# Dummy Variables

Susan Thomas
IGIDR, Bombay

24 November, 2008

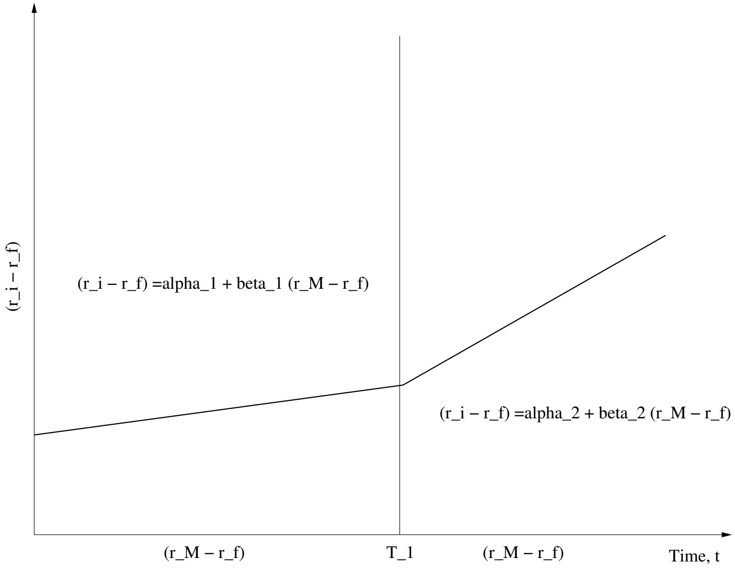
# The problem of structural change

- Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$
- Structural change, type 1: change in parameters in time.

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 X_i + e_{1i} \text{ for period 1} \\ Y_i &= \alpha_2 + \beta_2 X_i + e_{2i} \text{ for period 2} \end{aligned}$$

  Solution: for a given "break point" $\tau$,
  $\sigma^2_{\text{unrestricted}} = \sum^{n1} e^2_{1i} + \sum^{n2} e^2_{2i}$ vs.$\sigma^2_{\text{restricted}} = \sum^{N=n1+n2} \epsilon^2_i$
  Critical value: F(k, N - 2k)
- Other types of structural change
  - Type 2: change in constant terms (dummy variables)
  - Type 3: change in distribution of errors
  - Type 4: change in sets of coefficients

# Testing for change in parameters in the sample

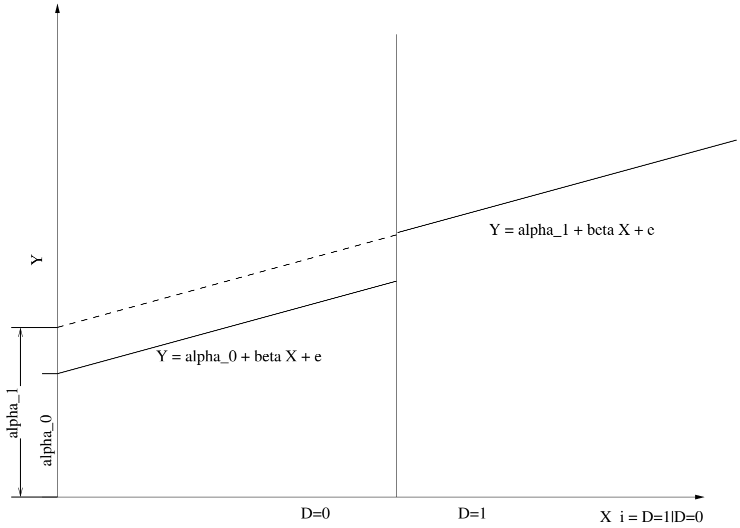


# Type 2: change in constant terms

- A dummy variable, $D_k$ is a binary variable which takes the value of 0 or 1 when the condition is "false" or "true".
  Example: $D_k = 0$ if a boy child is born, $D_k = 1$ if a girl child is born.
- Dummies are useful in changing the structure of the model depending upon the value of some conditioning variable.
- The simplest is to change the "intercept" term of the regression model.
  Example: $Y_i =$ weight, $X_i =$ height

$$
\begin{aligned}
Y_i &= \alpha_1 + \beta X_i + e_{1i}, i = \text{ female} \\
Y_i &= \alpha_2 + \beta X_i + e_{2i}, i = \text{ male}
\end{aligned}
$$

# Type 2: change in constant terms



# Change in intercept: test of mean

- Data: Group 1 $Y_i = \mu + \epsilon_i$ Group 2 $Y_i = (\mu + \delta) + \epsilon_i$ $\mu = \mu_{G_1}, \mu + \delta = \mu_{G_2}$ or $\delta = \mu_{G_2} - \mu_{G_1}$
- The regression can be estimated as:

$$Y_i = \mu + \delta D_i + \epsilon_i$$

  where $D_i = 0$ for Group 1, $D_i = 1$ for Group 2.
- Alternative model: $Y_i = \mu_1 G_1 + \mu_2 G_2 + e_i$

$$\begin{aligned} G_1 &= 1, ifi = \text{Group 1, otherwise } G_1 = 0 \\ G_2 &= 1, ifi = \text{Group 2, otherwise } G_2 = 0 \end{aligned}$$

- However, $H_0$ in model 2 cannot ask whether $\alpha_1, \alpha_2$ = 0. Advantage of the dummy variable model: $H_0 : \delta = \mu_{G_1} - \mu_{G_2}$ is a well posed test of whether the mean of $G_1, G_2$ are different.

# Change in intercept: test of mean

- Data frame for model 1

$$
\begin{aligned}
[y \;\; x] &= \begin{bmatrix} Y_1 & 1 & 0 \\ Y_2 & 1 & 0 \\ \ldots & \ldots & \ldots \\ Y_{n_1} & 1 & 0 \\ Y_{n_1+1} & 1 & 1 \\ Y_{n_1+2} & 1 & 1 \\ \ldots & \ldots & \ldots \\ Y_N & 1 & 1 \end{bmatrix} \\
Y &= \begin{bmatrix} I_{n1} & 0 \\ I_{n2} & I_{n2} \end{bmatrix} + \epsilon
\end{aligned}
$$

- OLS solution:

$$
\begin{bmatrix} \hat{\mu} \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} N & n_2 \\ n_2 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} n_1 \bar{y}_1 + n_2 \bar{y}_2 \\ n_2 \bar{y}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \end{bmatrix}
$$

# HW

- Use the normal equations of the OLS optimisation to show that

$$\begin{bmatrix} \hat{\mu} \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \end{bmatrix}$$

- What is the standard error of $\hat{\mu}, \hat{\delta}$?

# Change in intercept: test of mean

- Data frame for model 2

$$Y = \begin{bmatrix} I_{n1} & 0 \\ 0 & I_{n2} \end{bmatrix} + \epsilon$$

- OLS solution:

$$\begin{bmatrix} \hat{\mu}_{G_1} \\ \hat{\mu}_{G_2} \end{bmatrix} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} n_1 \bar{y}_1 \\ n_2 \bar{y}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix}$$

$$\begin{bmatrix} \sigma_{\hat{\mu}_{G_1}} \\ \sigma_{\hat{\mu}_{G_2}} \end{bmatrix} = \begin{bmatrix} \sigma_\epsilon / \sqrt{n1} \\ \sigma_\epsilon / \sqrt{n2} \end{bmatrix}$$

# Model choices when dealing with dummy variables

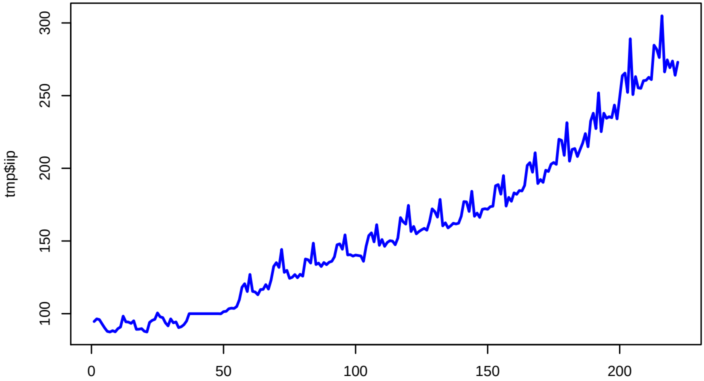
- Model 1: $Y_i = \mu + \delta D_i + \epsilon_i$
- Model 2: $Y_i = \mu_{G_1} D_{G_1} + \mu_{G_2} D_{G_2} + e_i$
- Incorrect model: $Y_i = \alpha + \mu_{G_1} D_{G_1} + \mu_{G_2} D_{G_2} + e_i$
- Data matrix for the models:

$$\begin{array}{ccc} \text{Model 1} & \text{Model 2} & \text{Incorrect model} \\ \begin{bmatrix} I_{n1} & 0 \\ I_{n2} & I_{n2} \end{bmatrix} & \begin{bmatrix} I_{n1} & 0 \\ 0 & I_{n2} \end{bmatrix} & \begin{bmatrix} I_{n1} & I_{n1} & 0 \\ I_{n2} & 0 & I_{n2} \end{bmatrix} \end{array}$$

- In the third data matrix, the sum of the second and third columns add up to the first.
  This means the inverse of $(X'X)^{-1}$ cannot be calculated.
  Which in turn means that three coefficents cannot be estimated.
- Problem of multicollinearity: with dummy variables, coefficients for a "comprehensive" set of dummies cannot be estimated simultaneously with an intercept.
  Model can either contain a comprehesive set of dummy variables or an intercept.

# Modelling the index of industrial production, IIP

# Seasonality in the IIP data



# Features of IIP

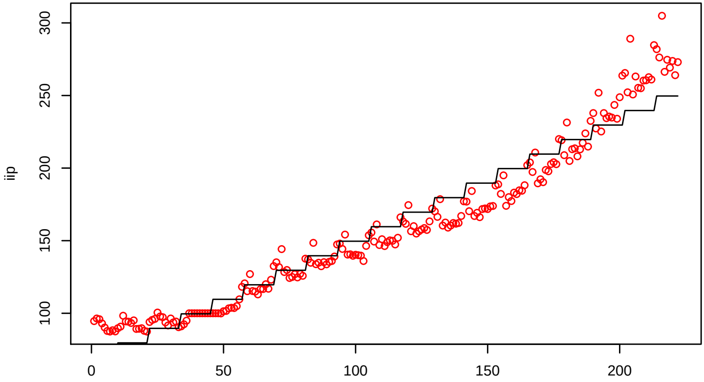
- Data has monthly frequency from April 1990 to Sep 2008
- Appears to have an annual trend – linear? non-linear?
- Appears to have "seasonality". Expected patterns at regular intervals.
- Model suggestions:
  - A different level for different years: year trend term Captures a level of IIP for a given year. For example,trend is denoted as "1" for 1990, "2" for 1991, "3" for 1992, etc.
  - A different level for different months: month dummies Captures a level of IIP for a given month in a year. Each month has a dummy. For example, $\mathrm{Jan}_t$ is "1" for January in any month, and "0" otherwise.
- Model 1:
  $$IIP_t = \alpha_0 + \alpha_1 Y_t + \beta_1 Jan_t + \beta_2 Feb_t + \ldots + \beta_{11} Nov_t + \epsilon_t$$

- Regression results

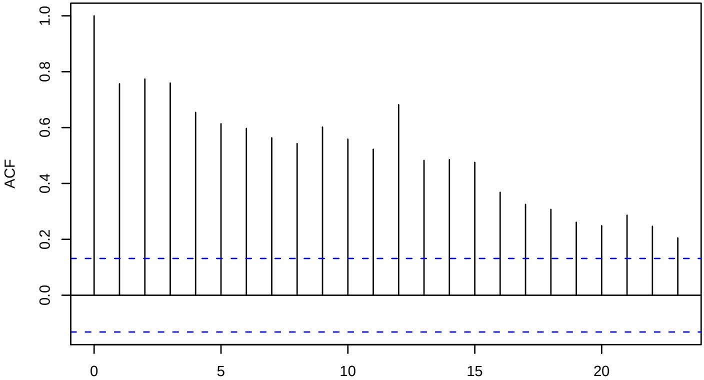
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 59.5827 | 1.9685 | 30.27 | 0.0000 |
| year | 10.0038 | 0.1736 | 57.63 | 0.0000 |

Residual SE = 0.0530
F-stat(1, 220) = 3322
prob value = 2.2e-16
R-squared = 0.9379
Adjusted R-squared: 0.9376

# Explained vs. Actual data



# Behaviour of serial dependence in residuals



# Full model 1

- Regression results

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 59.4567 | 2.7854 | 21.35 | 0.0000 |
| year | 10.0058 | 0.1677 | 59.68 | 0.0000 |
| jan | -3.6010 | 3.8334 | -0.94 | 0.3486 |
| feb | 10.1878 | 3.8334 | 2.66 | 0.0085 |
| mar | -4.3148 | 3.7649 | -1.15 | 0.2531 |
| may | -3.5201 | 3.7649 | -0.93 | 0.3509 |
| jun | -1.9253 | 3.7649 | -0.51 | 0.6096 |
| jul | -2.3411 | 3.7649 | -0.62 | 0.5347 |
| aug | -0.5201 | 3.7649 | -0.14 | 0.8903 |
| sep | -2.2230 | 3.8352 | -0.58 | 0.5628 |
| oct | 0.2770 | 3.8352 | 0.07 | 0.9425 |
| nov | 9.9826 | 3.8352 | 2.60 | 0.0099 |

Residual SE = 0.04730
F-stat(11, 210) = 3327.3
prob value = 2.2e-16
R-squared = 0.9449
Adjusted R-squared: 0.9420

# Interpreting the model

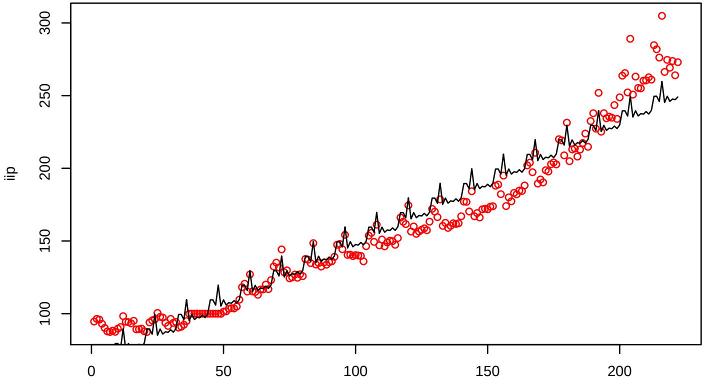
- The omitted dummy is “Dec”.
- Therefore, the ’jan” coefficient value of -3.601 is the additional shift for January in addition to the value for December. In this model, the January effect is:

$$59.4567 - 3.601 = 55.8557$$

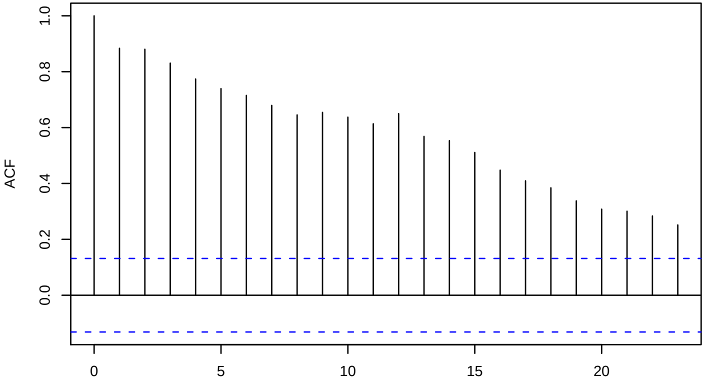
- From the model, the IIP level for January 1991 is

$$IIP_{\text{jan 1991}} = 59.4567 + 10.0058 * 2 - 3.601 = 75.8673$$

# Explained vs. Actual data



# Behaviour of serial dependence in residuals



# Model 2

- The trend and seasonality is non-linear
- Model suggestions:
  - Fit the model on log(IIP)
- Model 2:
  $\log IIP_t = \alpha_0 + \alpha_1 Y_t + \beta_1 Jan_t + \beta_2 Feb_t + \ldots + \beta_{11} Nov_t + \epsilon_t$

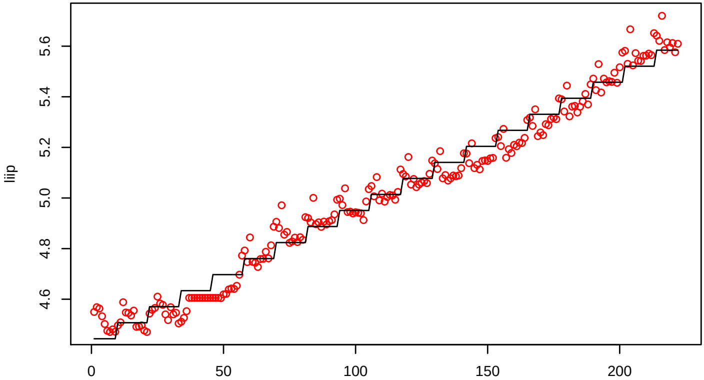
# Simple model 2

- Regression results:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.3804 | 0.0075 | 580.80 | 0.0000 |
| year | 0.0634 | 0.0007 | 95.27 | 0.0000 |

Residual SE = 0.05301
F-stat(1, 220) = 9077
prob value = 2.2e-16
R-squared = 0.9763
Adjusted R-squared: 0.9762

# Explained vs. Actual data



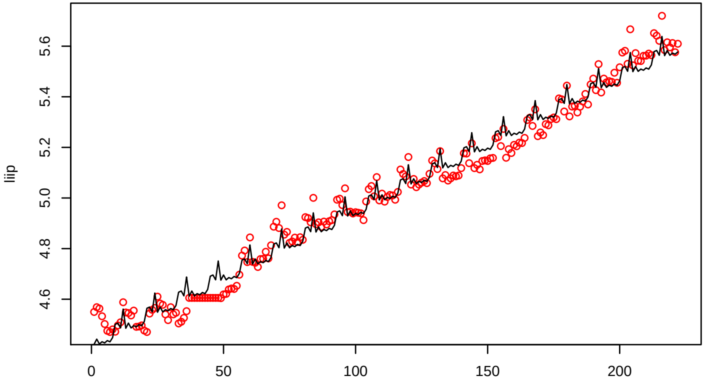
# Full model 2

- Regression results:

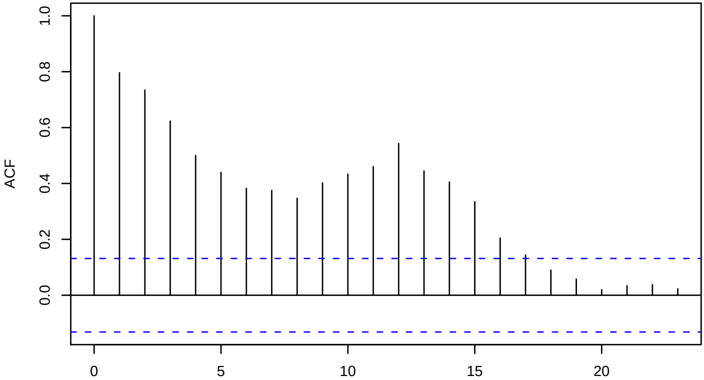
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.3788 | 0.0099 | 443.15 | 0.0000 |
| year | 0.0634 | 0.0006 | 106.56 | 0.0000 |
| jan | -0.0191 | 0.0136 | -1.40 | 0.1621 |
| feb | 0.0554 | 0.0136 | 4.07 | 0.0001 |
| mar | -0.0205 | 0.0134 | -1.53 | 0.1267 |
| may | -0.0193 | 0.0134 | -1.44 | 0.1502 |
| jun | -0.0103 | 0.0134 | -0.77 | 0.4429 |
| jul | -0.0149 | 0.0134 | -1.11 | 0.2664 |
| aug | -0.0057 | 0.0134 | -0.43 | 0.6681 |
| sep | -0.0106 | 0.0136 | -0.78 | 0.4376 |
| oct | 0.0063 | 0.0136 | 0.46 | 0.6436 |
| nov | 0.0587 | 0.0136 | 4.31 | 0.0000 |

Residual SE = 0.04732
F-stat(11, 210) = 1042
prob value = 2.2e-16
R-squared = 0.9820
Adjusted R-squared: 0.9811

# Explained vs. Actual data



# Residual data



# Results, inference

- There is still a lot of serial dependence in the residuals of the model.
  This means that there is yet a lot of variance about the IIP which is to be captured.

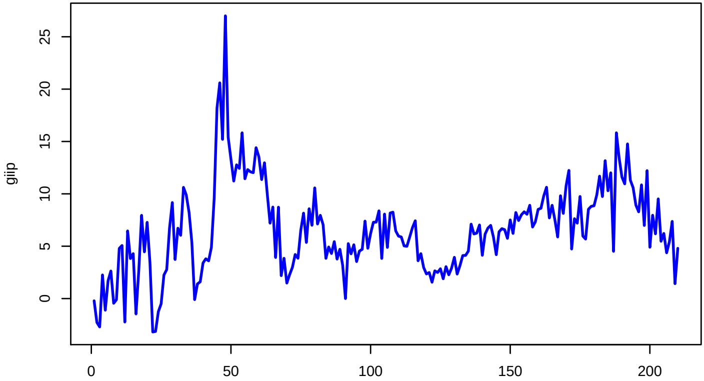
# Model 3

- The dummy variables capture a
- The linearity is better captured by log changes in IIP from the previous year.

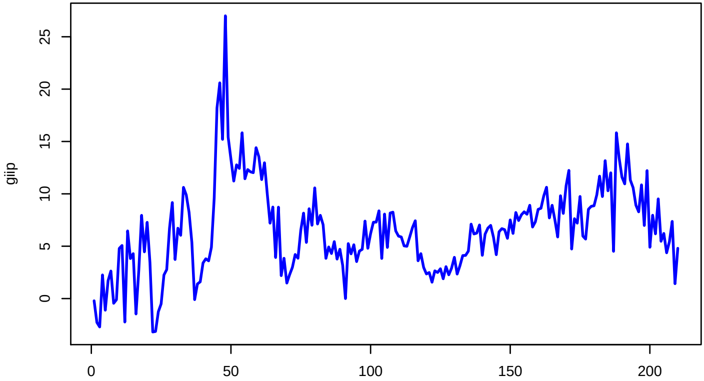
  $$y_t = log(IIP_t, y_1) - log(IIP_t, y_0)$$

  This is a standard data transformation used in the econometric literature for seasonally adjusting macro-economic data.
- Model suggestions:
  - There is a trend.
  - There is seasonality.
- Model 2: log $IIP_{t,y1}/IIP_{t,y0} = \alpha_0 + \alpha_1 Y_t + \beta_1 Jan_t + \beta_2 Feb_t + \ldots + \beta_{11} Nov_t + \epsilon_t$

# IIP – YoY growth series



# IIP – YoY growth series



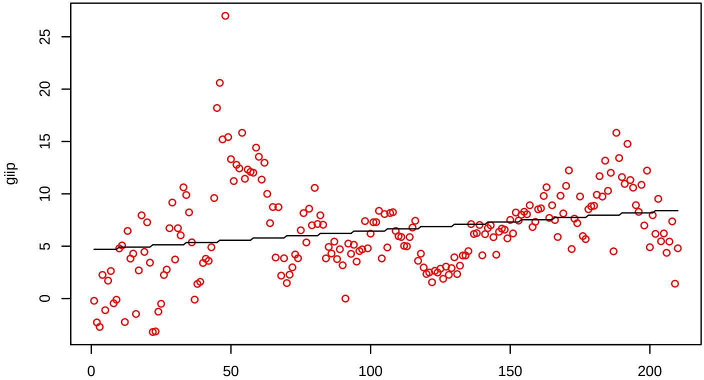
# Simple model 3

- Regression results:

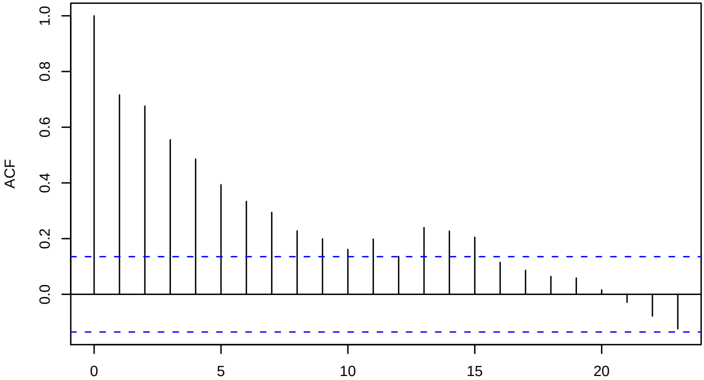
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2634 | 0.6553 | 6.51 | 0.0000 |
| gyear | 0.2176 | 0.0562 | 3.87 | 0.0001 |

Residual SE = 4.124
F-stat(1, 208) = 14.98
prob value = 1.45e-4
R-squared = 0.06718
Adjusted R-squared: 0.06269

# Explained vs. Actual data



# Residual data

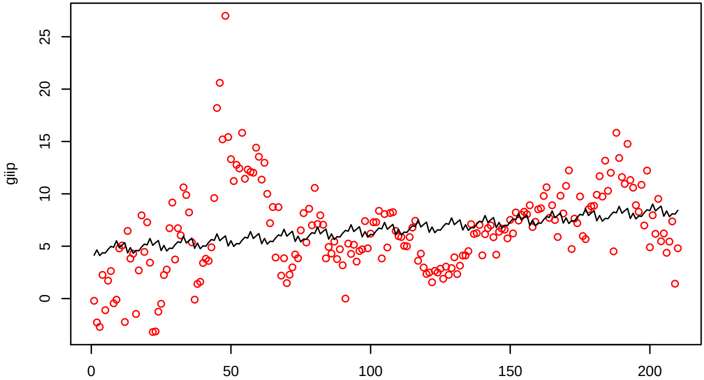


# Full model 3

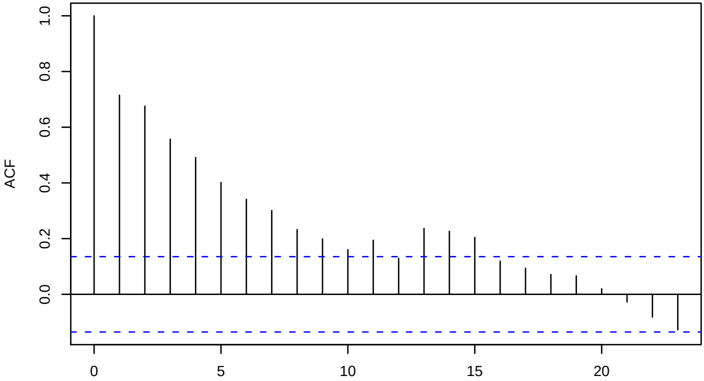
- Regression results:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2099 | 0.9417 | 4.47 | 0.0000 |
| year | 0.2188 | 0.0575 | 3.81 | 0.0002 |
| jan | 0.2393 | 1.2437 | 0.19 | 0.8476 |
| feb | 0.4629 | 1.2437 | 0.37 | 0.7102 |
| mar | -0.5066 | 1.2202 | -0.42 | 0.6785 |
| may | -0.5438 | 1.2202 | -0.45 | 0.6563 |
| jun | -0.2688 | 1.2202 | -0.22 | 0.8259 |
| jul | -0.3038 | 1.2202 | -0.25 | 0.8036 |
| aug | 0.0545 | 1.2202 | 0.04 | 0.9644 |
| sep | 0.3346 | 1.2443 | 0.27 | 0.7883 |
| oct | 0.2540 | 1.2443 | 0.20 | 0.8385 |
| nov | 0.8752 | 1.2443 | 0.70 | 0.4827 |

Residual SE = 4.207
F-stat(11, 198) = 1.479
prob value = 0.1415
R-squared = 0.0759
Adjusted R-squared: 0.0246


giip
0
5
10
15
20
25
0
50
100
150
200

# Residual data



# Model 4: an autoregressive model for IIP

- Time series models use information from previous periods of own data to explain the next.
- Example, an autoregressive model for IIP would take the form:

  $$IIP_t = \alpha + \beta_1 IIP_{t-1} + \epsilon_t$$

  This is called the Autoregressive model (AR) of order 1 because it has only one previous period variable as the explanatory variable for $IIP_t$.
- More generic forms of AR models are:

  $$IIP_t = \alpha + \beta_1 IIP_{t-1} + \ldots + \beta_k IIP_{t-k} + \epsilon_t$$

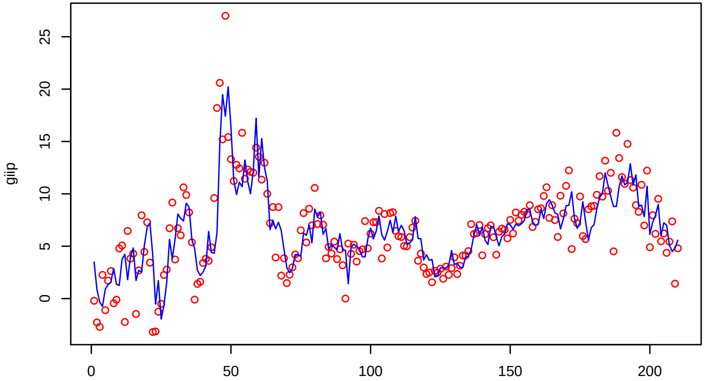
  This is an AR(k) model with IIP from “k” previous periods to explain $IIP_t$.

# Model 4: an autoregressive model for IIP

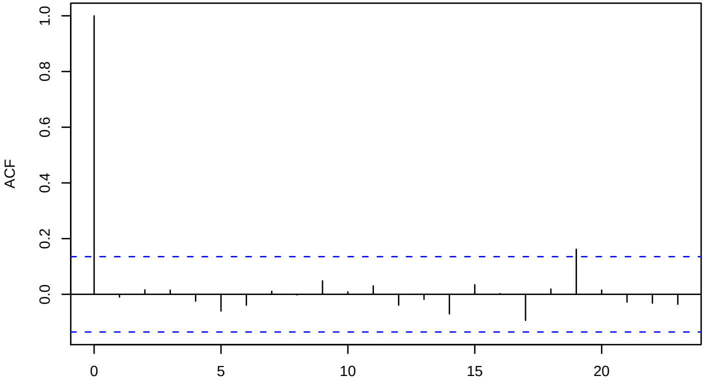
- Model for yoy-growth in IIP:

$$
\begin{aligned}
giip_t &= 6.2819 + 0.5174 giip_{t-1} + 0.3524 giip_{t-2} + 0.0217 giip_{t-3} \\
&\quad -0.0052 giip_{t-4} - 0.0125 giip_{t-5} - 0.0351 giip_{t-6} + 0.0452 giip_{t-7} \\
&\quad -0.0499 giip_{t-8} - 0.0318 giip_{t-9} - 0.0291 giip_{t-10} + 0.1101 giip_{t-11} \\
&\quad -0.2506 giip_{t-12} + 0.2932 giip_{t-13} + 0.1812 giip_{t-14} - 0.0402 giip_{t-15} \\
&\quad -0.2257 giip_{t-16} + \epsilon_t \\
\sigma_\epsilon &= 2.4385 \\
\sigma_{giip} &= 4.2594
\end{aligned}
$$

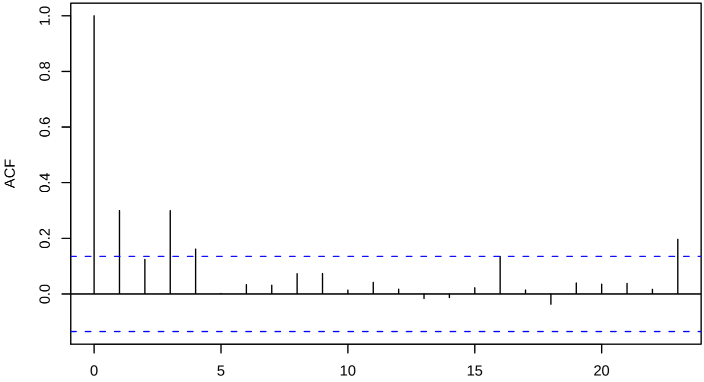
# Explained vs. Actual data



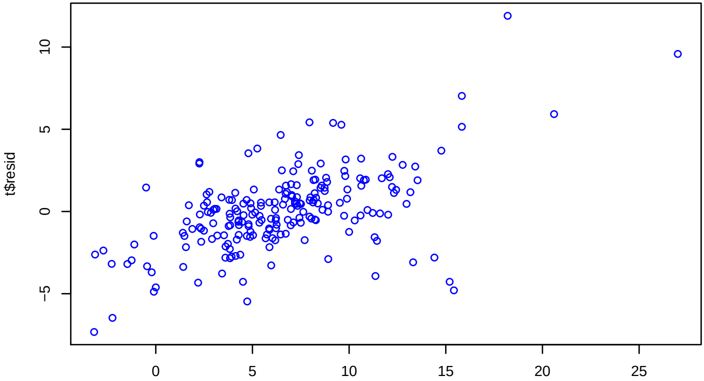
# Dependence in residual data



# Dependence in variance of residual data



# Cross-plot of giip vs. residuals



# Cross-plot of giip vs. residuals-squared