```
## (C) (cc by-sa) Wouter van Atteveldt, file generated mei 24 2016
```

> Note on the data used in this howto: This data can be downloaded from http://piketty.pse.ens.fr/files/capital21c/en/xls/, but the excel format is a bit difficult to parse at it is meant to be human readable, with multiple header rows etc. For that reason, I've extracted csv files for some interesting tables that I've uploaded to https://github.com/vanatteveldt/learningr/tree/master/data. If you're accessing this tutorial from the githup project, these files should be in your 'data' sub folder automatically.

Basic Modeling

In this hands-on we continue with the `capital` variable created in the transforming data howto. You can also download this variable from the course pages:

```r
load("data/capital.rdata")
head(capital)
```

```
##   Year   Country Public Private Total
## 1 1970 Australia   0.61    3.30  3.91
## 2 1970    Canada   0.37    2.47  2.84
## 3 1970    France   0.41    3.10  3.51
## 4 1970   Germany   0.88    2.25  3.13
## 5 1970     Italy   0.20    2.39  2.59
## 6 1970     Japan   0.61    2.99  3.60
```

T-tests

First, let's split our countries into two groups, anglo-saxon countries and european countries (plus Japan): We can use the `ifelse` command here combined with the `%in%` operator

```r
anglo = c("U.S.", "U.K.", "Canada", "Australia")
capital$Group = ifelse(capital$Country %in% anglo, "anglo", "european")
table(capital$Group)
```

```
## 
##    anglo european 
##      164      205
```

Now, let's see whether capital accumulation is different between these two groups.

```r
t.test(capital$Private ~ capital$Group)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  capital$Private by capital$Group
## t = -4.6664, df = 289.34, p-value = 4.692e-06
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.7775339 -0.3162154
## sample estimates:
##    mean in group anglo mean in group european
##               3.748232               4.295106
```

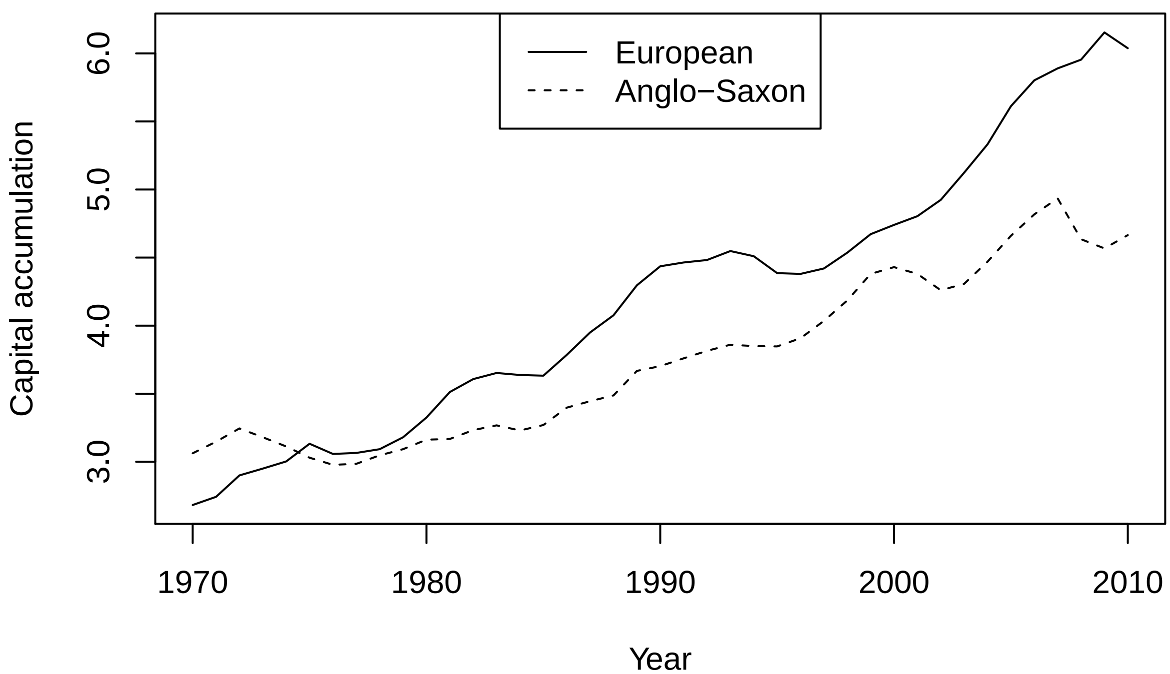
So, according to this test capital accumulation is indeed significantly higher in European countries than in Anglo-Saxon countries. Of course, the data here are not independently distributed since the data in the same year in different countries is related (as are data in subsequent years in the same country, but let's ignore that for the moment) We could also do a paired t-test of average accumulation per year per group by first using the cast command to aggregate the data. Note that we first remove the NA values (for Spain).

```
library(reshape2)
capital = na.omit(capital)
pergroup = dcast(capital, Year ~ Group, value.var="Private", fun.aggregate=mean)
head(pergroup)
```

```
##   Year  anglo european
## 1 1970 3.0625   2.6825
## 2 1971 3.1475   2.7425
## 3 1972 3.2450   2.9000
## 4 1973 3.1800   2.9500
## 5 1974 3.1125   3.0025
## 6 1975 3.0300   3.1325
```

Let's plot the data to have a look at the lines:

```
plot(pergroup$Year, pergroup$european, type="l", xlab="Year", ylab="Capital accumulation")
lines(pergroup$Year, pergroup$anglo, lty=2)
legend("top", lty=c(1,2), legend=c("European", "Anglo-Saxon"))
```



So initially capital is higher in the Anglo-Saxon countries, but the European countries overtake quickly and stay higher.

Now, a paired-sample t-test again shows a significant difference between the two:

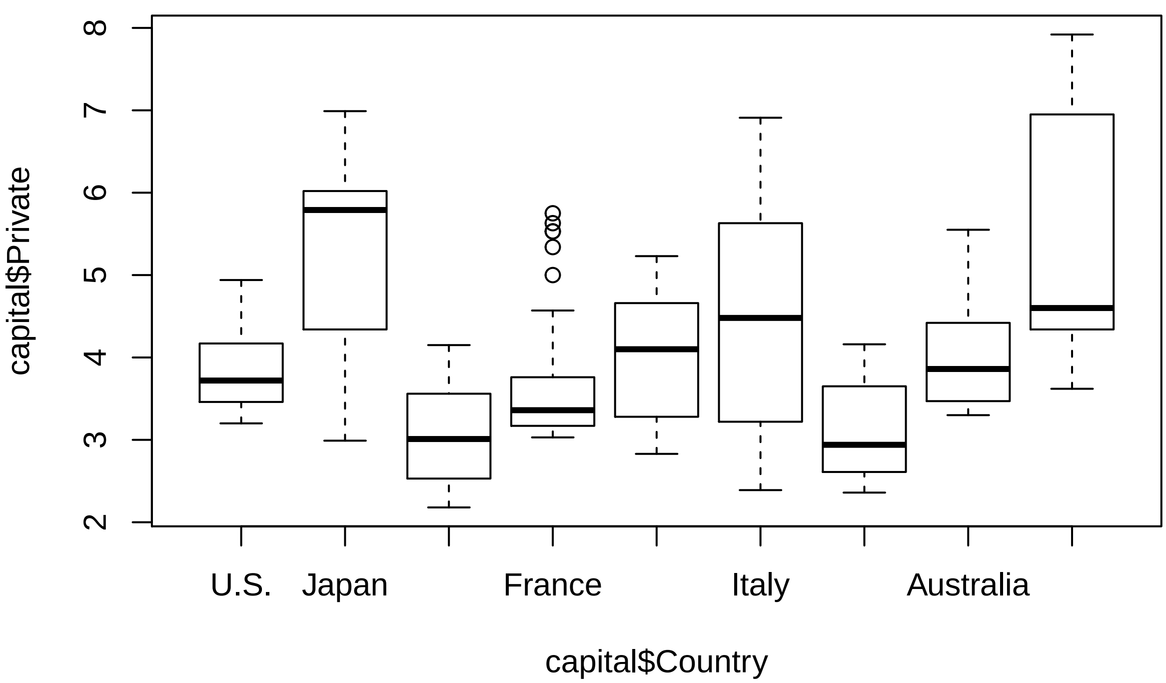
```
t.test(pergroup$anglo, pergroup$european, paired=T)
```

```
##
##  Paired t-test
##
## data:  pergroup$anglo and pergroup$european
## t = -6.5332, df = 40, p-value = 8.424e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6007073 -0.3168537
## sample estimates:
## mean of the differences
##              -0.4587805
```

Anova

We can also use a one-way Anova to see whether accumulation differs per country. Let's first do a box-plot to see how different the countries are. Plot by default gives a box plot of a formula with a nominal independeny variable

```
plot(capital$Private ~ capital$Country)
```



So, it seems that in fact a lot of countries are quite similar, with some extreme cases of high capital accumulation. (also, it seems that including Japan in the Europeaqn countries might have been a mistake). We use the `aov` function for this, the `anova` function is meant to analyze already fitted models, as will be shown below.

```r
m = aov(capital$Private ~ capital$Country)
summary(m)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## capital$Country   8  201.3  25.158   30.78 <2e-16 ***
## Residuals       343  280.3   0.817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So in fact there is a significant difference. We can use `pairwise.t.test` to perform

```r
posthoc = pairwise.t.test(capital$Private, capital$Country, p.adj = "bonf")
round(posthoc$p.value, 2)
```

```
##           U.S. Japan Germany France U.K. Italy Canada Australia
## Japan     0.00    NA      NA     NA   NA    NA     NA        NA
## Germany   0.00  0.00      NA     NA   NA    NA     NA        NA
## France    1.00  0.00    0.06     NA   NA    NA     NA        NA
## U.K.      1.00  0.00    0.00   1.00   NA    NA     NA        NA
## Italy     0.02  0.01    0.00   0.00 0.31    NA     NA        NA
## Canada    0.02  0.00    1.00   0.27 0.00  0.00     NA        NA
## Australia 1.00  0.00    0.00   1.00 1.00  0.46      0        NA
## Spain     0.00  1.00    0.00   0.00 0.00  0.01      0         0
```

Linear models

A more generic way of fitting models is using the `lm` command. In fact, `aov` is a wrapper around `lm`. Let's model private capital as a function of country and public capital:

```r
m = lm(Private ~ Country + Public, data=capital)
summary(m)
```

```
##
## Call:
## lm(formula = Private ~ Country + Public, data = capital)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4457 -0.4091 -0.1076  0.2601  2.8346
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.7241     0.1468  32.184  < 2e-16 ***
## CountryJapan       1.8183     0.1753  10.374  < 2e-16 ***
## CountryGermany    -0.7776     0.1721  -4.518 8.63e-06 ***
## CountryFrance     -0.3337     0.1729  -1.930 0.054420 .
## CountryU.K.        0.3910     0.1733   2.257 0.024643 *
## CountryItaly      -0.7911     0.2198  -3.600 0.000366 ***
## CountryCanada     -1.6928     0.1949  -8.685  < 2e-16 ***
## CountryAustralia   0.6421     0.1768   3.633 0.000323 ***
## CountrySpain       0.8331     0.2120   3.930 0.000103 ***
```

```
## Public            -1.8144     0.1660 -10.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7793 on 342 degrees of freedom
## Multiple R-squared:  0.5687, Adjusted R-squared:  0.5573
## F-statistic:  50.1 on 9 and 342 DF,  p-value: < 2.2e-16
```

As you can see, R automatically creates dummy values for nominal values, using the first value (U.S. in this case) as reference category. An alternative is to remove the intercept and create a dummy for each country:

```
m = lm(Private ~ Country + Public - 1, data=capital)
summary(m)
```

```
##
## Call:
## lm(formula = Private ~ Country + Public - 1, data = capital)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4457 -0.4091 -0.1076  0.2601  2.8346
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## CountryU.S.        4.7241     0.1468   32.18   <2e-16 ***
## CountryJapan       6.5424     0.1676   39.05   <2e-16 ***
## CountryGermany     3.9465     0.1474   26.77   <2e-16 ***
## CountryFrance      4.3904     0.1383   31.74   <2e-16 ***
## CountryU.K.        5.1151     0.1587   32.23   <2e-16 ***
## CountryItaly       3.9330     0.1334   29.48   <2e-16 ***
## CountryCanada      3.0313     0.1221   24.83   <2e-16 ***
## CountryAustralia   5.3662     0.1725   31.11   <2e-16 ***
## CountrySpain       5.5572     0.1596   34.82   <2e-16 ***
## Public            -1.8144     0.1660  -10.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7793 on 342 degrees of freedom
## Multiple R-squared:  0.9666, Adjusted R-squared:  0.9657
## F-statistic: 991.2 on 10 and 342 DF,  p-value: < 2.2e-16
```

(`- 1` removes the intercept because there is an implicit +1 constant for the intercept in the regression formula)

You can also introduce interaction terms by using either the `:` operator (which only creates the interaction term) or the `*` (which creates a full model including the main effects). To keep the model somewhat parsimonious, let's use the country group rather than the country itself

```
m1 = lm(Private ~ Group + Public, data=capital)
m2 = lm(Private ~ Group + Public + Group:Public, data=capital)
```

A nice package to display multiple regression results side by side is the `screenreg` function from the `texreg` package:

```r
library(texreg)
```

```
## Version:  1.36.4
## Date:     2016-02-16
## Author:   Philip Leifeld (Eawag & University of Bern)
##
## Please cite the JSS article in your publications -- see citation("texreg").
```

```r
screenreg(list(m1, m2))
```

```
##
## ============================================
##                       Model 1     Model 2
## --------------------------------------------
## (Intercept)             3.97 ***    3.75 ***
##                        (0.11)      (0.13)
## Groupeuropean           0.47 ***    0.78 ***
##                        (0.12)      (0.16)
## Public                 -0.49 ***   -0.01
##                        (0.14)      (0.22)
## Groupeuropean:Public               -0.83 **
##                                    (0.28)
## --------------------------------------------
## R^2                     0.09        0.11
## Adj. R^2                0.08        0.10
## Num. obs.             352         352
## RMSE                    1.12        1.11
## ============================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

So, there is a significant interaction effect which displaces the main effect of public wealth.

Comparing and diagnosing models

A relevant question can be whether a model with an interaction effect is in fact a better model than the model without the interaction. This can be investigated with an anova of the model fits of the two models:

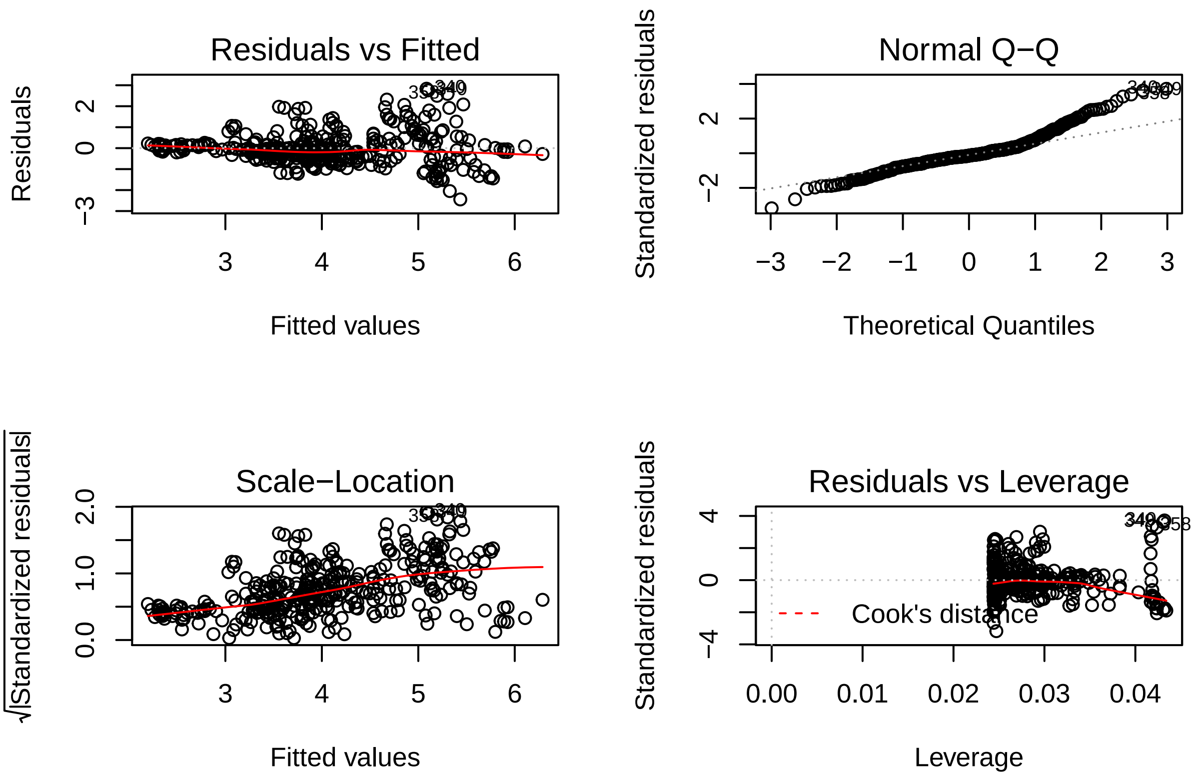
```r
m1 = lm(Private ~ Group + Public, data=capital)
m2 = lm(Private ~ Group + Public + Group:Public, data=capital)
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: Private ~ Group + Public
## Model 2: Private ~ Group + Public + Group:Public
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1    349 440.02
## 2    348 429.36  1    10.661 8.641 0.003506 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the interaction term is in fact a significant improvement of the model. Apparently, in European countries private capital is accumulated faster in those times that the government goes into depth.

After doing a linear model it is a good idea to do some diagnostics. We can ask R for a set of standard plots by simply calling `plot` on the model fit. We use the parameter (`par`) `mfrow` here to put the four plots this produces side by side.

```
old.settings = par(mfrow=c(2,2))
plot(m)
```




```
par(old.settings)
```

See http://www.statmethods.net/stats/rdiagnostics.html for a more exhausitve list of model diagnostics.