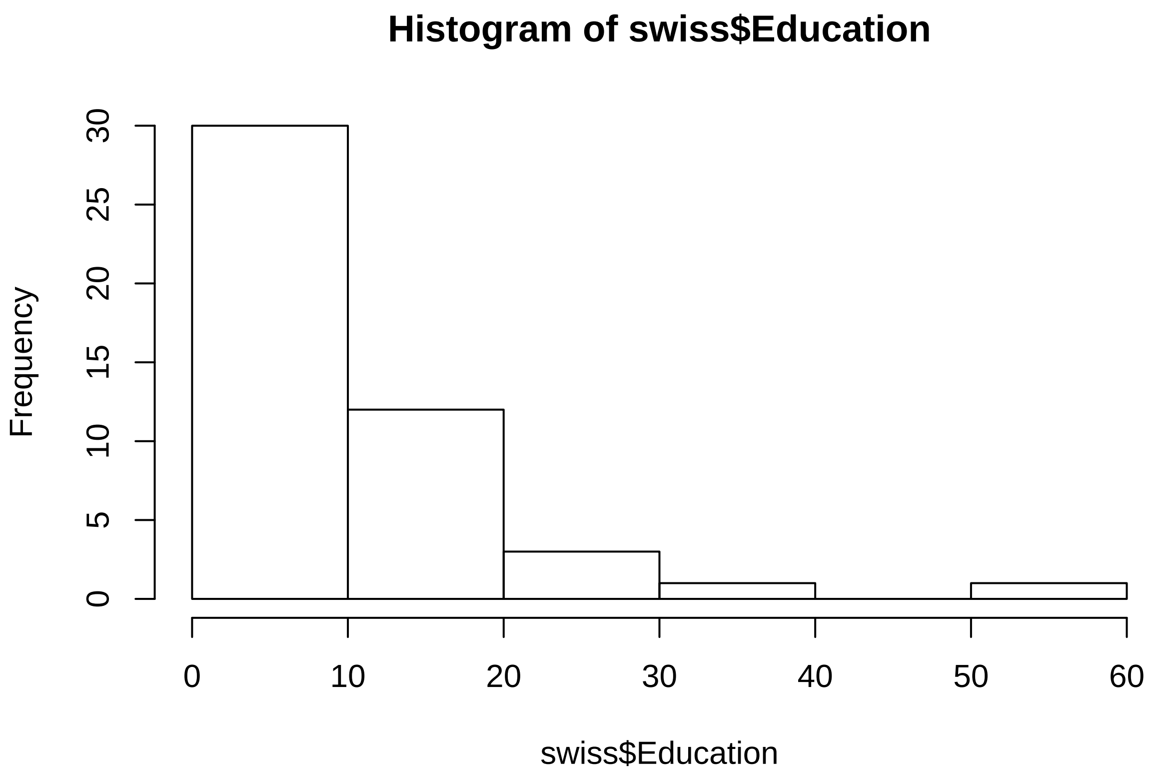
HW 8 notes

Chapter 8, Exercise 3

Some people noticed that Education was skewed and tried to do something about it

```
hist(swiss$Education)
```




```
library(e1071)
skewness(swiss$Education)
```

```
## [1] 2.268439
```

```
#install.packages("bestNormalize")
library(bestNormalize)
bestNormalize(swiss$Education)
```

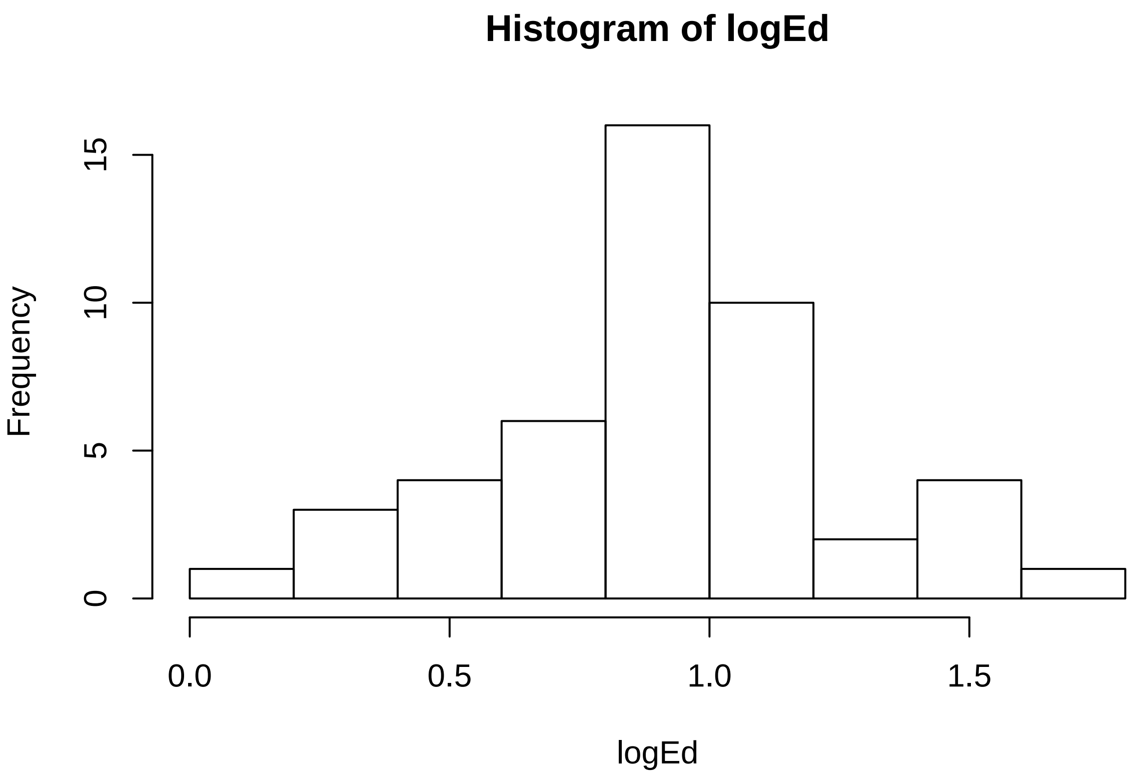
```
## Warning in orderNorm(standardize = TRUE, warn = TRUE, x = c(12L, 9L, 5L, : Ties in data, Normal dist

## Warning in get_oos_estimates(x, standardize, method_names, k, r, cluster, :
## fold_size is 4 (< 20), therefore P/df estimates may be off

## Best Normalizing transformation with 47 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - No transform: 2.668
##  - Box-Cox: 2.1
##  - Log_b(x+a): 2.036
```

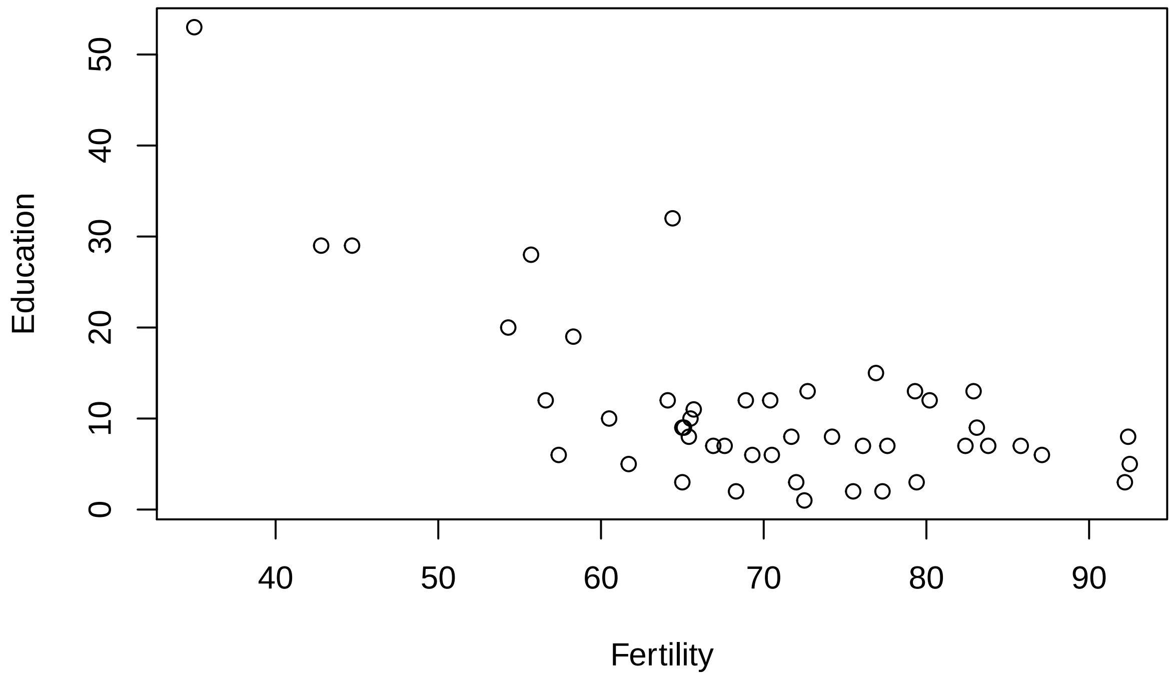
```
##  - sqrt(x+a): 2.084
##  - exp(x): 6.988
##  - arcsinh(x): 1.996
##  - Yeo-Johnson: 2.1
##  - orderNorm: 1.9
## Estimation method: Out-of-sample via CV with 10 folds and 5 repeats
##
## Based off these, bestNormalize chose:
## orderNorm Transformation with 47 nonmissing obs and ties
##  - 19 unique values
##  - Original quantiles:
##   0%  25%  50%  75% 100%
##    1    6    8   12   53
```

```
logEd<-log10(swiss$Education)
hist(logEd)
```



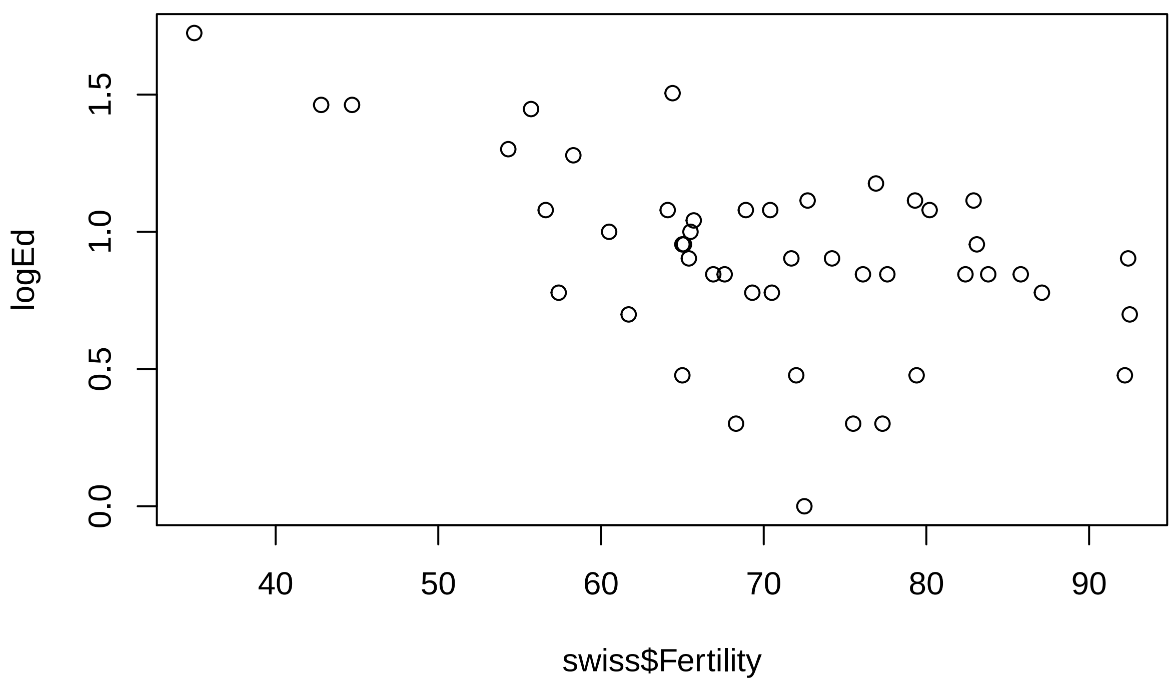
But keep in mind that normality of the independent variables is not an assumption of linear regression (normality of residuals is). The reason for a transform is to make the relationship linear (an assumption of linear regression). As it happens, Education and Fertility are linearly related without the transformation.

```
with(swiss, plot(Fertility, Education))
```



The transform does not make the relationship more linear.

```r
plot(swiss$Fertility, logEd)
```



And it does reduces the correlation:

```r
cor(with(swiss, cbind(Fertility, Education, logEd)))
```

```
##            Fertility  Education      logEd
## Fertility  1.0000000 -0.6637889 -0.5242985
## Education -0.6637889  1.0000000  0.8617851
## logEd     -0.5242985  0.8617851  1.0000000
```

So, the regression with the untransformed variable actually does better.

```
summary(lm(Fertility~Education+Agriculture, data=swiss))
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Agriculture, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3072  -6.6157  -0.9443   8.7028  20.5291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.08005    5.78180  14.542  < 2e-16 ***
## Education   -0.96276    0.18906  -5.092  7.1e-06 ***
## Agriculture -0.06648    0.08005  -0.830    0.411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.479 on 44 degrees of freedom
## Multiple R-squared:  0.4492, Adjusted R-squared:  0.4242
## F-statistic: 17.95 on 2 and 44 DF,  p-value: 2e-06
```

```
summary(lm(Fertility~logEd+Agriculture, data=swiss))
```

```
##
## Call:
## lm(formula = Fertility ~ logEd + Agriculture, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.715  -6.325  -2.834   8.680  22.116
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.591948   9.848586   8.894 2.18e-11 ***
## logEd       -19.066459   6.314965  -3.019  0.00421 **
## Agriculture  -0.001318   0.095808  -0.014  0.98909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.88 on 44 degrees of freedom
## Multiple R-squared:  0.2749, Adjusted R-squared:  0.2419
## F-statistic:  8.34 on 2 and 44 DF,  p-value: 0.0008489
```

Looking at the plot though, you might note that there's a seeming outlier in the upper left.

```
which.max(swiss$Education)
```

```
## [1] 45
```

```
swiss[which.max(swiss$Education),]
```

```
##              Fertility Agriculture Examination Education Catholic
## V. De Geneve        35         1.2          37        53    42.34
##              Infant.Mortality
```

```
## V. De Geneve                18
```

Interestingly, dropping this point doesn't change the results much, but it does weaken the correlation and the R^2.

```r
with(swiss[-45,], cor(Fertility, Education))
```

```
## [1] -0.5671542
```

```r
summary(lm(Fertility~Education+Agriculture, data=swiss[-45,]))
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Agriculture, data = swiss[-45,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.425  -7.055  -1.819   8.731  20.546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.64934    6.17437  13.710  < 2e-16 ***
## Education   -1.00476    0.24119  -4.166 0.000147 ***
## Agriculture -0.07016    0.08192  -0.856 0.396491
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.579 on 43 degrees of freedom
## Multiple R-squared:  0.333,  Adjusted R-squared:  0.302
## F-statistic: 10.74 on 2 and 43 DF,  p-value: 0.0001652
```

Chapter 8, Exercise 5

Some people wondered why the Bayes Factor for the model with Education and Agriculture was high when Agriculture wasn't significant. You could think of the Bayes Factor more like the F test: it's testing the whole model.

The regressionBF command will test multiple models, not just one (it explodes if the model is complicated and there are too many variations).

```r
library(BayesFactor)
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```r
regBFout<-regressionBF(formula = Fertility ~ Education + Agriculture, data = swiss)
regBFout
```

```
## Bayes factor analysis
## --------------
```

```
## [1] Education               : 32071.51 ±0.01%
## [2] Agriculture             : 3.579893 ±0%
## [3] Education + Agriculture : 8927.474 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

The results show that there's evidence for a model with Education + Agriculture but even stronger evidence for just Education and only weak evidence for just Agriculture.

You can directly compare two models by dividing them:

```
regBFout[3]/regBFout[1]
```

```
## Bayes factor analysis
## --------------
## [1] Education + Agriculture : 0.2783615 ±0.01%
##
## Against denominator:
##   Fertility ~ Education
## ---
## Bayes factor type: BFlinearModel, JZS
```

So, there's no evidence to favour adding Agriculture to the model.