01 Statistical data and their models

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

Statistical Data

The most common economic data comes in thee types:

- Cross-sectional data
- Time series data
- Panel data

Cross-sectional data

This type of data is characterized by individual units (e.g. companies, people, countries).

- The ordering of data points usually does not matter;
- Observations run from unit $i = 1$ to N;
- Data notations for the pair (X_i, Y_i) are used to indicate an observation for the i-th individual.

```
data(Salaries, package = "car")
head(Salaries)
```

rank	discipline	yrs.since.phd	yrs.service	sex	salary
Prof	B	19	18	Male	139750
Prof	B	20	16	Male	173200
AsstProf	B	4	3	Male	79750
Prof	B	45	39	Male	115000
Prof	B	40	41	Male	141500
AssocProf	B	6	6	Male	97000

In many cases a researcher is interested in establishing a relationship between two or more cross-sectional variables. For example, it may be interesting to check whether Y, the nine-month academic salary, depends on variables like years of service - we would expect that longer years of service would increase the salary. To digitize this belief one can create a regression model:

$$salary_i = \alpha + \beta \cdot yrs.service_i + \epsilon_i$$

We want to choose the estimates of α and β to minimize the distance between the data points and the fitted value - we want to minimize the error sum of squares:

$$SSE = \sum_{i=1}^{N} (salary_i - (\alpha + \beta \cdot yrs.service_i))^2$$

```
mdl <- lm(salary ~ yrs.service, data = Salaries)
summary(mdl)
```

```
##
## Call:
## lm(formula = salary ~ yrs.service, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81933 -20511  -3776  16417 101947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  99974.7     2416.6   41.37  < 2e-16 ***
## yrs.service    779.6      110.4    7.06 7.53e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 28580 on 395 degrees of freedom
## Multiple R-squared:  0.1121, Adjusted R-squared:  0.1098
## F-statistic: 49.85 on 1 and 395 DF,  p-value: 7.529e-12
```

After estimating our coefficients:

$$\widehat{salary_i} = 99974.7 + 779.6 \cdot yrs.service_i$$

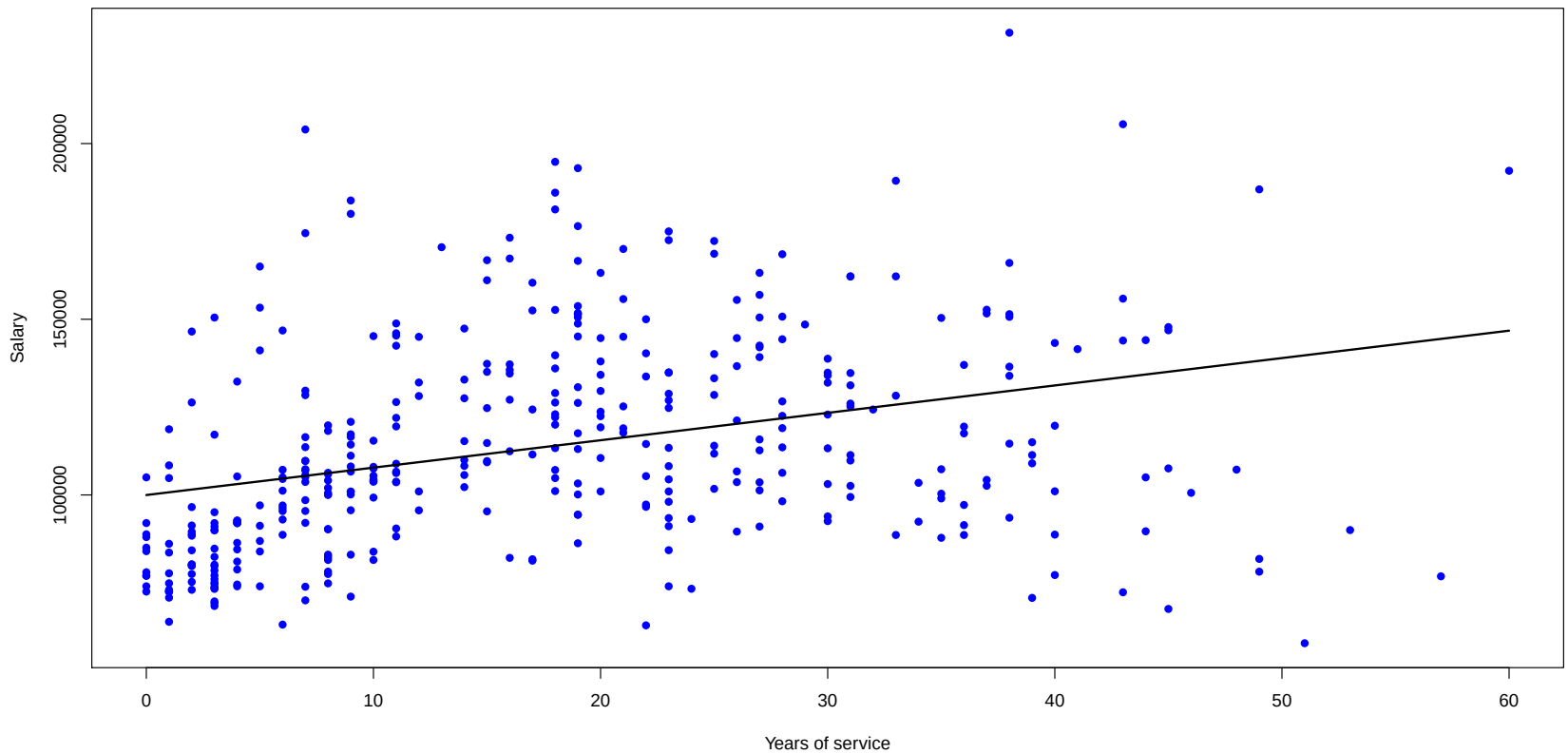
In this case it means that for 1 additional years of service the *average* nine-month academic salary increases by 779.6 dollars. The intercept, α, can **only** be interpreted if $yrs.service_i$ can be zero. Let's check:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   16.00   17.61   27.00   60.00
```

Because $yrs.service_i$ can attain a zero value, we can interpret it that 99974.7 dollars is the average salary of someone starting in academia without any previous experience.

In case that $yrs.service_i$ never attains a zero value, the intercept has no meaningful interpretation. The same applies if any other predictor variables, besides $yrs.service_i$ are included in the regression - if at least one of them never attains a zero value, then the intercept has no meaningful interpretation.

The constant term is in part estimated by the omission of predictors from a regression analysis. In essence, it serves as a garbage bin for any bias that is not accounted for by the terms in the model. The constant guarantees that the residuals don't have an overall positive or negative bias, but also makes it harder to interpret the value of the constant because it absorbs the bias - it guarantees that the model residuals have a mean of zero.



However, salary cannot be explained by years of experience alone - we can extend the model by including additional explanatory variables - we will estimate a multiple regression model:

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   65955.2324  4588.6009 14.373713 6.810626e-38
## rankAssocProf 12907.5879  4145.2783  3.113805 1.983251e-03
## rankProf      45065.9987  4237.5233 10.634985 2.296130e-23
## disciplineB   14417.6256  2342.8753  6.153817 1.878412e-09
## yrs.since.phd   535.0583   240.9941  2.220213 2.697855e-02
## yrs.service    -489.5157   211.9376 -2.309717 2.142543e-02
## sexMale        4783.4928  3858.6684  1.239675 2.158412e-01
```

We are testing the hypothesis $H_0 : \beta_j = 0$, thus the t-statistic $t = \frac{\widehat{\beta}_j}{std.err\beta_j}$. If $p - value < 0.05$ (i.e. if $t - statistic$ is 'big'), we reject the null hypothesis and X_j influences the dependent variable, $yrs.service_i$. The rule of thumb says - if the modulus of t exceeds 2, **reject** H_0.

We remove insignificant coefficients and re-estimate the model:

```
mdl <- lm(salary ~ rank + discipline +
            yrs.since.phd + yrs.service, data = Salaries)
summary(mdl)$coefficients
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   69869.0110  3332.1448 20.968180 5.828862e-66
## rankAssocProf 12831.5375  4147.6685  3.093675 2.118982e-03
## rankProf      45287.6890  4236.6534 10.689496 1.436047e-23
## disciplineB   14505.1514  2343.4181  6.189741 1.523745e-09
## yrs.since.phd   534.6313   241.1593  2.216922 2.720308e-02
## yrs.service    -476.7179   211.8312 -2.250461 2.497485e-02
```

We note that from our model, having **more** experience **lowers** the salary but the more time since PhD - the higher the salary.

Sometimes explanatory variables are tightly connected (e.g. linear relationship) and it is impossible to disentangle the *individual* influences of explanatory variables. A popular measure of multicollinearity is the Variance Inflation Factor (VIF): $VIF = \frac{1}{1 - R_k^2}$, where R_k^2 is the R^2 from regressing the variable x_k on all the remaining regressors.

```
mdl <- lm(salary ~ ., data = Salaries)
car::vif(mdl)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## rank           2.013193  2        1.191163
## discipline     1.064105  1        1.031555
## yrs.since.phd  7.518936  1        2.742068
## yrs.service    5.923038  1        2.433729
## sex            1.030805  1        1.015285
```

The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction. We can check that the correlation between *yrs.service* and *yrs.since.phd* is `0.9096491`, so we can try to remove either one of them and also include a polynomial term.

The resulting model includes *yrs.since.phd* with its quadratic term:

```
mdl <- lm(salary ~ rank + discipline
          + yrs.since.phd + I(yrs.since.phd^2), data = Salaries)
summary(mdl)$coefficients[, c(1,4)]
```

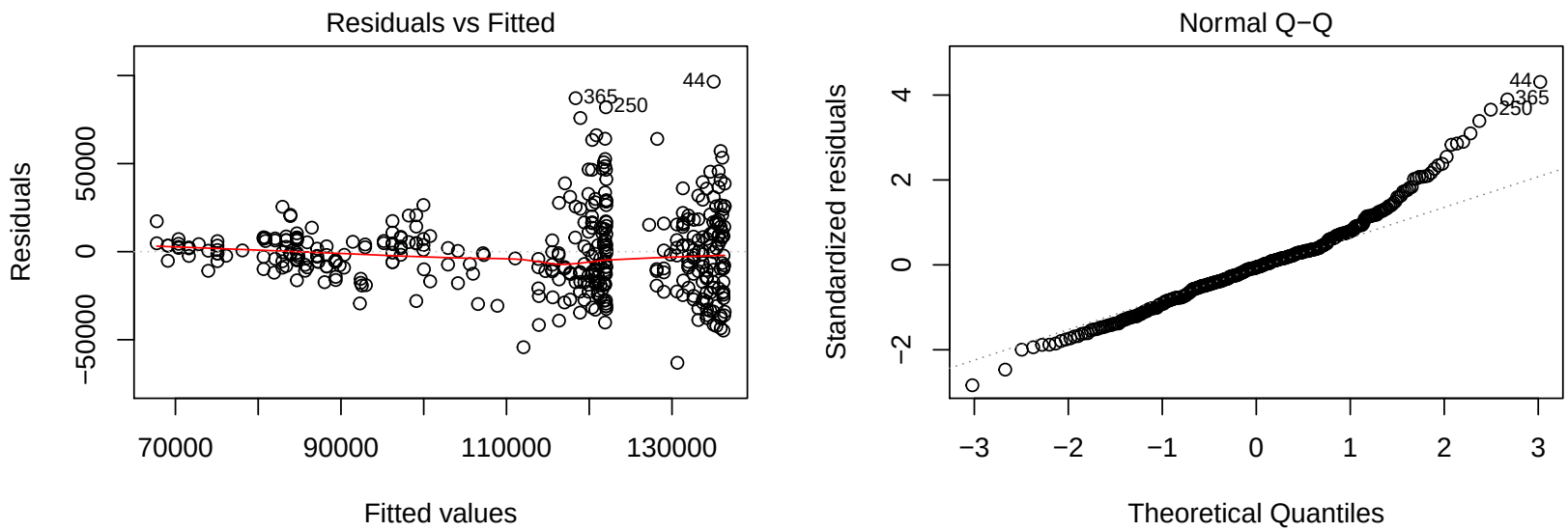
```
##                              Estimate     Pr(>|t|)
## (Intercept)              64919.67653 3.386603e-43
## rankAssocProf             5800.88482 2.500129e-01
## rankProf                 34848.87453 3.064268e-08
## disciplineB              14297.08049 2.153495e-09
## yrs.since.phd             1460.65638 1.006698e-02
## I(yrs.since.phd^2)         -23.91633 1.205955e-02
```

Note: multicollinearity inflates all the variances of $\widehat{\beta}_j$, thus deflates all respective $t - values$ and makes the coefficients **insignificant**. One of the solutions is to use the F-test: if seemingly insignificant parameters are truly zero, the F-test **should not reject** the joint null hypothesis: $H_0 : \beta_i = 0, \beta_j = 0$. If it rejects H_0, we have an indication that the low $t - values$ are due to multicollinearity.

Our final model states that the nine-month academic salary depends on the academic rank as well as the discipline as well as the years since PhD with its quadratic form. While the coefficient of *yrs.since.phd* is positive, the coefficient of $yrs.since.phd^2$ is negative, which indicates that as a person gets older, the effect of *yrs.since.phd* is lessened.

We now test the residuals of our model. The model form is accepted as correct if there are no changes in the variance of the residuals (residual variance must be constant) and there is no pattern to the residuals with respect to the predicted values.

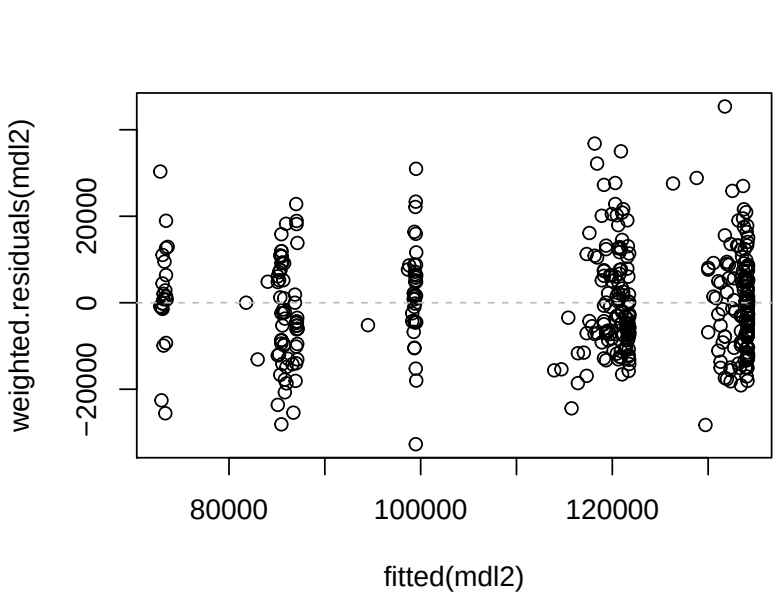
```
mdl <- lm(salary ~ rank + discipline
          + yrs.since.phd + I(yrs.since.phd^2),
          data = Salaries)
par(mfrow = c(1,2))
plot(mdl, which= c(1, 2))
```

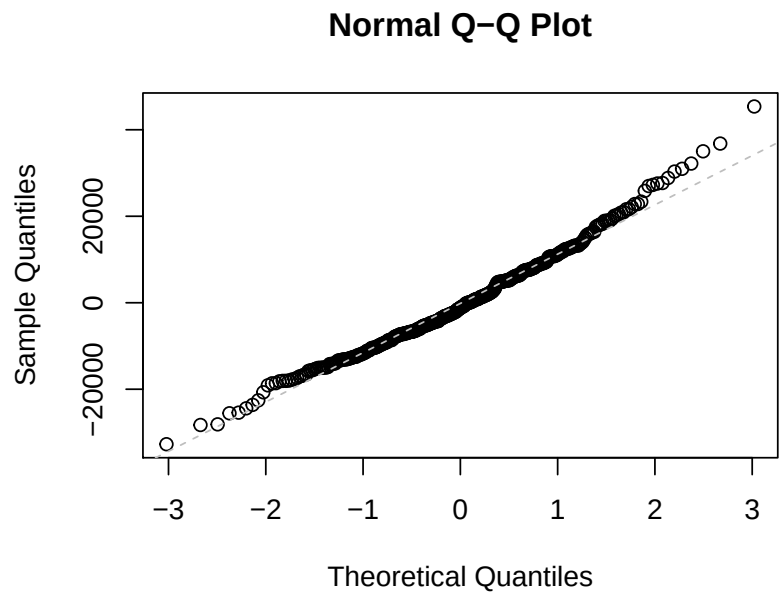


The residual plot shows increasing residual variance with increased value of predicted salary. We need to correct the model form. We will use the assumption that there is a known variance for each rank and discipline groups.

```
wGrp <- aggregate(list(var = Salaries$salary),
          by = list(rank       = Salaries$rank,
                    discipline = Salaries$discipline),
          FUN = var
          )
SalariesW      <- merge(Salaries, wGrp)
SalariesW$wght <- (1 / SalariesW$var) /
                          mean((1/SalariesW$var))
#Weighted Least Squares (minimizing sum(w*e^2))
mdl2 <- lm(salary ~ rank + discipline,
           weights = wght,
           data = SalariesW)
summary(mdl2)$coefficients
```

```
##                          Estimate      Pr(>|t|)
## (Intercept)          72553.795441 2.617142e-149
## rankAssocProf        13453.649288  2.183407e-09
## rankProf             48070.051889  5.024442e-38
## disciplineB          12397.584733  1.543305e-21
## yrs.since.phd          147.822620  6.117016e-01
## I(yrs.since.phd^2)      -4.773308  3.839775e-01
```



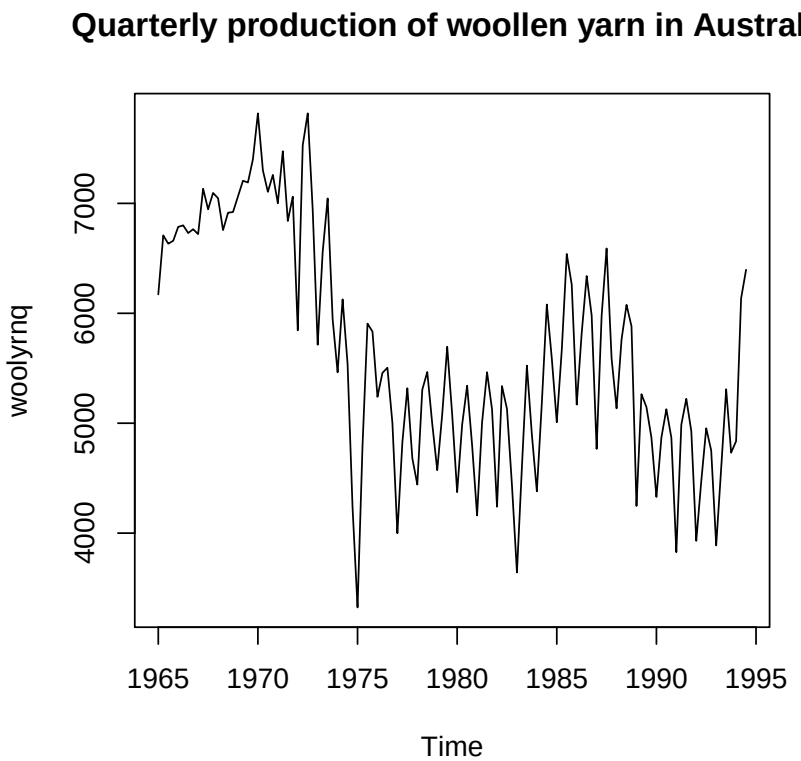


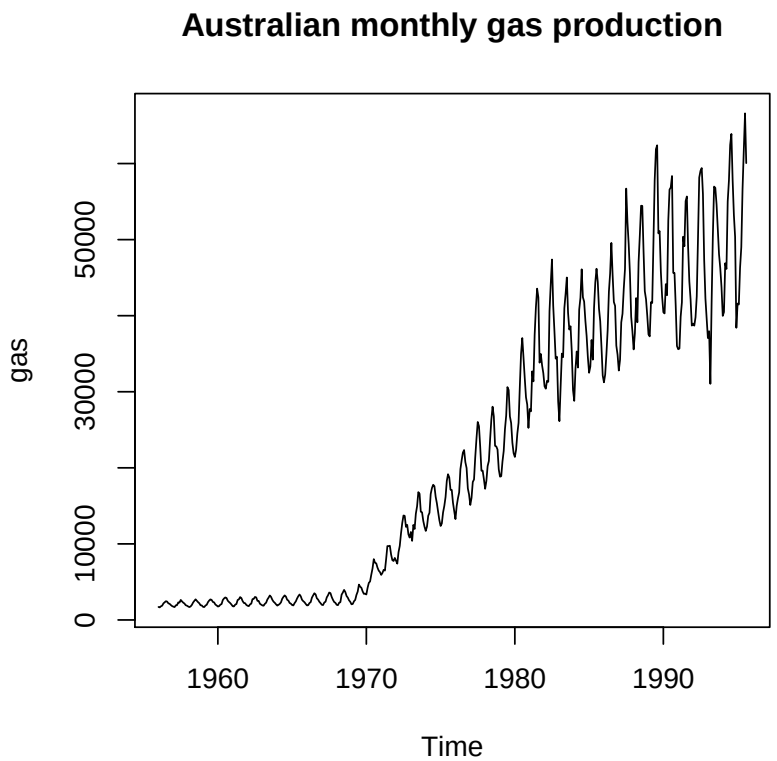
We see that the weighted residual variance is more consistent. From the quantile-quantile plot the residuals appear to be normal.

Time series data

A time series is a sequence of observations that are arranged according to the time of their outcome. Time series data can be observed at many frequencies: annual crop yield, quarterly financial reports, daily stock prices, hourly wind speeds, etc.

The characteristic property of a time series is the fact that the data are not generated independently, their dispersion varies in time, they are often governed by a trend and they might have cyclic components.





In this course, we will use the notation Y_t to indicate an observation on variables Y at time $t = 1, ..., T$.

One objective of analyzing economic data is to predict the future values of economic variables. One approach to do this is to build an econometric model, describing the relationship between the variable of interest and other economic quantities, then estimate the model using sample data and use it as a basis for forecasting. However, this approach is not always useful.

For example, it may be possible to adequately model the contemporaneous relationship between unemployment and the inflation rate, but as long as we cannot predict future inflation rates we are also unable to forecast future unemployment.

In the first part of this course we will follow a pure time series approach - we will assume that the current values of an economic variable are related to its past values only. The emphasis is purely on making use of the information in past values of a variable for forecasting its future. In addition to producing forecasts, time series models also produce the distribution of future values, conditional upon the past, and can thus be used to evaluate the likelihood of certain events.

In the second part of this course we shall get to know different variants of regressions with time series variables.

Finally, the most interesting results in econometrics are obtained in the intersection of cross-sectional and time series methods...

Panel data

A dataset containing observations on multiple phenomena observed over multiple time periods. Panel data aggregates all individuals and analyses them in a period of time. Whereas time series and cross-sectional data are one-dimensional, panel data sets are two dimensional.

```
require(plm)
#data(package = "plm")
data(Grunfeld)
```

firm	year	inv	value	capital
1	1937	410.6	5387.1	156.9
1	1938	257.7	2792.2	209.2
2	1937	469.9	2676.3	118.1
2	1938	262.3	1801.9	260.2
3	1937	77.2	2803.3	118.0
3	1938	44.6	2039.7	156.2

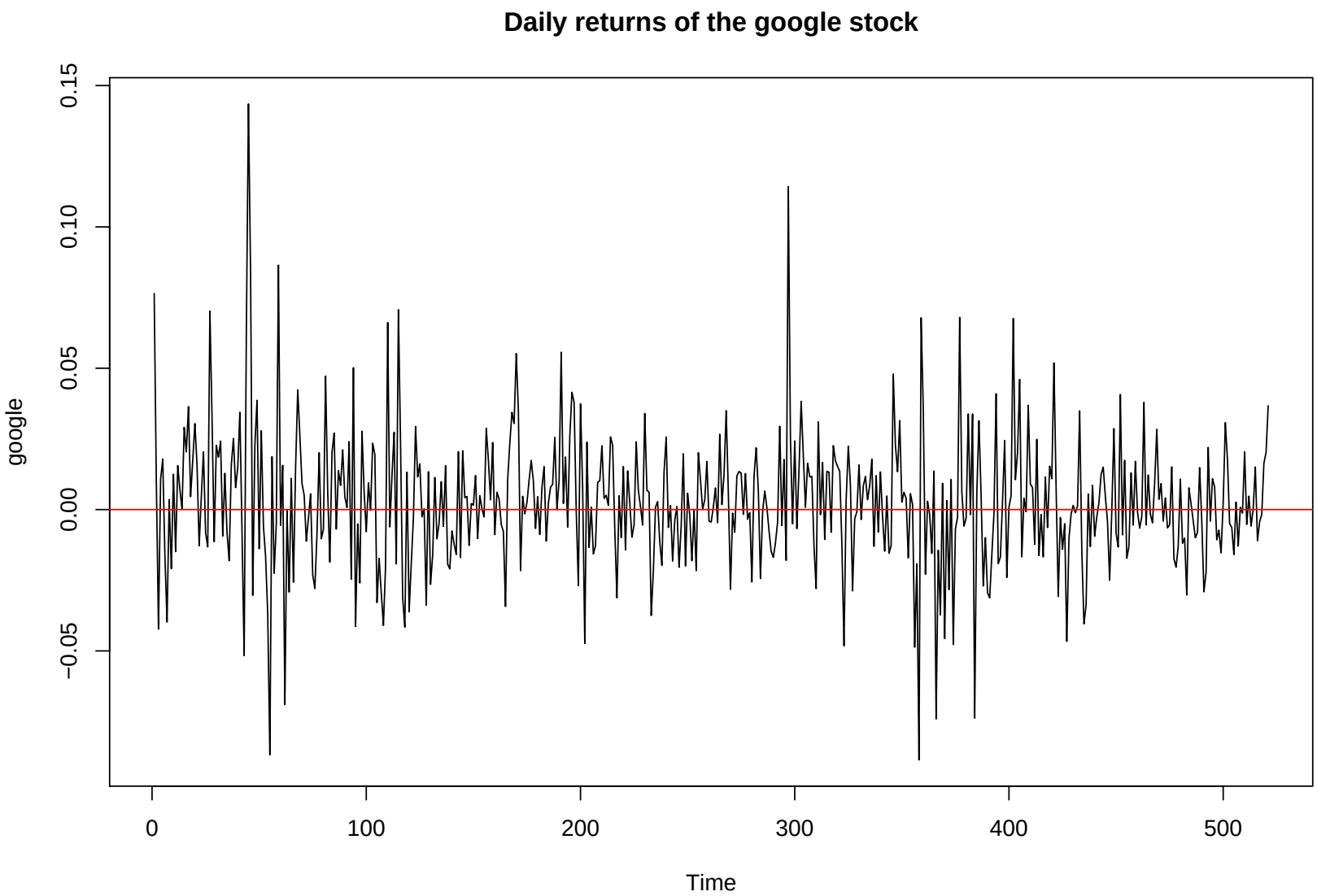
Time Series Examples

Stock returns

Let P_t be the price of an asset at time t. Then, the one-period return is:

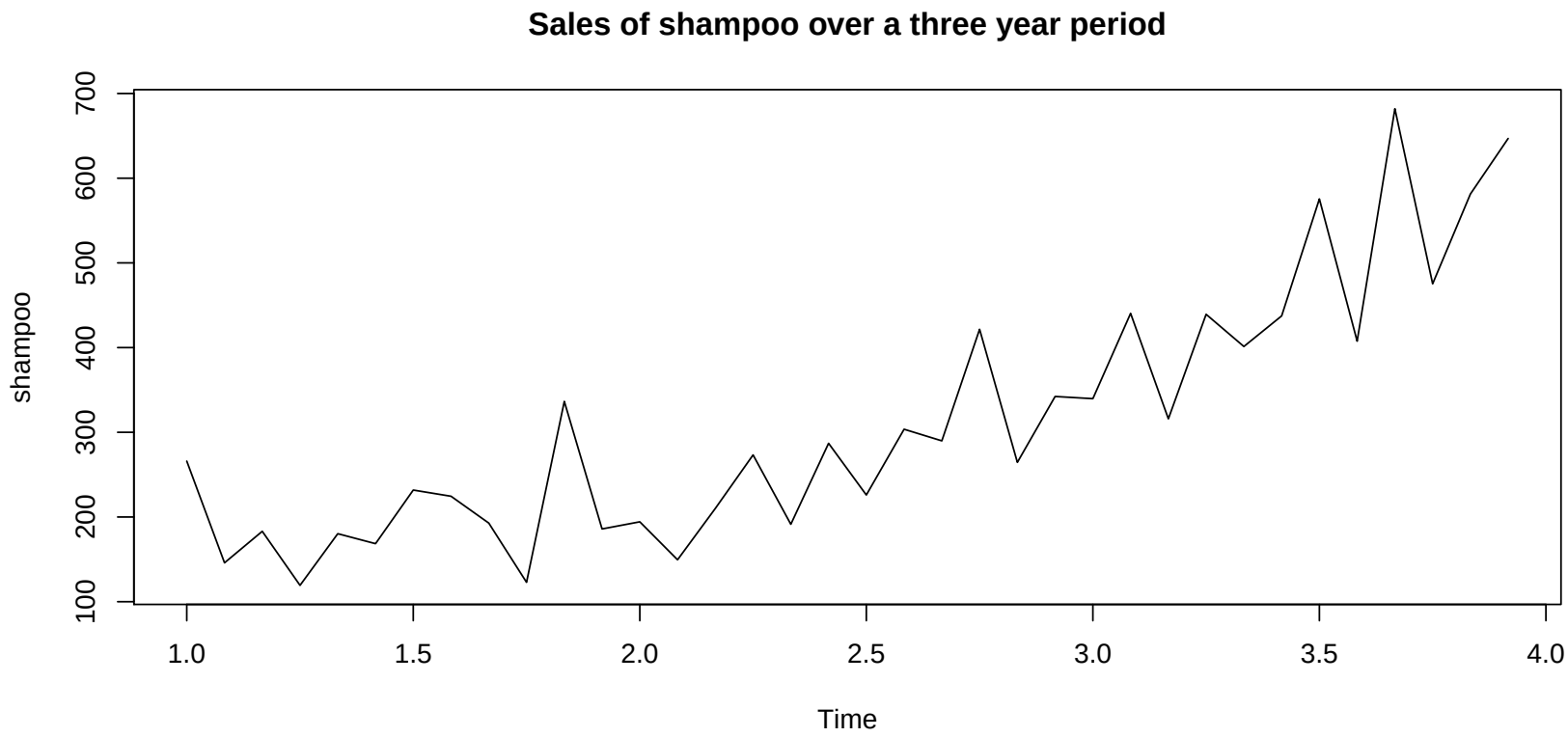
$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

```
suppressPackageStartupMessages({require("TSA")})
data("google")
plot.ts(google, main = "Daily returns of the google stock")
abline(0, 0, col = "red")
```


Daily returns of the google stock
google
0.15
0.10
0.05
0.00
−0.05
0
100
200
300
400
500
Time

Sales of shampoo

```
suppressPackageStartupMessages({require("fma")})
data(shampoo)
plot.ts(shampoo,
        main = "Sales of shampoo over a three year period")
```

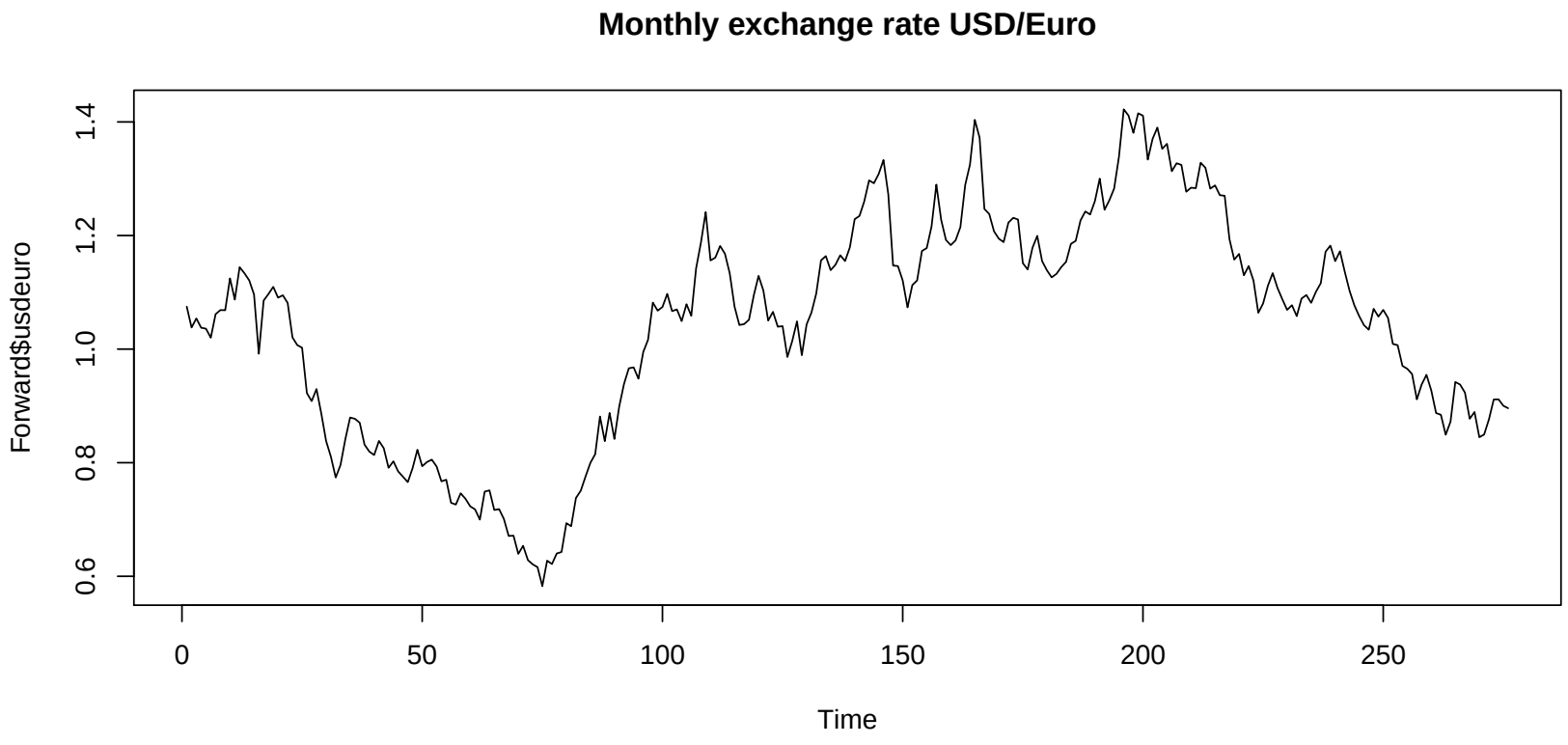


Air passenger numbers 1949 - 1960

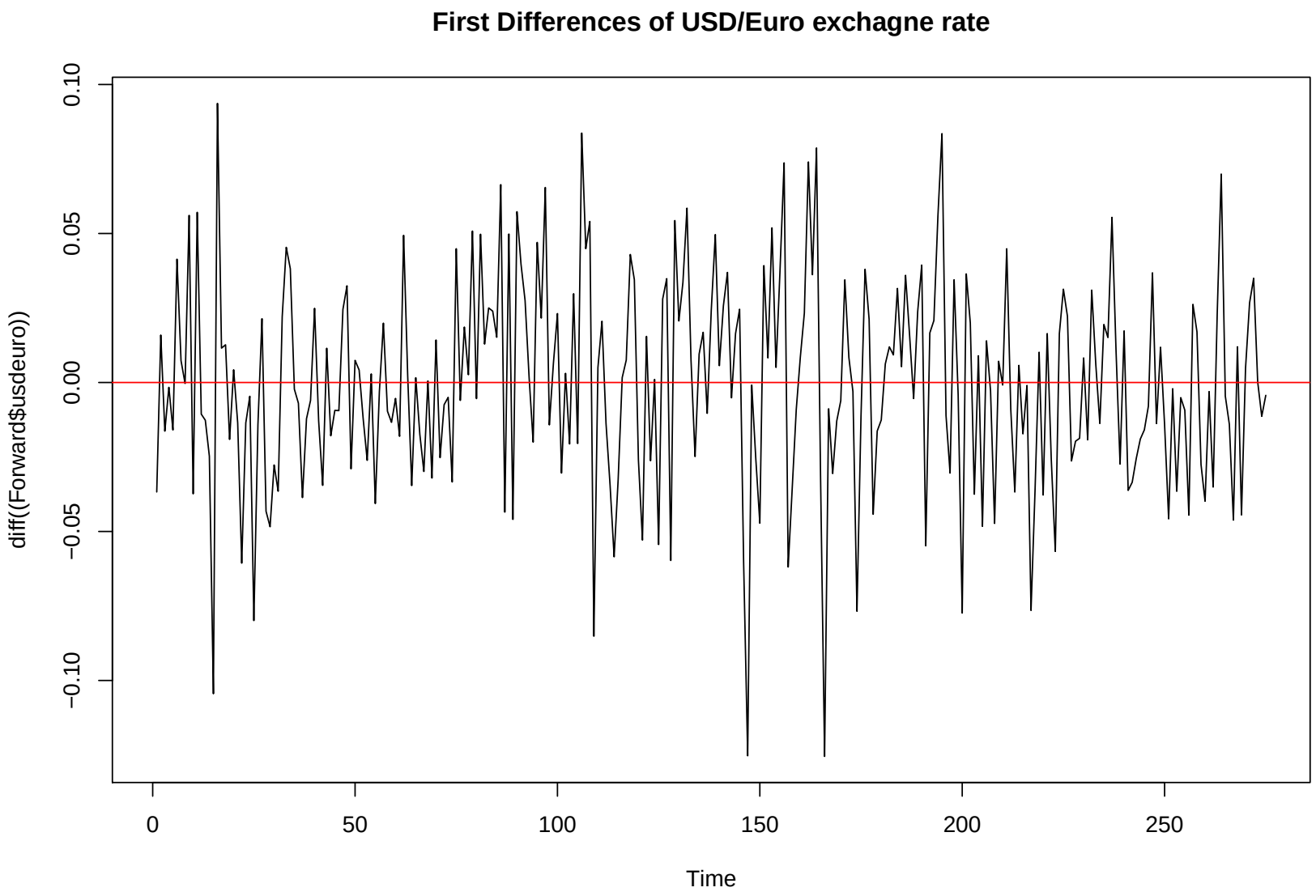
```
require("datasets")
data("AirPassengers")
plot.ts(AirPassengers,
        main = "Monthly totals of international airline passenge
```

Exchange rates

```
suppressPackageStartupMessages({require("Ecdat")})
data(Forward)
plot.ts(Forward$usdeuro,
        main = "Monthly exchange rate USD/Euro")
```



```
plot.ts(diff((Forward$usdeuro)),
        main = "First Differences of USD/Euro exchagne rate")
abline(0, 0, col = "red")
```



Forecasting

One of the main applications of time series theory is prediction. If we consider our examples, we can:

- Forecast the stock return using its mean;
- Forecast shampoo sales using its trend;
- Forecast air passenger numbers using its trend and a seasonal component;
- The trends of the USD/Euro exchange rates seem to change directions at unpredictable times. They are also known as stochastic (random) trends whereas some of the previous examples exhibit deterministic trends. A random walk or a Difference stationary time series can sometimes provide a good fit for this kind of data.