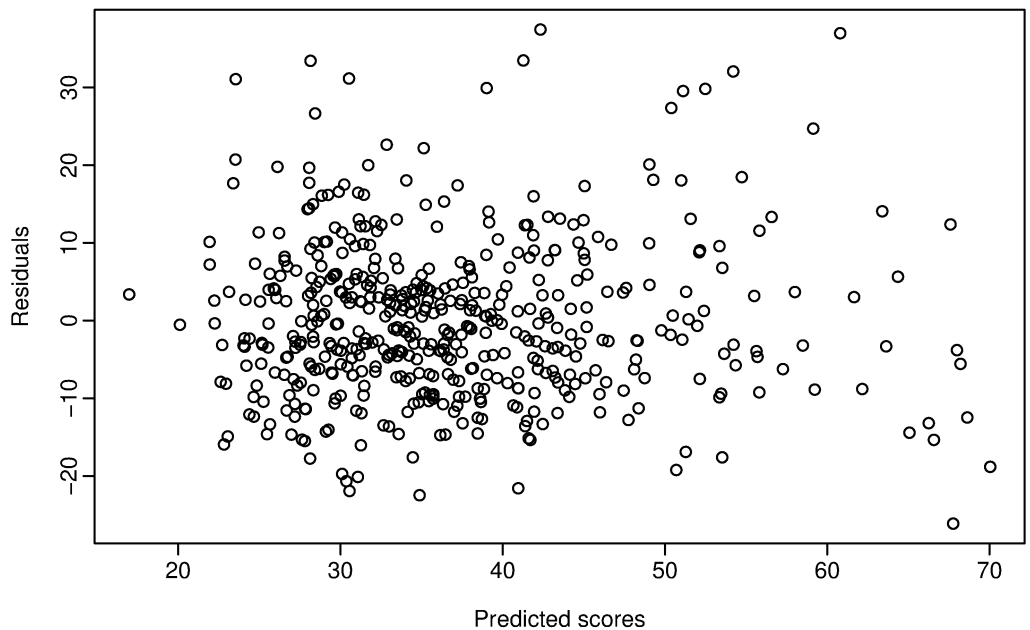


- It is also necessary to plot the residuals against each predictor NOT in the model. If these show a pattern, then the predictor may need to be added to the model.

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 autocorrelation

Autocorrelation in the residuals

- 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.
- 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

Autocorrelation in the residuals

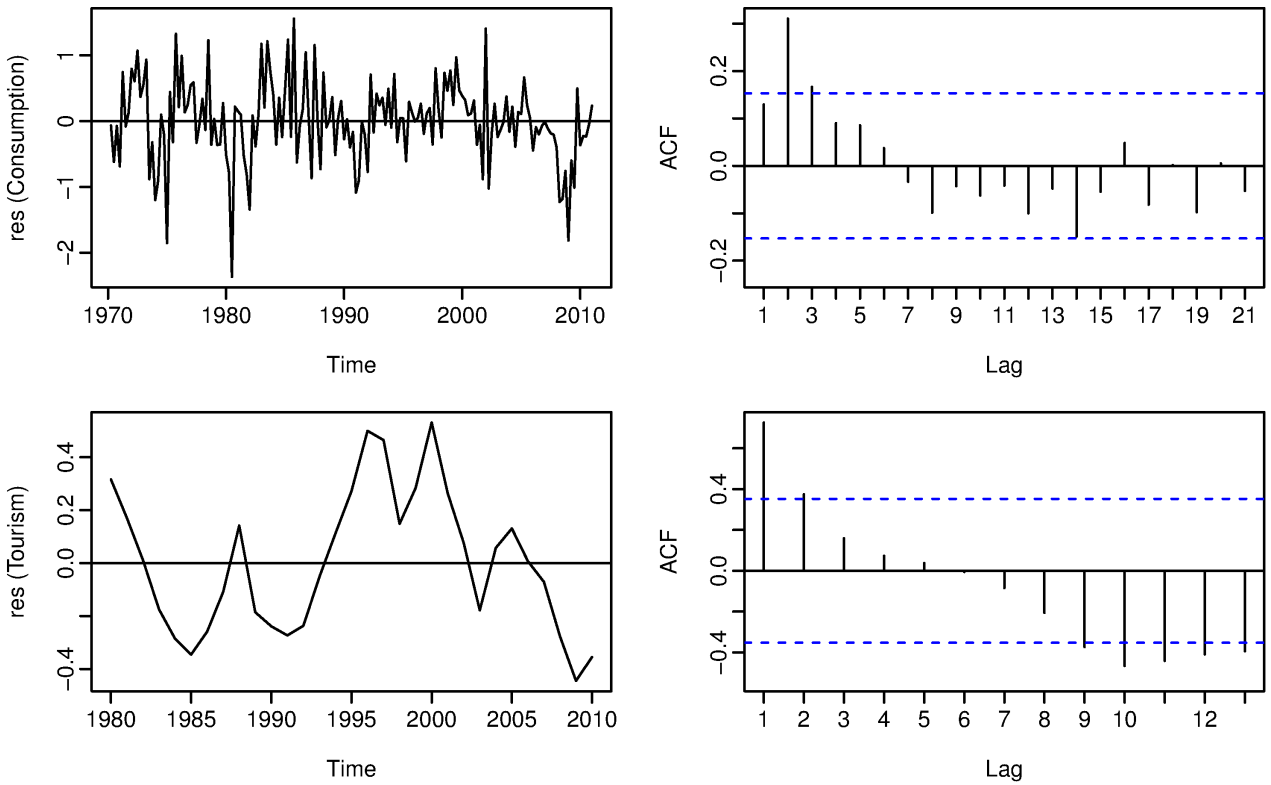


Figure: Residuals from the regression models for Consumption and Tourism. Because these involved time series data, it is important to look at the ACF of the residuals to see if there is any remaining information not accounted for by the model.

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.

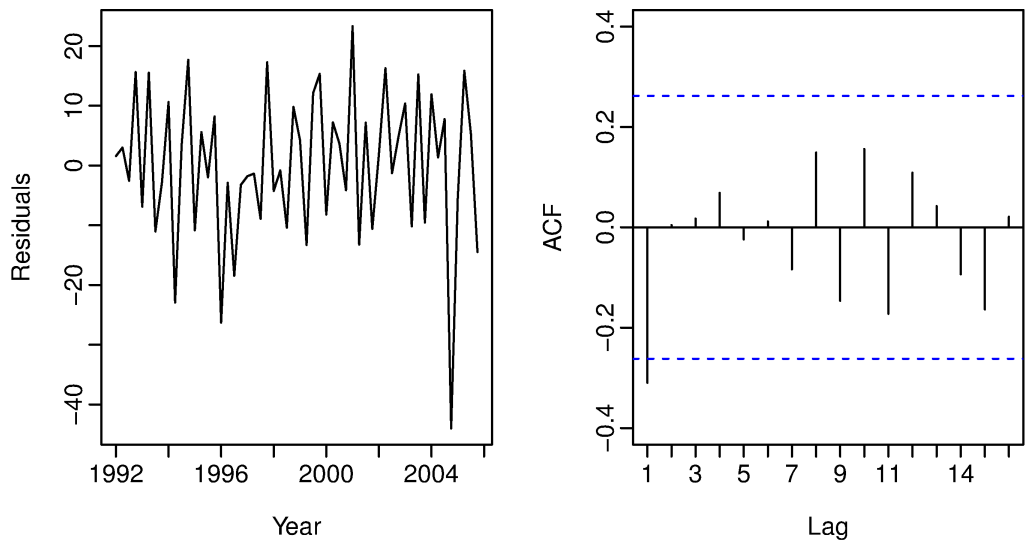


Figure: Residuals from the regression model for beer production.

Residual diagnostics

Histogram of residuals

- Finally, it is a good idea to check if the residuals are normally distributed.

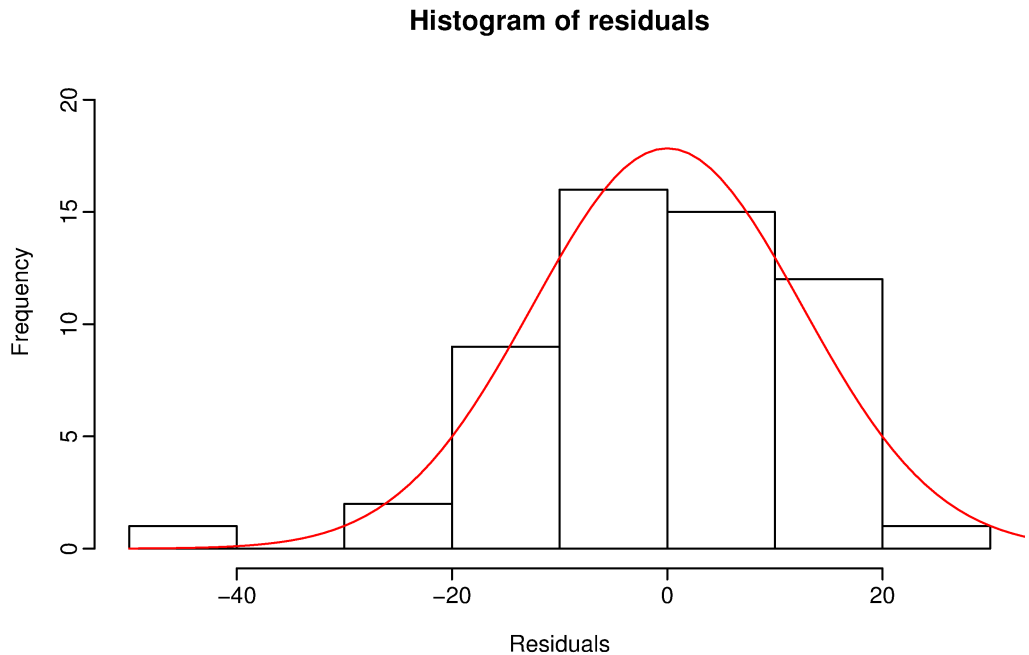


Figure: Histogram of residuals from regression model for beer production: The residuals seem to be slightly negatively skewed, although that is probably due to the outlier.

Key takeaways

- Time series observations are not independent, so that there are particular issues that arise in this context.
- 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.