Multiple regresion for time series

Xiaodong Lin

Outline

- Multiple Regression Model
- Use of predictor variables
- Variable selection
- Model diagnostics

Multiple regression model

- The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \cdots, x_p$ and an error term is called the multiple regression model.
- $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon.$$

  where $\beta_0, \beta_1, \cdots, \beta_p$ are parameters and ϵ is the error.
- The coefficients measure the marginal effects of the predictor variables.

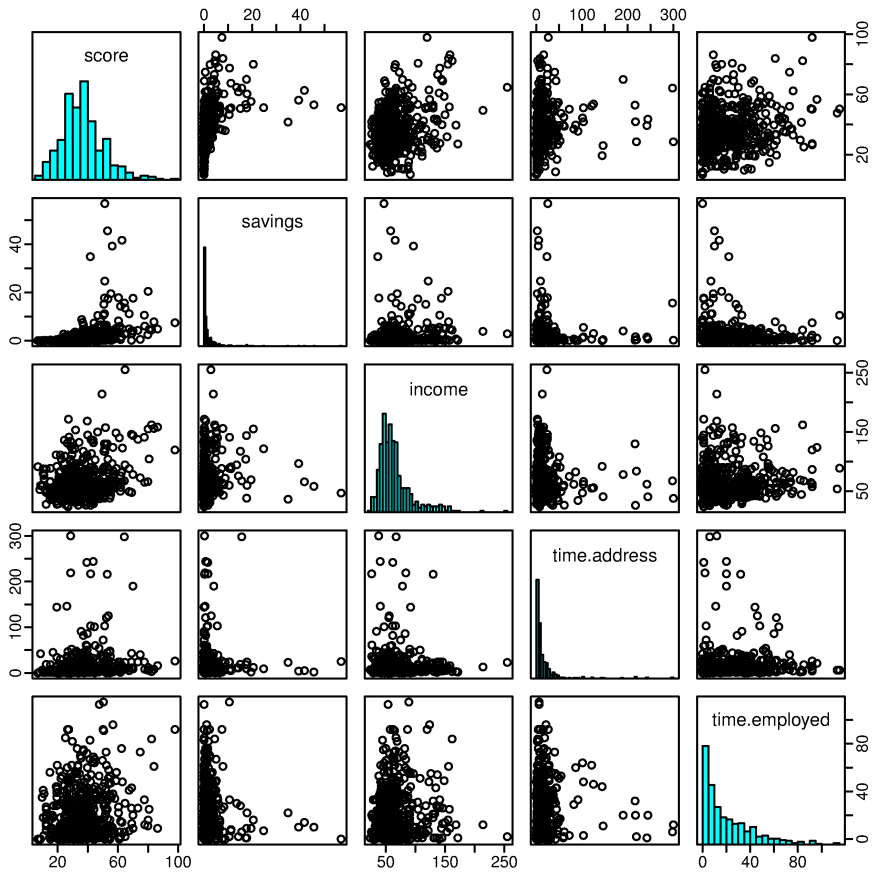
Credit scores data

Banks score loan customers based on a lot of personal information. A sample of 500 customers from an Australian bank provided the following information.

Score	Savings	Income	Time at add	time at job
39.40	0.01	111.17	27	8
51.79	0.65	56.40	29	33
32.82	0.75	36.74	2	16
57.31	0.62	55.99	14	7
37.17	4.13	62.04	2	14

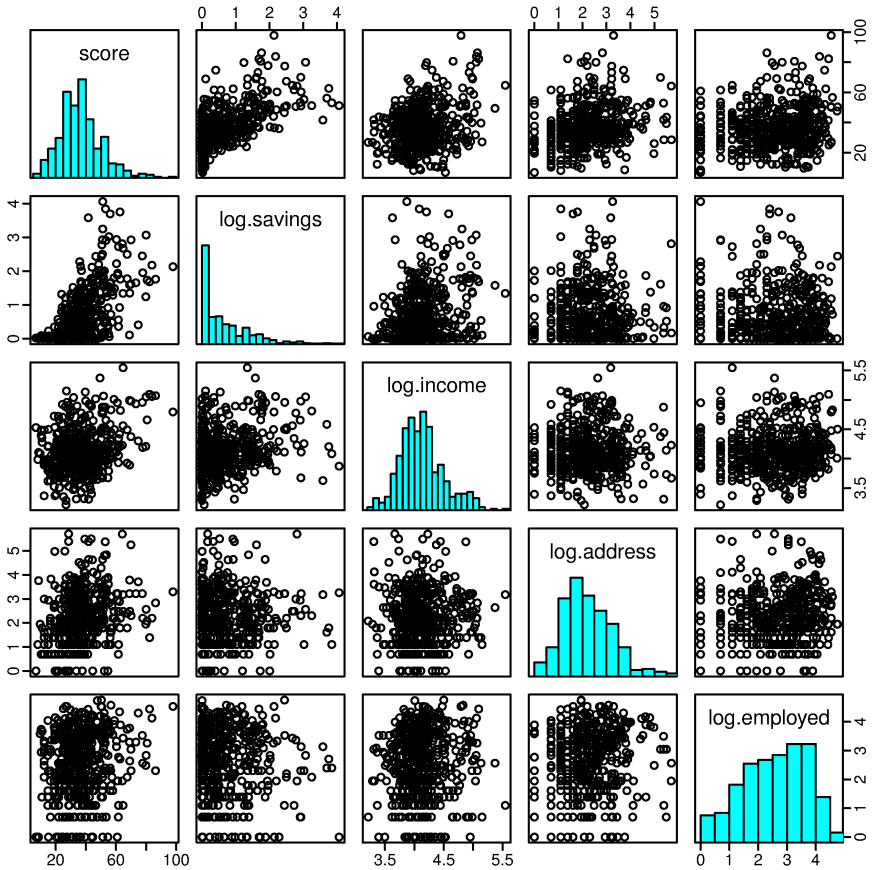
Credit scores

- Pairwise scatter plots.



Credit scores

- The predictor variables are highly skewed, suggesting taking the log transformation $\log(x+1)$.



Credit scores with multiple regression

- Now we can fit a multiple regression

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_4 x_4 + \epsilon$$

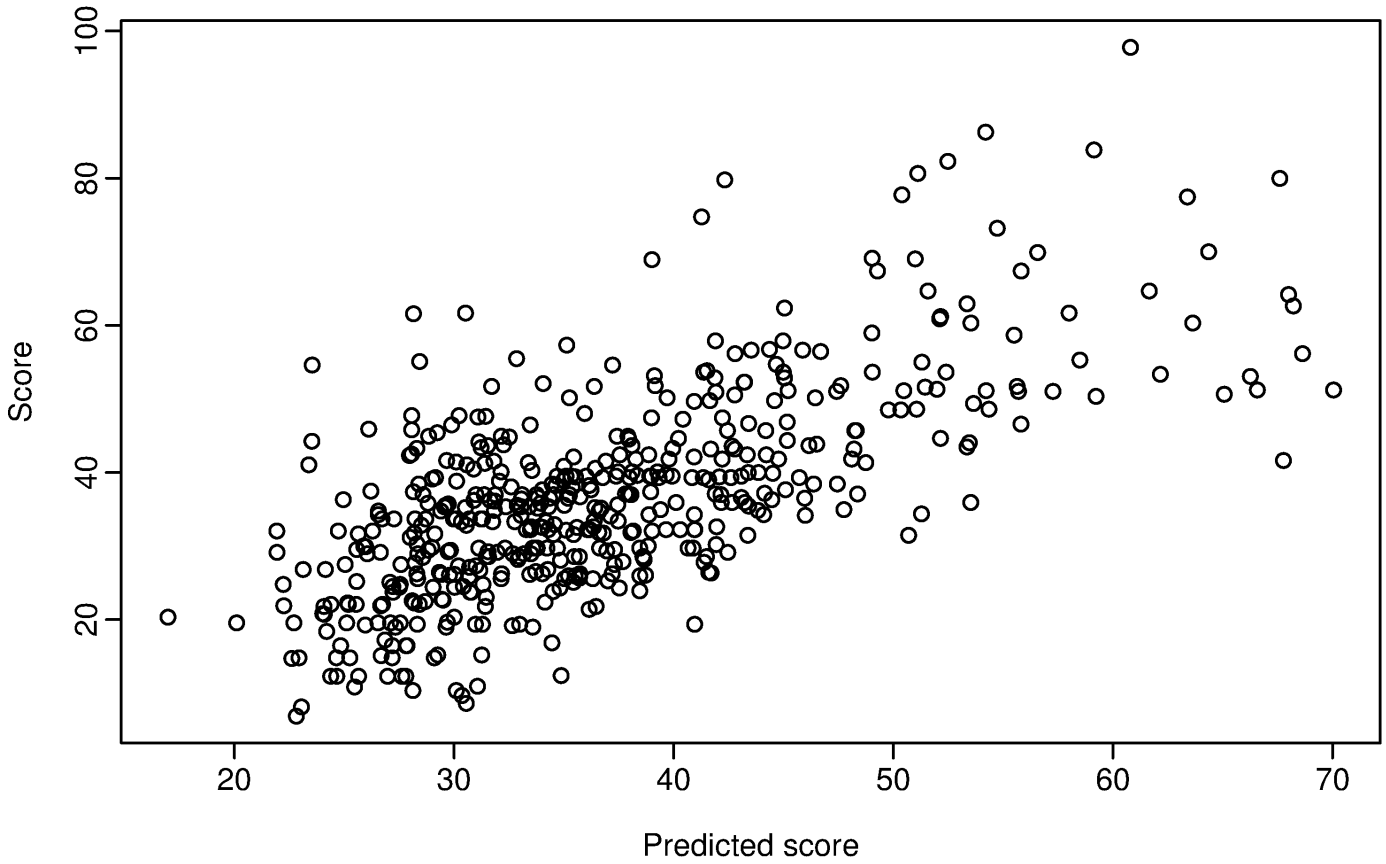
  where y is the credit score, $x_1 = \log(\textit{savings} + 1)$, $x_2 = \log(\textit{income} + 1)$, $x_3 = \log(\textit{time at address} + 1)$ and $x_4 = \log(\textit{time at job} + 1)$.
- The $\hat{\beta}_i$, $0 \leq i \leq 4$ can be obtained via minimizing the sum of squares errors.

Regression outputs

```
Call:
lm(formula = score ~ log.savings + log.income + log.address +
    log.employed, data = creditlog)
Residuals:
    Min      1Q  Median      3Q     Max
-26.133  -6.966  -1.125   5.379  37.446
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.2186     5.2309  -0.042  0.96668
log.savings   10.3526     0.6124  16.904  < 2e-16 ***
log.income     5.0521     1.2579   4.016 6.83e-05 ***
log.address    2.6666     0.4345   6.137 1.72e-09 ***
log.employed   1.3138     0.4094   3.209  0.00142 **
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 10.16 on 495 degrees of freedom
Multiple R-squared: 0.4701, Adjusted R-squared: 0.4658
F-statistic: 109.8 on 4 and 495 DF,  p-value: < 2.2e-16
```

Predicted v.s. actual

- Actual credit scores plotted against fitted credit scores using the multiple regression model. The correlation is 0.6856, so the squared correlation is $(0.6856)^2 = 0.4701$.



Dummy variables

- If we are forecasting daily electricity demand and would like to use day of the week as a predictor. Then we can add $7 - 1 = 6$ dummy variables.

Day	D1	D2	D3	D4	D5	D6
Monday	1	0	0	0	0	0
Tuesday	0	1	0	0	0	0
Wednesdays	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

- The coefficients measure the effect of that category relative to the omitted category. The coefficient associated with Monday will measure the effect of Monday compared to Sunday on the forecast variable.

Trend

- A linear trend can be accounted for using $x_{1,t} = t$.
- A piecewise linear trend

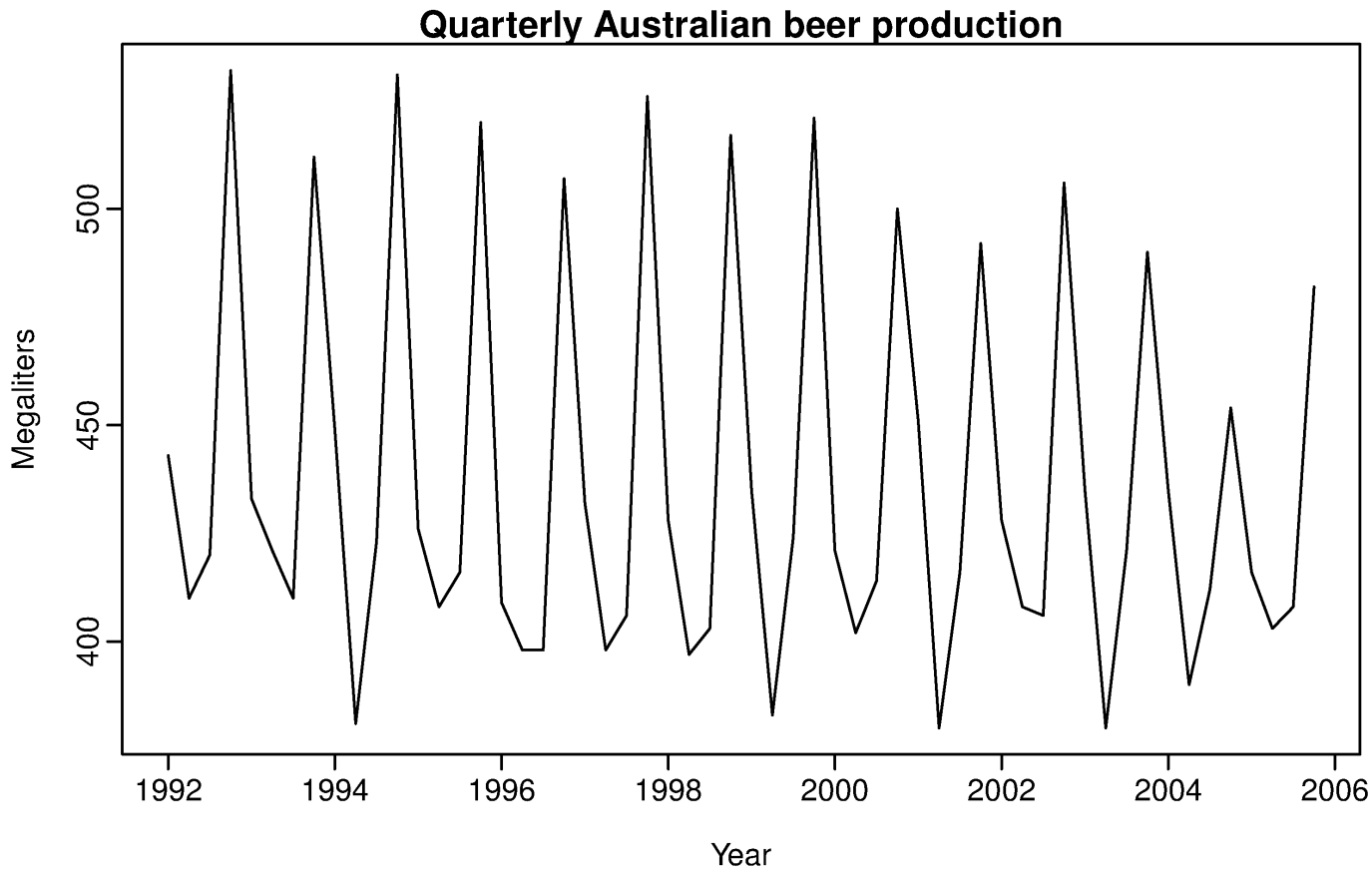
$$x_{1,t} = t$$

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

- A polynomial trend: $x_{1,t} = t$, $x_{2,t} = t^2$, $\cdots$

Australian quarterly beer production

- Fit a seasonal multiple regression model for beer production forecasting



Codes

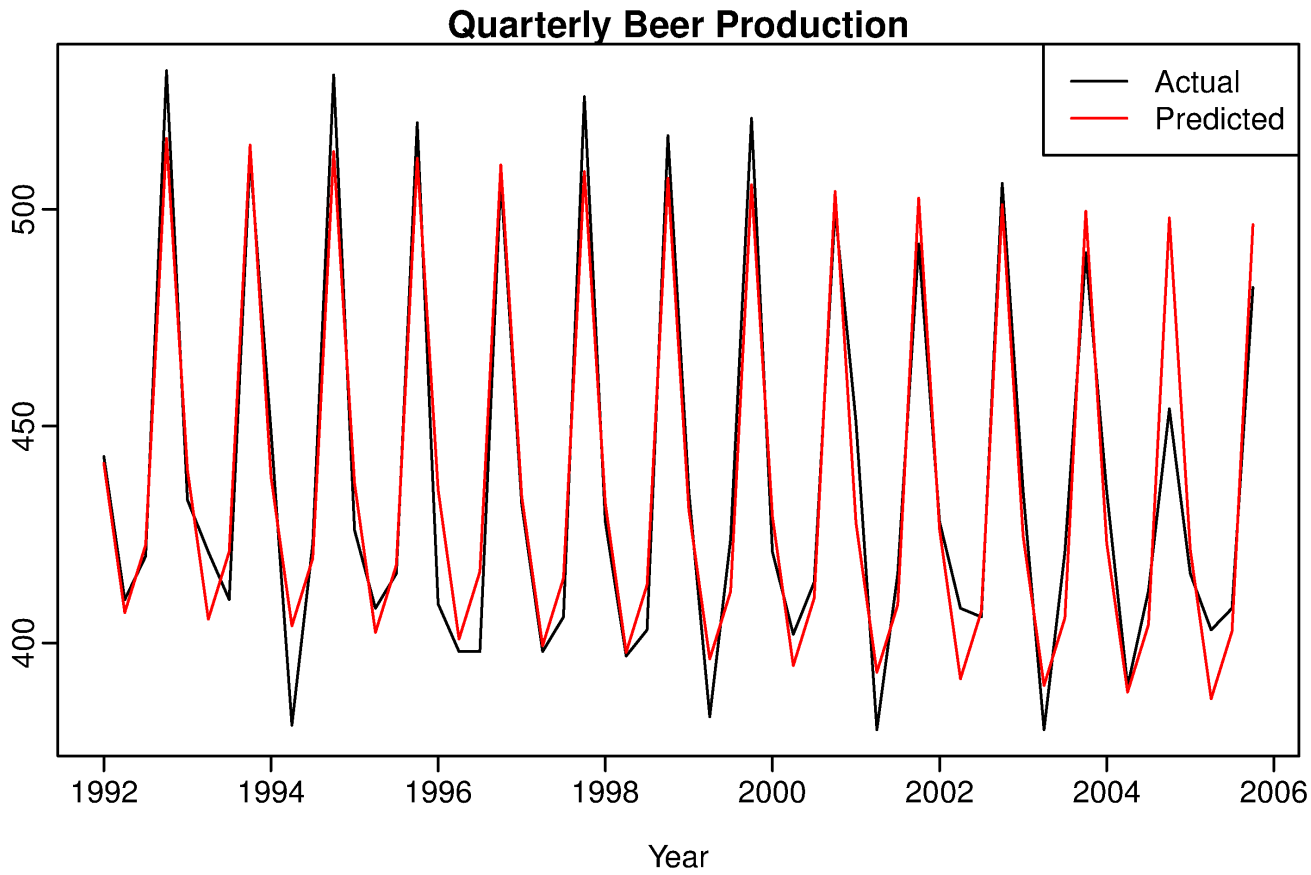
```
## Multiple regression using tslm.

beer2 <- window(ausbeer,start=1992,end=2006-.1)
fit <- tslm(beer2 ~ trend + season)
summary(fit)

## multiple regression with dummy quaterly variables.

x1=1:56
x2=rep(c(0,1,0,0),14)
x3=rep(c(0,0,1,0),14)
x4=rep(c(0,0,0,1),14)
f1=lm(beer2~x1+x2+x3+x4)
summary(f1)
```

Outputs

```
lm(formula = beer2 ~ x1 + x2 + x3 + x4)
Residuals:
    Min      1Q  Median      3Q     Max
-44.024  -8.390   0.249   8.619  23.320
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 441.8141     4.5338  97.449  < 2e-16 ***
x1           -0.3820     0.1078  -3.544 0.000854 ***
x2          -34.0466     4.9174  -6.924 7.18e-09 ***
x3          -18.0931     4.9209  -3.677 0.000568 ***
x4           76.0746     4.9268  15.441  < 2e-16 ***
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 13.01 on 51 degrees of freedom
Multiple R-squared: 0.921, Adjusted R-squared: 0.9149
F-statistic: 148.7 on 4 and 51 DF,  p-value: < 2.2e-16
```

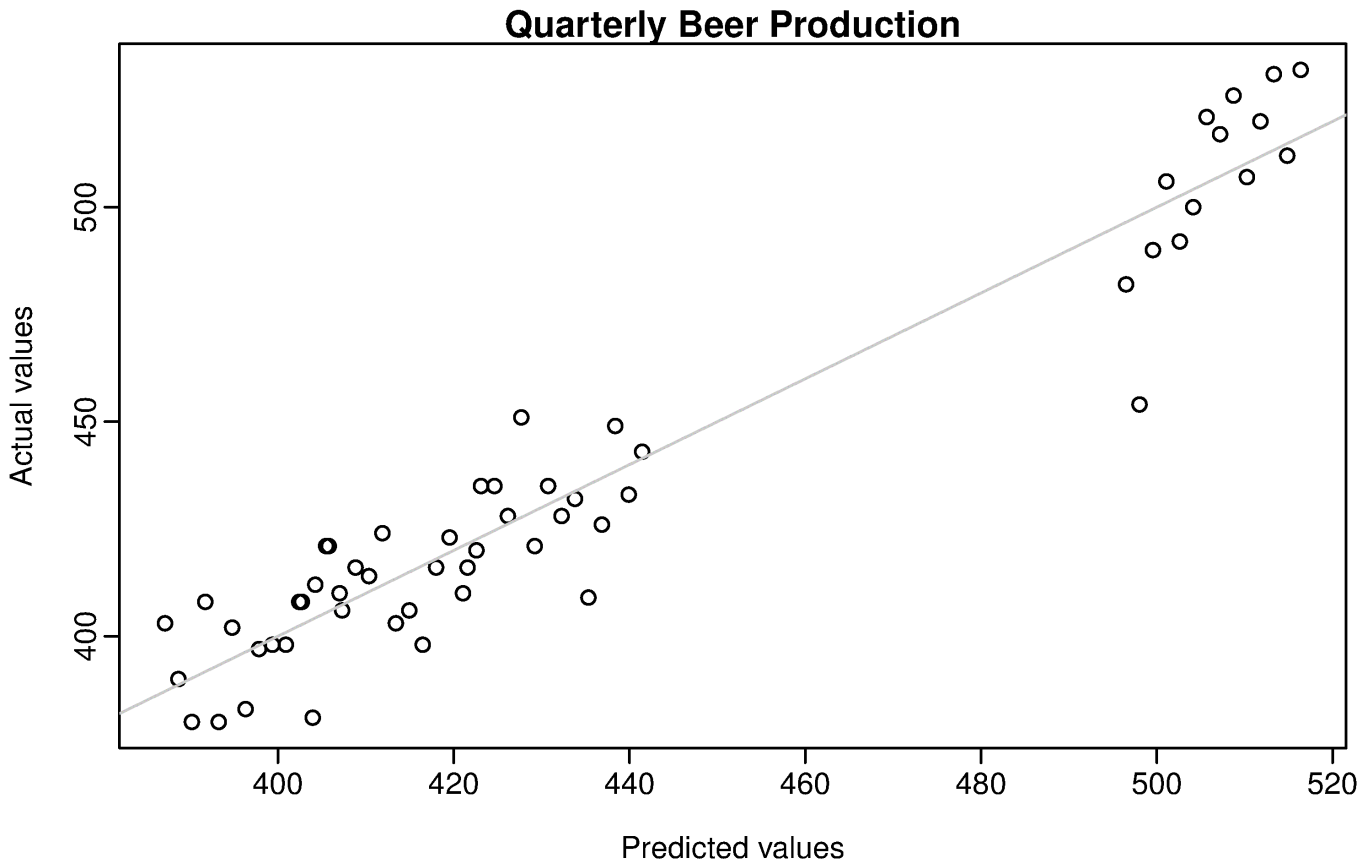
Actual v.s. predicted

- The actual values compared to the predicted values.



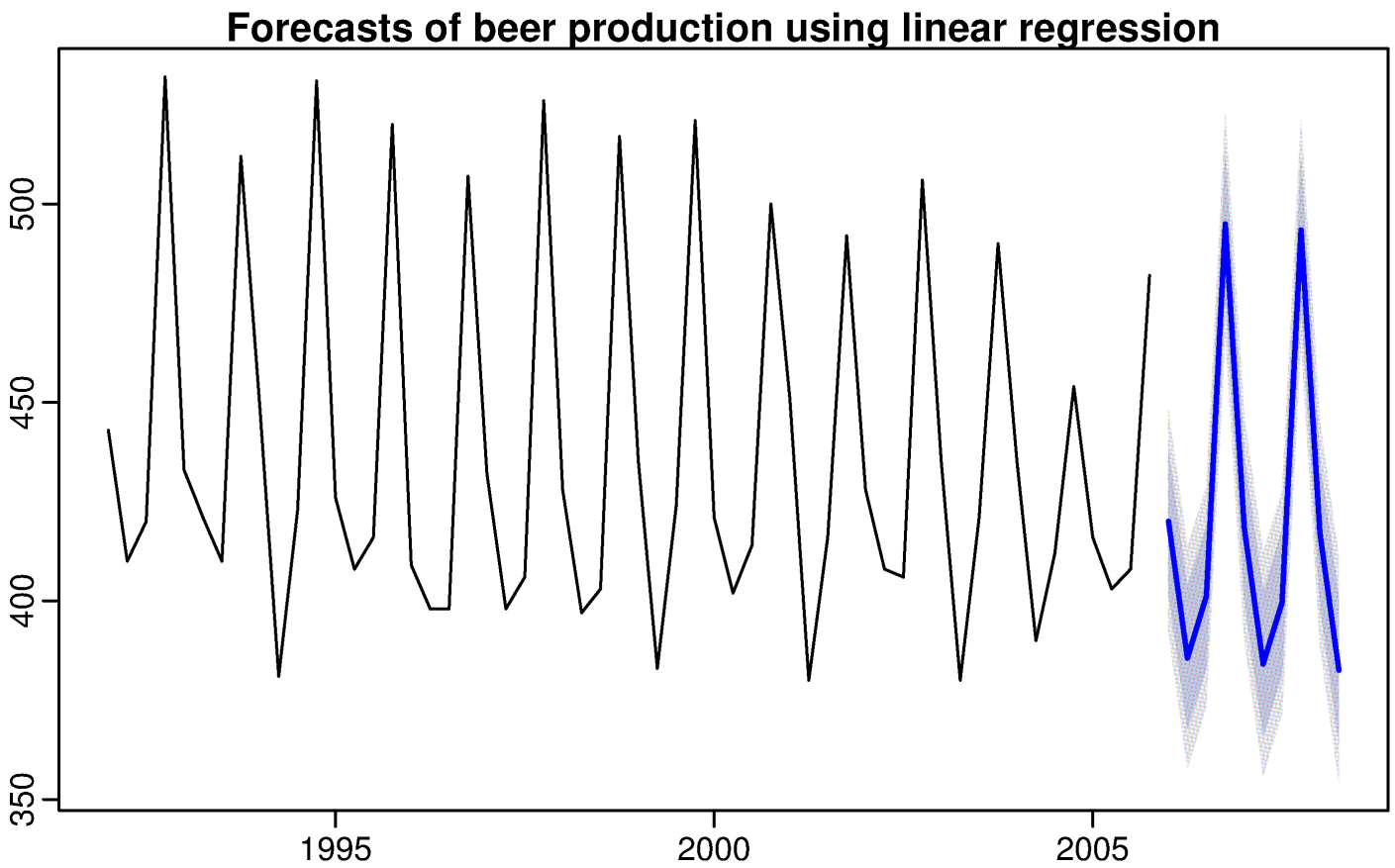
Actual v.s. predicted in a scatter plot

- The actual values compared to the predicted values in a scatter plot.



Predictions

- Predicted values for the next few quarters.



Some variable selection measures

- The adjusted R^2 is

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

  Maximizing $\bar{R}^2$ tends to lead to over complicated models.
- Balance between model fitting and model complexity.
- AIC:

$$n\log(\frac{SSE}{n}) + 2(k+2)$$

- BIC:

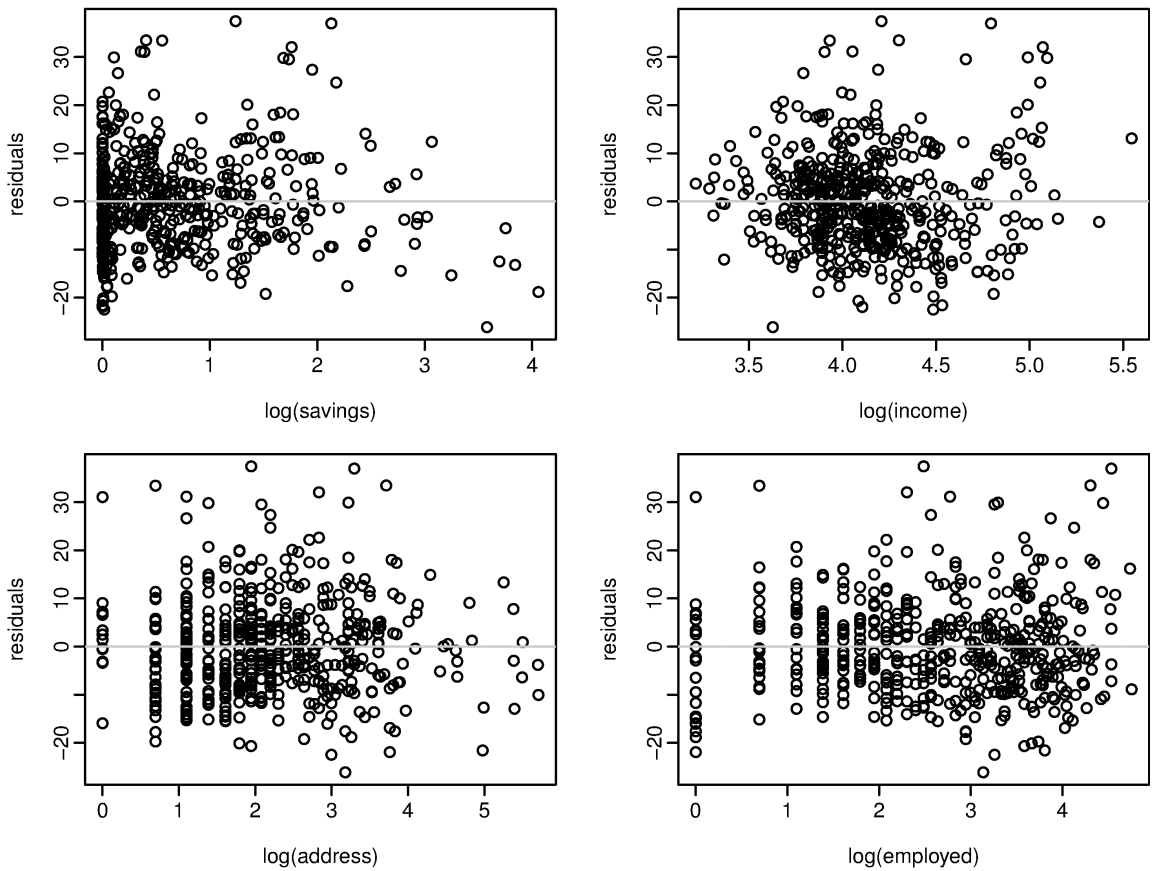
$$n\log(\frac{SSE}{n}) + (k+2)\log(n)$$

Selection procedures

- Best subset regression
  Fit all potential models, and select the one with the smallest model selection criteria score.
- Backwards stepwise regression
  1. Start with the full model
  2. Subtracting the least significant variable, check if the predictive accuracy improves.
  3. Iterative until no further improvement can be achieved.

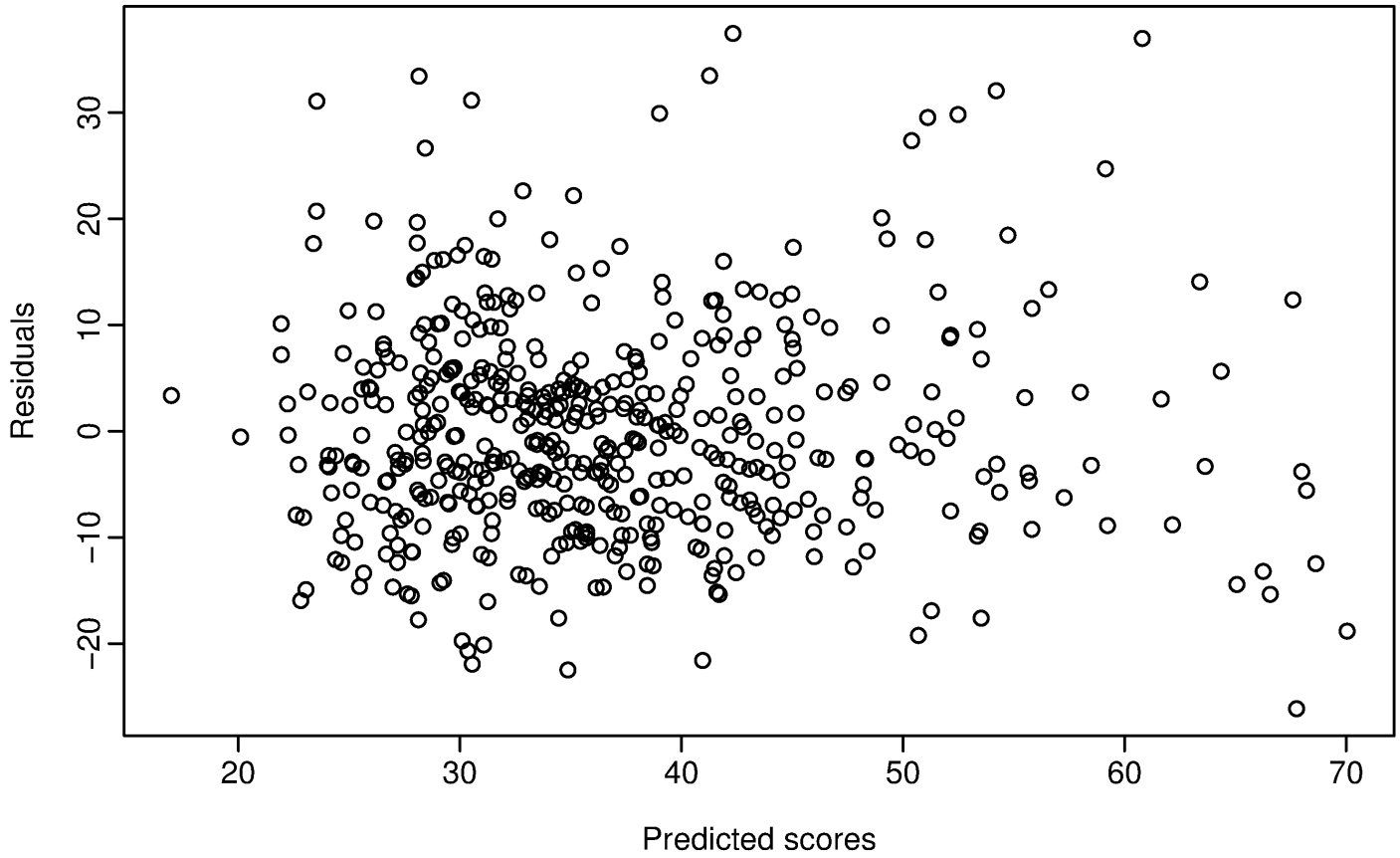
Residual plots

- Plot residuals against the predictor variables to check if there exists systemetic patterns.



Residual v.s. fitted values

- Plot residuals against the fitted values to check if there exists systemetic patterns.



Autocorrelation in the residuals

- When the data is a time series, we need to check time dependencies on the residuals.

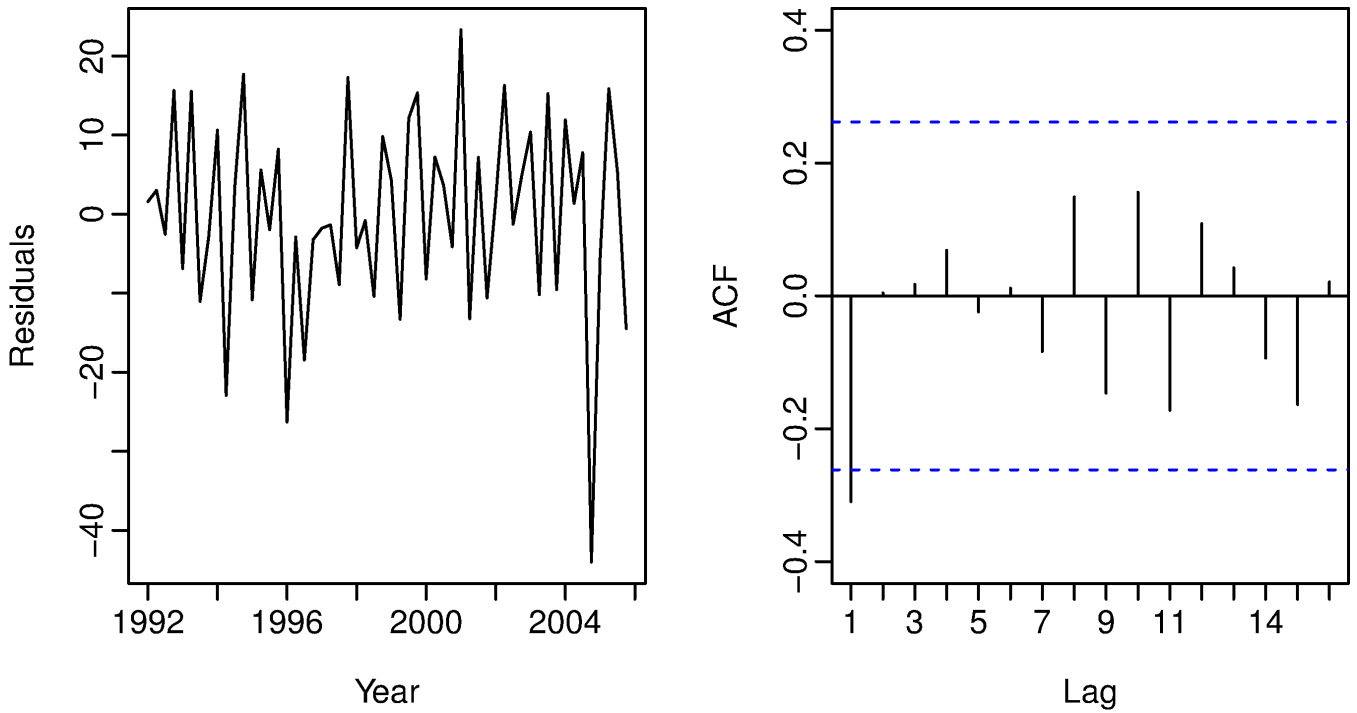


Figure: Residuals from beer production example. Note there is an outlier in 4th quarter of 2004.

Normality in the residuals

- Check if the residuals are normally distributed.

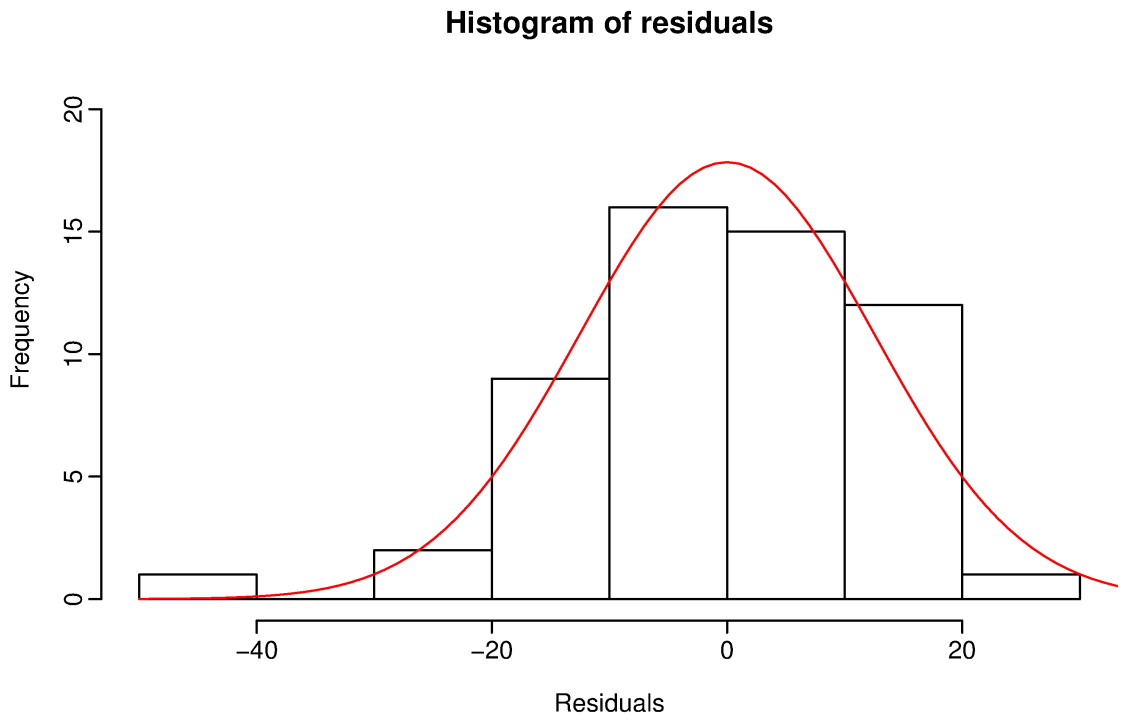


Figure: Residuals from beer production example.