5. Multiple regression

QBUS6840 Predictive Analytics

https://www.otexts.org/fpp/5

Outline

Outline

Example: Australian quarterly beer production I

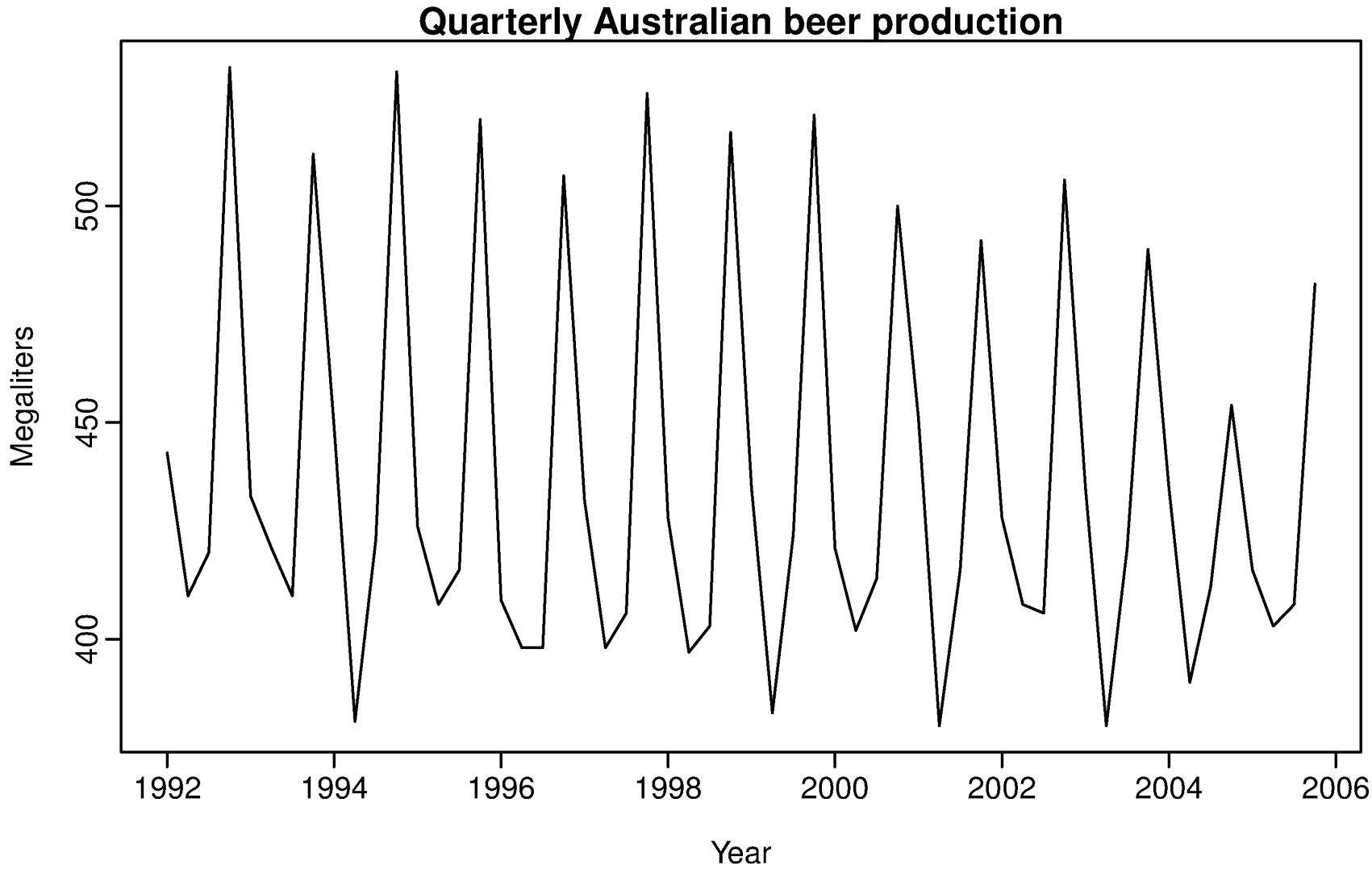

Figure 5.1: Australian quarterly beer production.

Introduction to multiple linear regression I

- 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 $\beta_1, \ldots, \beta_k$ measure the effect of each predictor after taking account of the effect of all other predictors in the model. Thus, the coefficients measure the marginal effects of the predictor variables.
- As for simple linear regression, when forecasting we require the following assumptions for the errors $(e_1, \ldots, e_N)$:
  1. the errors have mean zero;
  2. the errors are uncorrelated with each other;
  3. the errors are uncorrelated with each predictor xj,i.
- It is also useful to have the errors normally distributed with constant variance in order to produce prediction intervals, but this is not necessary for forecasting.

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

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

- 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 $\hat{\beta}_0, \ldots, \hat{\beta}_k$. The equations for these will be given in Section 5/5.

Fitted values, forecast values, residuals and R^2 I

- 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.
- When this calculation is done using values of the predictors from the data that were used to estimate the model, we call the resulting values of $\hat{y}$ the "fitted values". These are "predictions" of the data used in estimating the model. They are not genuine forecasts as the actual value of y for that set of predictors was used in estimating the model, and so the value of $\hat{y}$ is affected by the true value of y.
- When the values of $x_1, \ldots, x_k$ are new values (i.e., not part of the data that were used to estimate the model), the resulting value of $\hat{y}$ is a genuine forecast.

Fitted values, forecast values, residuals and R^2 II

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

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_k x_k.$$

- As with simple regression (see Section 4/2), the residuals have zero mean and are uncorrelated with any of the predictors.
- The R^2 value was introduced in Section 4/4. 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

Seasonal Dummy variables I

- For example, suppose we are forecasting daily electricity demand and we want to account for the day of the week as a predictor. Then the following dummy variables can be created.

Day	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 II

- 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.
- Many beginners will try to add a seventh dummy variable for the seventh category. This is known as the "dummy variable trap" because it will cause the regression to fail. There will be too many parameters to estimate. 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. In the above example, the coefficient associated with Monday will measure the effect of Monday compared to Sunday on the forecast variable.

Other dummy predictors

- 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.
- Easter: Easter is different from most holidays because it not held on the same date each year and the effect can last for several days. In this case, a dummy variable can be used with value one where any part of the holiday falls in the particular time period and zero otherwise.

Trend I

- A linear trend is easily accounted for by including the predictor $x_{1,t} = t$. A piecewise linear trend with a bend at time can be specified by including the following predictors in the model.

$$x_{1,t} = t$$
$$x_{2,t} = \begin{cases} 0 & t < \tau \\ (t-\tau) & t \geq \tau \end{cases}$$

- 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.

Trend II

- 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 $\beta_1 + \beta_2$.

Outline

Example: Australian quarterly beer production I

- 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.
- Estimation results: There is a strong downward trend of 0.382 megalitres per quarter. On average, the second quarter has production of 34.0 megalitres lower than the first quarter, the third quarter has production of 18.1 megalitres lower than the first quarter, and the fourth quarter has production 76.1 megalitres higher than the first quarter. The model explains 92.1% of the variation in the beer production data.
- The following plots show the actual values compared to the predicted values.

Example: Australian quarterly beer production II

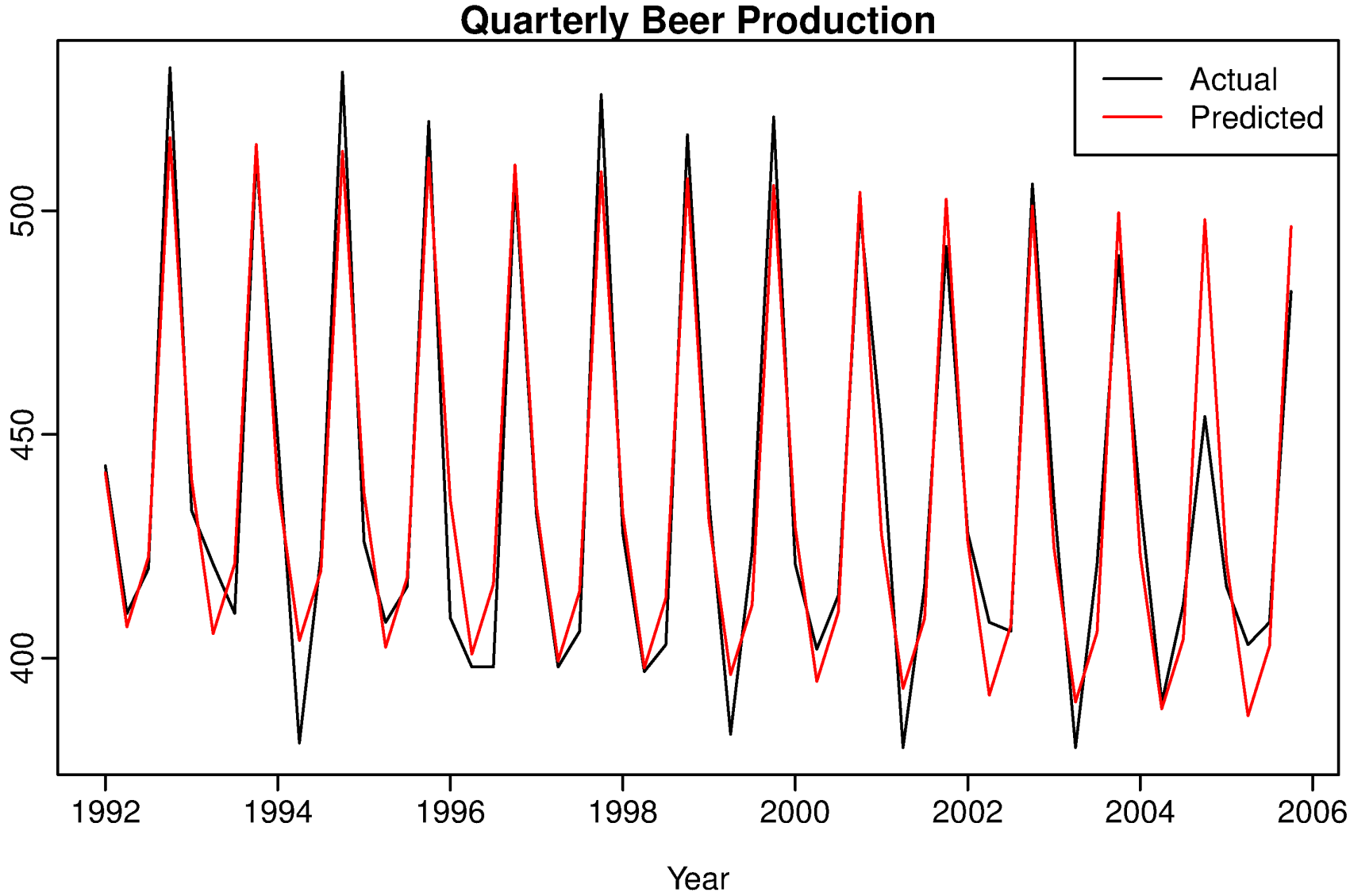

Figure 5.5: Time plot of beer production and predicted beer production.

Example: Australian quarterly beer production III

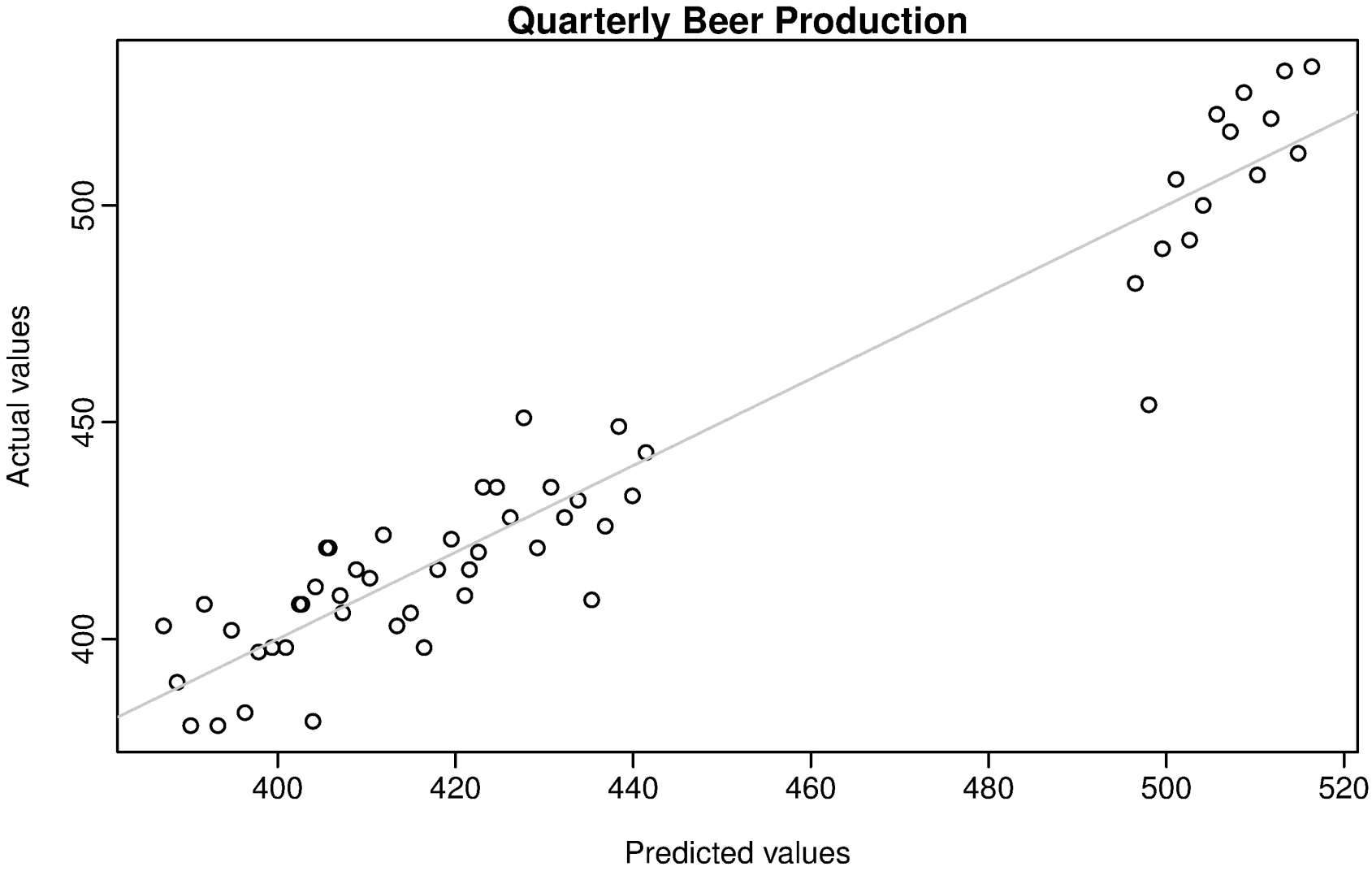

Figure 5.6: Actual beer production plotted against predicted beer production.

Example: Australian quarterly beer production IV

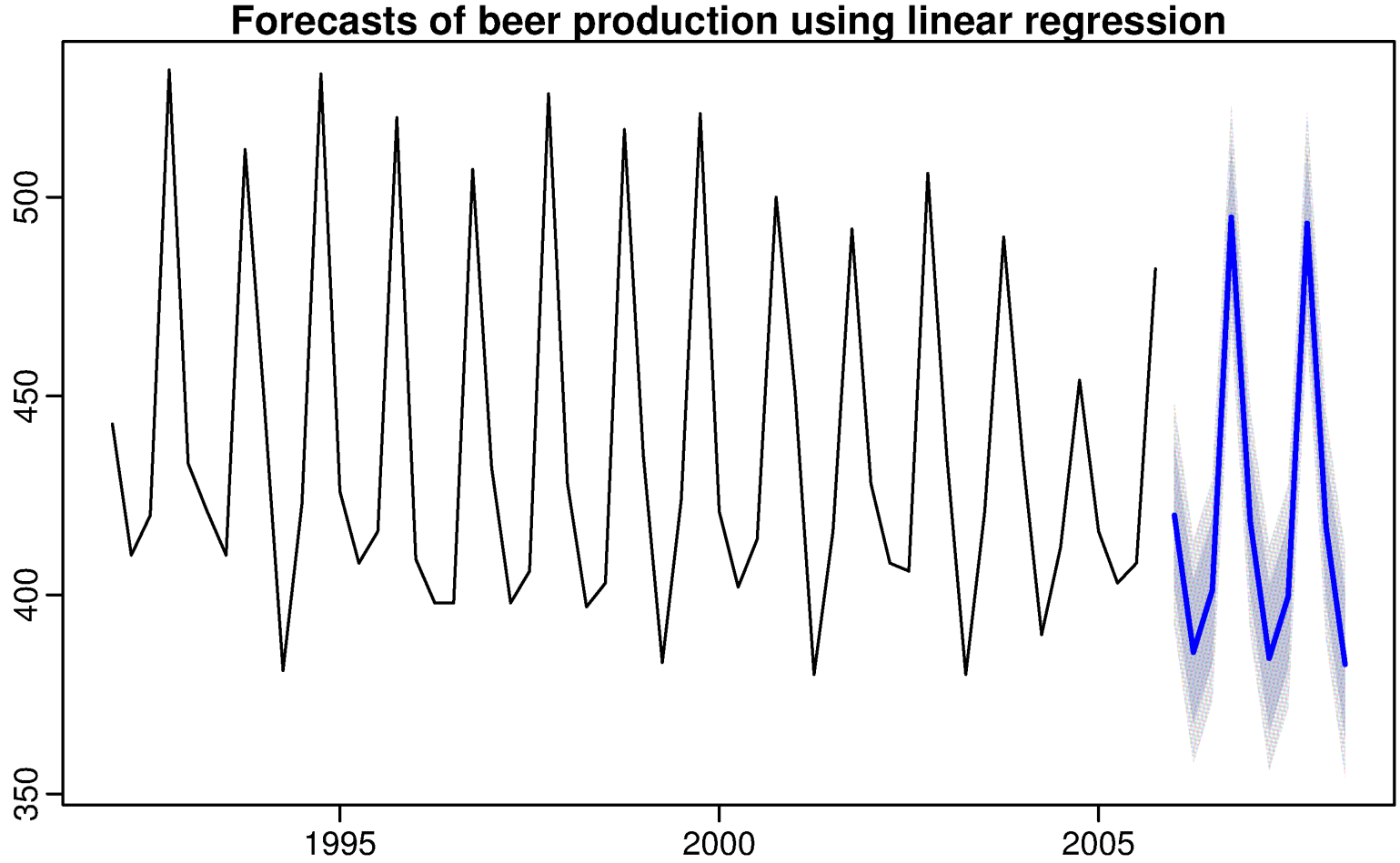

Figure 5.7: Forecasts from the regression model for beer production. The dark blue region shows 80% prediction intervals and the light blue region shows 95% prediction intervals.

Outline

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.
- 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. A step variable takes value zero before the intervention and one from the time of intervention onwards.
- Another form of permanent effect is a change of slope. Here the intervention is handled using a piecewise linear trend as discussed earlier (where is the time of intervention).

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.

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

- It is common to require the coefficients to decrease as the lag increases. In Chapter 9 we discuss methods to allow this constraint to be implemented.

Outline

Adjusted R^2 I

- Computer output for regression will always give the R^2 value, discussed in Section 5/1. However, it is not a good measure of the predictive ability of a model. Imagine a model which produces forecasts that are exactly 20% of the actual values. In that case, the R^2 value would be 1 (indicating perfect correlation), but the forecasts are not very close to the actual values.
- In addition, R^2 does not allow for "degrees of freedom". 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

$$\text{SSE} = \sum_{i=1}^{N} e_i^2.$$

Adjusted R^2 II

- 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.
- An alternative, designed to overcome these problems, is the adjusted R^2 (also called "R-bar-squared"):

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

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. Using this measure, the best model will be the one with the largest value of $\bar{R}^2$.
Maximizing $\bar{R}^2$ is equivalent to minimizing the following estimate of the variance of the forecast errors:

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

Adjusted R^2 III

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

Cross-validation

- As discussed in Section 2/5, cross-validation is a very useful way of determining the predictive ability of a model. 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. (This is not the same as the residual because the ith observation was not used in estimating the value of $\hat{y}_i$.)
  2. Repeat step 1 for $i = 1, \ldots, N$.
  3. Compute the MSE from $e_1^*, \ldots, e_N^*$. 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 5/5.
- Under this criterion, the best model is the one with the smallest value of CV.

Akaike's Information Criterion

- A closely-related method is Akaike's Information Criterion, which we define 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, although they should all lead to the same model being selected. 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.

Corrected Akaike's Information Criterion

- For small values of N, the AIC tends to select too many predictors, and so a bias-corrected version of the AIC has been developed.

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2(k+2)(k+3)}{N-k-3}.$$

- As with the AIC, the AICc should be minimized.

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. For large values of N, minimizing BIC is similar to leave-v-out cross-validation when $v = N[1 - 1/(\log(N) - 1)]$.
- 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. However, in reality there is rarely if ever a true underlying model, and even if there was a true underlying model, selecting that model will not necessarily give the best forecasts (because the parameter estimates may not be accurate).

Best subset regression

- Where possible, all potential regression models can be fitted (as was done in the above example) 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. Most software packages will at least produce AIC, although CV and AICc will be more accurate for smaller values of N.
- 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 CV, AIC or AICc. Also, the tendency of BIC to select too few variables makes it less suitable for forecasting than either CV, AIC or AICc.

Stepwise regression

- If there are a large number of predictors, it is not possible to fit all possible models. For example, 40 predictors leads to $2^{40} > 1$ trillion possible models! Consequently, a strategy is required to limit the number of models to be explored.
- An approach that works quite well is **backwards stepwise regression**:
  1. Start with the model containing all potential predictors.
  2. Try subtracting one predictor at a time. Keep the model if it improves the measure of predictive accuracy.
  3. Iterate until no further improvement.
- It is important to realise that a stepwise approach is not guaranteed to lead to the best possible model. But it almost always leads to a good model.
- If the number of potential predictors is too large, then this backwards stepwise regression will not work and the starting model will need to use only a subset of potential predictors. In this case, an extra step needs to be inserted in which predictors are also added one at a time, with the model being retained if it improves the measure of predictive accuracy.

Outline

Scatterplots of residuals against predictors

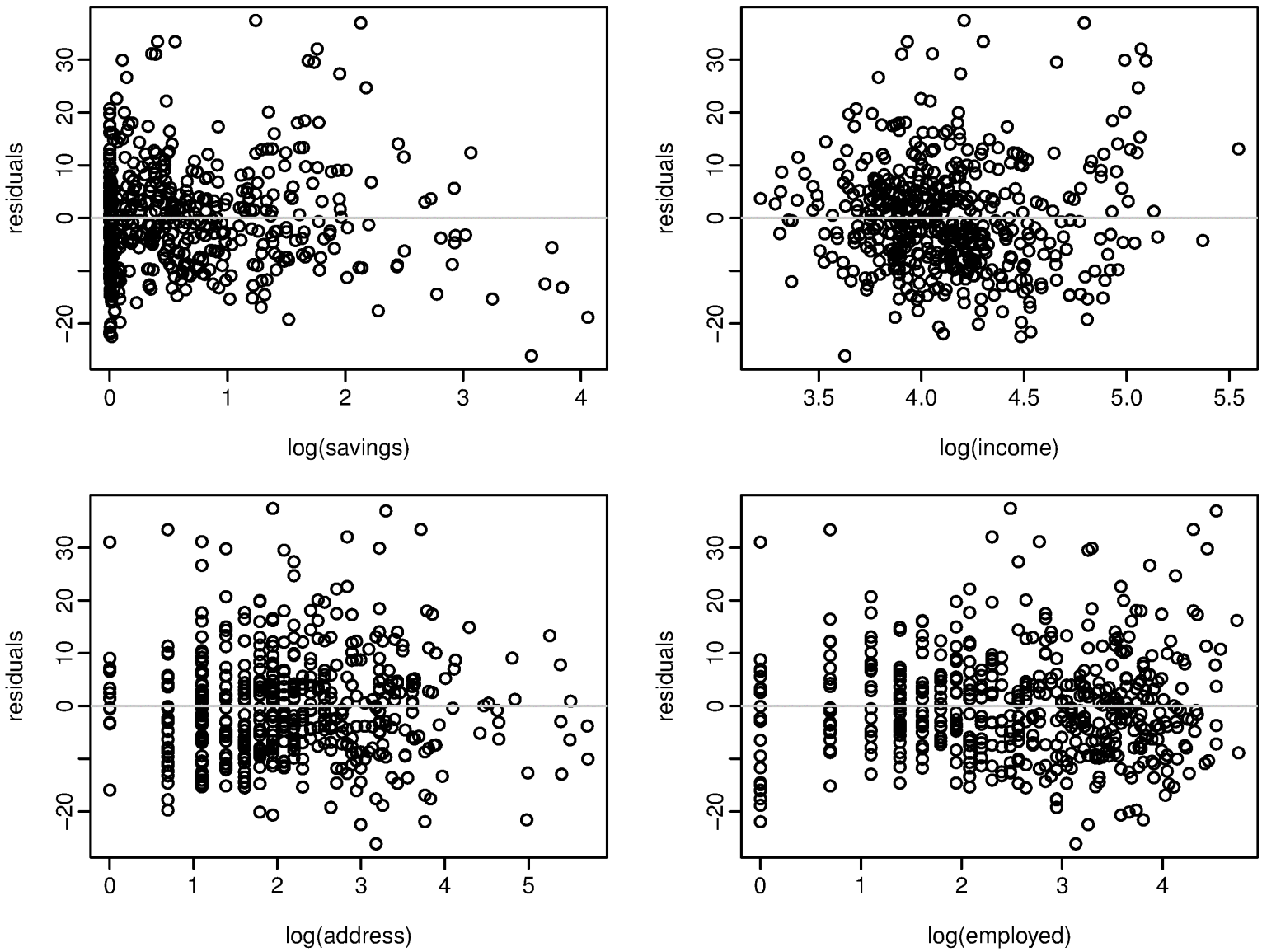

Figure 5.8: The residuals from the regression model for credit scores plotted against each of the predictors in the model.

Scatterplot of residuals against fitted values

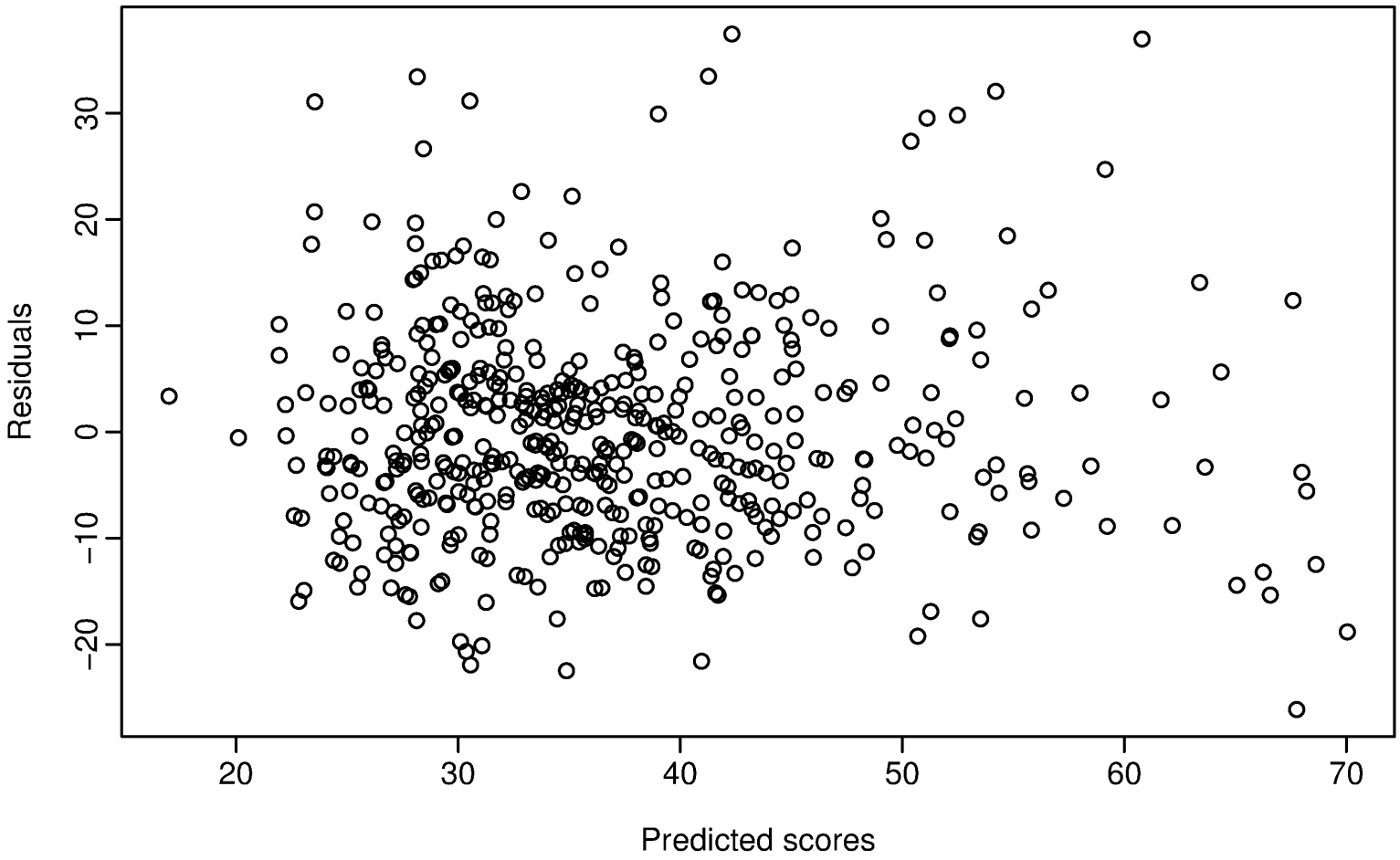

Figure 5.9: The residuals from the credit score model plotted against the fitted values obtained from the model.

Autocorrelation in the residuals

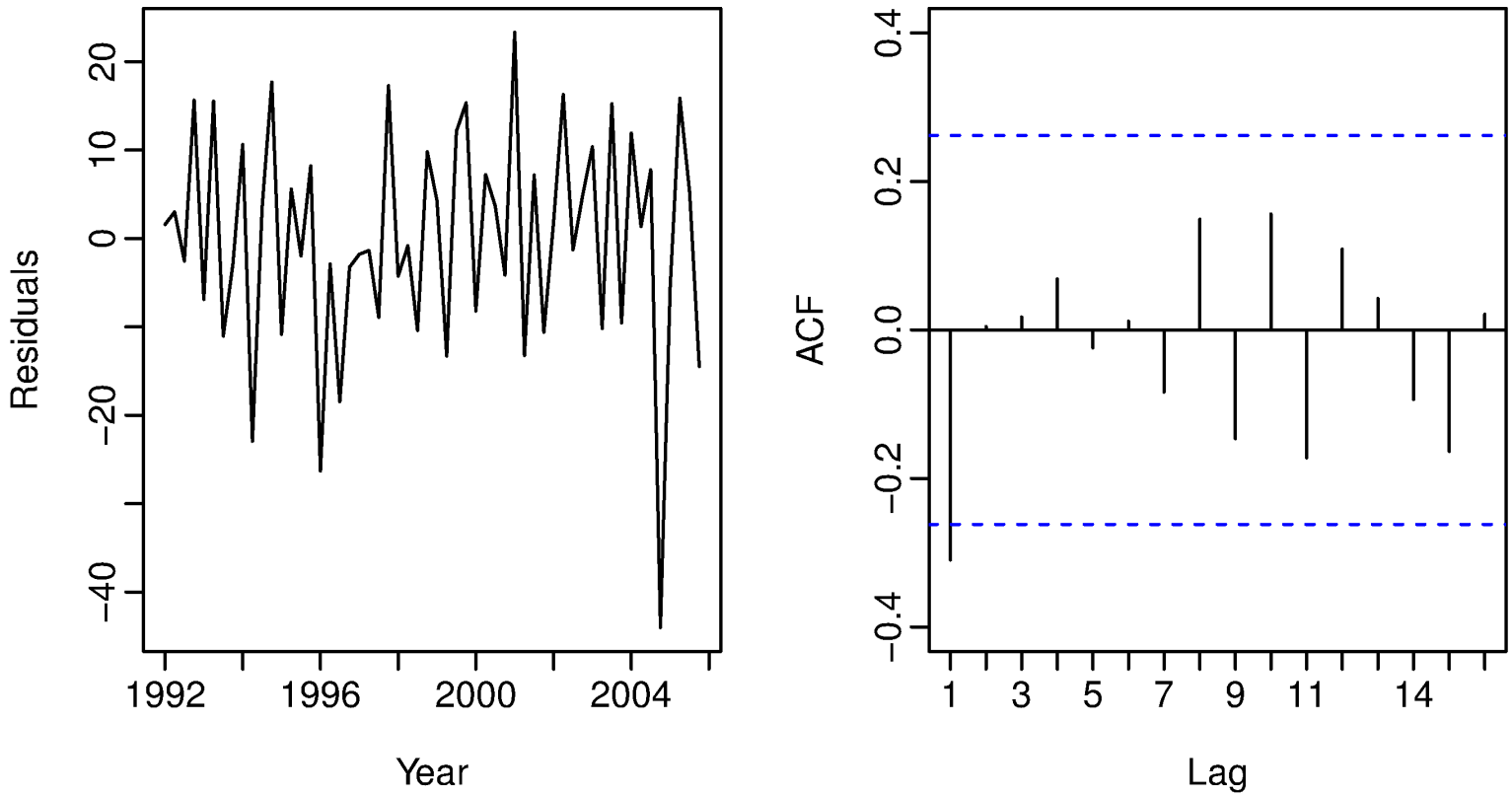

Figure 5.10: Residuals from the regression model for beer production.

Histogram of residuals

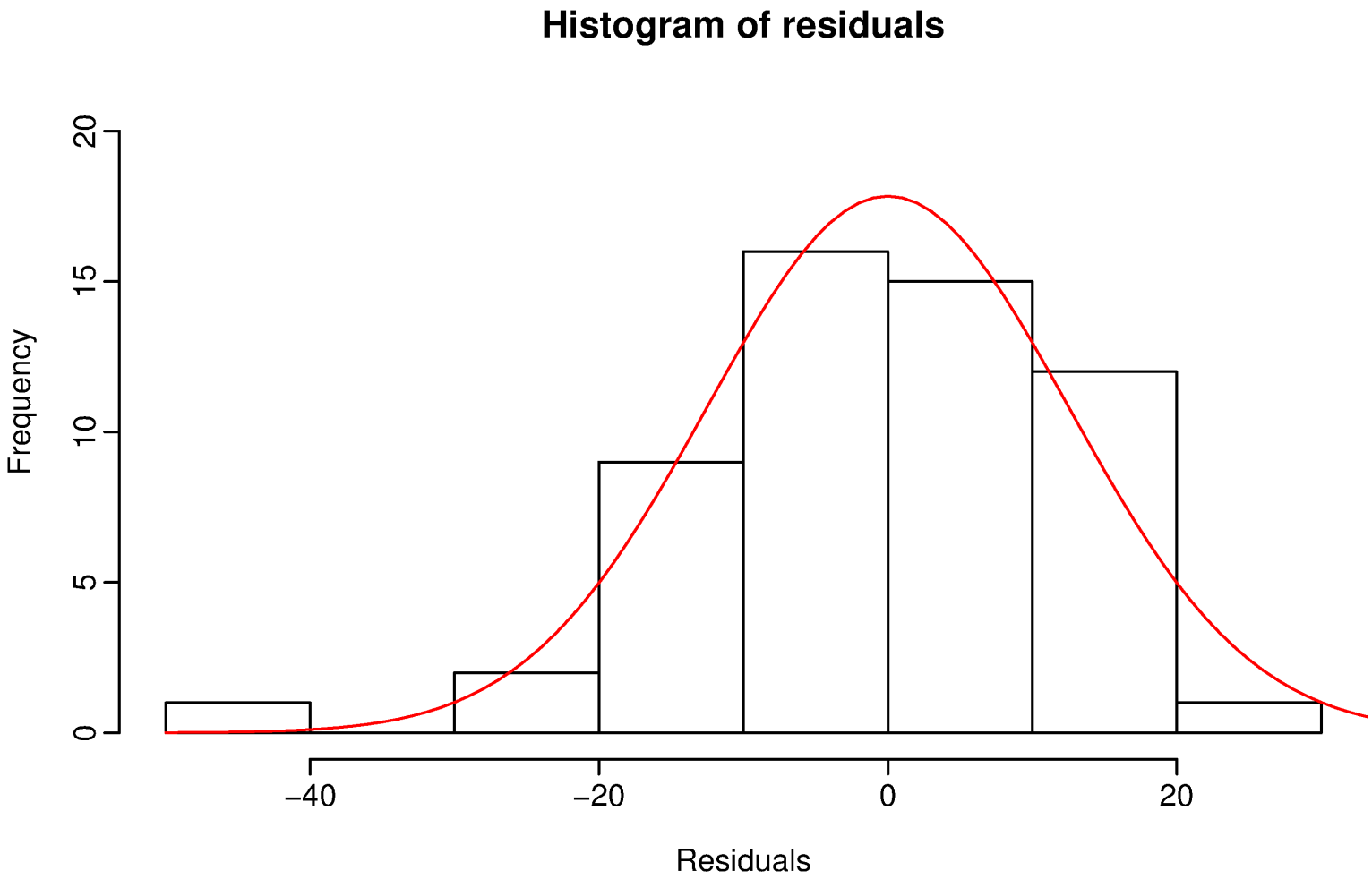

Figure 5.11: Histogram of residuals from regression model for beer production.

Outline

Introduction to multiple linear regression

Some useful predictors
- Seasonal Dummy variables
- Other dummy predictors
- Trend

Example: Australian quarterly beer production

Other things to keep in mind

Selecting predictors

Residual diagnostics