https://workshop.f4sg.org/africast/

Africast-Time Series Analysis & Forecasting Using R

6. Forecasting with regression, how to represent temporal structure with regressors

tsibble feasts Fable

Outline

1. The linear model with time series
2. Evaluating the regression model
3. Selecting predictors
4. Forecasting with regression
5. Correlation, causation and forecasting
6. Some useful predictors for regression models

Outline

Regression models

- To **explain**
- To **forecast**

- Simple linear regression model(SLR)
- Multiple linear regression model (MLR)

SLR model in thoery

Regression model allows for a linear relationship between the forecast variable y and a single predictor variable x.

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

- y_t is the variable we want to predict: the `response` variable
- Each x_t is numerical and is called a `predictor`
- β_0 and β_1 are regression coefficients

SLR model in practice

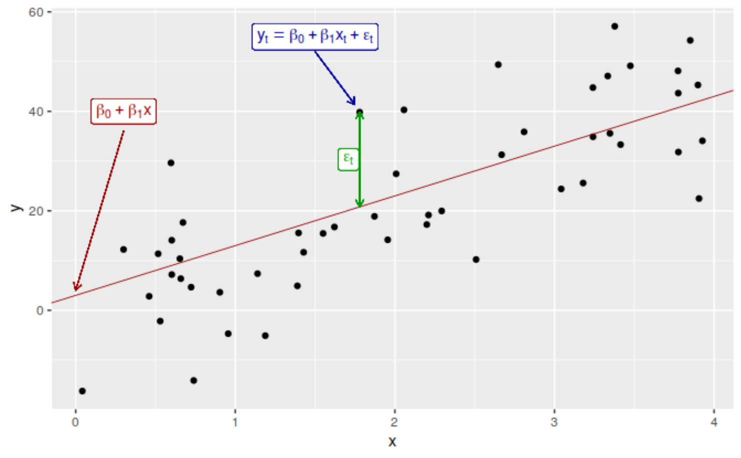
In practice, of course, we have a collection of observations but we do not know the values of the coefficients $\hat{\beta}_0, \hat{\beta}_1$. These need to be estimated from the data.

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_t.$$

- y_t is the `response` variable
- Each x_t is a `predictor`
- $\hat{\beta}_0$ is the estimated intercept
- $\hat{\beta}_1$ is the estimated slope

What is the best fit

- There are many ways that a straight line can be laid on the scatter
- Best known criterion is called Ordinary Least Squares(OLS)



Estimation of the model

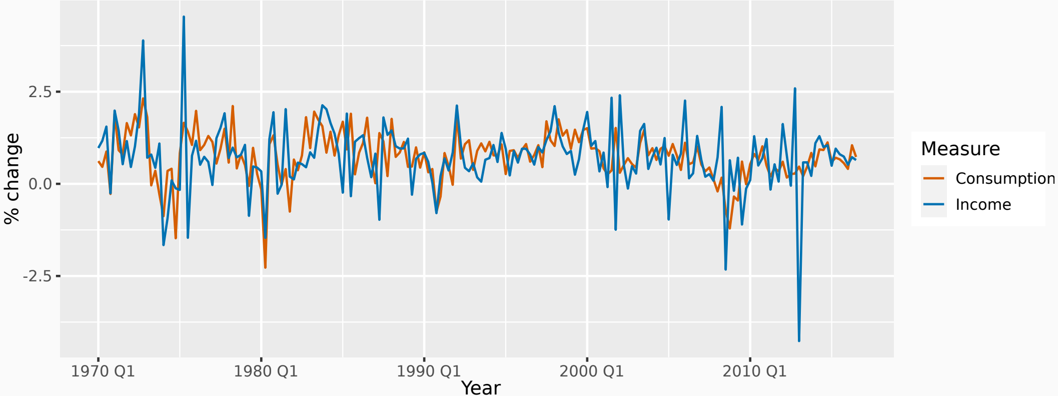
That is, we find the values of β_0 and β_1 which minimize

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2.$$

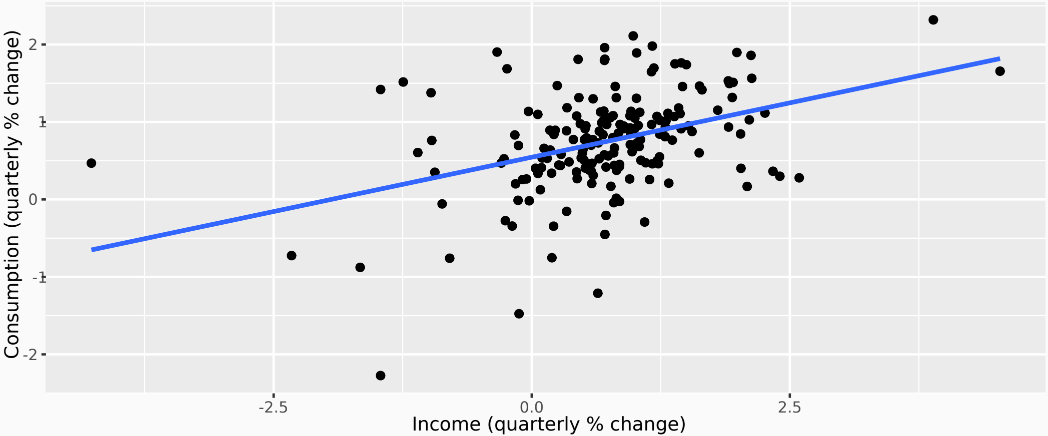
- This is called *least squares* estimation because it gives the least value of the sum of squared errors.
- Finding the best estimates of the coefficients is often called *fitting* the model to the data.
- We refer to the *estimated* coefficients using the notation $\hat{\beta}_0, \hat{\beta}_1$.

Example: US consumption expenditure

```
us_change %>%
  gather("Measure", "Change", Consumption, Income) %>%
  autoplot(Change) +
  ylab("% change") + xlab("Year")
```



Example: US consumption expenditure



Example: US consumption expenditure

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

```
Series: Consumption
Model: TSLM

Residuals:
   Min     1Q Median     3Q    Max
-2.408 -0.318  0.026  0.300  1.452

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5451     0.0557    9.79  < 2e-16 ***
Income        0.2806     0.0474    5.91  1.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.603 on 185 degrees of freedom
Multiple R-squared: 0.159,   Adjusted R-squared: 0.154
```

Multiple regression

- In multiple regression there is one variable to be forecast and several predictor variables.
- The basic concept is that we forecast the time series of interest y assuming that it has a linear relationship with other time series x_1, x_2, ..., x_K
- We might forecast daily A&E attendnace y using temperature x_1 and GP visits x_2 as predictors.

How many variable can we add?

You can add as many as you want but be aware of:

- Overfitting
- Multicollinearity

Multiple regression and forecasting

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

- 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**.

- ε_t is a white noise error term

Estimation of the model

We find the values of $\hat{\beta}_0, \dots, \hat{\beta}_k$ which minimize

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{1,i} - \cdots - \beta_k x_{k,i})^2.$$

- This is called *least squares* estimation because it gives the least value of the sum of squared errors
- Finding the best estimates of the coefficients is often called *fitting* the model to the data
- We refer to the *estimated* coefficients using the notation $\hat{\beta}_0, \dots, \hat{\beta}_k$.

Useful predictors in linear regression

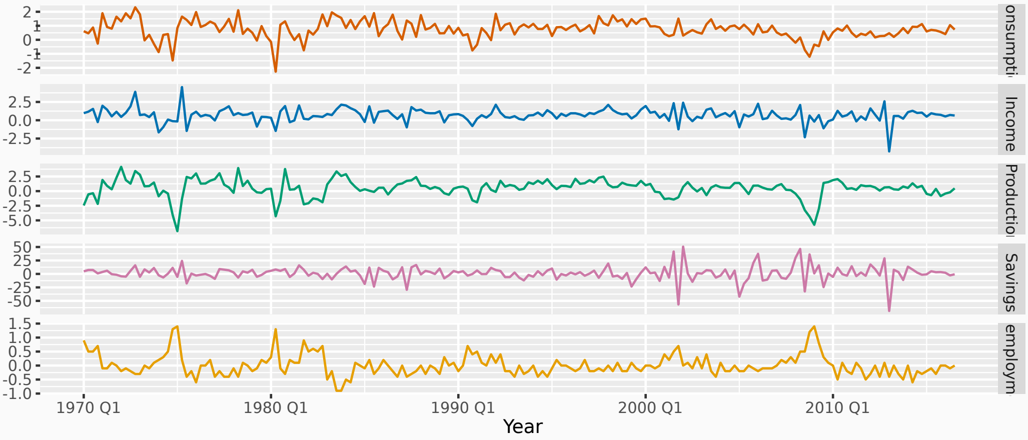
Linear trend [x_t = t]

- $t = 1, 2, \ldots, T$
- Strong assumption that trend will continue.
- use special function `trend()`

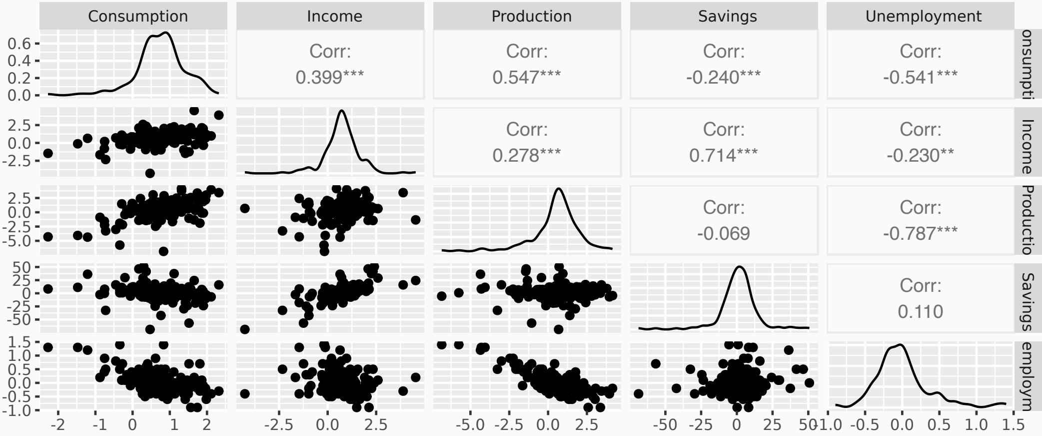
Seasonality

- Seasinality will be considered based on the interval of index
- use special fucntion `season()`

Example: US consumption expenditure



Example: US consumption expenditure

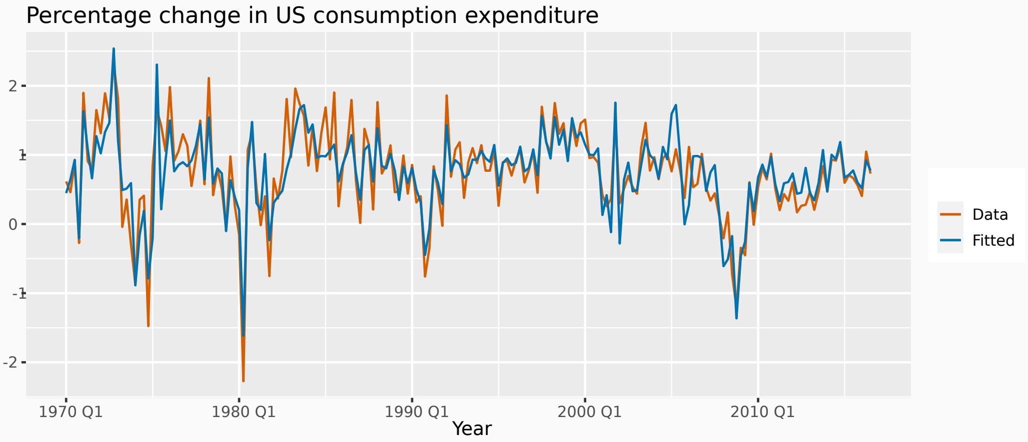


Example: US consumption expenditure

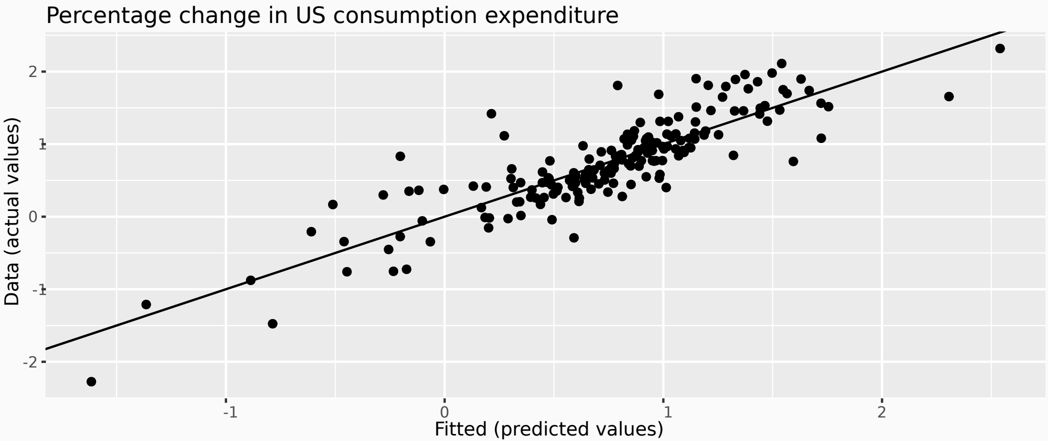
```
fit_consMR <- us_change %>%
  model(lm = TSLM(Consumption ~ Income + Production + Unemployment + Savings))
report(fit_consMR)
```

```
Series: Consumption
Model: TSLM

Residuals:
   Min     1Q Median     3Q    Max
-0.883 -0.176 -0.037  0.153  1.206

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.26729    0.03721    7.18  1.7e-11 ***
Income        0.71448    0.04219   16.93  < 2e-16 ***
Production    0.04589    0.02588    1.77    0.078 .
Unemployment -0.20477    0.10550   -1.94    0.054 .
Savings      -0.04527    0.00278  -16.29  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.329 on 182 degrees of freedom
Multiple R-squared: 0.754,  Adjusted R-squared: 0.749
```

Example: US consumption expenditure

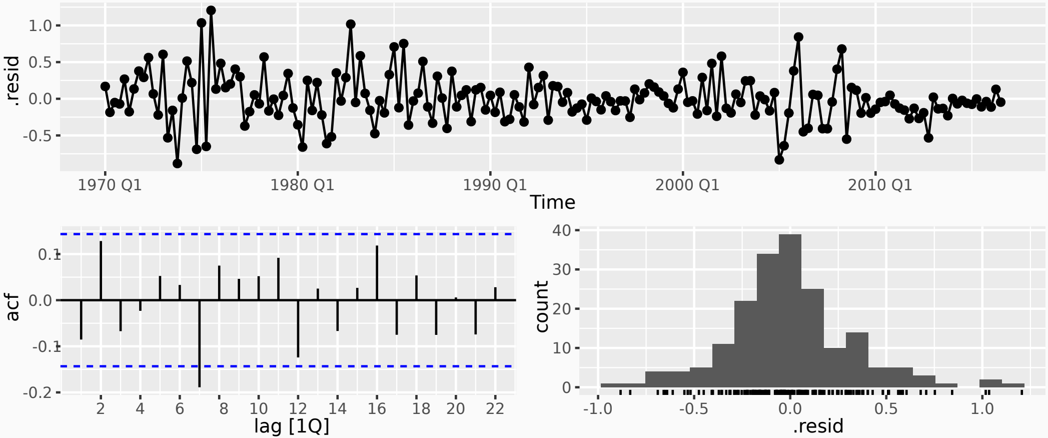


Example: US consumption expenditure



Example: US consumption expenditure

```r
augment(fit_consMR) %>%
  gg_tsdisplay(.resid, plot_type="hist")
```



Outline

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

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

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

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

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

Residual diagnostics

There are a series of plots that should be produced in order to check different aspects of the fitted model and the underlying assumptions.

1. check if residuls are uncorrelated using ACF
2. Check if residuals are normally distributed

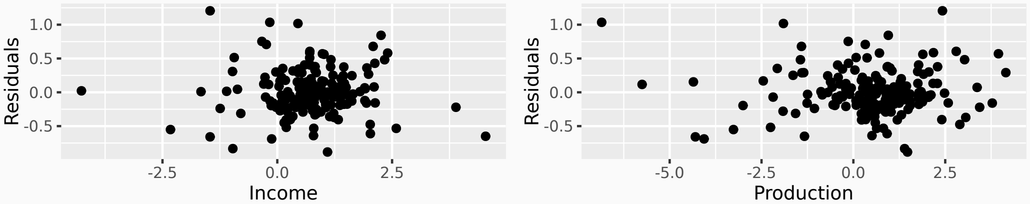
Residual scatterplots

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

- Scatterplot of residuals ε_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.

Example: US consumption expenditure

```
df <- left_join(us_change, residuals(fit_consMR), by = "Time")
p1 <- ggplot(df, aes(x=Income, y=.resid)) +
  geom_point() + ylab("Residuals")
p2 <- ggplot(df, aes(x=Production, y=.resid)) +
  geom_point() + ylab("Residuals")
p3 <- ggplot(df, aes(x=Savings, y=.resid)) +
  geom_point() + ylab("Residuals")
p4 <- ggplot(df, aes(x=Unemployment, y=.resid)) +
  geom_point() + ylab("Residuals")
(p1 | p2) / (p3 | p4)
```



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.)

Outline

1. The linear model with time series
2. Evaluating the regression model
3. **Selecting predictors**
4. Forecasting with regression
5. Correlation, causation and forecasting
6. Some useful predictors for regression models

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.

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.

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 $\hat{\sigma}^2$.

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

Cross-validation

1. Remove observation t from the data set, and fit the model using the remaining data. Then compute the error for the omitted observation
2. Repeat step 1 for $t = 1, ..., T$
3. Compute the MSE from errors obtained in 1. We shall call this the CV

Akaike's Information Criterion

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

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.

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

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

Comparing regression models

```
glance(fit_consMR) %>%
  select(r_squared, adj_r_squared, AIC, AICc, CV)
```

```
# A tibble: 1 x 5
  r_squared adj_r_squared   AIC  AICc    CV
      <dbl>         <dbl> <dbl> <dbl> <dbl>
1     0.754         0.749 -409. -409. 0.116
```

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

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.
- You can also do forward stepwise

Outline

Ex-ante versus ex-post forecasts

- *Ex ante forecasts* are made using only information available in advance.
  - require forecasts of predictors
- *Ex post forecasts* are made using later information on the predictors.
  - 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.

Scenario based forecasting

- Assumes possible scenarios for the predictor variables
- Prediction intervals for scenario based forecasts do not include the uncertainty associated with the future values of the predictor variables.

Building a predictive regression model

- If getting forecasts of predictors is difficult, you can use lagged predictors instead.

$$beta_0 + \beta_1 x_{1,t-h} + \cdots + \beta_k x_{k,t-h} + \varepsilon_t.$$

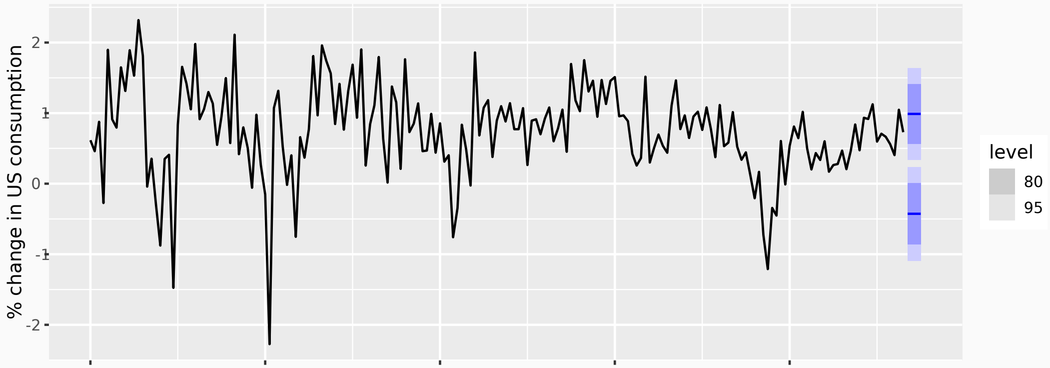
- A different model for each forecast horizon h.

US Consumption

```
fit_consBest <- us_change %>%
  model(
    TSLM(Consumption ~ Income + Savings + Unemployment)
  )

down_future <- new_data(us_change, 4) %>%
  mutate(Income = -1, Savings = -0.5, Unemployment = 0)
fc_down <- forecast(fit_consBest, new_data = down_future)

up_future <- new_data(us_change, 4) %>%
  mutate(Income = 1, Savings = 0.5, Unemployment = 0)
fc_up <- forecast(fit_consBest, new_data = up_future)
```

US Consumption

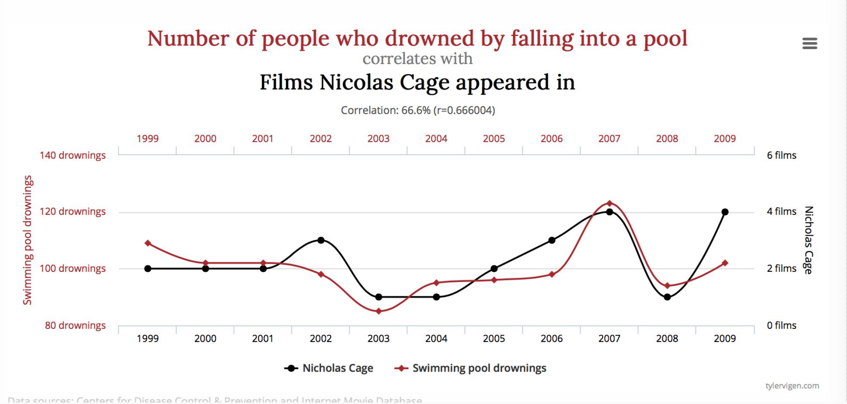
```
us_change %>% autoplot(Consumption) +
  ylab("% change in US consumption") +
  autolayer(fc_up, series = "increase") +
  autolayer(fc_down, series = "decrease") +
  guides(colour = guide_legend(title = "Scenario"))
```



Outline

Correlation does not imply causation

Check out https://www.tylervigen.com/spurious-correlations



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

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 predictors 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 that produce large residuals.
- *Influential observations*: removing them would markedly change the coefficients. (Often outliers in the x variable).
- *Lurking variable*: a predictor not included in the regression but which has an important effect on the response.
- Points should not normally be removed without a good explanation of why they are different.

Modern regression models

- Suppose instead of 3 regressor we had 44.
  - For example, 44 predictors leads to 18 trillion possible models!
- Stepwise regression cannot solve this problem due to the number of variables.
- We need to use the family of Lasso models: lasso, ridge, elastic net

Outline

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.

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies

Outliers

- If there is an outlier, you can use a dummy variable to remove its effect.

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies

Outliers

- If there is an outlier, you can use a dummy variable to remove its effect.

Public holidays

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

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 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.
- this could be also handled using `trend()`

Include special event using dummies

- Christmas Eve: if Christmas Eve, $v_t = 1$, $v_t = 0$ otherwise
- New year's Day: if New year's Day, $v_t = 1$, $v_t = 0$ otherwise.
- and more: Ramadan and Chinese new year, school holiday, etc

Interactions

For example, sometimes the effect of a partiucluar event might be different if it is on a weekend or a week day or its efefct might be different in each shift:

- you need to introduce an interaction variable
- you can use a new dummy as : `v1*v2`

Lagged predictors

The model include present and past values of predictor:

$x_t, x_{t-1}, x_{t-2}, \dots$.

[y_{t} = a + �{0} x{t} + �{1} x{t-1} + ...+ �{k} x{t-k} + �_{t}]

- x can influence y, but y is not allowed to influence x.

Lagged predictors

- Lagged values of a predictor:
  - Create new variables by shifting the existing variable backwards

Example: x is advertising which has a delayed effect

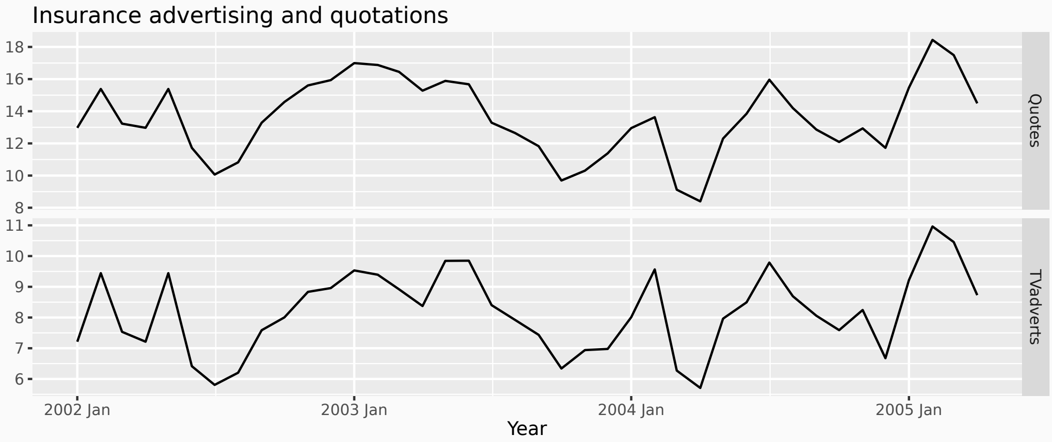
$$\begin{aligned}
x_1 &= \text{advertising for previous month;} \\
x_2 &= \text{advertising for two months previously;} \\
&\vdots \\
x_m &= \text{advertising for } m \text{ months previously.}
\end{aligned}$$

Example: Insurance quotes and TV adverts

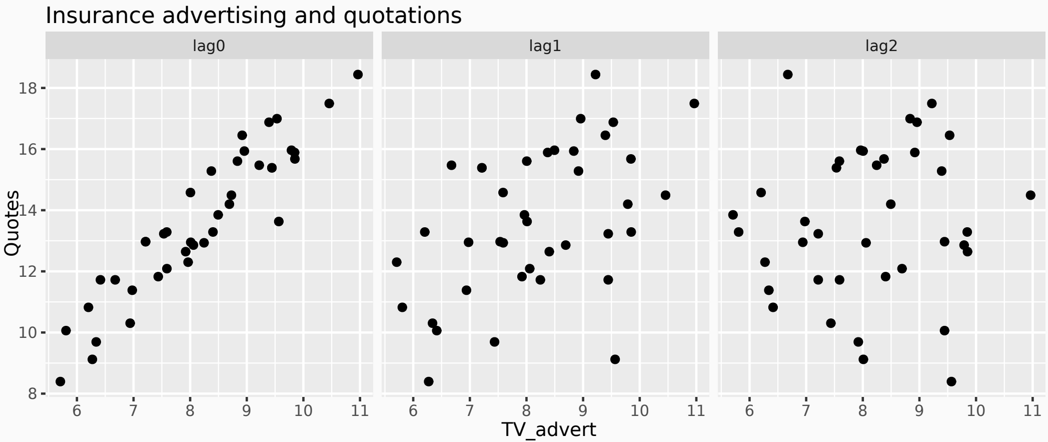
```
insurance
```

```
# A tsibble: 40 x 3 [1M]
      Month Quotes TVadverts
      <mth>  <dbl>     <dbl>
 1 2002 Jan   13.0      7.21
 2 2002 Feb   15.4      9.44
 3 2002 Mar   13.2      7.53
 4 2002 Apr   13.0      7.21
 5 2002 May   15.4      9.44
 6 2002 Jun   11.7      6.42
 7 2002 Jul   10.1      5.81
 8 2002 Aug   10.8      6.20
 9 2002 Sep   13.3      7.59
10 2002 Oct   14.6      8.00
# ℹ 30 more rows
```

Example: Insurance quotes and TV adverts



Example: Insurance quotes and TV adverts



Example: Insurance quotes and TV adverts

```
fit <- insurance |>
  # Restrict data so models use same fitting period
  mutate(Quotes = c(NA, NA, NA, Quotes[4:40])) |>
  model(
    ARIMA(Quotes ~ pdq(d = 0) + TVadverts),
    ARIMA(Quotes ~ pdq(d = 0) + TVadverts +
      lag(TVadverts)),
    ARIMA(Quotes ~ pdq(d = 0) + TVadverts +
      lag(TVadverts) +
      lag(TVadverts, 2)),
    ARIMA(Quotes ~ pdq(d = 0) + TVadverts +
      lag(TVadverts) +
      lag(TVadverts, 2) +
      lag(TVadverts, 3))
  )
```

Example: Insurance quotes and TV adverts

```
glance(fit)
```

Lag order	sigma2	log_lik	AIC	AICc	BIC
0	0.265	-28.3	66.6	68.3	75.0
1	0.209	-24.0	58.1	59.9	66.5
2	0.215	-24.0	60.0	62.6	70.2
3	0.206	-22.2	60.3	65.0	73.8

Example: Insurance quotes and TV adverts

```r
# Re-fit to all data
fit <- insurance |>
  model(ARIMA(Quotes ~ TVadverts + lag(TVadverts) + pdq(d = 0)))
report(fit)
```

```
Series: Quotes
Model: LM w/ ARIMA(1,0,2) errors

Coefficients:
        ar1    ma1    ma2  TVadverts  lag(TVadverts)  intercept
      0.512  0.917  0.459     1.2527          0.1464       2.16
s.e.  0.185  0.205  0.190     0.0588          0.0531       0.86

sigma^2 estimated as 0.2166:  log likelihood=-23.9
AIC=61.9   AICc=65.4   BIC=73.7
```

Example: Insurance quotes and TV adverts

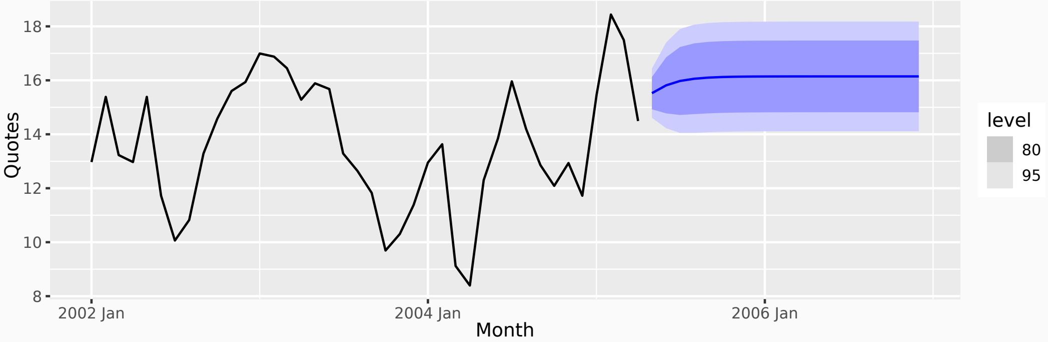
```r
# Re-fit to all data
fit <- insurance |>
  model(ARIMA(Quotes ~ TVadverts + lag(TVadverts) + pdq(d = 0)))
report(fit)
```

```
Series: Quotes
Model: LM w/ ARIMA(1,0,2) errors

Coefficients:
        ar1    ma1    ma2  TVadverts  lag(TVadverts)  intercept
      0.512  0.917  0.459     1.2527          0.1464       2.16
s.e.  0.185  0.205  0.190     0.0588          0.0531       0.86

sigma^2 estimated as 0.2166:  log likelihood=-23.9
AIC=61.9   AICc=65.4   BIC=73.7
```

$$y_t = 2.16 + 1.25x_t + 0.15x_{t-1} + \eta_t,$$
$$\eta_t = 0.512\eta_{t-1} + \varepsilon_t + 0.92\varepsilon_{t-1} + 0.46\varepsilon_{t-2}.$$

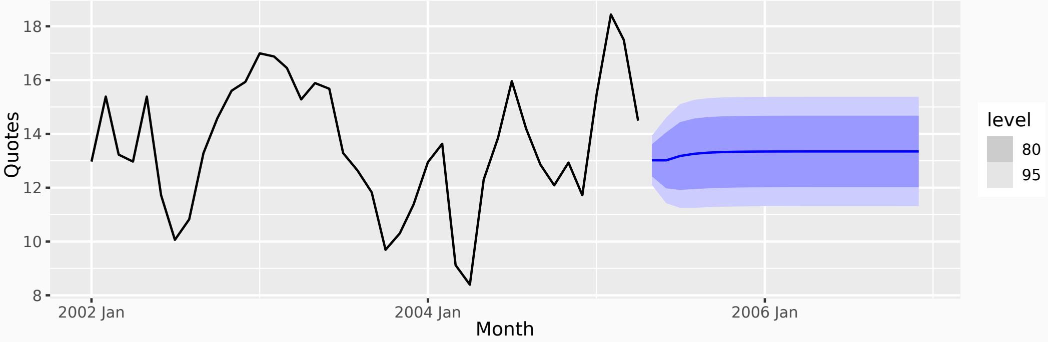
Example: Insurance quotes and TV adverts

```
advert_a <- new_data(insurance, 20) |>
  mutate(TVadverts = 10)
forecast(fit, advert_a) |> autoplot(insurance)
```



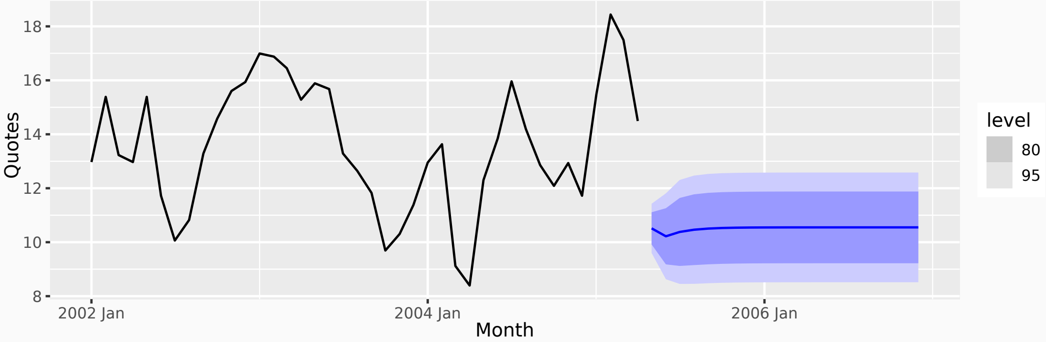
Example: Insurance quotes and TV adverts

```
advert_b <- new_data(insurance, 20) |>
  mutate(TVadverts = 8)
forecast(fit, advert_b) |> autoplot(insurance)
```



Example: Insurance quotes and TV adverts

```
advert_c <- new_data(insurance, 20) |>
  mutate(TVadverts = 6)
forecast(fit, advert_c) |> autoplot(insurance)
```



Lead predictors

Sometimes a change in the predictor x_t that will happen in the future will affect the value of y_t in the past. We say x_t is a leading indicator.

- Lead values of a predictor:
  - Create new variables by shifting the existing variable forwards
- $y_t =$ sales, $x_t =$ tax policy announcement