Financial Time Series

Course Content

- Introduction to time series, review of basic probability and statistics methods.
- Regression, decomposition and smoothing techniques.
- Basic stochastic models.
- ARIMA models and seasonal ARIMA models.
- Seasonal ARIMA models, regression with TS errors.
- ARCH and GARCH models.
- Value at Risk.
- Classification techniques.
- High frequency models and market microstructure.
- Vector AR models and pairs trading.

R

- https://cran.r-project.org/ download and install R
- https://www.rstudio.com/ download and install R studio
- https://www.statmethods.net/r-tutorial/index.html R tutorial

Grading Scheme

Method	Weight %
Midterm Exam	30%
Final Exam	40%
Homework	20%
Attendance	10%

Attendance

- Your attendance will be evaluated through quizzes and participation.
- There will be quizzes throughout the semester.
- Grading of each quiz will be
  - You didn’t attend
  - You attended
- Grades will be available on blackboard 2 days after the quiz.
- If a student cannot attend, he/she should provide a document. The weights of the final and midterm exam will be raised accordingly to compensate for attendance.

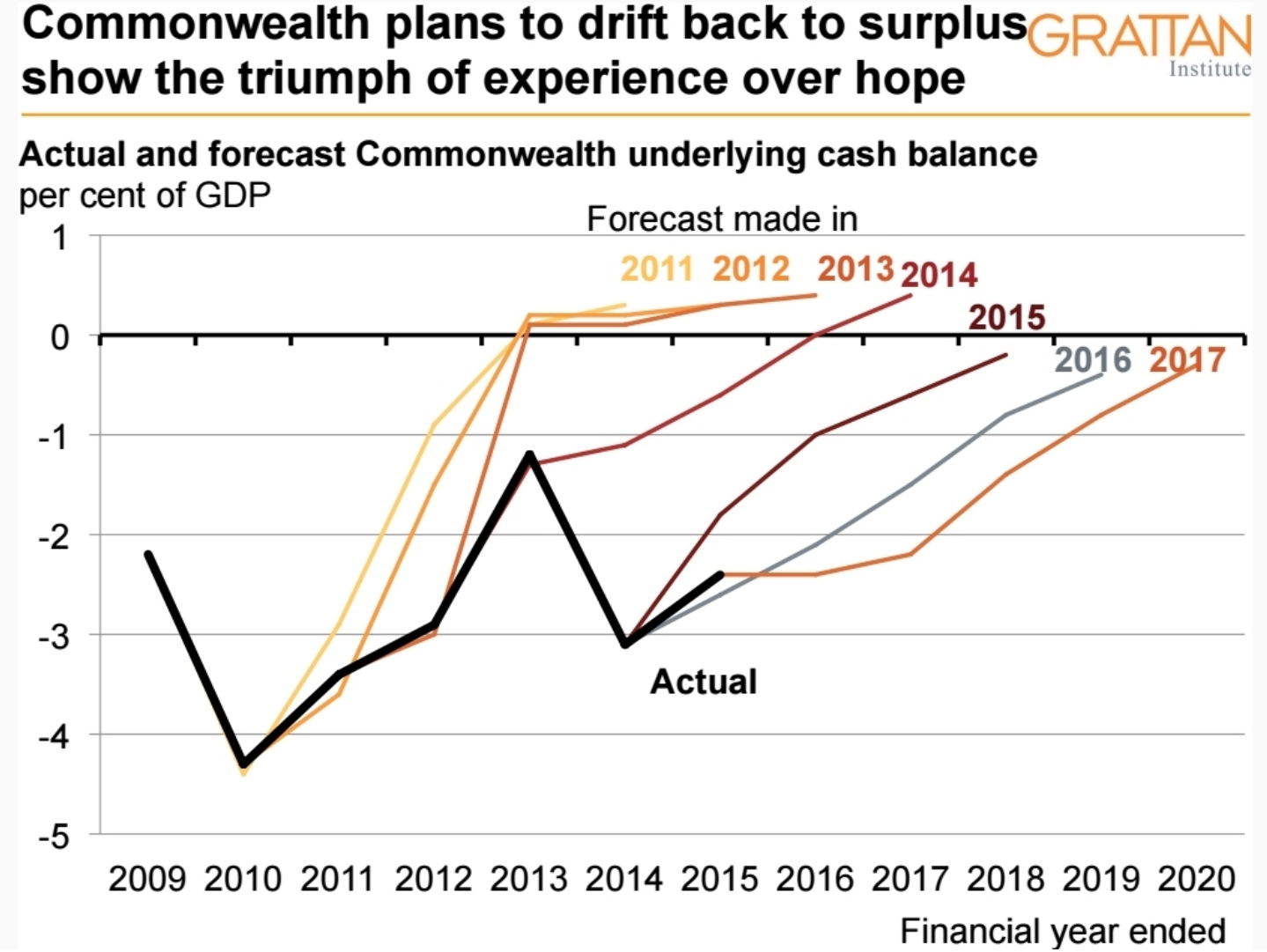
Homework

- There will be homework assignments throughout the semester.
- You should type your answers or scan your handwritten answers (no photos!) and upload the files to blackboard.
- Submissions after the due time will not be accepted.
- Instructor and TA’s will grade the homework.
- Grades will be available on blackboard one week after the submission deadline.

Midterm & Final Exam

- Both exams will be in-class and closed book.
- No makeup exam will be given in general.
- Grades will be available on blackboard.

Forecasting: Is it easy?



Forecasting is difficult

A Timeline of

Very Bad Future Predictions

1800

"Rail travel at high speed is not possible, because passengers, unable to breathe, would die of asphyxia."

Dr. Dionysys Larder, Professor of Natural Philosophy & Astronomy, University College London

1859

"Drill for oil? You mean drill into the ground to try and find oil? You're crazy!"

Associates of Edwin L. Drake refusing his suggestion to drill for oil in 1859 (Later that year, Drake succeeded in drilling the first oil well.)

1876

"This telephone has too many shortcomings to be seriously considered as a means of communication."

Western Union internal memo

1880

"Everyone acquainted with the subject will recognize it as a conspicuous failure."

Henry Morton, president of the Stevens Institute of Technology, on Edison's light bulb

1902

"Flight by machines heavier than air is unpractical and insignificant, if not utterly impossible."

Simon Newcomb, Canadian-American astronomer and mathematician, 18 months before the Wright Brothers' flight at Kittyhawk

1903

"The horse is here to stay, but the automobile is only a novelty, a fad."

The president of the Michigan Savings Bank, advising Henry Ford's lawyer not to invest in the Ford Motor Company

1916

"The idea that cavalry will be replaced by these iron coaches is absurd. It is little short of treasonous."

Comment of Aide-de-camp to Field Marshal Haig, at tank demonstration

1916

"The cinema is little more than a fad. It's canned drama. What audiences really want to see is flesh and blood on the stage."

Charlie Chaplin, actor, producer, director, and studio founder

1921

"The wireless music box has no imaginable commercial value. Who would pay for a message sent to no one in particular?"

Associates of commercial radio and television pioneer, David Sarnoff, responding to his call for investment in the radio

1946

"Television won't last because people will soon get tired of staring at a plywood box every night."

Darryl Zanuck, movie producer, 20th Century Fox

1977

"There is no reason for any individual to have a computer in his home."

Ken Olson, president, chairman and founder of Digital Equipment Corporation

1995

"The truth is no online database will replace your daily newspaper…"

Clifford Stoll, Newsweek article entitled ***The Internet? Bah!***

What can we forecast?

- Daily electricity demand in 3 days time
- Google stock price tomorrow
- Google stock price in 6 months time
- Maximum temperature tomorrow
- Exchange rate of $ US next week

Factors affecting forecastability

- Something is easy to forecast if:
  - we have good understanding of the factors that contribute to it
  - there is lots of data available
  - the forecasts cannot affect the thing we are trying to forecast
  - there is relatively low natural/unexplainable random variation
  - the future is somewhat similar to the past

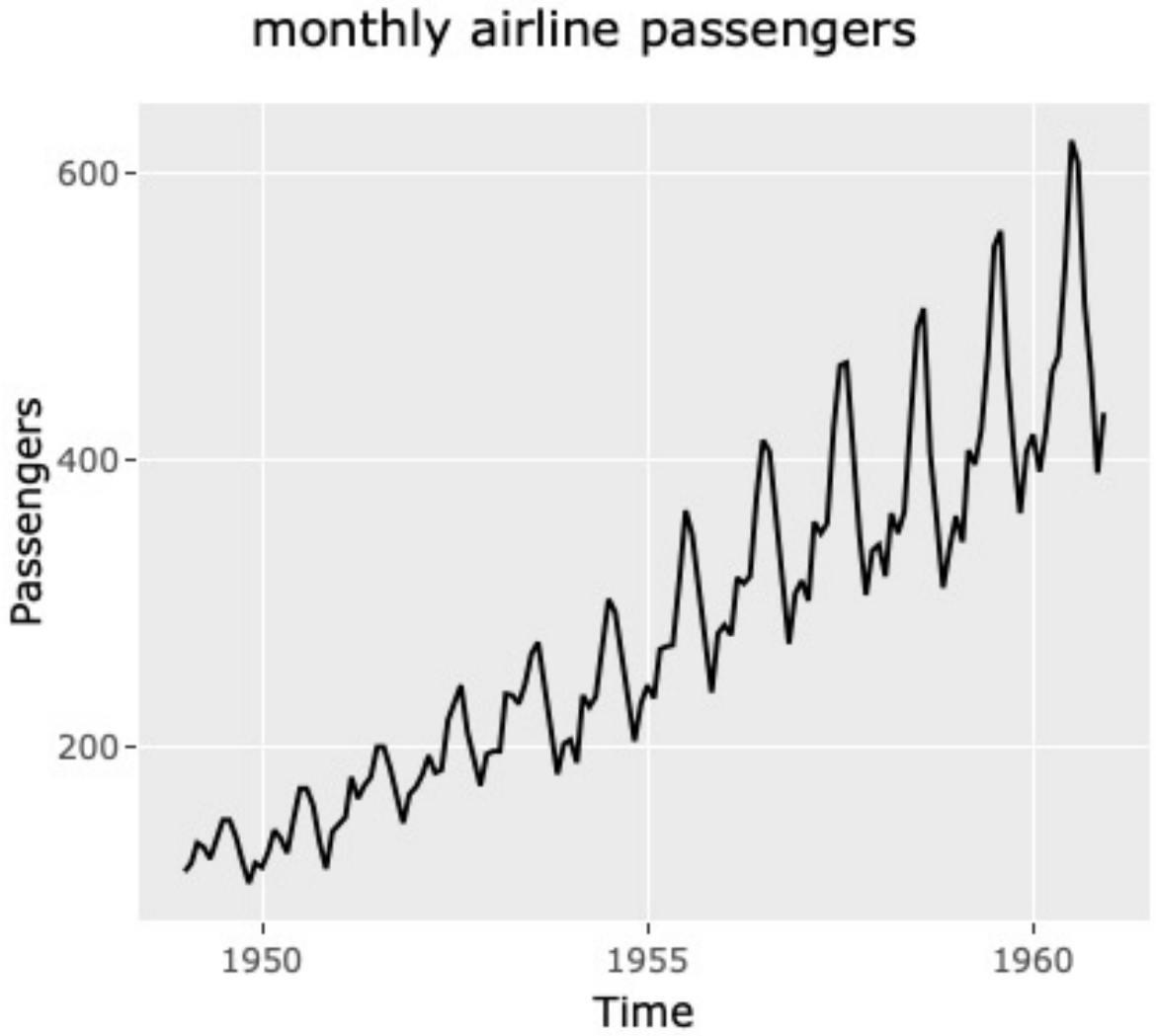
What is time series?

- A time series is a set of observations x_t each one being recorded at a specific time t.
- A time series is a collection of observations made sequentially through time.
- There are correlations among successive observations: Autocorrelated
- Fundamentally different from i.i.d. observations.

Application areas

- Financial investment
  - Daily S&P stocks
  - Johnson and Johnson daily stock closing prices
- Macro Economics
  - Monthly percent changes in US wages and salaries
  - U.S. Annual industrial production
- Micro Economics
  - Daily morning gold prices
  - Annual Copper prices
- Sales
  - Monthly car sales
  - Quarterly sales of toys

TS model for passengers from 1949 to 1960



R code for the air-passengers example

- my_input<-AirPassengers
- gb<-autoplot(my_input) + labs(title="Time Series Model For AirPassengers from 1949 to 1960",x="Time",y="Passengers")
- ggplotly(gb)

Classical decomposition

$$X_t = T_t + S_t + C_t + E_t$$

- Trend (T_t) – Long term movement in the mean
- Seasonal variation (S_t) – Cyclical fluctuations due to calendar
- Cycles (C_t) – Cyclical fluctuations of larger period (e.g. business cycles)
- Residuals (E_t) – random and all other unexplained variations

General approach to TS modeling

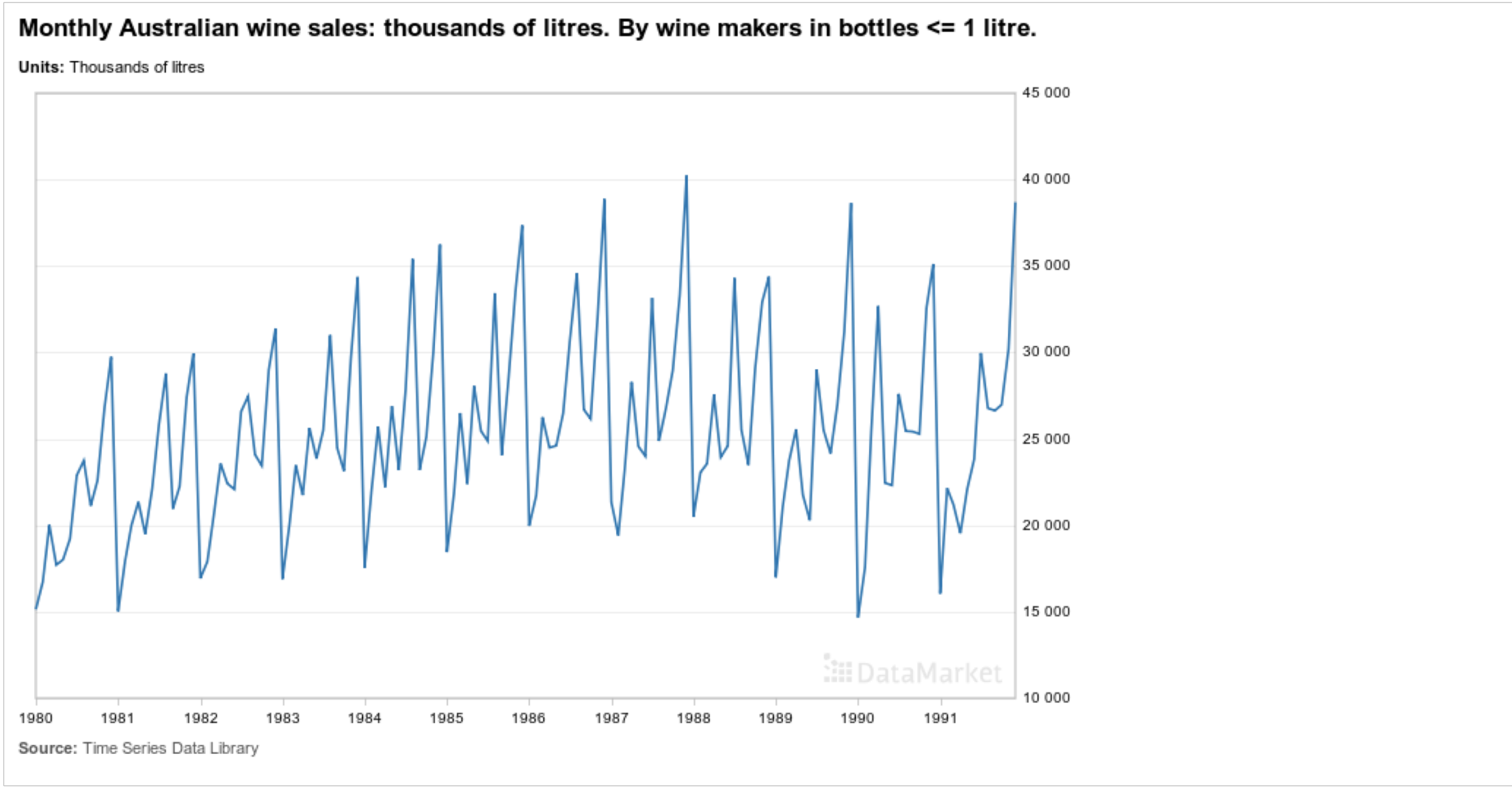
- Plot the series and examine the main features
  - Is there a trend? Is there a seasonality effect? Are there cycles?
  - Is variation apparently time-dependent?
  - Are any apparent sharp changes in behavior?
  - Are there any outliers?
- Perform a transformation of the data if necessary
  - Logarithmic transformation:

    replace $\{X_1, X_2, \dots, X_n\}$ with $\{logX_1, logX_2, \dots, logX_n\}$ if fluctuations appear to grow linearly with the level of the series
  - Box-Cox transformation

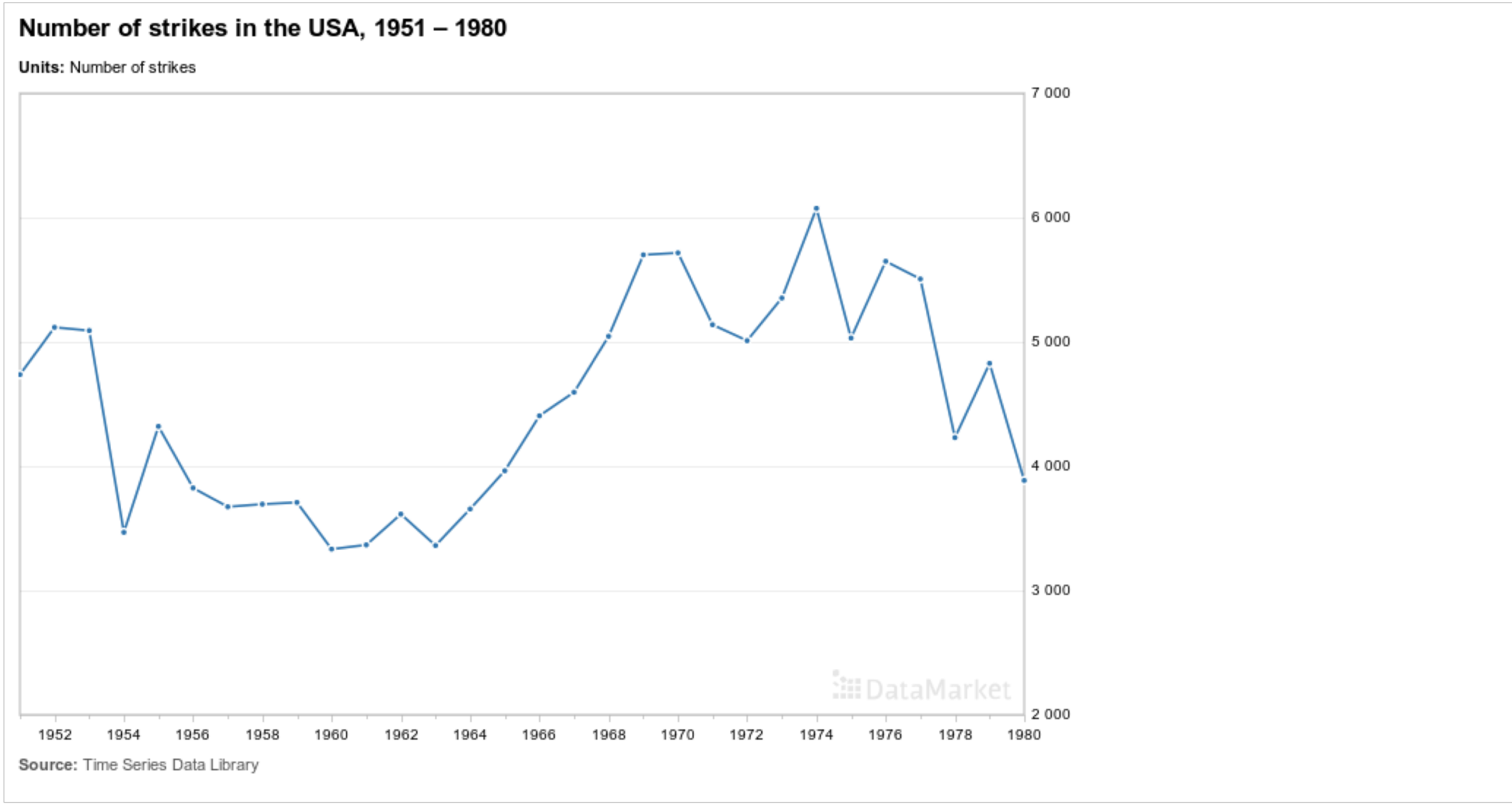
General approach to TS modeling

- Remove the trend, seasonal components and cycles to get stationary residuals
  - Regression analysis
  - Differencing the data: Replacing the original series $\{X_t\}$ by $\{Y_t = X_t - X_{t-1}\}$
- Choose a model to fit the residuals
  - Models will be discussed at length later
- Do the forecasting
  - Be sure to invert the transformations in order to obtain forecasts of the original $\{X_t\}$

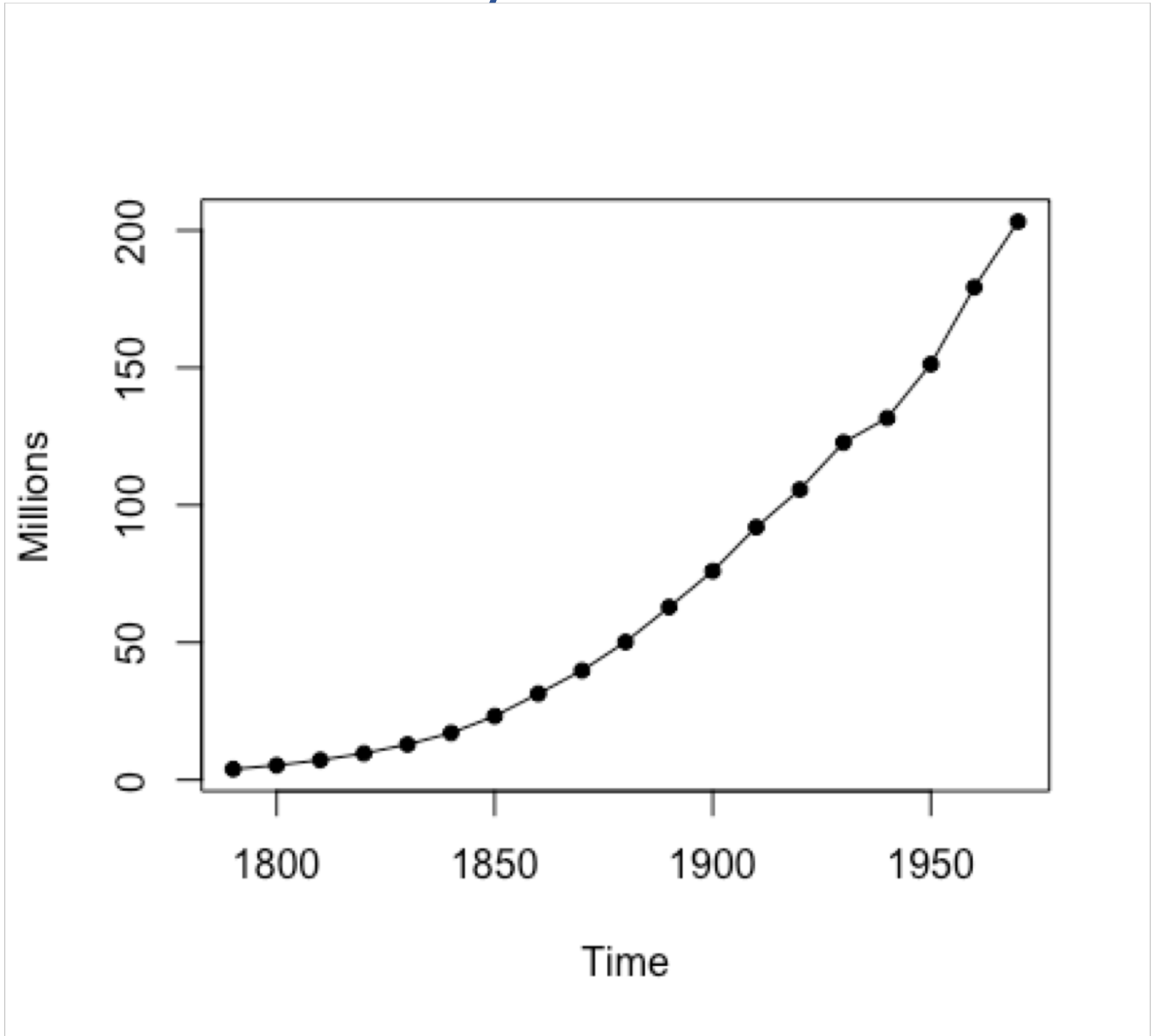
Australian wine (Jan 1980 to Dec 1991)



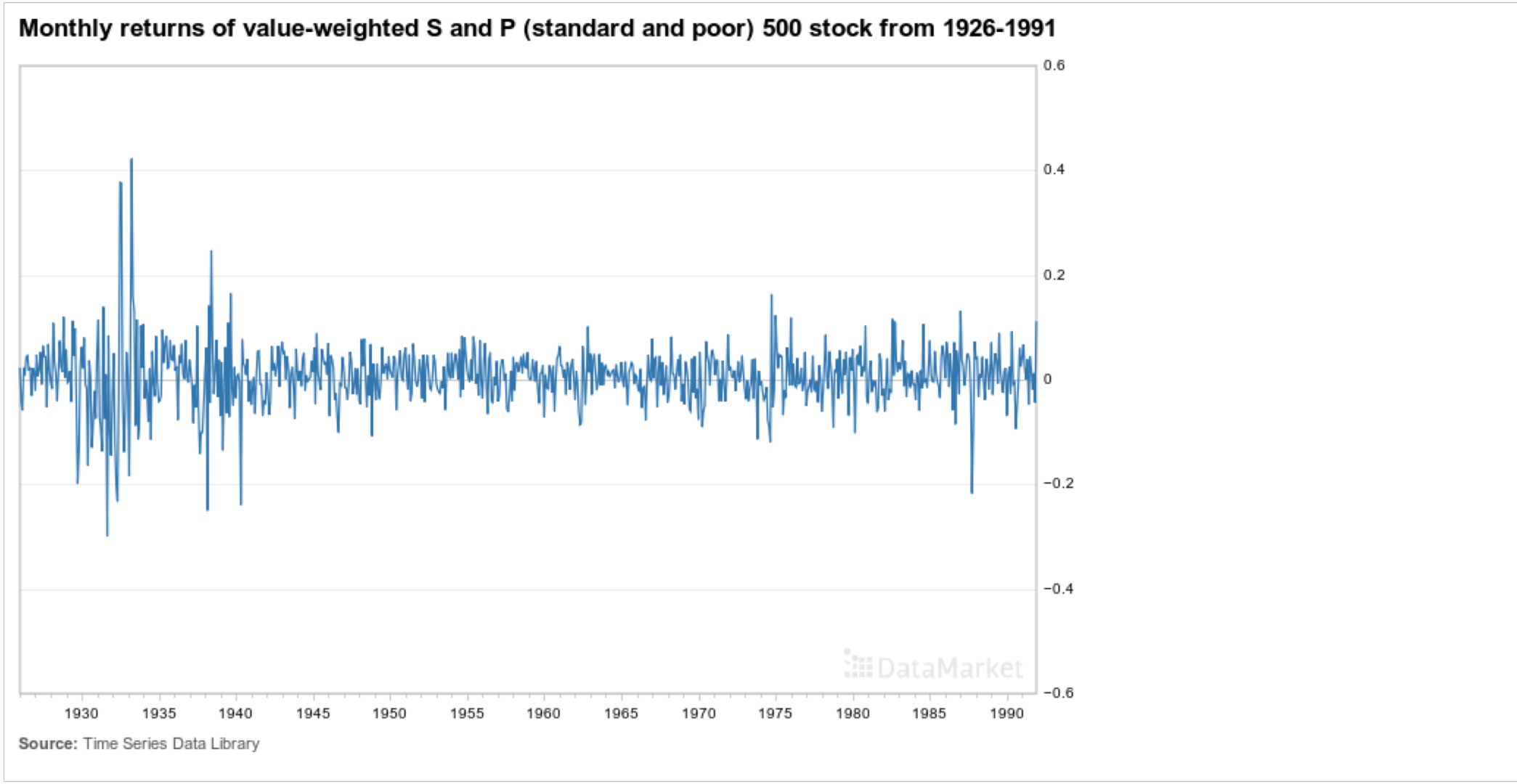
Strikes in the USA, 1951-1980



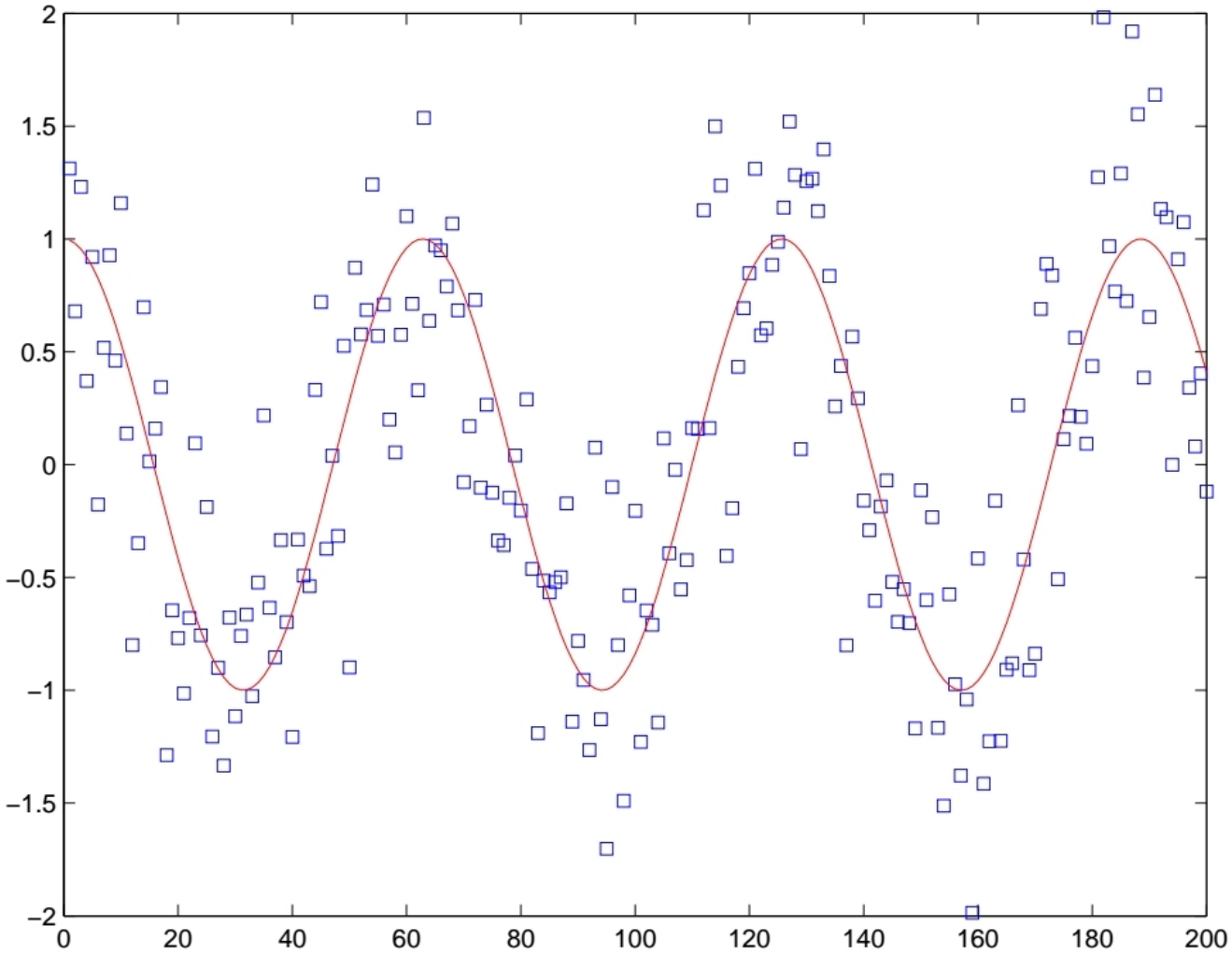
US population at 10-year intervals



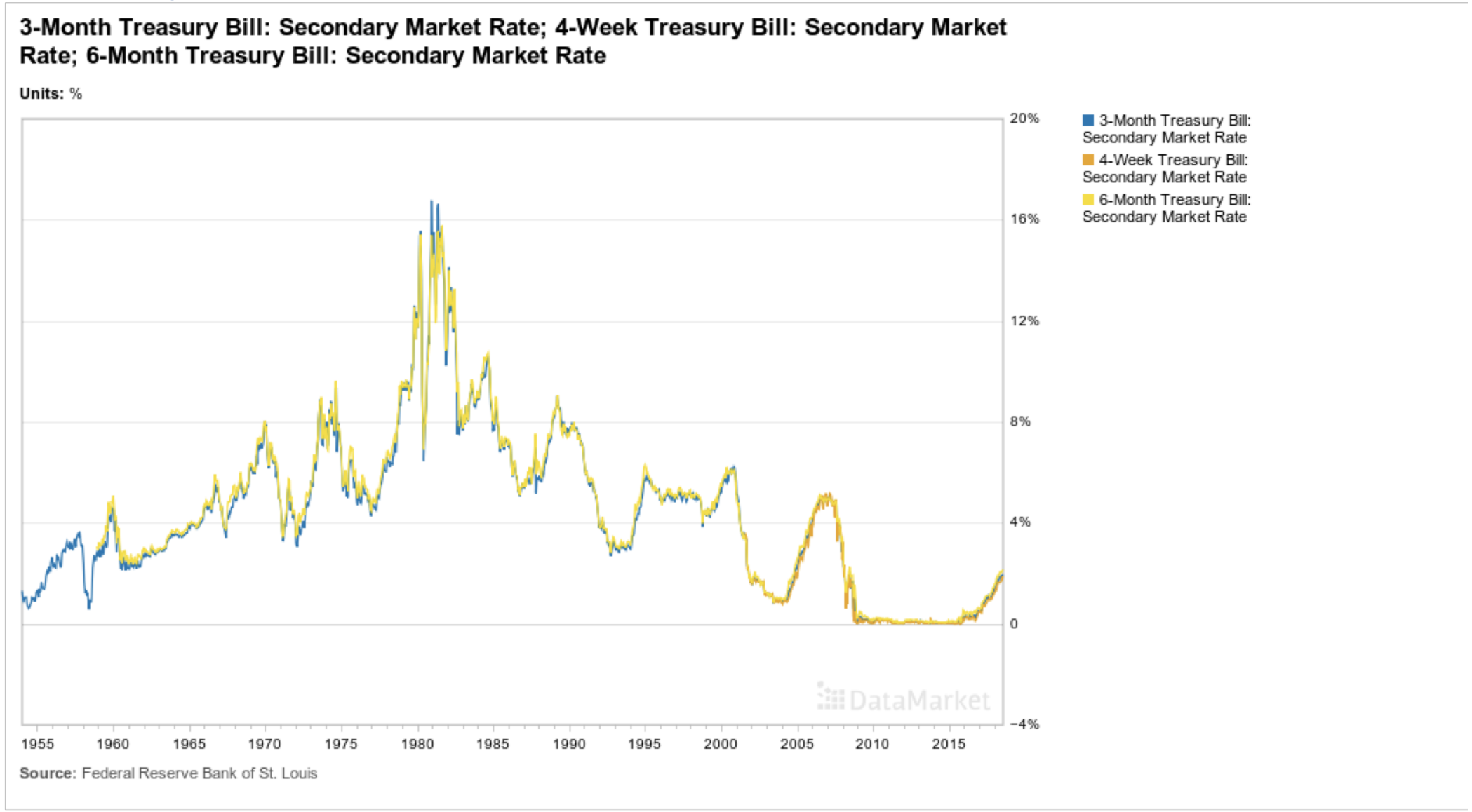
S&P monthly returns



TS from cosine



Monthly T-bill rate, 3 and 6 month



Formal definition

- The time series is denoted by $\{X_t, t \in \mathcal{T}\}$, where $\mathcal{T}$ is the time index
- If $\mathcal{T}$ is continuous, we have a continuous time series
- If $\mathcal{T}$ is discrete, we have a discrete time series. For simplicity, we will drop the index set and write $\{X_t\}$
- A realization of $\{X_t\}$ will be denoted by $\{x_t\}$ or $\{x_1, x_2, x_3, \ldots\}$ to indicate that they are observations
- In practice, the time interval for collection of time series could be: seconds, minutes, hours, days, weeks, months, years or function of these, or any reasonable regular time intervals

Objective of Time Series analysis

- Modeling or model building
  - Set up a model to represent the data generating mechanism
- Estimation
  - Estimate the parameters of the model
- Model checking
  - Check the goodness of fit of the model to the data
- Forecasting
  - Predict the future price of a share of a given stock
- Intervention
  - Understand the impact of intervention

Descriptive statistics

- Measures of central tendency: mean, median, …

Month	Sales	Month	Sales	Month	Sales
1	3	10	8	19	5
2	4	11	1	20	7
3	5	12	13	21	4
4	1	13	4	22	5
5	5	14	4	23	2
6	3	15	7	24	6
7	6	16	3	25	4
8	2	17	4		
9	7	18	2		

- Median = 4, Mean= 4.6

Dispersion, variance and risk

- You have two choices, which one would you choose?
  - Get 1000$ with probability 1
  - Get 1 million with prob. 1/1000, and 0 otherwise
- Stock return
  - Stock A: 3% return with prob. 0.7; -1% return with prob. 0.3: Mean=1.8%, std=1.83
  - Stock B: 6% return with prob. 0.7; -8% return with prob. 0.3: Mean=1.8%, std=6.42
  - Stock C: 45.42% return with prob. 0.7; -100% return with prob. 0.3: Mean=1.8%, std=66.64

Basic probability

- Joint, marginal and conditional

$$P(X,Y), P(X), P(X|Y)$$

- Bayes Theorem

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- Independence, dependence, mutual exclusive, correlation, correlation in time domain
- Dependence, causality, prediction. Even without a causal relationship, the association can still be used for prediction.

Linear combination of r.v.

$$E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2).$$

$$VAR(a_1X_1 + a_2X_2) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\sigma_1\sigma_2\rho$$

$$E\left(\sum_{i=1}^{n} a_iX_i\right) = \left(\sum_{i=1}^{n} a_iE(X_i)\right)$$

$$VAR\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_ia_j\sigma_i\sigma_j\rho$$

Confidence interval

- Sample statistics is a point estimate for the population parameter.
- It is very likely to be wrong, so we use interval estimate.
- Consider a large bank that wants to estimate the average amount owed by delinquent debtors μ. A random sample of size 100 is selected and found the sample mean is $230. Suppose it's known that the standard deviation of the amount owed for all delinquent accounts is σ = 90$. Get a 95% confidence interval for μ.

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = 230 \pm 17.64$$

Confidence interval

- Assume the sample size is 10, and the sampled population is approximately normal. If we only know the sample std, and the value is 80. We have student t with n-1 degree of freedom.

$$\bar{x} \pm t_{10-1,0.025}\sigma_{\bar{x}} = 230 \pm 2.262 \times 80/\sqrt{10} = 230 \pm 57.225$$

- Note that t-distribution comes with a heavier tail

Hypothesis testing

- Concepts of hypothesis testing, p-value, χ^2 test and F-test.
  type I, type II errors.
- Suppose a portfolio manager is interested in whether the variance of the daily return of a specific index is greater than 0.00015. Thus we are testing:

$$H_0 : 0 < \sigma^2 \leq 0.00015 \quad v.s. \quad \sigma^2 > 0.00015$$

The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Suppose in the past 50 days, the sample variance of the index is 0.0001, thus the test statistic is 32.67. Do not reject the null hypothesis.

The simple linear regression

The regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is usually assumed to have mean 0, standard deviation σ.

Given a random sample (X_i, Y_i), we have

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

with ϵ_i i.i.d. with mean 0 and std σ.

- Y_i follows a distribution with mean $\beta_0 + \beta_1 X_i$ and std σ
- The mean of Y lies on a straight line of X
- Variance of Y is constant across X
- The slope β_1 is the amount of increase in the mean of Y when X increased by one unit

The simple linear regression

- The parameters β_0 and β_1 can be estimated based on the least square criterion given data (x_i, y_i)

$$(\hat{\beta}_0, \hat{\beta}_1) = argmin \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

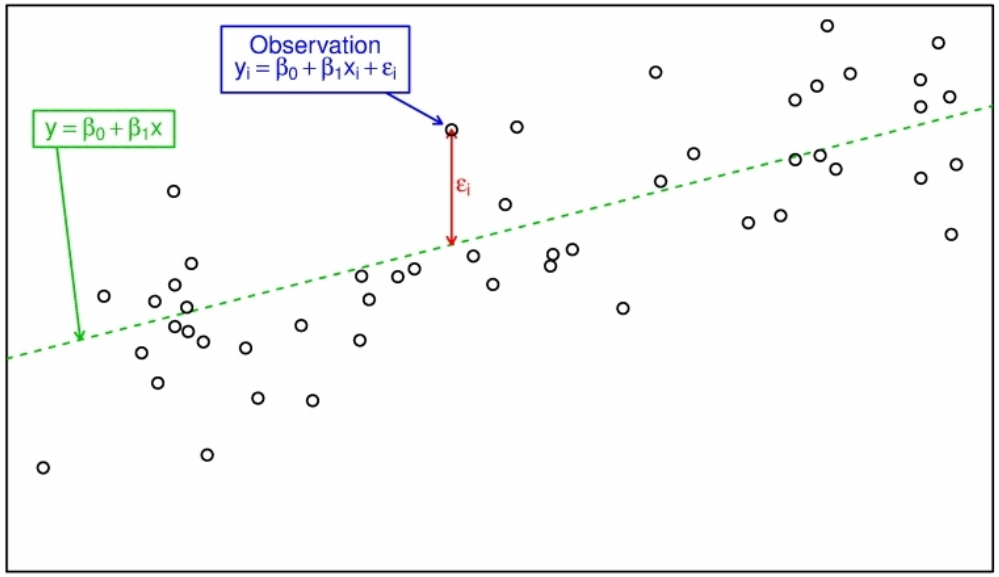
- The estimators $(\hat{\beta}_0, \hat{\beta}_1)$ are random variables with certain sampling distribution. Based on these we can perform hypothesis testing and compute confidence intervals of the parameters.
- Sampling distribution, test statistics, confidence intervals involved in simple regression.

The simple linear regression

- Consider the following simple regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

where x is the predictor variable and y is the response variable

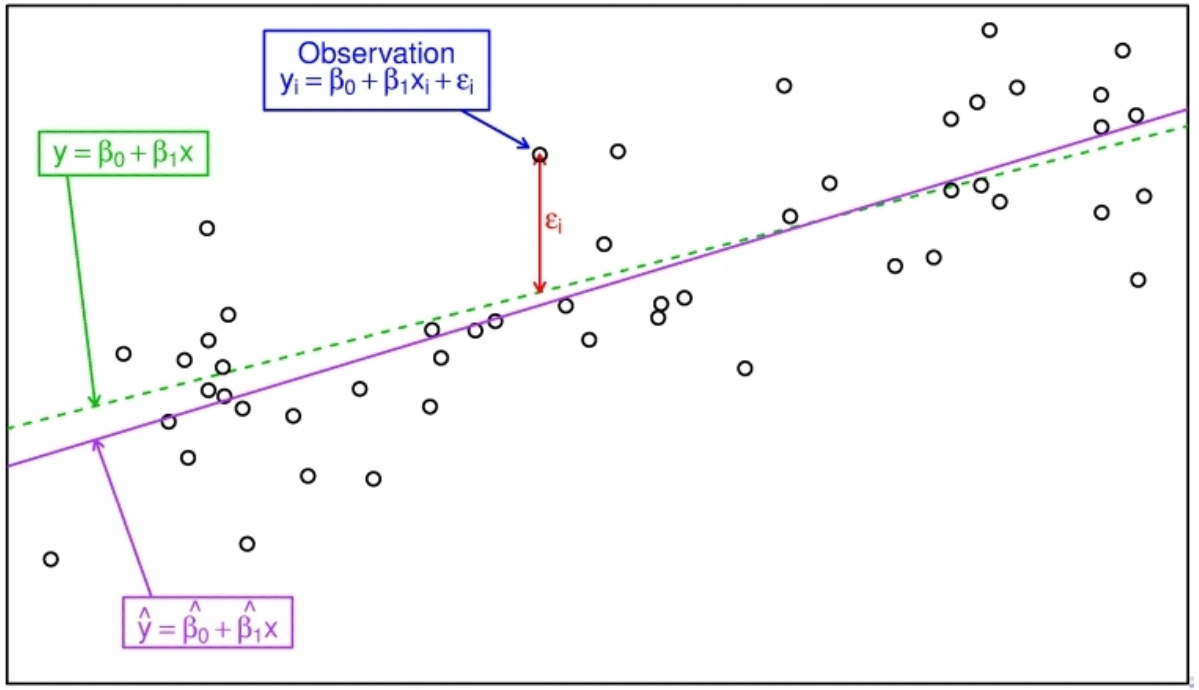


- The errors have mean 0, are uncorrelated and uncorrelated to the predictor variable

Least square estimate

- How do we define the regression line? What is ``best”?
- Minimize the sum of the squared errors

$$\sum_{i=1}^{n} (y - \beta_0 - \beta_1 x)^2$$



Estimates and residuals

- The true line

$$y = \beta_0 + \beta_1 x$$

- The fitted line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Thus for each individual x_i, we obtain the estimate $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \dots, n$.

- Residual is defined as $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. It is used to estimate the unknown ϵ.
- The residuals are centered around 0, and the correlation with the observations is 0

$$\sum_{i=1}^{n} e_i = 0 \text{ and } \sum_{i=1}^{n} x_i e_i = 0$$

Correlation coefficients and regression

- Recall the correlation coefficient

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

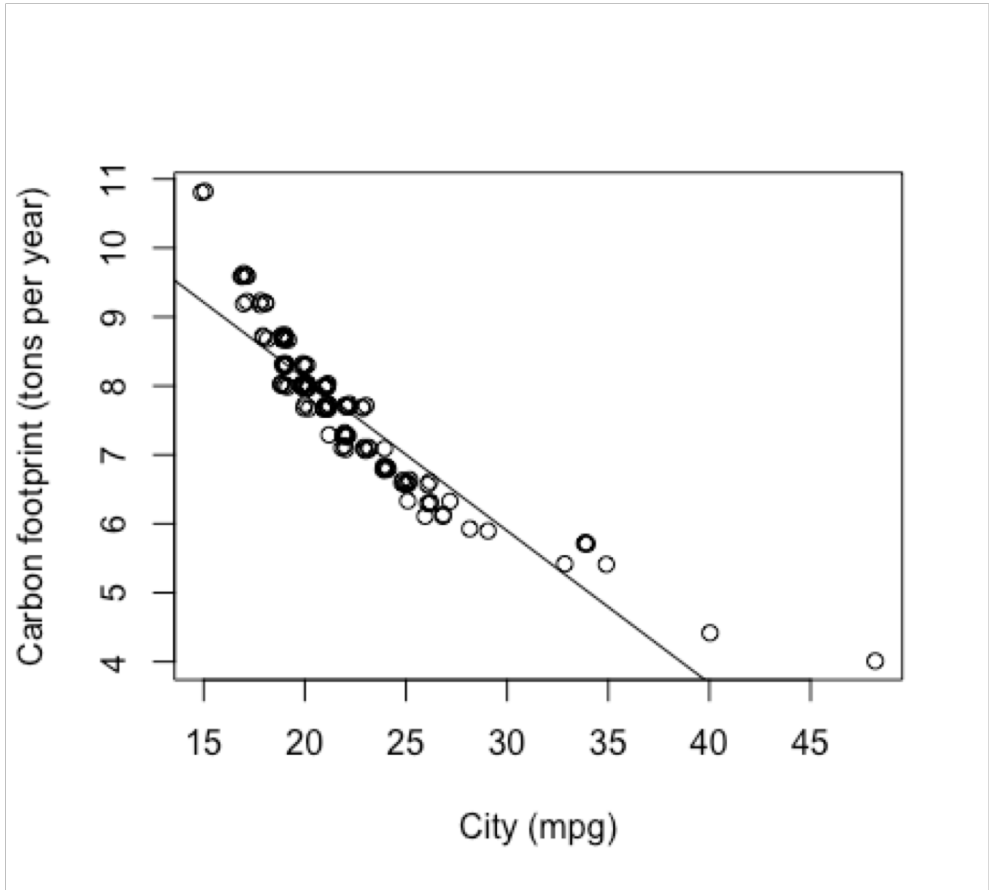
- The slope coefficient $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

where s_x and s_y are the standard deviation of the x and y observations respectively

Carbon footprint

- This is a regression between city based fuel economy (mpg) and the carbon footprint of 134 different car models



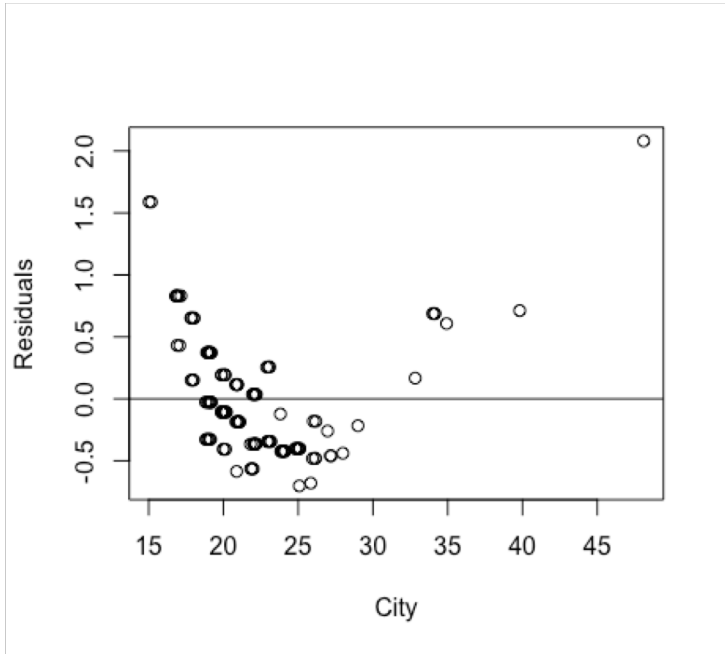
R code and outputs

- plot(jitter(Carbon) ~ jitter(City),xlab="City (mpg)",ylab="Carbon footprint (tons per year)",data=fuel)
- fit<-lm(Carbon~City,data=fuel)
- abline(fit)
- summary(fit)

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.7014 -0.3643 -0.1062  0.1938  2.0809
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.525647   0.199232   62.87   <2e-16 ***
City        -0.220970   0.008878  -24.89   <2e-16 ***
Residual standard error: 0.4703 on 132 degrees of freedom
Multiple R-squared:  0.8244,            Adjusted R-squared:  0.823
F-statistic: 619.5 on 1 and 132 DF,  p-value: < 2.2e-16
```

Residual analysis

- We expect the residuals to scatter around 0 and do not show systematic patterns.
- res<-residuals(fit)
- plot(jitter(res) ~ jitter(City),xlab="City",ylab="Residuals",data=fuel)
- abline(0,0)



Goodness of fit

- The concept of R^2: the proportion of variation in the forecast variable that is accounted for (or explained) by the regression model.
- A high R^2 does not always indicate a good model for estimation and forecasting.
- For instance, in the car example, R^2 = 82%, which is quite high. But from the residual analysis, we know that the linear regression model is not a good fit for the data.
- For simple regression, the R^2 equals the square of the correlation coefficient between x and y.

Residual sum of square

- SS residual

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

- The standard error is related to the size of the average error that the model produces
- This quantity is scale dependent. It’s also used for generating forecasting intervals

Forecasting

- Forecasts from a simple regression model for a specific 'new' x:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The prediction interval for this forecast is

$$\hat{y} \pm z_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}$$

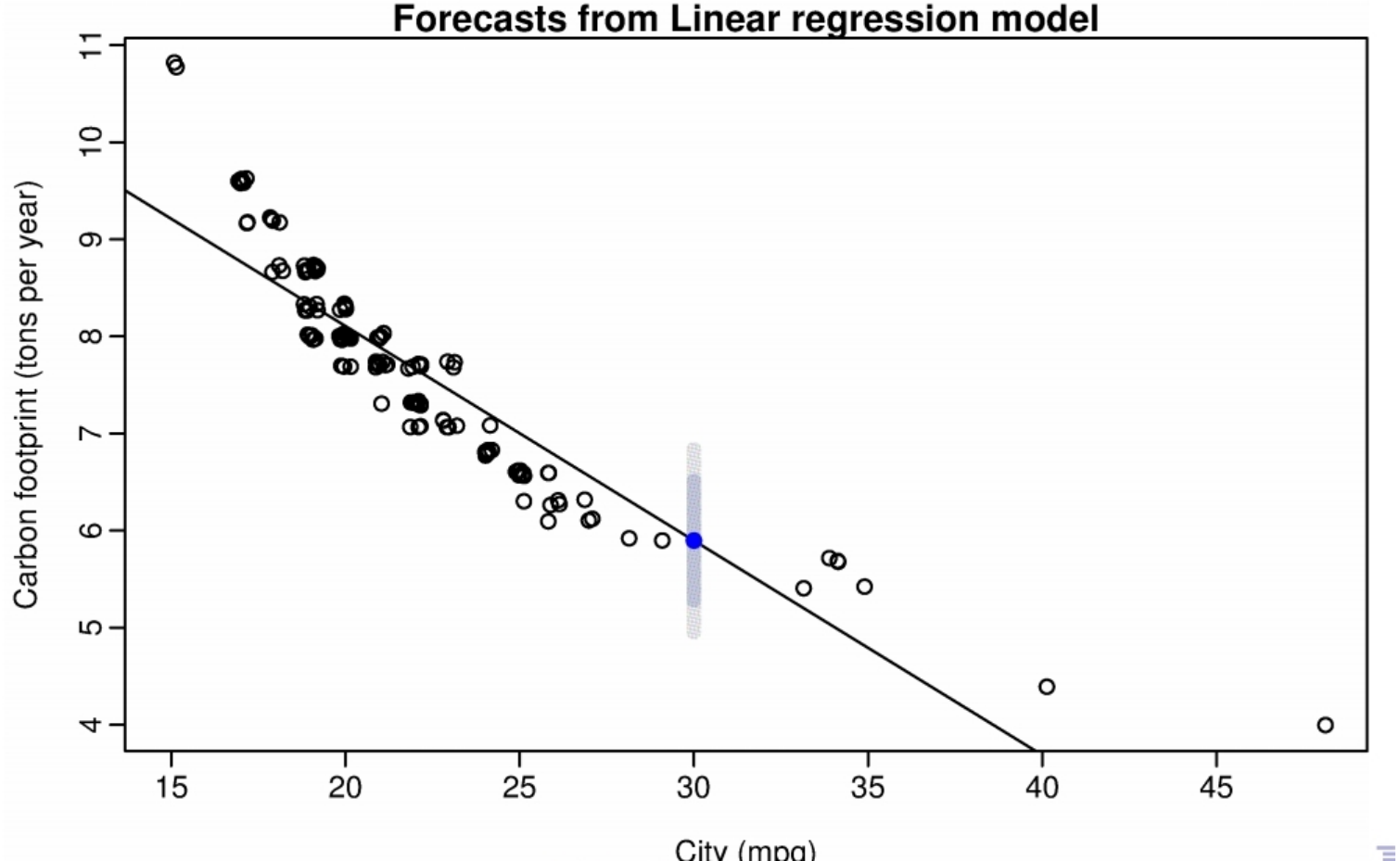
- The estimated regression line for the car example is

$$\hat{y} = 12.53 - 0.22x$$

- For a new car model with city mpg 30, the forecast carbon footprint is $\hat{y} = 5.9$ tons of CO2/year. We can also compute the corresponding forecasting intervals.

Forecasting

- Forecast with 80% and 95% forecast intervals for a car with 30 city mpg



Inferences

- You may be interested in testing whether the variable x has had a significant effect on y
- If x and y are unrelated, then the slope parameter $\beta_1 = 0$. We can construct a test to see if it is plausible given the observed data.

$$H_0 : \beta_1 = 0$$

- It is also some times useful to provide an interval estimate for β_1 and β_0.

confint(fit,level=0.95)

	2.5 %	97.5 %
(Intercept)	12.1315464	12.9197478
City	-0.2385315	-0.2034092

Nonlinear model

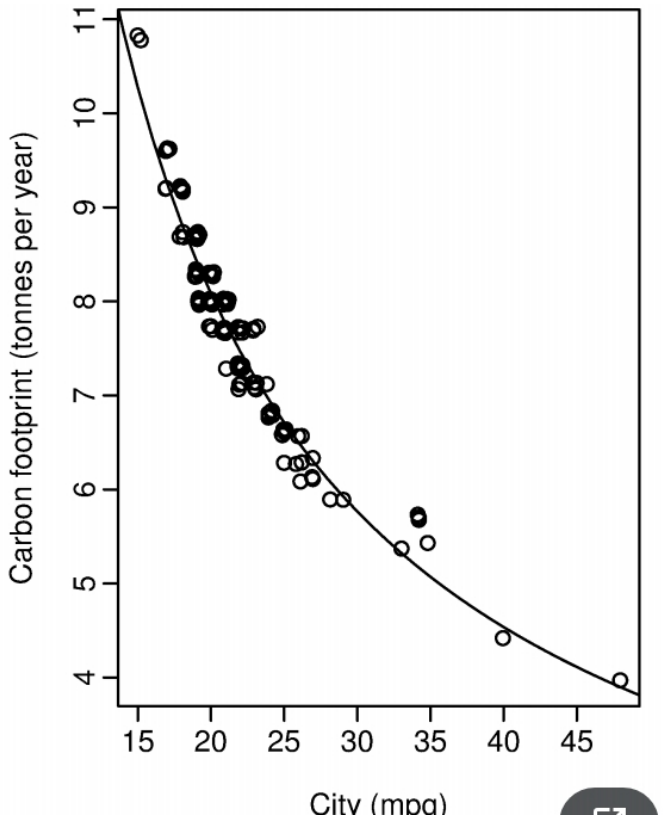
- One simple way to estimate a nonlinear model is to transform the variables
- The simplest way is log-log transform

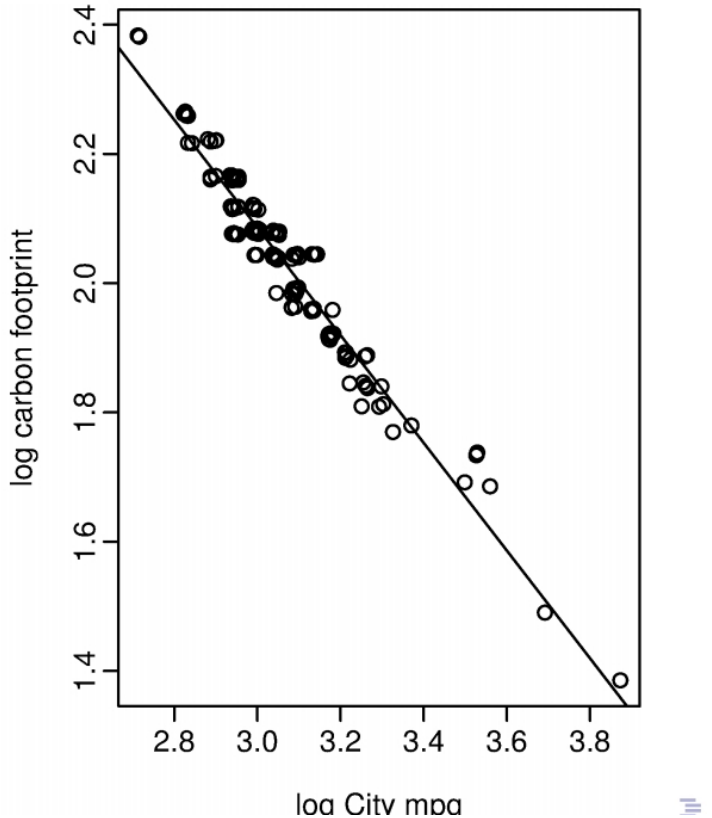
$$logy_i = \beta_0 + \beta_1 logx_i + \epsilon_i$$

- Interpretation: average percentage change in y resulting from 1% change in x
- Other forms: log-linear and linear-log

Car example

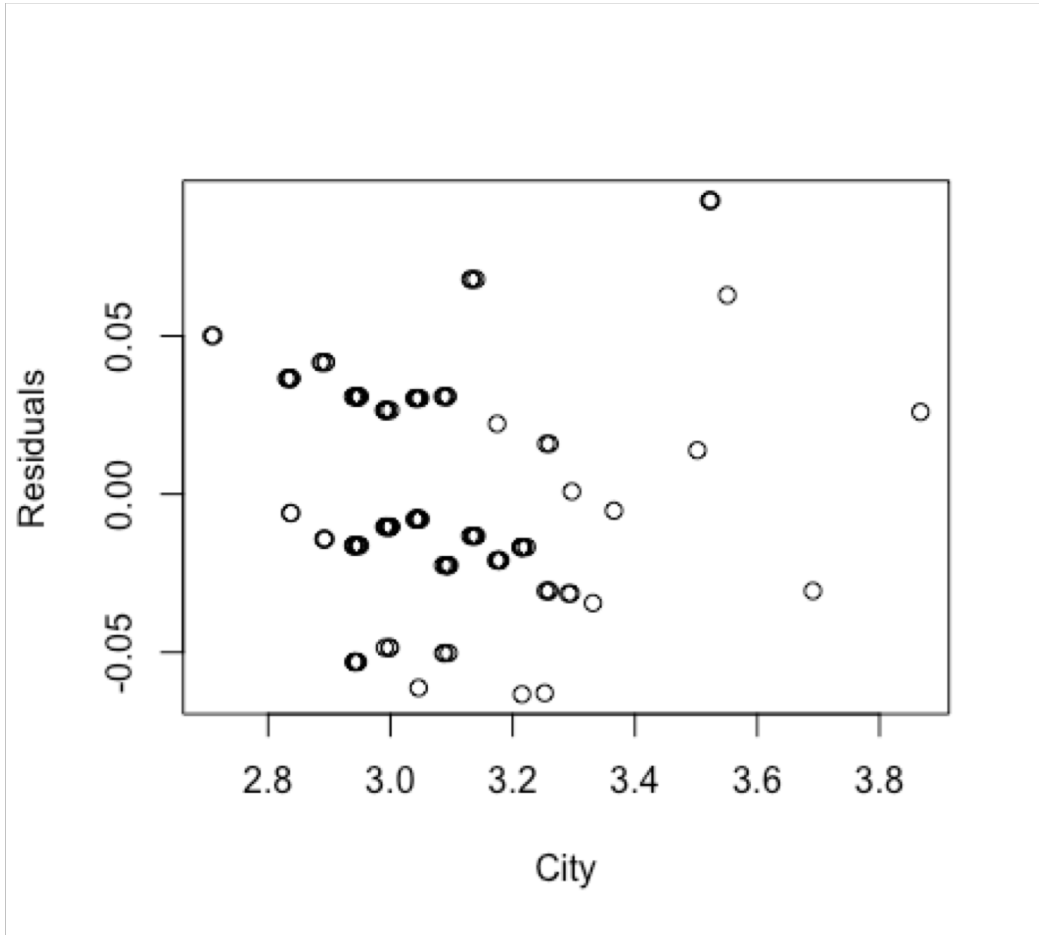
- Fitting of a log-log functional to the car data example





Car example

- Residual of the log-log fit



R code

- fit2<-lm(log(Carbon) ~ log(City), data = fuel)
- plot(jitter(log(Carbon)) ~ jitter(log(City)),xlab="City (mpg)",ylab="Carbon footprint (tons per year)",data=fuel)
- abline(fit2)
- summary(fit2)
- res2<-residuals(fit2)
- plot(jitter(res2) ~jitter(log(City)),xlab="City",ylab="Residuals",data=fuel)