ØKA201 Data analysis and programming for economists

Time series (BKK ch. 9, GK ch. 9-12)

Bjørnar Karlsen Kivedal
bjornark@hiof.no

March 21, 2024

Plan for today

1. Problem set 5 (non-linear regression)
2. Time series
   - What is time series?
   - Transforming variables
   - Autocorrelation and Autoregressive models
   - Seasonal effects
   - Forecasting
   - Testing for stationarity
   - Autocorrelation in the residuals

Regression with time series

In the simple regression model, we had

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i was the dependent variable, and X_i was the independent variable. This is a model for cross-sectional data (data collected at a single point in time) measured, for example, for each individual i.

Regression with time series

In the simple regression model, we had

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i was the dependent variable, and X_i was the independent variable. This is a model for cross-sectional data (data collected at a single point in time) measured, for example, for each individual i.

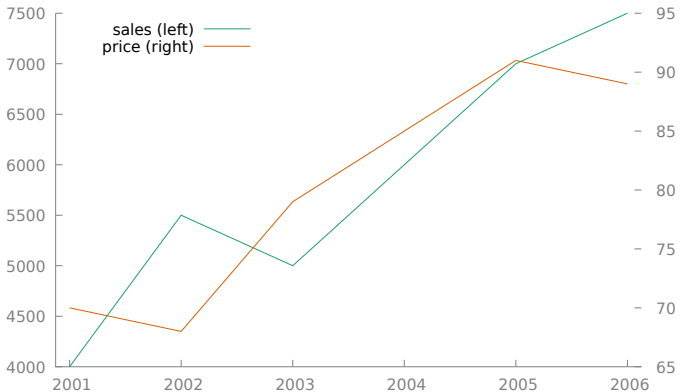
If we have time series data, we use **subscript** t since we have data for period t and not for an individual or entity i. We then have

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

For example, Y_t can be the quantity of a product sold in year t, and X_t is the price of the product in year t.

Regression with time series

Year	t	sales	price
2001	1	4000	70
2002	2	5500	68
2003	3	5000	79
2004	4	6000	85
2005	5	7000	91
2006	6	7500	89



Spurious regression

High R^2 and significant effect.
But it may just be by coincidence.

Trend as an explanatory variable

We can add the trend variable as an explanatory variable in the regression model to estimate the model:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + u_t$$

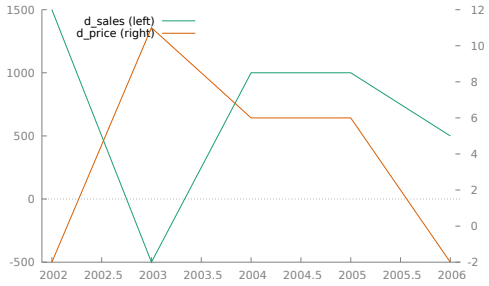
$\hat{\beta}_2$ can be interpreted as the expected increase in sales from one year to the next, given that the price does not change.
$\hat{\beta}_1$ can tell us what a price increase of 1 within a given year is expected to have on sales.

First differences

$$\Delta Y_t = Y_t - Y_{t-1}$$

$$dsales_t = sales_t - sales_{t-1}$$

Year	t	sales	price	dsales	dprice
2001	1	4 000	70		
2002	2	5 500	67	1 500	−3
2003	3	5 000	79	−500	12
2004	4	6 000	85	1 000	6
2005	5	7 000	91	1 000	6
2006	6	7 500	89	500	−2



Growth

Percentage growth in a variable for period t is measured as

$$X_{\text{growth}} = \frac{X_t - X_{t-1}}{X_{t-1}} \cdot 100$$

Often approximated with (if the change is small)

$$X_{\text{growth}} = \frac{X_t - X_{t-1}}{X_{t-1}} \cdot 100 \approx (\ln X_t - \ln X_{t-1}) \cdot 100$$

Seasonal variation

Year	t	sales	lsales	salesgrowth	dlsales
2001	1	4 000	8.29		
2002	2	5 500	8.61	37.50	31.85
2003	3	5 000	8.52	−9.09	−9.53
2004	4	6 000	8.70	20.00	18.23
2005	5	7 000	8.85	16.67	15.42
2006	6	7 500	8.92	7.14	6.90

Lags

$$X_t,\ X_{t-1}, X_{t-2}$$

Year	t	Sales	Sales_1	Sales_2
2001	1	4 000		
2002	2	5 500	4 000	
2003	3	5 000	5 500	4 000
2004	4	6 000	5 000	5 500
2005	5	7 000	6 000	5 000
2006	6	7 500	7 000	6 000

Autocorrelation

Autocorrelation is present when a variable is correlated with its past values, meaning that Y_t depends on Y_{t-1}. This is called **first-order autocorrelation**. Additionally, we can have situations where Y_t depends on a value of itself p periods back in time, Y_{t-p}, which is autocorrelation of order p.

Autoregressive model

An autoregressive model of the first order (an AR(1) model) is a model in which a variable depends on its value one period earlier. In other words, we can express it as follows:

$$X_t = \beta_0 + \beta_1 X_{t-1} + u_t$$

We can also include multiple lagged terms in an autoregressive model. This results in an AR(p) model, where X depends on X p periods earlier. In other words, we have the model:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + u_t$$

Seasonal variation

We estimate (**new data set**)

$$sales_t = \beta_0 + \beta_1 sales_{t-1} + \gamma t + u_t$$

	sales	fitted	residual
2000:2	6000.00	7237.37	−1237.37
2000:3	12000.00	8800.67	3199.33
2000:4	8000.00	8956.57	−956.57
2001:1	10000.00	12630.98	−2630.98
2001:2	13000.00	14194.28	−1194.28
2001:3	20000.00	15405.72	4594.28
2001:4	15000.00	15209.76	−209.76
2002:1	17000.00	19236.03	−2236.03
2002:2	20000.00	20799.33	−799.33
2002:3	25000.00	22010.77	2989.23
2002:4	21000.00	22518.52	−1518.52

Do you see a seasonal pattern?

Forecasting

Consider the three following models:

An AR(1) model:

$$\widehat{\text{sales}}_t = 5347.7 + 0.72\text{sales}_{t-1} \tag{1}$$

An AR(1) model with drift:

$$\widehat{\text{sales}}_t = 4110.8 - 0.35\text{sales}_{t-1} + 2267.0 \cdot t \tag{2}$$

An AR(1) model with drift and a dummy variable for 3rd quarter:

$$\widehat{\text{sales}}_t = 2573.7 - 0.19\text{sales}_{t-1} + 1967.3 \cdot t + 5086.3 \cdot dq3_t \tag{3}$$

Forecasting

We do not have observations for the first quarter of 2003, but we can create forecasts for the first quarter of 2001 using each of the three models. Here, we have $t = 13$, $\text{sales}_{t-1} = \text{sales}_{12} = 21,000$, and $dq3_{13} = 0$. We then get:

$$\widehat{\text{sales}}_{13} = 5347.7 + 0.72\text{sales}_{12}$$
$$\widehat{\text{sales}}_{13} = 4110.8 - 0.35\text{sales}_{12} + 2267.0 \cdot 13$$
$$\widehat{\text{sales}}_{13} = 2573.7 - 0.19\text{sales}_{12} + 1967.3 \cdot 13 + 5086.3 \cdot 0$$

which results in:

$$\widehat{\text{sales}}_{13} = 5347.7 + 0.72 \cdot 21,000 = 20,468$$
$$\widehat{\text{sales}}_{13} = 4110.8 - 0.35 \cdot 21,000 + 2267.0 \cdot 13 = 26,232$$
$$\widehat{\text{sales}}_{13} = 2573.7 - 0.19 \cdot 21,000 + 1967.3 \cdot 13 + 0 = 24,159$$

Forecasting

$$\widehat{\text{sales}}_{13} = 5347.7 + 0.72 \cdot 21,000 = 20,468$$
$$\widehat{\text{sales}}_{13} = 4110.8 - 0.35 \cdot 21,000 + 2267.0 \cdot 13 = 26,232$$
$$\widehat{\text{sales}}_{13} = 2573.7 - 0.19 \cdot 21,000 + 1967.3 \cdot 13 + 0 = 24,159$$

But what is the best forecast?
Hard to say because we have done **out-of sample forecasting**

Forecasting

We instead need to do **in-sample forecasting**
For the AR(1) model, we get, a predicted value for the second quarter of 2000 ($t = 2$):

$$\widehat{\text{sales}}_2 = 5347.7 + 0.72\text{sales}_1 = 5347.7 + 0.72 \cdot 6000 = 8227.7$$

In-sample Forecasting

t	Sales	(1)	(2)	(3)
1	4 000			
2	6 000	8 232.36	7 237.37	5760.27
3	12 000	9 674.70	8 800.67	12 439.90
4	8 000	14 001.72	8 956.57	8 198.75
5	10 000	11 117.04	12 630.98	10 914.12
6	13 000	12 559.38	14 194.28	12 507.39
7	20 000	14 722.89	15 405.72	19 000.00
8	15 000	19 771.08	15 209.76	14 571.83
9	17 000	16 165.23	19 236.03	17 474.21
10	20 000	17 607.57	20 799.33	19 067.48
11	25 000	19 771.08	22 010.77	25 560.10
12	21 000	23 376.94	22 518.52	21 505.96

In-sample Forecasting, residuals ($\hat{u}_t$)

t	(1)	(2)	(3)
2	−2 232.36	−1 237.37	239.73
3	2 325.3	3 199.33	−439.90
4	−6 001.72	−956.57	−198.75
5	−1 117.04	−2 630.98	−914.12
6	440.62	−1 194.28	492.61
7	5 277.11	4 594.28	1000.00
8	−4 771.08	−209.76	428.17
9	834.77	−2 236.03	−474.21
10	2 392.43	−799.33	932.52
11	5 228.92	2 989.23	−560.10
12	−2 376.94	−1 518.52	−505.96

Forecast evaluation

Root Mean Squared Error (TT is the number of predictions made (here: 11)):

$$RMSE = \sqrt{\frac{1}{TT}\sum \hat{u}_t^2}$$

Mean Absolute Deviation (MAD):

$$MAE = \frac{1}{TT}\sum |\hat{u}_t|$$

Mean Percentage Error (MPE):

$$MPE = \frac{1}{TT}\sum \frac{\hat{u}_t}{Y_t} \cdot 100$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{TT}\sum \left|\frac{\hat{u}_t}{Y_t}\right| \cdot 100$$

Should be as small as possible (in absolute value)

Forecast evaluation of AR(1) model

t	sales	$\hat{u}$	$\hat{u}^2$	$\lvert\hat{u}\rvert$	$\hat{u}$/sales	$\lvert\hat{u}/\text{sales}\rvert$
2	6 000	−2 232.4	4 983 431.2	2 232.4	−37.2	37.2
3	12 000	2 325.3	5 407 020.1	2 325.3	19.4	19.4
4	8 000	−6 001.7	36 020 643.0	6 001.7	−75.0	75.0
5	10 000	−1 117.0	1 247 778.4	1 117.0	−11.2	11.2
6	13 000	440.6	194 146.0	440.6	3.4	3.4
7	20 000	5 277.1	27 847 890.0	5 277.1	26.4	26.4
8	15 000	−4 771.1	22 763 204.4	4 771.1	−31.8	31.8
9	17 000	834.8	696 841.0	834.8	4.9	4.9
10	20 000	2 392.4	5 723 721.3	2 392.4	12.0	12.0
11	25 000	5 228.9	27 341 604.4	5 228.9	20.9	20.9
12	21 000	−2 376.9	5 649 843.8	2 376.9	−11.3	11.3
Sum		0.0	137 876 123.3	32 998.3	−79.6	253.5
Mean			12 534 193.0	2 999.8	−7.2	23.0

Forecast evaluation

AR(1) model:

$$RMSE = \sqrt{12534193.0} = 3540.4$$
$$MAE = 2999.8$$
$$MPE = -7.2$$
$$MAPE = 23.0$$

AR(1) model with drift:

$$RMSE = 2317.2$$
$$MAE = 1960.5$$
$$MPE = -2.9$$
$$MAPE = 14.0$$

AR(1) model with drift and a dummy variable for the third quarter:

$$RMSE = 619.11$$
$$MAE = 562.37$$
$$MPE = -0.22$$
$$MAPE = 3.91$$

Test for stationarity

To test whether a variable is stationary or not, we can use a **Dickey-Fuller test**.
If we have an AR(1) model for Y_t as follows:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

Y_t is non-stationary if $\beta_1 \geq 1$ or $\beta_1 \leq -1$.
However, this cannot be formally tested using the t-distribution because, under the null hypothesis $H_0 : \beta_1 = 1$, we have non-stationarity.

Test for stationarity

To make it possible to test for both a coefficient equal to -1 and 1, we need to rewrite the AR(1) model slightly. By subtracting Y_{t-1} from both sides of the equation in the AR(1) model, we get:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + u_t \\ Y_t - Y_{t-1} &= \beta_0 + \beta_1 Y_{t-1} - Y_{t-1} + u_t \\ \Delta Y_t &= \beta_0 + (\beta_1 - 1) Y_{t-1} + u_t \\ \Delta Y_t &= \beta_0 + \delta Y_{t-1} + u_t \end{aligned}$$

where $\Delta Y_t = Y_t - Y_{t-1}$ and $\delta = \beta_1 - 1$. We now want to test the hypotheses:

$$\begin{aligned} H_0 &: \delta = 0 \ (Y_t \text{ is non-stationary}) \\ H_A &: \delta \neq 0 \ (Y_t \text{ is stationary}) \end{aligned}$$

Test for stationarity

However, it is also common to extend it with lagged variables of ΔY_t, resulting in:

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_p \Delta Y_{t-p} + u_t$$

This model is used for conducting an **Augmented Dickey-Fuller (ADF)** test. The number of lagged variables (p) can be determined based on the type of data or using information criteria.
We can also include a trend in the model, allowing us to test whether the variable Y_t is **trend-stationary**. This means that the variable is stationary around a deterministic trend, such as a trend variable trend $= t$.

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_p + \gamma_t t + \Delta Y_{t-p} + u_t$$

Test for autocorrelated residuals

- The error term must be both stationary and non-autocorrelated.
- This means that the correlation between u_t and the errors from previous periods must be zero, i.e., $Corr(u_t, u_{t-1}) = 0$, $Corr(u_t, u_{t-2}) = 0$, ...
- Autocorrelation can often be addressed by introducing more dynamics into the model, such as including additional lagged values of the dependent variable or other explanatory variables.

Test for autocorrelated residuals

$$\hat{u}_t = \beta_0 + \gamma_1 Y_{t-1} + \cdots + \gamma_p Y_{t-p} + \beta_1 \hat{u}_{t-1} + \beta_2 \hat{u}_{t-2} + \cdots + \beta_3 \hat{u}_{t-q} + \varepsilon_t$$