Cardiff University
Prifysgol Caerdydd

Cardiff Business School
Ysgol Busnes Caerdydd

Forecasting in R

Regression models

Bahman Rostami-Tabar

Outline

Outline

Learning objectives

- Describe linear associations between variables
- Explain regression model assumptions
- Construct a regression model
- Forecast using regression models
- Check residual diagnostics

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

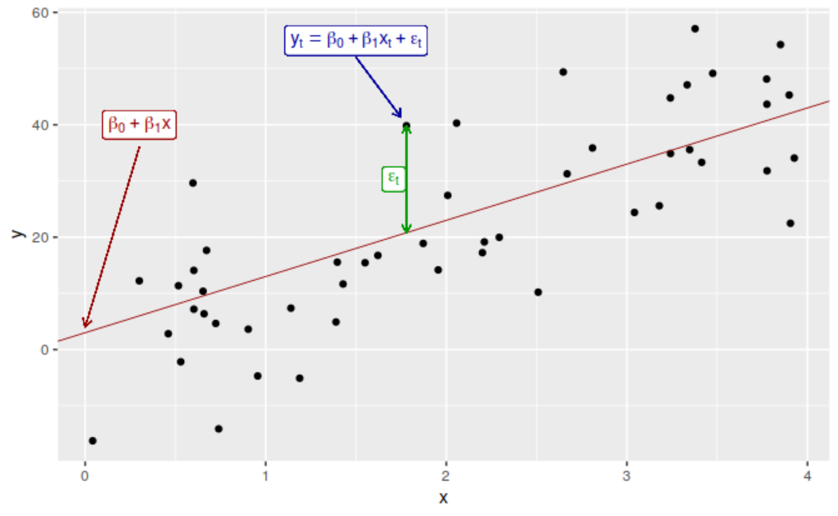
SLR model in practice

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



Estimation of the model

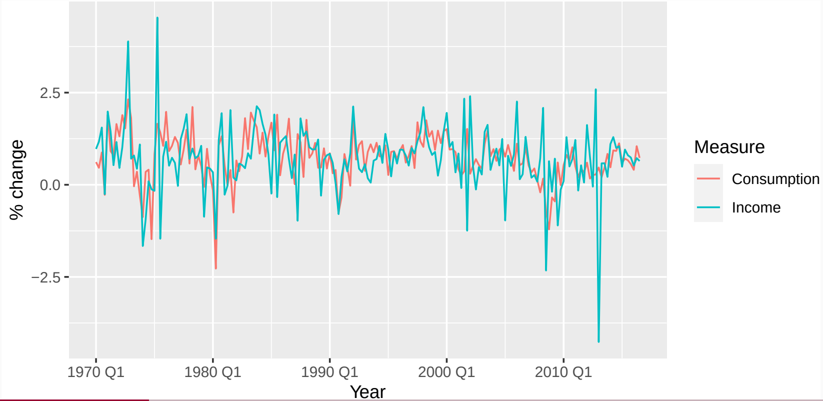
That is, we find the values of β_0 and β_1 which minimize

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

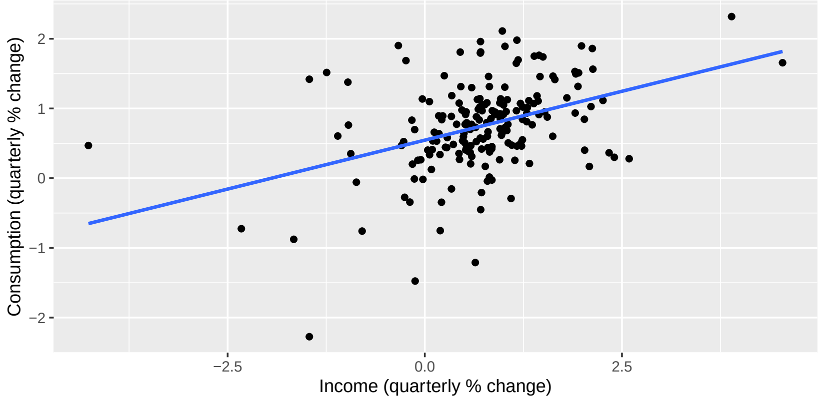
- This is called *least squares* estimation because it gives the least value of the sum of squared errors.
- Finding the best estimates of the coefficients is often called *fitting* the model to the data.
- We refer to the *estimated* coefficients using the notation $\hat{\beta}_0, \hat{\beta}_1$.

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

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

```
## Series: Consumption
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4084 -0.3182  0.0256  0.2998  1.4516
##
## 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 $x_1, x_2, \ldots., x_K$
- 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.
- 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

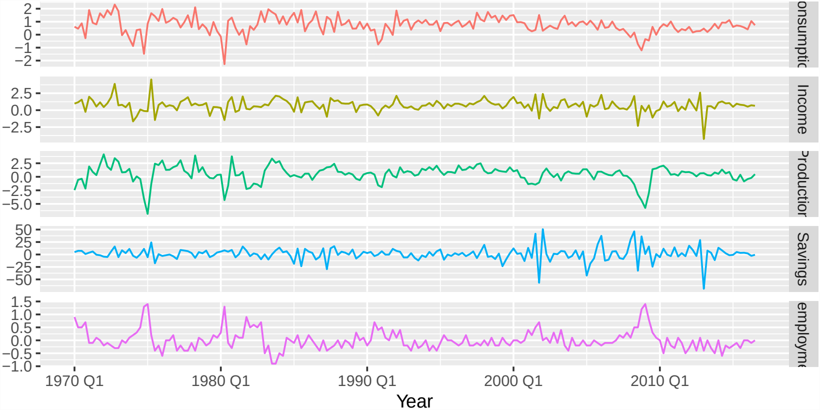
$$x_t = t$$

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

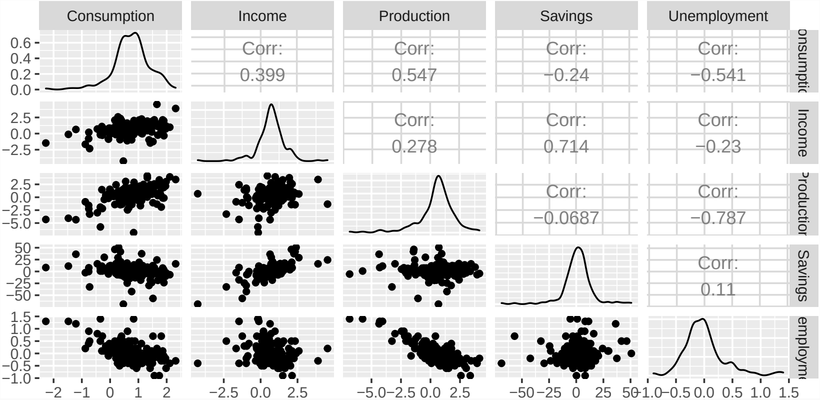
Seasonality

- 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

```r
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
## -0.8830 -0.1764 -0.0368  0.1525  1.2055
##
## 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
## F-statistic:  139 on 4 and 182 DF, p-value: <2e-16
```

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.

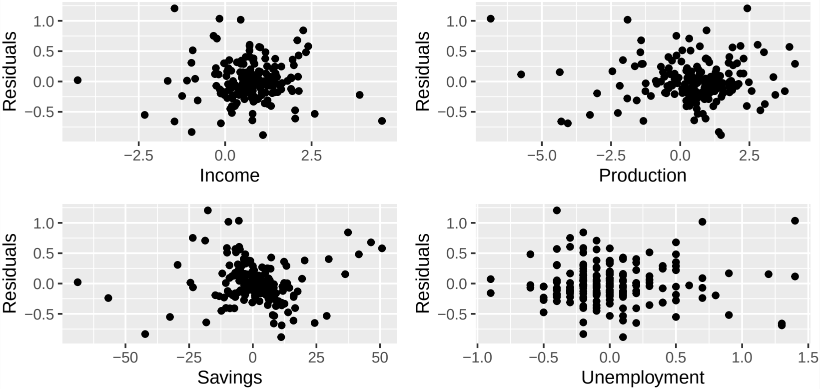
1. check if residuls are uncorrelated using ACF
2. Check if residuals are normally distributed

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



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

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 $t = 1, \ldots, T$
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

```r
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.

US Consumption

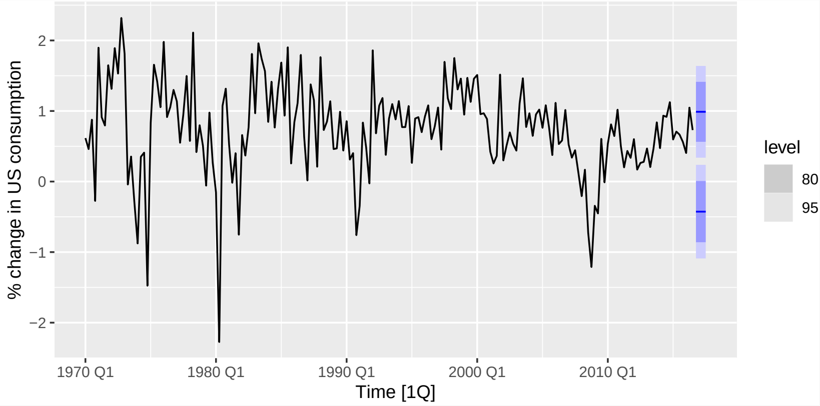
```r
fit_consBest <- us_change %>%
  model(
    TSLM(Consumption ~ Income + Savings + Unemployment)
  )

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)
```

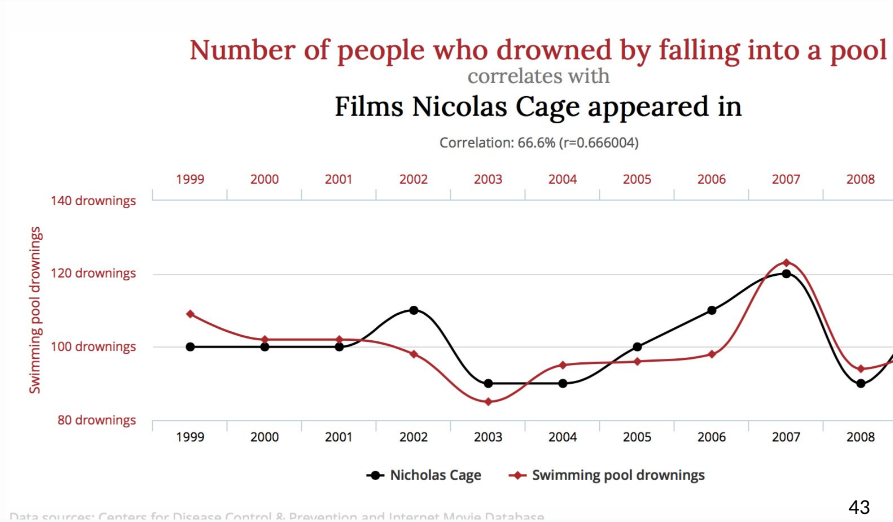
US Consumption

```
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



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.

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
  - watch out for a series of blogs on this in coming weeks

Outline

Lab Session 8

Given the daily A&E data, we want to develop a regression model that takes into account temperature, and daily sesonality:

1. Import the temeperature data `temp` from the project directory
2. Join them to daily data set you have created before
3. Check the linear relationshiop between daily attendance and temperature
4. Split the data into train and test
5. Train data using two regression models 5.1. using temperature and seasonality 5.2. using only seasonality
6. Produce forecast
7. Calculate point forecast accuracy