Handout 7. Regression and correlation

1. Simple linear regression

The linear regression model is given by

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

in which the ε_i are assumed independent and $N(0, \sigma^2)$, The parameters α, β, and σ^2 can be estimated using the *method of least squares*. In particular, the values of α and β can be obtained by minimizing the sum of squared residuals, and σ^2 can be estimated via the sum of squared residuals.

It is usually of prime interest to test the null hypothesis $\beta = 0$, for which we can use a *t* test.

```
library(ISwR)
data(thuesen)
attach(thuesen)
lm(short.velocity~blood.glucose)
summary(lm(short.velocity~blood.glucose))

plot(blood.glucose,short.velocity)
abline(lm(short.velocity~blood.glucose))
```

2. Residuals and fitted values

We have seen how *summary* can be used to extract information about the results of a regression analysis. Two further extraction functions are *fitted* and *resid*.

```
lm.velo<-lm(short.velocity~blood.glucose)
fitted(lm.velo)
resid(lm.velo)
plot(blood.glucose,short.velocity)
abline(lm.velo)

plot(blood.glucose,short.velocity)
lines(blood.glucose,fitted(lm.velo))

plot(blood.glucose,short.velocity)
lines(blood.glucose[!is.na(short.velocity)],fitted(lm.velo))

# blood.glucose[!is.na(short.velocity) & !is.na(blood.glucose)]
#lines(blood.glucose[!is.na(short.velocity) & !is.na(blood.glucose)],fitted(lm.velo))
```

```
options(na.action=na.exclude)
lm.velo<-lm(short.velocity~blood.glucose)
fitted(lm.velo)


  plot(blood.glucose,short.velocity)
  abline(lm.velo)
  segments(blood.glucose,fitted(lm.velo),blood.glucose,short.velocity)

  plot(fitted(lm.velo),resid(lm.velo))
  qqnorm(resid(lm.velo))
```

```
> logret<- read.table("~/Desktop/d_logret_6stocks.txt",
header=T)
> names(logret)
[1] "Date"     "Pfizer"    "Intel"     "Citigroup" "AmerExp"
[6] "Exxon"       "GenMotor"
> attach(logret)
> (fit1<-lm(Pfizer~Intel))

  Call:   lm(formula = Pfizer ~ Intel)

  Coefficients:
  (Intercept)         Intel
    -0.003903      0.023078

  > summary(fit1)

  Call:   lm(formula = Pfizer ~ Intel)

  Residuals:
         Min          1Q      Median          3Q         Max
  -0.0559197  -0.0138454   0.0008506   0.0172455   0.0456926

  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) -0.003903   0.002913  -1.340    0.185
  Intel        0.023078   0.043112   0.535    0.594

  Residual standard error: 0.02321 on 62 degrees of freedom
```

```
Multiple R-squared: 0.0046,         Adjusted R-squared: -0.01145
    F-statistic: 0.2865 on 1 and 62 DF,   p-value: 0.5944
```

```
> names(fit1)
[1] "coefficients"  "residuals"     "effects"     "rank"
[5] "fitted.values" "assign"        "qr"          "df.residual"
[9] "xlevels"       "call"          "terms"       "model"

> fit1$coeff
> plot(Intel, Pfizer)
> abline(lm(Pfizer~Intel))


> fit2<-lm(Pfizer~-1+Intel)    ### regression without intercept
> summary(fit2)


> fitted(fit1)
> resid(fit1)

> plot(Intel, Pfizer)
> lines(Intel, fitted(fit1))
```

3. Correlation

The function *cor* can be used to compute the correlation between two or more vectors.

```
attach(thuesen)
cor(blood.glucose,short.velocity)
cor(blood.glucose,short.velocity,use="complete.obs")
cor(thuesen,use="complete.obs")
cor.test(blood.glucose,short.velocity)
```

```
> cor(Intel, Pfizer)
> cor.test(Intel, Pfizer)
```