MONASH University

MONASH BUSINESS SCHOOL

ETF3231/5231
Business forecasting

* We are interested in using these for forecasting.

Ch7. Regression models

* Pure time series models
v
models with predictors

https://bf.numbat.space/

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION AMBA ACCREDITED

Outline

1 The linear model with time series

2 Some useful predictors for linear models

3 Residual diagnostics

4 Selecting predictors and forecast evaluation

5 Forecasting with regression

6 Matrix formulation

7 Correlation, causation and forecasting

Outline

1 The linear model with time series

2 Some useful predictors for linear models

3 Residual diagnostics

4 Selecting predictors and forecast evaluation

5 Forecasting with regression

6 Matrix formulation

7 Correlation, causation and forecasting

REVISION

Multiple regression and forecasting

- response
- dependent
- regressand
- outcome

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t.$$

- predictors
- Indep. variables
- regressors
- explanatory

- y_t is the variable we want to predict: the "response" variable
- Each $x_{j,t}$ is numerical and is called a "predictor". They are usually assumed to be known for all past and future times.
  not always the case.
- 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.

That is, the coefficients measure the marginal effects.

- ε_t is a white noise error term

Example: US consumption expenditure

has special features over lm()

```
fit_cons <- us_change %>%
  model(lm = TSLM(Consumption ~ Income))
report(fit_cons)
```

$E(y_t / x_t) = \beta_0 + \beta_1 x_t$ Economic model

$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ Statistical model

$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$

$= b_0 + b_1 x_t$ Estimated model

$= 0.54 + 0.27 x_t$

```
## Series: Consumption
## Model: TSLM
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.582 -0.278   0.019   0.323   1.422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5445     0.0540   10.08  < 2e-16 ***
## Income        0.2718     0.0467    5.82  2.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.591 on 196 degrees of freedom
## Multiple R-squared: 0.147,   Adjusted R-squared: 0.143
## F-statistic: 33.8 on 1 and 196 DF, p-value: 2e-08
```

Always included

Example: US consumption expenditure

```
fit_consMR <- us_change %>%
  model(lm = TSLM(Consumption ~ Income + Production + Savings + Unemployment))
report(fit_consMR)
```

y_t X_{1t} X_{2t} X_{3t} X_{4t}

* Intercept always included unless

$y \sim 0 + \ldots$

```
## Series: Consumption
## Model: TSLM
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -0.906 -0.158 -0.036  0.136  1.155
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.25311    0.03447    7.34  5.7e-12 ***
## Income        0.74058    0.04012   18.46  < 2e-16 ***
## Production    0.04717    0.02314    2.04    0.043 *
## Savings      -0.05289    0.00292  -18.09  < 2e-16 ***
## Unemployment -0.17469    0.09551   -1.83    0.069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.31 on 193 degrees of freedom
## Multiple R-squared: 0.768,   Adjusted R-squared: 0.763
## F-statistic:  160 on 4 and 193 DF, p-value: <2e-16
```

b_0 b_1 b_2 b_3 b_4

Compare to 2-way correlations (switch to R & show these)

◦ Forecasting v Inference

(we don't care about p-values)

Outline

1 The linear model with time series

2 Some useful predictors for linear models

3 Residual diagnostics

4 Selecting predictors and forecast evaluation

5 Forecasting with regression

6 Matrix formulation

7 Correlation, causation and forecasting

+ these are not always observed but you can create some that are useful.

Trend

Linear trend

$$x_t = t$$

$y_t = \beta_0 + \beta_1 t + \varepsilon_t$

- $t = 1, 2, \ldots, T$
- Strong assumption that trend will continue.

→ very strong assumption
→ possibly ok for short-term

Nonlinear trend

Piecewise linear trend with bend “knot” at τ

$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_t$

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

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

- β_1 trend slope before time τ
- $\beta_1 + \beta_2$ trend slope after time τ
- More knots can be added forming more $(t - \tau)_+$

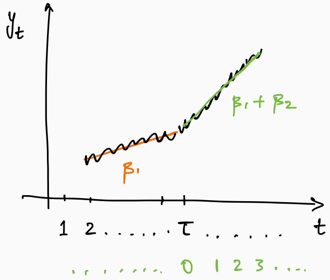
Nonlinear trend

Piecewise linear trend with bend "knot" at τ

$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_t$

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

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



- β_1 trend slope before time τ
- $\beta_1 + \beta_2$ trend slope after time τ
- More knots can be added forming more $(t-\tau)_+$

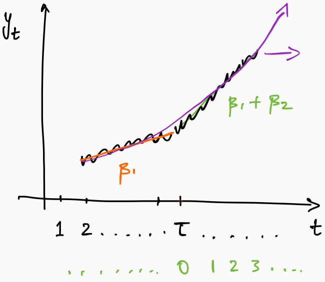
Nonlinear trend

Piecewise linear trend with bend “knot” at τ

$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_t$

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

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



- β_1 trend slope before time τ
- $\beta_1 + \beta_2$ trend slope after time τ
- More knots can be added forming more $(t-\tau)_+$

Quadratic or higher order trend

$$x_{1,t} = t, \quad x_{2,t} = t^2, \quad \ldots$$

NOT RECOMMENDED!

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies why? P.T.O.
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?
  - not exactly 52 weeks (365/7 = 52.14)
  - 51 dummy variables ?

Outliers

- If there is an outlier, you can use a dummy variable to remove its effect.
  - Sydney Olympics (spike in bus travel the quarter after)

DUMMY VARIABLE TRAP

C	Yes	No	Y + N = C
1	1	0	1
1	1	0	1
1	0	1	1
⋮	⋮		⋮
1	1	0	1

Holidays

For monthly data

- Christmas: always in December so part of monthly seasonal effect
- Easter: use a dummy variable $v_t = 1$ if any part of Easter is in that month, $v_t = 0$ otherwise.
- Ramadan and Chinese new year similar.

For daily data

- If it is a public holiday, dummy=1, otherwise dummy=0.

Weekend v working day

Fourier series

Periodic seasonality can be handled using pairs of Fourier terms:

Set up these: $s_k(t) = \sin\left(\frac{2\pi kt}{m}\right) \qquad c_k(t) = \cos\left(\frac{2\pi kt}{m}\right)$ seasonal period

$$y_t = \underbrace{a}_{\text{int}} + \underbrace{bt}_{\text{trend}} + \sum_{k=1}^{K} \underbrace{[\alpha_k s_k(t) + \beta_k c_k(t)]}_{\text{these are included as pairs}} + \varepsilon_t$$

- Every periodic function can be approximated by sums of sin and cos terms for large enough K. $(K \le \frac{m}{2})$
- Choose K by minimizing AICc.
- Called "harmonic regression" → as K increases we get harmonics of the first two Fourier terms.

```
TSLM(y ~ trend() + fourier(K))
```

specify K

* particularly useful for large m
  - hourly data m=24, weekly data m≃52

General form

for $k = 1$ $\quad s_1(t) = \sin\left(\frac{2\pi t}{m}\right)$ $\quad c_1(t) = \cos\left(\frac{2\pi t}{m}\right)$

$k = 2$ $\quad s_2(t) = \sin\left(\frac{2\pi 2t}{m}\right)$ $\quad c_2(t) = \cos\left(\frac{2\pi 2t}{m}\right)$

General form

for $k = 1$ $\quad s_1(t) = \sin\left(\frac{2\pi t}{m}\right)$ $\quad c_1(t) = \cos\left(\frac{2\pi t}{m}\right)$

Quarterly data

$m = 4$ $\quad = \sin\left(\frac{\pi}{2} t\right)$ $\quad = \cos\left(\frac{\pi}{2} t\right)$

$k = 2$ $\quad s_2(t) = \sin\left(\frac{2\pi 2t}{m}\right)$ $\quad c_2(t) = \cos\left(\frac{2\pi 2t}{m}\right)$

$= \underline{\sin(\pi t) = 0}$ $\quad = \cos(\pi t)$

true for $k = \frac{m}{2}$ in fact $\max(k) = \frac{m}{2}$ or $k \leq \frac{m}{2}$

Distributed lags

Lagged values of a predictor.

Example: x is advertising which has a delayed effect

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

Outline

1. The linear model with time series
2. Some useful predictors for linear models
3. Residual diagnostics — a couple of extra things
4. Selecting predictors and forecast evaluation
5. Forecasting with regression
6. Matrix formulation
7. Correlation, causation and forecasting

Multiple regression and forecasting

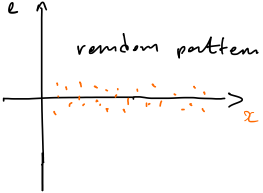
For forecasting purposes, we require the following assumptions:

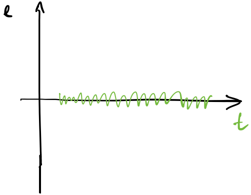
$\sum_{t=1}^{T} e_t = 0$ $\quad \sum_{t=1}^{T} x_{k,t} e_t = 0$

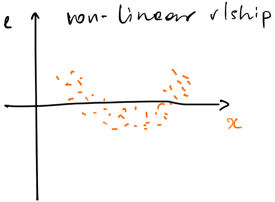
- ε_t are uncorrelated and zero mean
- Extra ε_t are uncorrelated with each $x_{j,t}$.

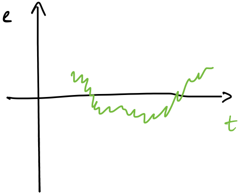
• normal equations as long as we have intercept. ε_t may not: endogeneity

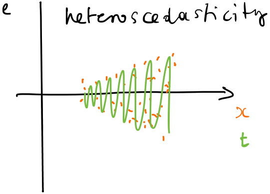
It is useful to also have $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$ when producing prediction intervals or doing statistical tests.

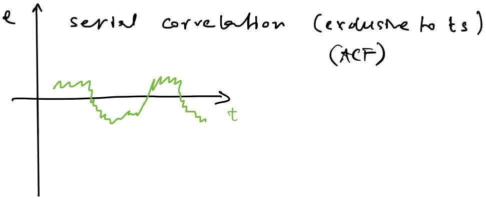

e
random pattern
x


e
t


e
non-linear rlship
x


e
t


e
heteroscedasticity
x
t


e
serial correlation (exclusive to ts)
(ACF)
t

Outline

1 The linear model with time series

2 Some useful predictors for linear models

3 Residual diagnostics

4 Selecting predictors and forecast evaluation

5 Forecasting with regression

6 Matrix formulation

7 Correlation, causation and forecasting

Comparing regression models

- 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.

Comparing regression models

- 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.

To overcome this problem, we can use *adjusted* R^2:

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

where k = no. predictors and T = no. observations.

Comparing regression models

- 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.

To overcome this problem, we can use *adjusted* R^2:

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

where k = no. predictors and T = no. observations.

Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.

estimated residual variance

$$\hat{\sigma}^2 = \frac{1}{T - k - 1}\sum_{t=1}^{T}\varepsilon_t^2$$

WE CAN DO BETTER

Akaike's Information Criterion

predictors

intercept + variance

$$\text{AIC} = -2\log(L) + 2(k+2)$$

- L = likelihood
- k = # predictors in model.
- AIC penalizes terms more heavily than $\bar{R}^2$. — smaller models

Akaike's Information Criterion

predictors

intercept + variance

$$\text{AIC} = -2\log(L) + 2(k+2)$$

- L = likelihood
- k = # predictors in model.
- AIC penalizes terms more heavily than $\bar{R}^2$. – smaller models

$$\text{AIC}_\text{C} = \text{AIC} + \frac{2(k+2)(k+3)}{T-k-3}$$

- Minimizing the AIC or AICc is asymptotically equivalent to minimizing MSE via **leave-one-out cross-validation** (for any linear regression).

will show soon, useful for prediction

Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+2)\log(T)$$

where L is the likelihood and k is the number of predictors in the model.

- BIC penalizes terms more heavily than AIC ← even smaller models if needed
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = T[1 - 1/(log(T) - 1)]$.

Leave-one-out cross-validation

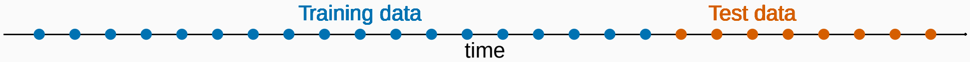
For regression, leave-one-out cross-validation is faster and more efficient than time-series cross-validation.

only with one regression
(will show this).

- Select one observation for test set, and use *remaining* observations in training set. Compute error on test observation.
- Repeat using each possible observation as the test set.
- Compute accuracy measure over all errors.

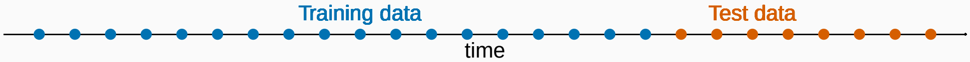
Cross-validation

Traditional evaluation

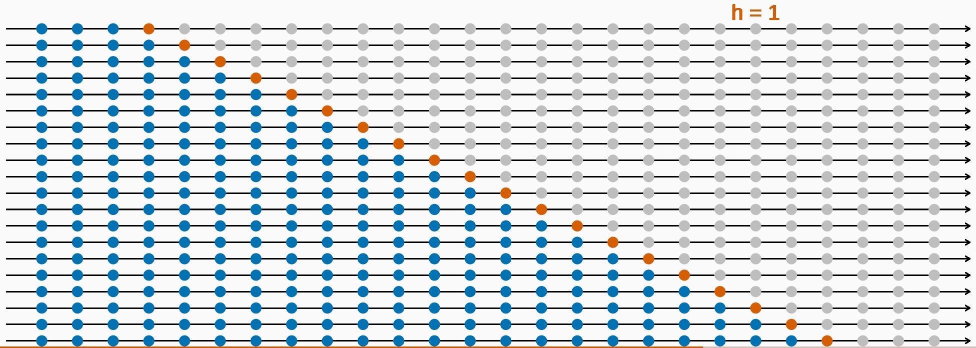


Cross-validation

Traditional evaluation

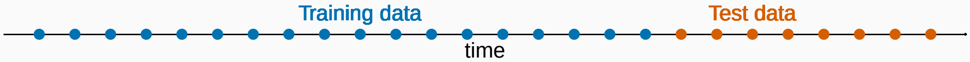


Time series cross-validation



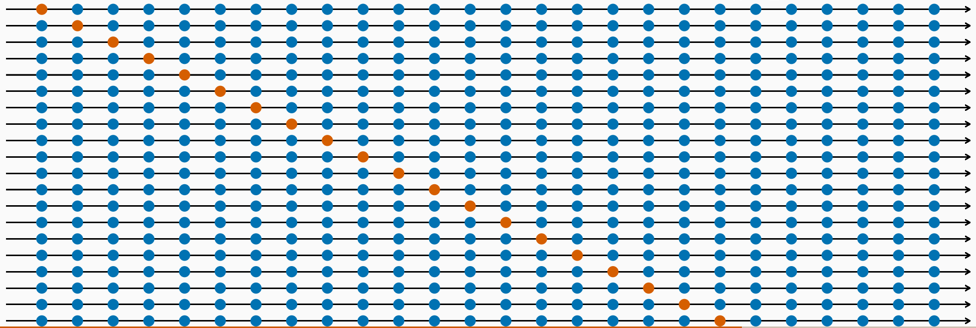
Cross-validation

Traditional evaluation



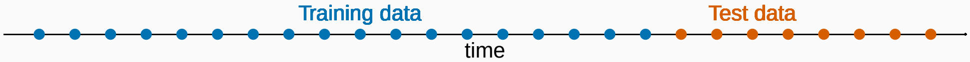
Leave-one-out cross-validation

• you are actually using some future information in forecasting.



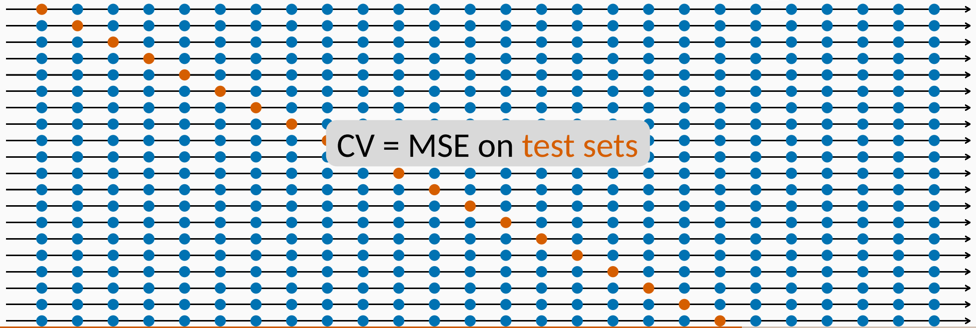
Cross-validation

Traditional evaluation



Leave-one-out cross-validation

* This is very fast - there is a trick



Choosing regression variables

Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on one of the measures of predictive ability (CV, AIC, AICc).

Choosing regression variables

Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on one of the measures of predictive ability (CV, AIC, AICc).

Warning!

- If there are a large number of predictors, this is not possible.
- For example, 44 predictors leads to 18 trillion possible models!

Choosing regression variables

Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV or AICc.
- Iterate until no further improvement.

Forwards stepwise regression

* useful when you cannot fit all variables $k > T$

- Start with a model containing only a constant.
- Add one variable at a time. Keep the model if it has lower CV or AICc.
- Iterate until no further improvement.

Hybrid backwards and forwards also possible.

- Stepwise regression is not guaranteed to lead to the best possible model.

What should you use?

Notes

- Stepwise regression is not guaranteed to lead to the best possible model.
- Inference on coefficients of final model will be wrong. * Common mistake

Choice: CV, AIC, AICc, BIC, $\bar{R}^2$

- BIC tends to choose models too small for prediction (however can be useful for large k).
- $\bar{R}^2$ tends to select models too large.
- AIC also slightly biased towards larger models (especially when T is small).
- Empirical studies in forecasting show AIC is better than BIC for forecast accuracy.

Choice between AICc and CV (double check AIC and BIC where possible).

Outline

1. The linear model with time series
2. Some useful predictors for linear models
3. Residual diagnostics
4. Selecting predictors and forecast evaluation
5. Forecasting with regression
6. Matrix formulation
7. Correlation, causation and forecasting

Ex-ante versus ex-post forecasts

- Ex ante forecasts are made using only information available in advance.
  - require forecasts of predictors * or future values are given to us.
- Ex post forecasts are made using later information on the predictors. * scenario based forecasting.
  - useful for studying behaviour of forecasting models.
- trend, seasonal and calendar variables are all known in advance, so these don't need to be forecast.

× In all cases prediction intervals are underestimated

Outline

1 The linear model with time series

2 Some useful predictors for linear models

3 Residual diagnostics

4 Selecting predictors and forecast evaluation

5 Forecasting with regression

6 Matrix formulation

7 Correlation, causation and forecasting

Multiple regression forecasts

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{k,1} \\ 1 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & \cdots & x_{k,T} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \underbrace{y}_{(T\times 1)} = \underbrace{X}_{(T\times k)}\underbrace{\beta}_{(k\times 1)} + \varepsilon$$

Fitted values

$$\underbrace{\hat{\boldsymbol{y}}}_{(T\times 1)} = \underbrace{\boldsymbol{X}}_{(T\times k)}\underbrace{\hat{\boldsymbol{\beta}}}_{(k\times 1)} = \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{(T\times T)}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat matrix".

Residual $e = y - \hat{y} = Ry$
where $R = I - H$

Projection – project y onto X
– gives the best linear approx of y using X

Leave-one-out residuals

Let $h_1, \dots, h_T$ be the diagonal values of $\boldsymbol{H}$, then the cross-validation statistic is

$$\text{CV} = \frac{1}{T}\sum_{t=1}^{T}[e_t/(1-h_t)]^2,$$

where e_t is the residual obtained from fitting the model to all T observations.

- So you can calculate the CV from only fitting one model

Outline

1 The linear model with time series

2 Some useful predictors for linear models

3 Residual diagnostics

4 Selecting predictors and forecast evaluation

5 Forecasting with regression

6 Matrix formulation

7 Correlation, causation and forecasting

Correlation is not causation

correlation ⇏ causation ; causation ⇒ correlation BUT

- When x is useful for predicting y, it is not necessarily causing y.
- e.g., predict number of drownings y using number of ice-creams sold x. Cyclists on St Kilda Rd do not cause rain BUT . . . .
- Correlations are useful for forecasting, even when there is no causality.
- Better models usually involve causal relationships (e.g., temperature x and people z to predict drownings y). → but you can still forecast without causation.