MONASH BUSINESS SCHOOL

ETC2450

Applied forecasting for business and economics

5: Multiple regression

OTexts.org/fpp/5/

Outline

1 Some useful predictors for linear models

2 Residual diagnostics

3 Selecting predictors and forecast evaluation

4 Matrix formulation

5 Correlation, causation and forecasting

Multiple regression and forecasting

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

- 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.
- The coefficients $\beta_1, \dots, \beta_k$ 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**.
- e_t is a white noise error term

Multiple regression and forecasting

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

- 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.
- The coefficients $\beta_1, \dots, \beta_k$ 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**.
- e_t is a white noise error term

Multiple regression and forecasting

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

- 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.
- The coefficients $\beta_1, \dots, \beta_k$ 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**.
- e_t is a white noise error term

Multiple regression and forecasting

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

- 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.
- The coefficients $\beta_1, \dots, \beta_k$ 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**.
- e_t is a white noise error term

Dummy variables

If a categorical variable takes only two values (e.g., 'Yes' or 'No'), then an equivalent numerical variable can be constructed taking value 1 if yes and 0 if no. This is called a **dummy variable**.

	A	B
1	Yes	1
2	Yes	1
3	No	0
4	Yes	1
5	No	0
6	No	0
7	Yes	1
8	Yes	1
9	No	0
10	No	0
11	No	0
12	No	0
13	Yes	1
14	No	0

Dummy variables

If there are more than two categories, then the variable can be coded using several dummy variables (one fewer than the total number of categories).

	A	B	C	D	E
1	Monday	1	0	0	0
2	Tuesday	0	1	0	0
3	Wednesday	0	0	1	0
4	Thursday	0	0	0	1
5	Friday	0	0	0	0
6	Monday	1	0	0	0
7	Tuesday	0	1	0	0
8	Wednesday	0	0	1	0
9	Thursday	0	0	0	1
10	Friday	0	0	0	0
11	Monday	1	0	0	0
12	Tuesday	0	1	0	0
13	Wednesday	0	0	1	0
14	Thursday	0	0	0	1
15	Friday	0	0	0	0

Beware of the dummy variable trap!

- Using one dummy for each category gives too many dummy variables!
- The regression will then be singular and inestimable.
- Either omit the constant, or omit the dummy for one category.
- The coefficients of the dummies are relative to the omitted category.

Beware of the dummy variable trap!

- Using one dummy for each category gives too many dummy variables!
- The regression will then be singular and inestimable.
- Either omit the constant, or omit the dummy for one category.
- The coefficients of the dummies are relative to the omitted category.

Beware of the dummy variable trap!

- Using one dummy for each category gives too many dummy variables!
- The regression will then be singular and inestimable.
- Either omit the constant, or omit the dummy for one category.
- The coefficients of the dummies are relative to the omitted category.

Beware of the dummy variable trap!

- Using one dummy for each category gives too many dummy variables!
- The regression will then be singular and inestimable.
- Either omit the constant, or omit the dummy for one category.
- The coefficients of the dummies are relative to the omitted category.

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

Public holidays

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

Public holidays

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable (taking value 1 for that observation and 0 elsewhere) to remove its effect.

Public holidays

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable (taking value 1 for that observation and 0 elsewhere) to remove its effect.

Public holidays

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable (taking value 1 for that observation and 0 elsewhere) to remove its effect.

Public holidays

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable (taking value 1 for that observation and 0 elsewhere) to remove its effect.

Public holidays

- For daily data: if it is a public holiday, dummy=1, otherwise dummy=0.

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable (taking value 1 for that observation and 0 elsewhere) to remove its effect.

Public holidays

- For daily data: if it is a public holiday, dummy=1, otherwise dummy=0.

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable (taking value 1 for that observation and 0 elsewhere) to remove its effect.

Public holidays

- For daily data: if it is a public holiday, dummy=1, otherwise dummy=0.

Linear trend

$$x_t = t$$

Piecewise linear trend with bend at τ

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

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

Quadratic or higher order trend

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

NOT RECOMMENDED!

Trend

Linear trend

$$x_t = t$$

Piecewise linear trend with bend at τ

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

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

Quadratic or higher order trend

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

NOT RECOMMENDED!

Trend

Linear trend

$$x_t = t$$

Piecewise linear trend with bend at τ

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

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

Quadratic or higher order trend

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

NOT RECOMMENDED!

Trend

Linear trend

$$x_t = t$$

Piecewise linear trend with bend at τ

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

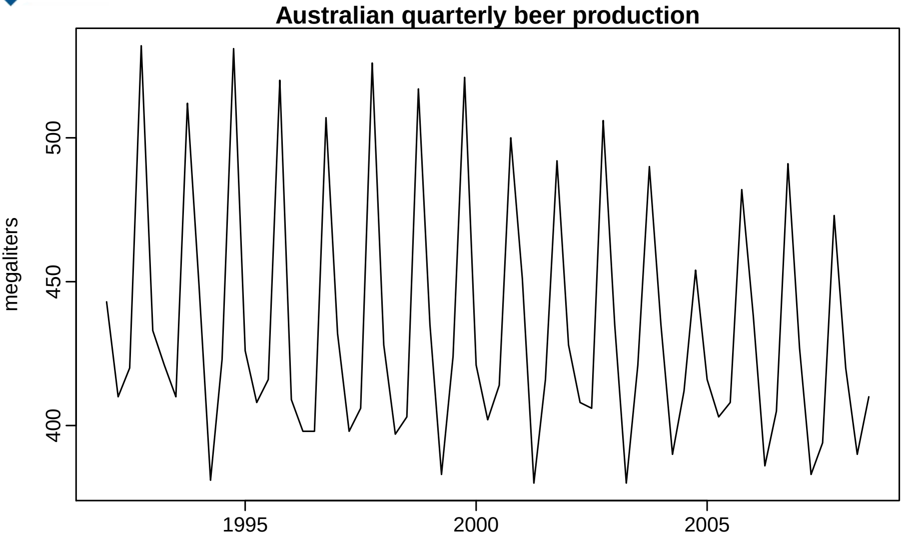
$$x_{2,t} = \begin{cases} 0 & t < \tau \\ (t - \tau) & t \geq \tau \end{cases}$$

Quadratic or higher order trend

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

NOT RECOMMENDED!

Beer production revisited



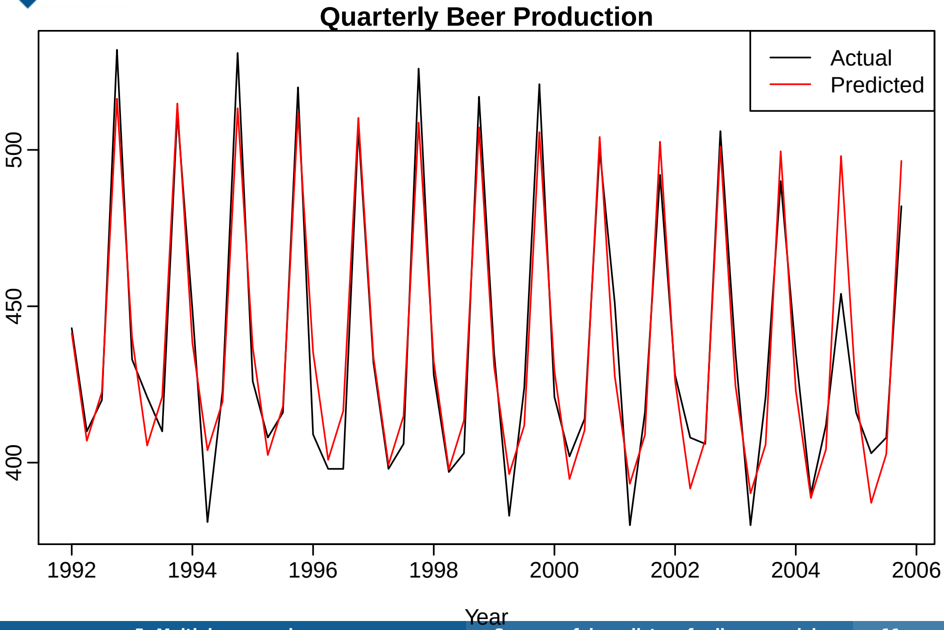
Beer production revisited

Regression model

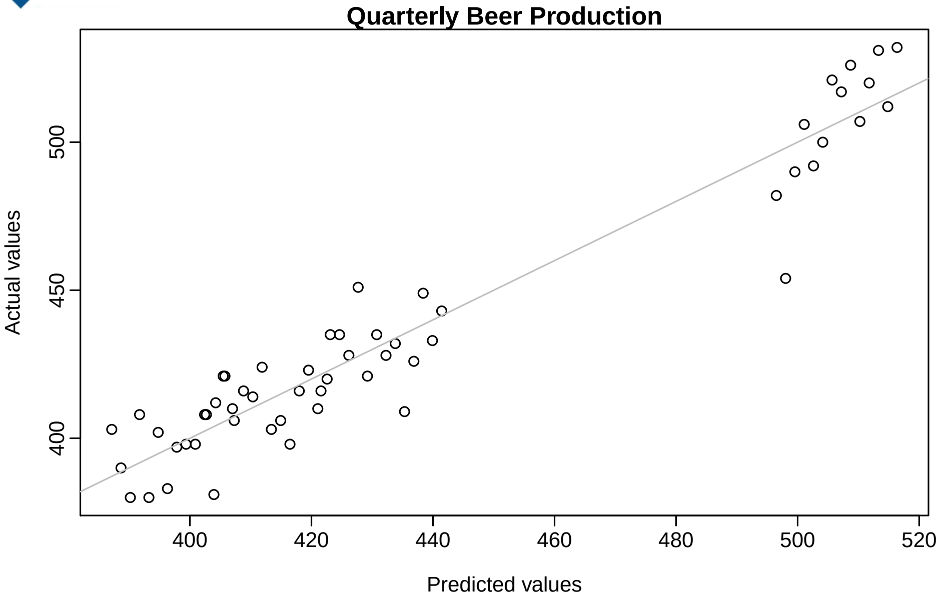
$$y_t = \beta_0 + \beta_1 t + \beta_2 d_{1,t} + \beta_3 d_{2,t} + \beta_4 d_{3,t} + e_t$$

- $d_{i,t} = 1$ if t is quarter i and 0 otherwise.

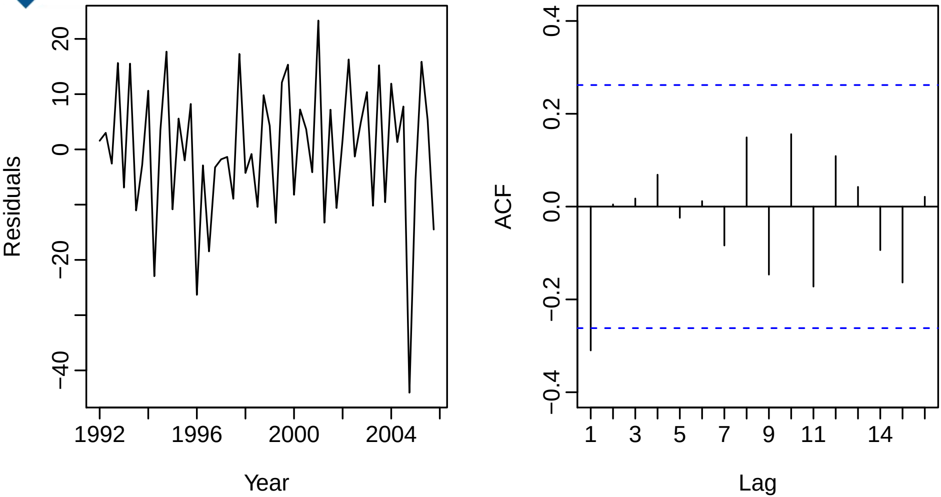
Beer production revisited



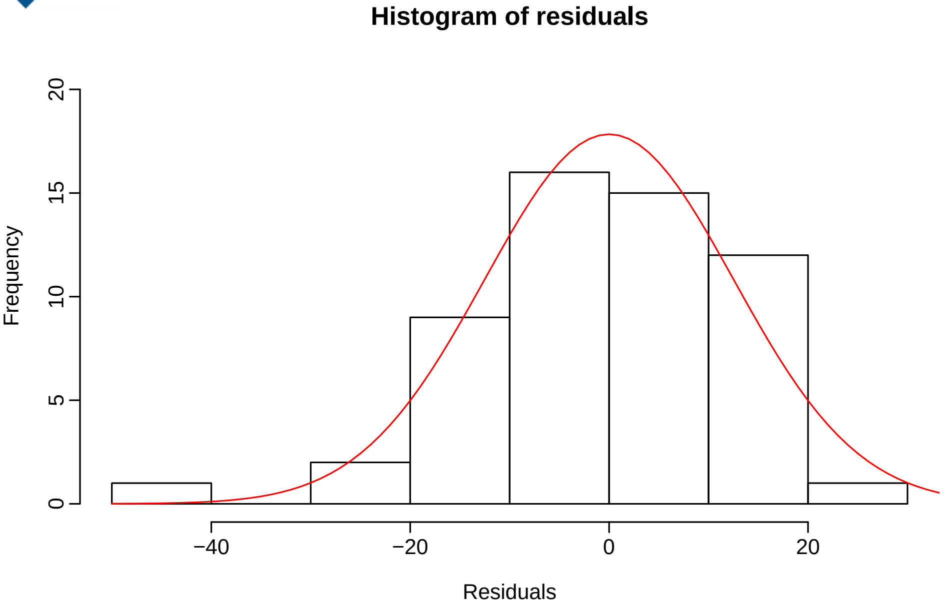
Beer production revisited



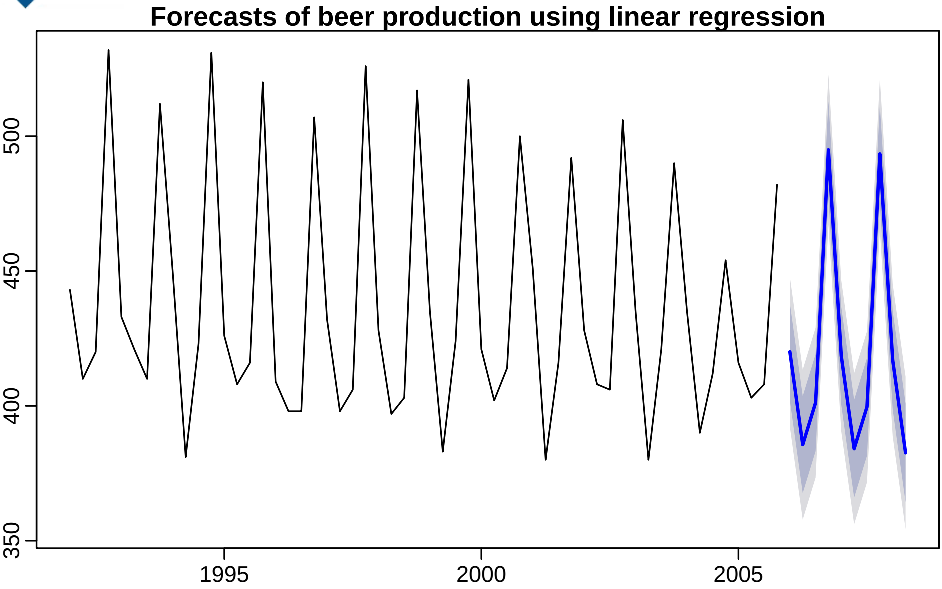
Beer production revisited



Beer production revisited



Beer production revisited



How to do this in R

```
beer2 <- window(ausbeer,start=1992,end=2006-.1)
plot(beer2, main="Quarterly Australian beer production",
 ylab="Megaliters", xlab="Year")

fit <- tslm(beer2 ~ trend + season)
plot(beer2, main="Quarterly Beer Production")
lines(fitted(fit),col=2)
legend("topright", lty=1, col=c(1,2),
  legend=c("Actual","Predicted"))

res <- residuals(fit)
par(mfrow=c(1,2))
plot(res,ylab="Residuals",xlab="Year")
Acf(res,main="ACF of residuals")
hist(res)

fcast <- forecast(fit)
plot(fcast, main="Forecasts of beer production")
```

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 afterwards.

Change of slope

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 afterwards.

Change of slope

- Variables take values 0 before the intervention and values $\{1, 2, 3, \dots\}$ afterwards.

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 afterwards.

Change of slope

- Variables take values 0 before the intervention and values $\{1, 2, 3, \dots\}$ afterwards.

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 afterwards.

Change of slope

- Variables take values 0 before the intervention and values $\{1, 2, 3, \dots\}$ afterwards.

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 afterwards.

Change of slope

- Variables take values 0 before the intervention and values $\{1, 2, 3, \ldots\}$ afterwards.

Holiday and trading day variations

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.
- Trading days:

$$\begin{aligned} z_1 &= \text{\# Mondays in month;} \\ z_2 &= \text{\# Tuesdays in month;} \\ &\vdots \\ z_7 &= \text{\# Sundays in month.} \end{aligned}$$

Holiday and trading day variations

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.
- Trading days:

$$z_1 = \#\text{ Mondays in month;}$$
$$z_2 = \#\text{ Tuesdays in month;}$$
$$\vdots$$
$$z_7 = \#\text{ Sundays in month.}$$

Holiday and trading day variations

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.
- Trading days:

$$\begin{aligned} z_1 &= \text{\# Mondays in month;} \\ z_2 &= \text{\# Tuesdays in month;} \\ &\vdots \\ z_7 &= \text{\# Sundays in month.} \end{aligned}$$

Fourier terms for seasonality

Periodic seasonality can be handled using pairs of Fourier terms:

$$s_k(t) = \sin\left(\frac{2\pi kt}{m}\right) \qquad c_k(t) = \cos\left(\frac{2\pi kt}{m}\right)$$

$$y_t = a + bt + \sum_{k=1}^{K} [\alpha_k s_k(t) + \beta_k c_k(t)] + e_t$$

- Every periodic function can be approximated by sums of sin and cos terms for large enough K.
- Choose K by minimizing AICc.

```
fit <- tslm(y ~ trend + fourier(y, K))
```

Fourier terms for seasonality

Periodic seasonality can be handled using pairs of Fourier terms:

$$s_k(t) = \sin\left(\frac{2\pi kt}{m}\right) \qquad c_k(t) = \cos\left(\frac{2\pi kt}{m}\right)$$

$$y_t = a + bt + \sum_{k=1}^{K}\left[\alpha_k s_k(t) + \beta_k c_k(t)\right] + e_t$$

- Every periodic function can be approximated by sums of sin and cos terms for large enough K.
- Choose K by minimizing AICc.

```
fit <- tslm(y ~ trend + fourier(y, K))
```

Fourier terms for seasonality

Periodic seasonality can be handled using pairs of Fourier terms:

$$s_k(t) = \sin\left(\frac{2\pi kt}{m}\right) \qquad c_k(t) = \cos\left(\frac{2\pi kt}{m}\right)$$

$$y_t = a + bt + \sum_{k=1}^{K} [\alpha_k s_k(t) + \beta_k c_k(t)] + e_t$$

- Every periodic function can be approximated by sums of sin and cos terms for large enough K.
- Choose K by minimizing AICc.

```
fit <- tslm(y ~ trend + fourier(y, K))
```

Fourier terms for seasonality

Periodic seasonality can be handled using pairs of Fourier terms:

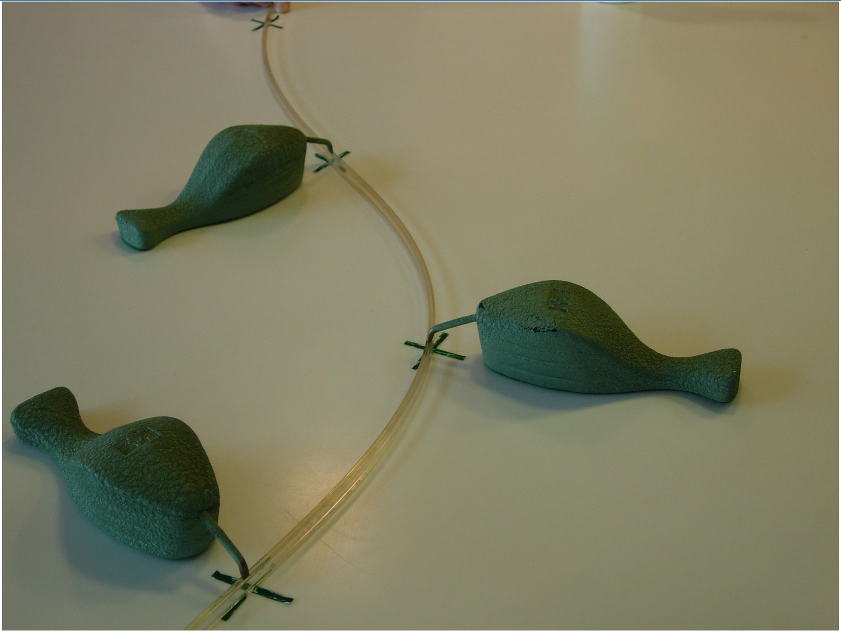
$$s_k(t) = \sin\left(\frac{2\pi k t}{m}\right) \qquad c_k(t) = \cos\left(\frac{2\pi k t}{m}\right)$$

$$y_t = a + bt + \sum_{k=1}^{K} \left[\alpha_k s_k(t) + \beta_k c_k(t)\right] + e_t$$

- Every periodic function can be approximated by sums of sin and cos terms for large enough K.
- Choose K by minimizing AICc.

```
fit <- tslm(y ~ trend + fourier(y, K))
```

Interpolating splines



Interpolating splines

Interpolating splines

Interpolating splines

Points: (κ_j, y_j) for $j = 1, \dots, K$.

A spline is a continuous function $f(x)$ interpolating all points and consisting of polynomials between each consecutive pair of 'knots' κ_j and κ_{j+1}.

- Parameters constrained so that $f(x)$ is continuous.
- Further constraints imposed to give continuous derivatives.

Interpolating splines

Points: (κ_j, y_j) for $j = 1, \dots, K$.

A spline is a continuous function $f(x)$ interpolating all points and consisting of polynomials between each consecutive pair of 'knots' κ_j and κ_{j+1}.

- Parameters constrained so that $f(x)$ is continuous.
- Further constraints imposed to give continuous derivatives.

Interpolating splines

Points: (κ_j, y_j) for $j = 1, \dots, K$.

A spline is a continuous function $f(x)$ interpolating all points and consisting of polynomials between each consecutive pair of 'knots' κ_j and κ_{j+1}.

- Parameters constrained so that $f(x)$ is continuous.
- Further constraints imposed to give continuous derivatives.

Interpolating splines

Points: (κ_j, y_j) for $j = 1, \dots, K$.

A spline is a continuous function $f(x)$ interpolating all points and consisting of polynomials between each consecutive pair of 'knots' κ_j and κ_{j+1}.

- Parameters constrained so that $f(x)$ is continuous.
- Further constraints imposed to give continuous derivatives.

General linear regression splines

- Let $\kappa_1 < \kappa_2 < \cdots < \kappa_K$ be "knots" in interval (a, b).
- Let $x_1 = x$, $x_j = (x - \kappa_{j-1})_+$ for $j = 2, \ldots, K+1$.
- Then the regression is piecewise linear with bends at the knots.

General linear regression splines

- Let $\kappa_1 < \kappa_2 < \cdots < \kappa_K$ be "knots" in interval (a, b).
- Let $x_1 = x$, $x_j = (x - \kappa_{j-1})_+$ for $j = 2, \dots, K + 1$.
- Then the regression is piecewise linear with bends at the knots.

General linear regression splines

- Let $\kappa_1 < \kappa_2 < \cdots < \kappa_K$ be "knots" in interval (a, b).
- Let $x_1 = x$, $x_j = (x - \kappa_{j-1})_+$ for $j = 2, \dots, K + 1$.
- Then the regression is piecewise linear with bends at the knots.

General cubic splines

- Let $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_j = (x - \kappa_{j-3})_+^3$ for $j = 4, \dots, K+3$.
- Then the regression is piecewise cubic, but smooth at the knots.
- Choice of knots can be difficult and arbitrary.
- Automatic knot selection algorithms very slow.

General cubic splines

- Let $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_j = (x - \kappa_{j-3})_+^3$ for $j = 4, \dots, K+3$.
- Then the regression is piecewise cubic, but smooth at the knots.
- Choice of knots can be difficult and arbitrary.
- Automatic knot selection algorithms very slow.

General cubic splines

- Let $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_j = (x - \kappa_{j-3})^3_+$ for $j = 4, \dots, K+3$.
- Then the regression is piecewise cubic, but smooth at the knots.
- Choice of knots can be difficult and arbitrary.
- Automatic knot selection algorithms very slow.

General cubic splines

- Let $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_j = (x - \kappa_{j-3})^3_+$ for $j = 4, \dots, K + 3$.
- Then the regression is piecewise cubic, but smooth at the knots.
- Choice of knots can be difficult and arbitrary.
- Automatic knot selection algorithms very slow.

Spline bases

Truncated power basis of degree p

$$1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p$$

- $p-1$ continuous derivatives
- In penalized regression splines, none of the polynomial coefficients is penalized.

Spline bases

Truncated power basis of degree p

$1, x, \dots, x^p, (x - \kappa_1)_+^p, \dots, (x - \kappa_K)_+^p$

- $p - 1$ continuous derivatives
- In penalized regression splines, none of the polynomial coefficients is penalized.

Spline bases

Truncated power basis of degree p

$1, x, \dots, x^p, (x-\kappa_1)_+^p, \dots, (x-\kappa_K)_+^p$

- $p-1$ continuous derivatives
- In penalized regression splines, none of the polynomial coefficients is penalized.

Spline bases

Truncated power basis of degree p

$1, x, \dots, x^p, (x-\kappa_1)_+^p, \dots, (x-\kappa_K)_+^p$

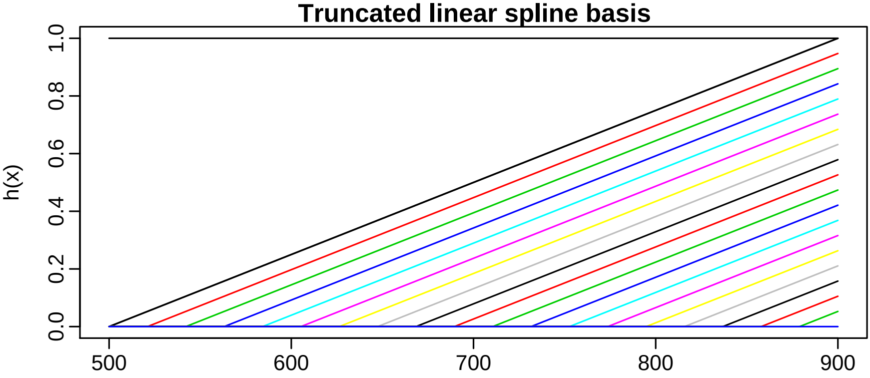
- $p-1$ continuous derivatives
- In penalized regression splines, none of the polynomial coefficients is penalized.

Spline bases

Truncated power basis of degree p

$1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p$

- $p-1$ continuous derivatives
- In penalized regression splines, none of the polynomial coefficients is penalized.

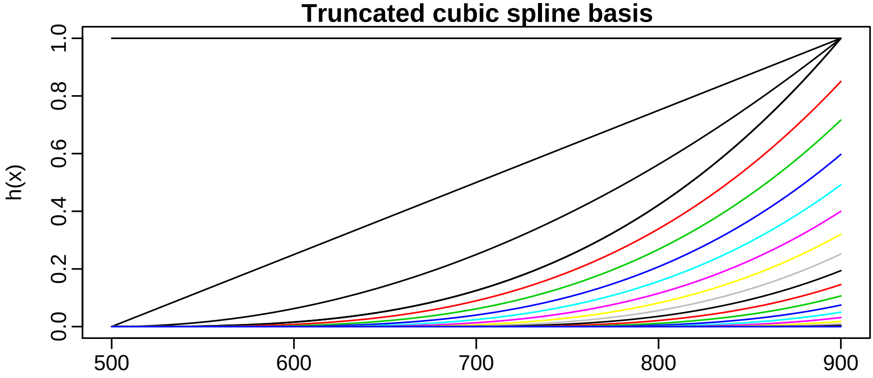


Spline bases

Truncated power basis of degree p

$1, x, \dots, x^p, (x-\kappa_1)_+^p, \dots, (x-\kappa_K)_+^p$

- $p-1$ continuous derivatives
- In penalized regression splines, none of the polynomial coefficients is penalized.



Outline

1 Some useful predictors for linear models

2 **Residual diagnostics**

3 Selecting predictors and forecast evaluation

4 Matrix formulation

5 Correlation, causation and forecasting

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

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

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

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

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

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

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

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

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

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

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

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

Residual plots

Useful for spotting outliers and whether the straight a linear model was appropriate.

- Scatterplot of residuals e_t against each predictor $x_{j,t}$.
- Scatterplot residuals against the fitted values $\hat{y}_t$
- Expect to see scatterplots resembling a horizontal band with no values too far from the band and no patterns such as curvature or increasing spread.

Residual plots

Useful for spotting outliers and whether the straight a linear model was appropriate.

- Scatterplot of residuals e_t against each predictor $x_{j,t}$.
- Scatterplot residuals against the fitted values $\hat{y}_t$
- Expect to see scatterplots resembling a horizontal band with no values too far from the band and no patterns such as curvature or increasing spread.

Residual plots

Useful for spotting outliers and whether the straight a linear model was appropriate.

- Scatterplot of residuals e_t against each predictor $x_{j,t}$.
- Scatterplot residuals against the fitted values $\hat{y}_t$
- Expect to see scatterplots resembling a horizontal band with no values too far from the band and no patterns such as curvature or increasing spread.

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Durbin-Watson test

If residuals from a linear model:

- DW tests hypothesis that there is no lag one autocorrelation present in the residuals.
- If there is no autocorrelation, the DW distribution is symmetric around 2, its mean value.
- R gives p-values.

Durbin-Watson test

If residuals from a linear model:

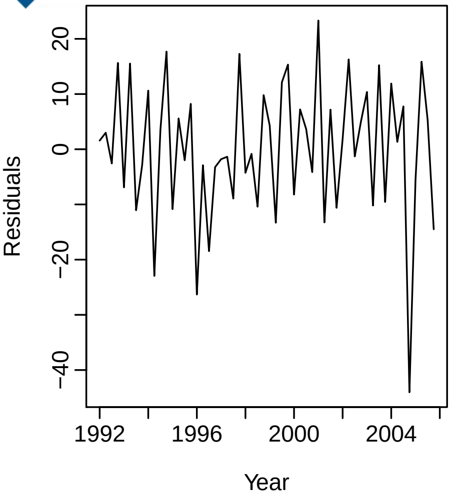
- DW tests hypothesis that there is no lag one autocorrelation present in the residuals.
- If there is no autocorrelation, the DW distribution is symmetric around 2, its mean value.
- R gives p-values.

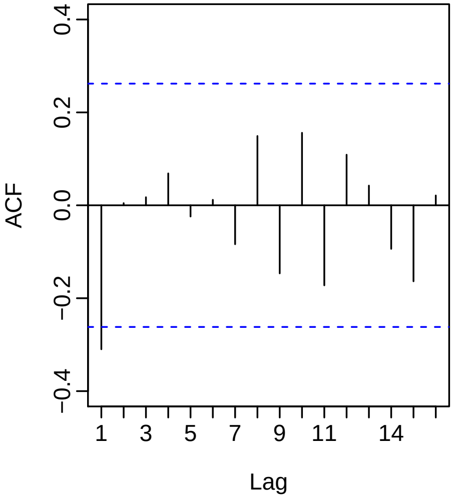
Durbin-Watson test

If residuals from a linear model:

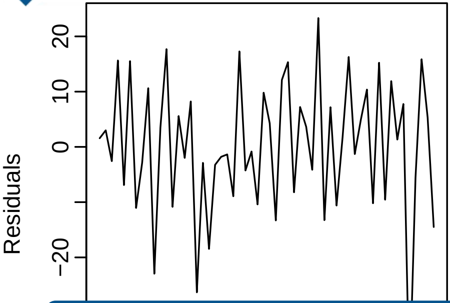
- DW tests hypothesis that there is no lag one autocorrelation present in the residuals.
- If there is no autocorrelation, the DW distribution is symmetric around 2, its mean value.
- R gives p-values.

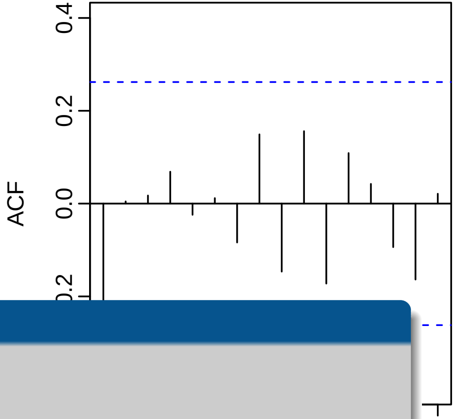
Beer production again





Beer production again





Durbin-Watson test

```
> dwtest(fit,alt="two.sided")

  Durbin-Watson test

data:  fit
DW = 2.5951, p-value = 0.02764
alternative hypothesis: true autocorrelation is not 0
```

Durbin-Watson test

If the model fails the Durbin-Watson test . . .

- The forecasts are not wrong, but have higher variance than they need to.
- There is information in the residuals that we should exploit.
- This is done with a regression model with ARMA errors which will be covered later in the course.

Durbin-Watson test

If the model fails the Durbin-Watson test . . .

- The forecasts are not wrong, but have higher variance than they need to.
- There is information in the residuals that we should exploit.
- This is done with a regression model with ARMA errors which will be covered later in the course.

Durbin-Watson test

If the model fails the Durbin-Watson test . . .

- The forecasts are not wrong, but have higher variance than they need to.
- There is information in the residuals that we should exploit.
- This is done with a regression model with ARMA errors which will be covered later in the course.

Outline

1 Some useful predictors for linear models

2 Residual diagnostics

3 Selecting predictors and forecast evaluation

4 Matrix formulation

5 Correlation, causation and forecasting

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose p values are greater than 0.05.
- Maximize R^2
- Minimize MSE

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose p values are greater than 0.05.
- Maximize R^2
- Minimize MSE

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose p values are greater than 0.05.
- Maximize R^2
- Minimize MSE.

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose p values are greater than 0.05.
- Maximize R^2
- Minimize MSE.

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose p values are greater than 0.05.
- Maximize R^2
- Minimize MSE.

Selecting predictors

- When there are many predictors, how should we choose which ones to use?
- We need a way of comparing two competing models.

What not to do!

- Plot y against a particular predictor (x_j) and if it shows no noticeable relationship, drop it.
- Do a multiple linear regression on all the predictors and disregard all variables whose p values are greater than 0.05.
- Maximize R^2
- Minimize MSE.

Comparing regression models

Computer output for regression will always give the R^2 value. This is a useful summary of the model.

- It is equal to the square of the correlation between y and $\hat{y}$.
- It is often called the "coefficient of determination".
- It can also be calculated as follows:

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2}$$

- It is the proportion of variance accounted for (explained) by the predictors.

Comparing regression models

Computer output for regression will always give the R^2 value. This is a useful summary of the model.

- It is equal to the square of the correlation between y and $\hat{y}$.
- It is often called the "coefficient of determination".
- It can also be calculated as follows:

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2}$$

- It is the proportion of variance accounted for (explained) by the predictors.

Comparing regression models

Computer output for regression will always give the R^2 value. This is a useful summary of the model.

- It is equal to the square of the correlation between y and $\hat{y}$.
- It is often called the "coefficient of determination".
- It can also be calculated as follows:

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2}$$

- It is the proportion of variance accounted for (explained) by the predictors.

Comparing regression models

Computer output for regression will always give the R^2 value. This is a useful summary of the model.

- It is equal to the square of the correlation between y and $\hat{y}$.
- It is often called the "coefficient of determination".
- It can also be calculated as follows:

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2}$$

- It is the proportion of variance accounted for (explained) by the predictors.

Comparing regression models

However . . .

- 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 $\hat{\sigma}^2$.

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

Comparing regression models

However . . .

- 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 $\hat{\sigma}^2$.

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

Comparing regression models

However . . .

- 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 $\hat{\sigma}^2$.

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

Comparing regression models

However ...

- 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 $\hat{\sigma}^2$.

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

Comparing regression models

However ...

- 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 $\hat{\sigma}^2$.

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

Beware of over-fitting

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.
- Problems can be overcome by measuring true *out-of-sample* forecast accuracy. That is, total data divided into “training” set and “test” set. Training set used to estimate parameters. Forecasts are made for test set.
- Accuracy measures computed for errors in test set only.

Beware of over-fitting

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.
- Problems can be overcome by measuring true *out-of-sample* forecast accuracy. That is, total data divided into “training” set and “test” set. Training set used to estimate parameters. Forecasts are made for test set.
- Accuracy measures computed for errors in test set only.

Beware of over-fitting

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.
- Problems can be overcome by measuring true *out-of-sample* forecast accuracy. That is, total data divided into “training” set and “test” set. Training set used to estimate parameters. Forecasts are made for test set.
- Accuracy measures computed for errors in test set only.

Beware of over-fitting

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.
- Problems can be overcome by measuring true *out-of-sample* forecast accuracy. That is, total data divided into "training" set and "test" set. Training set used to estimate parameters. Forecasts are made for test set.
- Accuracy measures computed for errors in test set only.

Beware of over-fitting

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.
- Problems can be overcome by measuring true *out-of-sample* forecast accuracy. That is, total data divided into “training” set and “test” set. Training set used to estimate parameters. Forecasts are made for test set.
- Accuracy measures computed for errors in test set only.

Cross-validation

Cross-validation for regression

- 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

Cross-validation for regression

- 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.

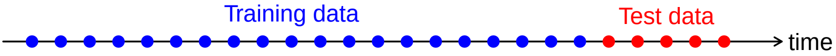
Cross-validation

Cross-validation for regression

- 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.

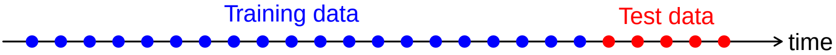
Cross-validation

Traditional evaluation

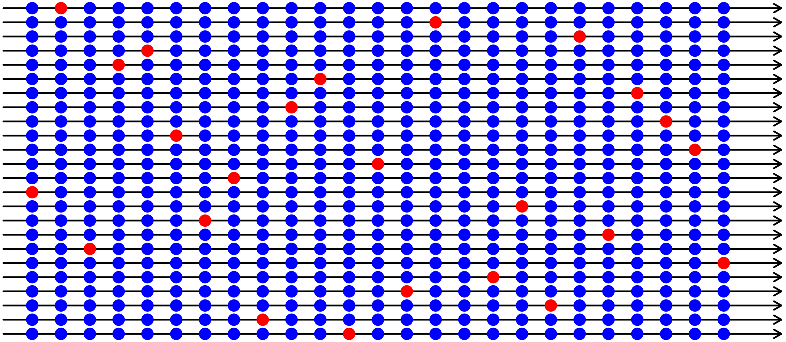


Cross-validation

Traditional evaluation

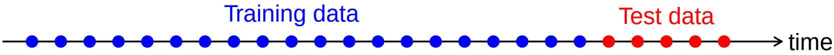


Leave-one-out cross-validation

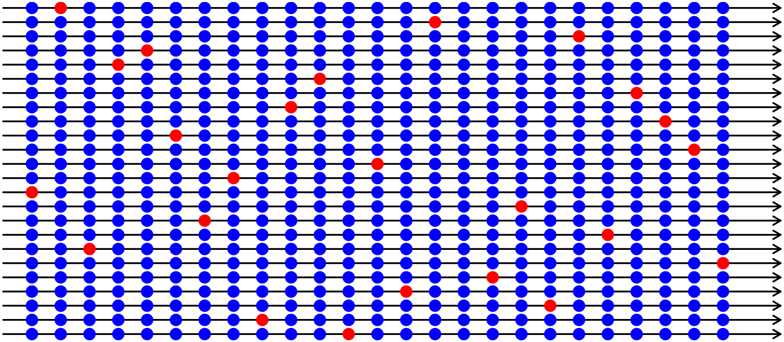


Cross-validation

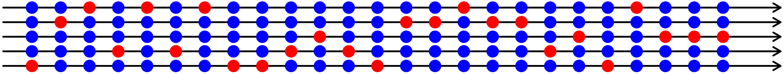
Traditional evaluation



Leave-one-out cross-validation



Five-fold cross-validation



Cross-validation

Leave-one-out cross-validation for regression can be carried out using the following steps.

1. Remove observation t from the data set, and fit the model using the remaining data. Then compute the error ($e_t^* = y_t - \hat{y}_t$) for the omitted observation.
2. Repeat step 1 for $t = 1, \dots, T$.
3. Compute the MSE from $\{e_1^*, \dots, e_T^*\}$. We shall call this the CV.

The best model is the one with minimum CV.

Cross-validation

Leave-one-out cross-validation for regression can be carried out using the following steps.

1. Remove observation t from the data set, and fit the model using the remaining data. Then compute the error ($e_t^* = y_t - \hat{y}_t$) for the omitted observation.
2. Repeat step 1 for $t = 1, \dots, T$.
3. Compute the MSE from $\{e_1^*, \dots, e_T^*\}$. We shall call this the CV.

The best model is the one with minimum CV.

Cross-validation

Leave-one-out cross-validation for regression can be carried out using the following steps.

1. Remove observation t from the data set, and fit the model using the remaining data. Then compute the error ($e_t^* = y_t - \hat{y}_t$) for the omitted observation.
2. Repeat step 1 for $t = 1, \dots, T$.
3. Compute the MSE from $\{e_1^*, \dots, e_T^*\}$. We shall call this the CV.

The best model is the one with minimum CV.

Cross-validation

Leave-one-out cross-validation for regression can be carried out using the following steps.

1. Remove observation t from the data set, and fit the model using the remaining data. Then compute the error ($e_t^* = y_t - \hat{y}_t$) for the omitted observation.
2. Repeat step 1 for $t = 1, \dots, T$.
3. Compute the MSE from $\{e_1^*, \dots, e_T^*\}$. We shall call this the CV.

The best model is the one with minimum CV.

Cross-validation

Leave-one-out cross-validation for regression can be carried out using the following steps.

1. Remove observation t from the data set, and fit the model using the remaining data. Then compute the error ($e_t^* = y_t - \hat{y}_t$) for the omitted observation.
2. Repeat step 1 for $t = 1, \dots, T$.
3. Compute the MSE from $\{e_1^*, \dots, e_T^*\}$. We shall call this the CV.

The best model is the one with minimum CV.

Cross-validation

Ten-fold cross-validation

- Randomly select ten observations for test set, and use remaining observations in training set. Compute accuracy measures on test observations.
- Repeat many times
- Average over all measures.

Cross-validation

Ten-fold cross-validation

- Randomly select ten observations for test set, and use remaining observations in training set. Compute accuracy measures on test observations.
- Repeat many times
- Average over all measures.

Cross-validation

Ten-fold cross-validation

- Randomly select ten observations for test set, and use remaining observations in training set. Compute accuracy measures on test observations.
- Repeat many times
- Average over all measures.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

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

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

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

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

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

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

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

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Akaike's Information Criterion

$$\text{AIC} = -2\log(L) + 2(k+1)$$

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

- This is a *penalized likelihood* approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes terms more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Corrected AIC

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

$$\text{AIC}_{\text{C}} = \text{AIC} + \frac{2(k+2)(k+3)}{T-k-1}$$

As with the AIC, the AIC_{C} should be minimized.

Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(T)$$

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

Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(T)$$

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

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = T[1 - 1/(log(T) - 1)]$.

Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(T)$$

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

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = T[1 - 1/(log(T) - 1)]$.

Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(T)$$

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

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = T[1 - 1/(log(T) - 1)]$.

Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (k+1)\log(T)$$

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

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes terms more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = T[1 - 1/(log(T) - 1)]$.

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, BIC, $\bar{R}^2$).

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

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, BIC, $\bar{R}^2$).

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

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, BIC, $\bar{R}^2$).

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

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, BIC, $\bar{R}^2$).

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.

Notes:

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.

Notes:

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

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.

Notes:

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

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.

Notes:

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

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.

Notes:

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

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.

Notes:

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

Outline

1 Some useful predictors for linear models

2 Residual diagnostics

3 Selecting predictors and forecast evaluation

4 **Matrix formulation**

5 Correlation, causation and forecasting

Matrix formulation

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

Let $\boldsymbol{y} = (y_1, \dots, y_T)'$, $\boldsymbol{e} = (e_1, \dots, e_T)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \dots & x_{k,T} \end{bmatrix}.$$

Then

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

Matrix formulation

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

Let $\boldsymbol{y} = (y_1, \dots, y_T)'$, $\boldsymbol{e} = (e_1, \dots, e_T)'$,
$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \dots & x_{k,T} \end{bmatrix}.$$

Then

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}.$$

Matrix formulation

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

Let $\boldsymbol{y} = (y_1, \dots, y_T)'$, $\boldsymbol{e} = (e_1, \dots, e_T)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \dots & x_{k,T} \end{bmatrix}.$$

Then

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}.$$

Matrix formulation

Least squares estimation

Minimize: $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$

Differentiate wrt β gives

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

(The “normal equation”.)

$$\hat{\sigma}^2 = \frac{1}{T-k-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\boldsymbol{X}'\boldsymbol{X})$ is a singular matrix.

Matrix formulation

Least squares estimation

Minimize: $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$

Differentiate wrt β gives

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

(The "normal equation".)

$$\hat{\sigma}^2 = \frac{1}{T-k-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\boldsymbol{X}'\boldsymbol{X})$ is a singular matrix.

Matrix formulation

Least squares estimation

Minimize: $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$

Differentiate wrt β gives

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

(The "normal equation".)

$$\hat{\sigma}^2 = \frac{1}{T-k-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\boldsymbol{X}'\boldsymbol{X})$ is a singular matrix.

Matrix formulation

Least squares estimation

Minimize: $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$

Differentiate wrt β gives

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

(The “normal equation”.)

$$\hat{\sigma}^2 = \frac{1}{T-k-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\boldsymbol{X}'\boldsymbol{X})$ is a singular matrix.

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)$ is minimized.

So **MLE = OLS**.

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$ is minimized.

So **MLE = OLS**.

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$ is minimized.

So **MLE = OLS**.

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\right)$$

which is maximized when $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$ is minimized.

So **MLE = OLS**.

Multiple regression forecasts

Optimal forecasts

$$\hat{y}^* = \mathrm{E}(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*) = \boldsymbol{x}^*\hat{\beta} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{x}^*$ is a row vector containing the values of the regressors for the forecasts (in the same format as $\boldsymbol{X}$).

Forecast variance

$$\mathrm{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*) = \sigma^2\left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right]$$

- This ignores any errors in $\boldsymbol{x}^*$.
- 95% prediction intervals assuming normal errors: $\hat{y}^* \pm 1.96\sqrt{\mathrm{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*)}$.

Multiple regression forecasts

Optimal forecasts

$$\hat{y}^* = \mathrm{E}(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*) = \boldsymbol{x}^*\hat{\beta} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{x}^*$ is a row vector containing the values of the regressors for the forecasts (in the same format as $\boldsymbol{X}$).

Forecast variance

$$\mathrm{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*) = \sigma^2\left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right]$$

- This ignores any errors in $\boldsymbol{x}^*$.
- 95% prediction intervals assuming normal errors: $\hat{y}^* \pm 1.96\sqrt{\mathrm{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*)}$.

Multiple regression forecasts

Optimal forecasts

$$\hat{y}^* = \text{E}(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*) = \boldsymbol{x}^*\hat{\beta} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{x}^*$ is a row vector containing the values of the regressors for the forecasts (in the same format as $\boldsymbol{X}$).

Forecast variance

$$\text{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*) = \sigma^2 \left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right]$$

- This ignores any errors in $\boldsymbol{x}^*$.
- 95% prediction intervals assuming normal errors: $\hat{y}^* \pm 1.96\sqrt{\text{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*)}$.

Multiple regression forecasts

Optimal forecasts

$$\hat{y}^* = \mathrm{E}(y^*|\boldsymbol{y},\boldsymbol{X},\boldsymbol{x}^*) = \boldsymbol{x}^*\hat{\beta} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{x}^*$ is a row vector containing the values of the regressors for the forecasts (in the same format as $\boldsymbol{X}$).

Forecast variance

$$\mathrm{Var}(y^*|\boldsymbol{X},\boldsymbol{x}^*) = \sigma^2\left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right]$$

- This ignores any errors in $\boldsymbol{x}^*$.
- 95% prediction intervals assuming normal errors: $\hat{y}^* \pm 1.96\sqrt{\mathrm{Var}(y^*|\boldsymbol{X},\boldsymbol{x}^*)}$.

Multiple regression forecasts

Fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

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

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.

Multiple regression forecasts

Fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

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

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.

Outline

1 Some useful predictors for linear models

2 Residual diagnostics

3 Selecting predictors and forecast evaluation

4 Matrix formulation

5 Correlation, causation and forecasting

Correlation is not causation

- 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.
- 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).

Correlation is not causation

- 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.
- 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).

Correlation is not causation

- 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.
- 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).

Correlation is not causation

- 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.
- 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).

Multicollinearity

In regression analysis, multicollinearity occurs when:

- Two predictors are highly correlated (i.e., the correlation between them is close to ± 1).
- A linear combination of some of the predictors is highly correlated with another predictor.
- A linear combination of one subset of predictors is highly correlated with a linear combination of another subset of predictors.

Multicollinearity

In regression analysis, multicollinearity occurs when:

- Two predictors are highly correlated (i.e., the correlation between them is close to ± 1).
- A linear combination of some of the predictors is highly correlated with another predictor.
- A linear combination of one subset of predictors is highly correlated with a linear combination of another subset of predictors.

Multicollinearity

In regression analysis, multicollinearity occurs when:

- Two predictors are highly correlated (i.e., the correlation between them is close to ± 1).
- A linear combination of some of the predictors is highly correlated with another predictor.
- A linear combination of one subset of predictors is highly correlated with a linear combination of another subset of predictors.

Multicollinearity

If multicollinearity exists. . .

- the numerical estimates of coefficients may be wrong (worse in Excel than in a statistics package)
- don't rely on the *p*-values to determine significance.
- there is no problem with model *predictions* provided the regressors used for forecasting are within the range used for fitting.
- omitting variables can help.
- combining variables can help.

Multicollinearity

If multicollinearity exists. . .

- the numerical estimates of coefficients may be wrong (worse in Excel than in a statistics package)
- don't rely on the *p*-values to determine significance.
- there is no problem with model *predictions* provided the regressors used for forecasting are within the range used for fitting.
- omitting variables can help.
- combining variables can help.

Multicollinearity

If multicollinearity exists. . .

- the numerical estimates of coefficients may be wrong (worse in Excel than in a statistics package)
- don't rely on the p-values to determine significance.
- there is no problem with model *predictions* provided the regressors used for forecasting are within the range used for fitting.
- omitting variables can help.
- combining variables can help.

Multicollinearity

If multicollinearity exists. . .

- the numerical estimates of coefficients may be wrong (worse in Excel than in a statistics package)
- don't rely on the *p*-values to determine significance.
- there is no problem with model *predictions* provided the regressors used for forecasting are within the range used for fitting.
- omitting variables can help.
- combining variables can help.

Multicollinearity

If multicollinearity exists. . .

- the numerical estimates of coefficients may be wrong (worse in Excel than in a statistics package)
- don't rely on the *p*-values to determine significance.
- there is no problem with model *predictions* provided the regressors used for forecasting are within the range used for fitting.
- omitting variables can help.
- combining variables can help.

Outliers and influential observations

Things to watch for

- *Outliers:* observations which produce large residuals.
- *Influential observations:* An observation is influential if removing it would markedly change the position of the regression line. (Often outliers in the x variable).
- *Lurking variable:* a predictor which was not included in the regression but has an important effect on the response.

Points should not normally be removed without a good explanation of why they are different.

Outliers and influential observations

Things to watch for

- *Outliers:* observations which produce large residuals.
- *Influential observations:* An observation is influential if removing it would markedly change the position of the regression line. (Often outliers in the x variable).
- *Lurking variable:* a predictor which was not included in the regression but has an important effect on the response.

Points should not normally be removed without a good explanation of why they are different.

Outliers and influential observations

Things to watch for

- *Outliers:* observations which produce large residuals.
- *Influential observations:* An observation is influential if removing it would markedly change the position of the regression line. (Often outliers in the x variable).
- *Lurking variable:* a predictor which was not included in the regression but has an important effect on the response.

Points should not normally be removed without a good explanation of why they are different.

Outliers and influential observations

Things to watch for

- *Outliers:* observations which produce large residuals.
- *Influential observations:* An observation is influential if removing it would markedly change the position of the regression line. (Often outliers in the x variable).
- *Lurking variable:* a predictor which was not included in the regression but has an important effect on the response.

Points should not normally be removed without a good explanation of why they are different.

Outliers and influential observations

Things to watch for

- *Outliers:* observations which produce large residuals.
- *Influential observations:* An observation is influential if removing it would markedly change the position of the regression line. (Often outliers in the x variable).
- *Lurking variable:* a predictor which was not included in the regression but has an important effect on the response.

Points should not normally be removed without a good explanation of why they are different.