Project: Trading with ETF

Vilem Knap
02403 Introduction to Mathematical Statistics Jun 20
Technical University of Denmark

September 19, 2023

Data1

a)

According to the output from R, The dataset contains 454 observations (number of rows), collected from 2006-5-5 to 2015-5-8 with one-week period. The missing observations can be identified as as a zero values or NA, they can be found by using filtering commands. First I try to check if there are NA values by using the command is.na(wr) and which (is.na(wr)). The output - integer(0) means that there are no NA values in the dataset, but we can still find some zero values. By replacing zeros with NA wr[wr == 0] ¡- NA, we can than use buil in function is.na.data.frame(wr), to check for missing values. By using summary(wr), we can locate columns with NA values. It turns out that most indexes, contain some missing values. XLE, IWM, IWB, IJR, IVE, IJJ,IJT,IWW,ADRA, EZA,VPU,UDC,JKH,JKL,VDE,VIS,FXI indexes are free of zero values. Others may require dropping NA values or interpolation from neighbouring values for future tasks.

b)

Basic parameters for descriptive statistics are filled in the table 1.

EFT	(Number of obs.)	Sample mean	Sample variance	Std. dev.	Lower quartile	Median	Upper quartile
AGG	446	0.000265757	3.571068e-05	0.005975841	-0.0029997273	0.0002374461	0.0038943769
VAW	453	0.00179379	0.001301973	0.03608286	-0.016248480	0.004797925	0.019736842
IWN	453	0.001187679	0.00102499	0.03201547	-0.014348918	0.003119637	0.019060284
SPY	453	0.001360105	0.0006143463	0.02478601	-0.011356888	0.004215788	0.014522822

Table 1: Descriptive statistiscs

Boxplots and histograms

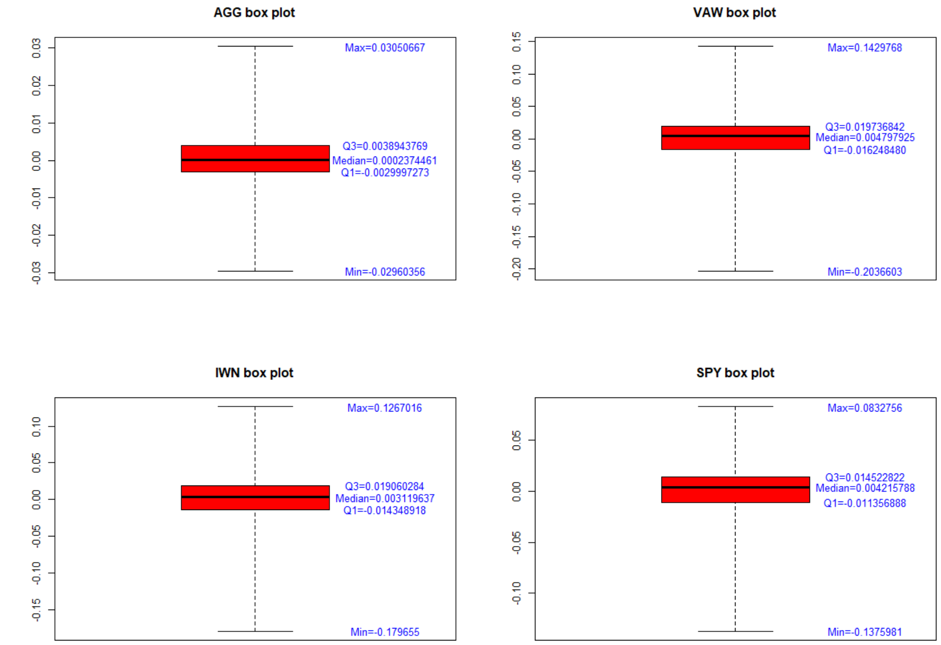


Figure 1: *Box plots*

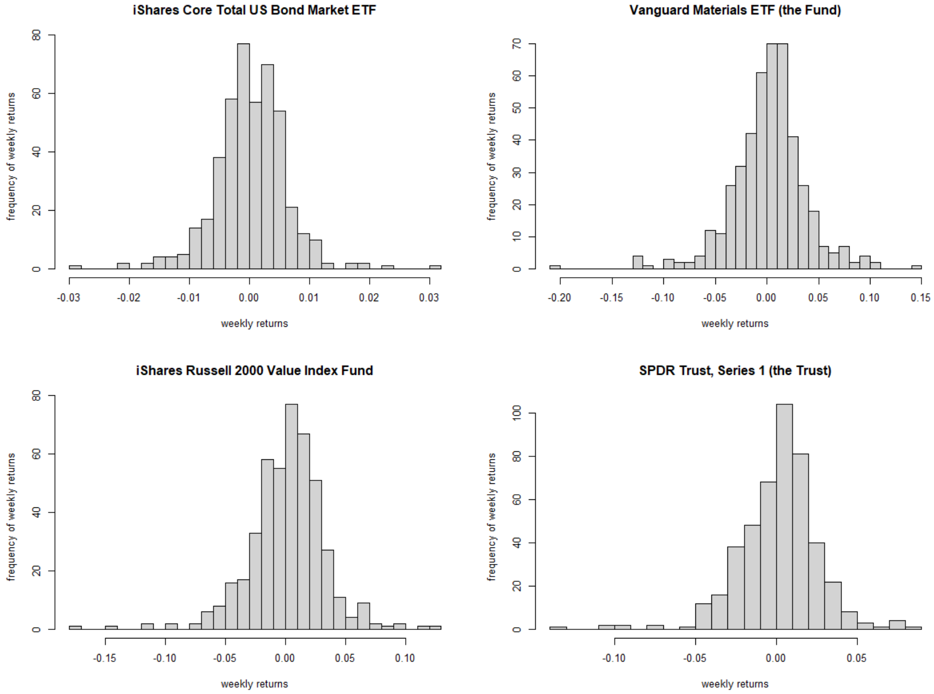


Figure 2: *Empirical densities plots*

c)

From Box-plots observation, most of the data are falling to the positive side of distribution, while remaining funds are skewed left. The iShares Core Total US Bond Market is also the most symmetrical, with the lowest difference between its mean and median. It is also most centered arround zero ,creating 52% positive observations with 232 out of 446 It's interesting to see that its most frequent value, which coresponds to its modus is negative. On the other hand, the US bond market tends to have great stability, resulting in the lowest variance. It doesn't provide big revenues with maximal weekly returns of 3 % but the greatest loss is -2,96 % , which is the lowest minimal value of these 4 indexes. Based on the examination of the Empirical densities plots, the Vanguard Materials ETF tends to have most normal distribution shape, thanks to symmetrical values of frequencies.This index has the lowest minimal value with -20% of weekly returns, but also the highest weekly return of more than 14 %. The high occurence of negative returns can correspond with the decrease of the US industrial sector. With the highest inter quartile range of 0.03578098, the Vanguard Materials fund has the most widely spread 50% of its data. The iShares Russel fund behaves very simmilarly as Vanguard. It does not have so large extremes and the variance is lower, compared to Vanguard. Although it has same amount of positive observation (253), the lower amount of negative values over 5%, could contribute to the better overall performance of the Fund. The SPDR trust is very interesting by the fact that is the most right skewed, having 265 (58%) positive observations. With the second lowest minimum of 13.8 % and the first quartile is close to zero from the left side, the Trust can provide a fair compromise between stability of AGG and larger but volatile revenues of VAW.

d)

The covariances between the following ETFs: AGG, VAW, IWN, SPY, EWG and EWW are shown in the table 2.

	AGG	VAW	IWN	SPY	EWG	EWW
AGG	3.571068e-05	-4.260068e-05	-2.587826e-05	-3.239567e-05	-5.084499e-05	-3.710732e-05
VAW	-4.260068e-05	1.301973e-03	9.838237e-04	7.927169e-04	1.109888e-03	1.184849e-03
IWN	-2.587826e-05	9.838237e-04	1.024990e-03	7.221941e-04	9.502014e-04	1.010141e-03
SPY	-3.239567e-05	7.927169e-04	7.221941e-04	6.143463e-04	8.046398e-04	8.152602e-04
EWG	-5.084499e-05	1.109888e-03	9.502014e-04	8.046398e-04	1.444157e-03	1.179610e-03
EWW	-3.710732e-05	1.184849e-03	1.010141e-03	8.152602e-04	1.179610e-03	1.659257e-03

Table 2: Covariance of the selected Variables

e)

From Remark 2.59 in the textbook [1] implies that the Variance can be obtain by getting Covariance of the Random Variable with itself, e.g

$$Var(P1) = Cov(P1, P1) \tag{1}$$

The probability of portfolio random variable is linear combination of two funds/indexes random variables

$$P_1 = \alpha E_{EWG} + (1-\alpha)E_{EWW} = 1 + \alpha E_{EWG} - \alpha E_{EWW} \tag{2}$$

Theorem 2.60:

$$Cov(a_0 + a_1X + a_2Y, b_0 + b_1X + b_2Y) = a_1b_1V(X) + a_2b_2V(Y) + (a_1b_2 + a_2b_1)Cov(X,Y) \tag{3}$$

To adjust this theorem to our case we set:

$$a_0 = b_0 = 1,\ a_1 = b_1 = \alpha,\ a_2 = b_2 = -\alpha \tag{4}$$

$$X = E_{EWG},\ Y = E_{EWW} \tag{5}$$

Than, we obtain:

$$\begin{aligned}
Cov(P_1, P_1) &= Var(P_1) && (6)\\
&= \alpha^2 Var(E_{EWG}) + \alpha^2 Var(E_{EWW}) + (-\alpha^2 - \alpha^2)Cov(E_{EWG}, E_{EWW}) && (7)\\
&= \alpha^2(Var(E_{EWG}) + Var(E_{EWW})) - 2\alpha^2 Cov(E_{EWG}, E_{EWW}) && (8)\\
&= \alpha^2(0.001444157 + 0.001659257) - 2\alpha^2 0.00117961 && (9)
\end{aligned}$$

The same expression will be used for other combinations of ETFs. On the figere 3 The Variance as a function of α is illustrated as a parabolic curve.

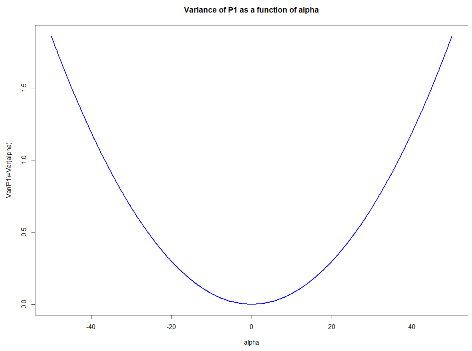


Figure 3: $V_1(\alpha)$

The minimum will be located around $\alpha = 0$. Since the α coefficient specify the proportion of the portfolio invested in E_{EWG}, if the result will be negative, it means that we should decrease the share invested in E_{EWG} and increase the share in E_{EWW}. The same rule applies for any other combination of ETFs. When is the minimalization solved analytically by finding, where is the first derivate of $V(\alpha)$ equals to zero, the equation for α_m is:

$$\alpha_m = \frac{1}{2(Var(E_{EWW}) + Var(E_{EWG}) - 4Cov(E_{EWG}, E_{EWW})} \tag{10}$$

It gives us result for $\alpha = 0$ so lets set constrains (-1000, -0.5) for $\alpha_m < 0$ and (1, 2000) for $\alpha_m > 1$ and solve it numerically in R, the results are shown in the following table.

ETF combination	$\alpha_m < 0$	$\sigma_{m,left}$	$E[P_{i,left}]$	$\alpha_m > 1$	$\sigma_{m,right}$	$E[P_{i,right}]$
(EWG,EWW)	-0.5000784	0.0001316298	0.0019373939	1.000056	0.0007442772	0.0012366604
(AGG,SPY)	-0.5000784	0.0001787681	0.0019073655	1.000056	0.0007149286	0.0002656957
(VAW,IWN)	-0.5000784	8.985713e-05	0.0008845763	1.000056	0.0003593562	0.0017938241
(VAW,EWG)	-0.5000784	0.0001316298	0.0012611940	1.000056	0.0005264133	0.0011876765
(VAW,EWW)	-0.5000784	0.0001479292	0.0016587984	1.000056	0.0005915978	0.0017937952
(IWN,EWG)	-0.5000784	0.0001422306	0.0012611940	1.000056	0.0005688082	0.0011876765

Table 3: Table of minimized alphas, variances and expected weekly returns

Overall it seems that this method tends to choose only one portfolio with 99%. That's because of parabolic shape of Variance so the minimal values of variance are centered arround 0. In this case I would prefer (EWG,EWW) variant with $\alpha_m = -0.0005688082$, because it provides the largest returns and the risk, modelled by variance is reasonable.

f)

Check for normality and independence of the observations

From Central limit theorem is implicated, that for sample with suffecient amount of observations, usually more than 30 observation, sample means tend to converge to normal distribution. Using Normal distribution for modelling is plausible. This assumption is fullfilled because our dataset contains more than 4 hundred observations. The assumption of independence of the observation is fullfilled enough by difference of the sectors, where these indexes primarily operates. AGG consists of US market's bonds, VAW is focused on materials and commodities, IWN targets small capitalization companies and SPY contains all of the common stocks. There can be some interweaving, due to connections in the global market but we can assume sufficient level of independence.

Models of selected ETFs

On the figure 4 models with empirical densities are illustrated.The first model fits very good. For the second, third and fourth it could be better to transform the weekly returns quantiles to logaritmic scales, because the data are skewed right. There are two problems - negative values in the dataset and the fact that values are centered arround zeros, resulting in a high amount of negative observation after logaritmic transformation. I tried to use log transformation, with geometric mea (see the R-code) but the fit wasn't much better.

Models validation

On the figure 5 Model validation is performed by comparing on quantile-quantile plot. The AGG portfolio has the best fit with the highest amount of quantiles placed on the line. The VAW model has the largest extremes The SPY dataset is the most left-skewed, therefore on its QQ plot is a slightly visible concavity arround zero. Overall the most of the models data fits the lines well.

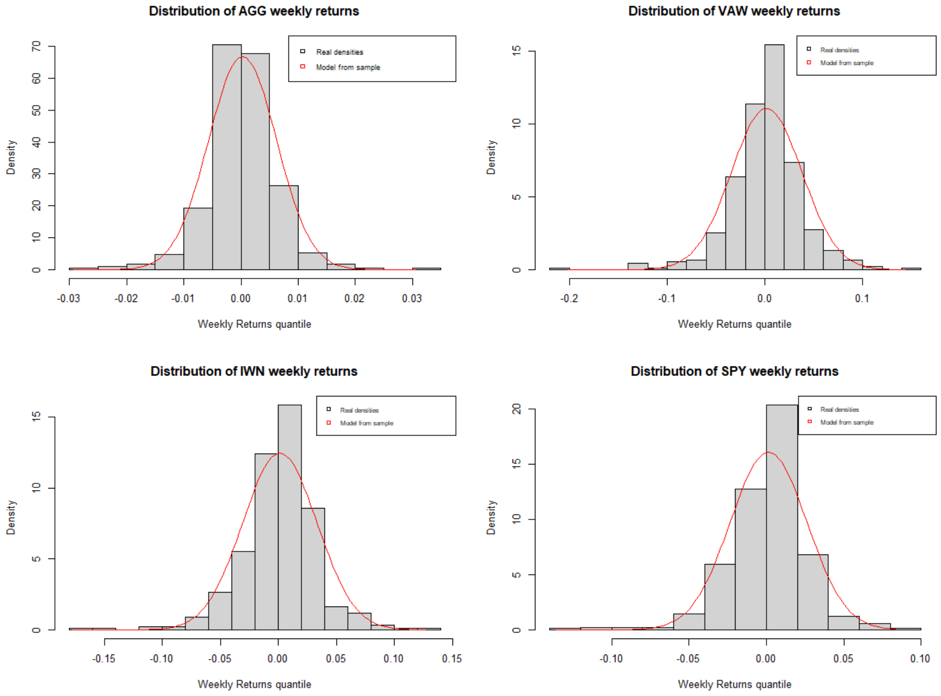


Figure 4: *Models and empirical densities of selected ETFs*

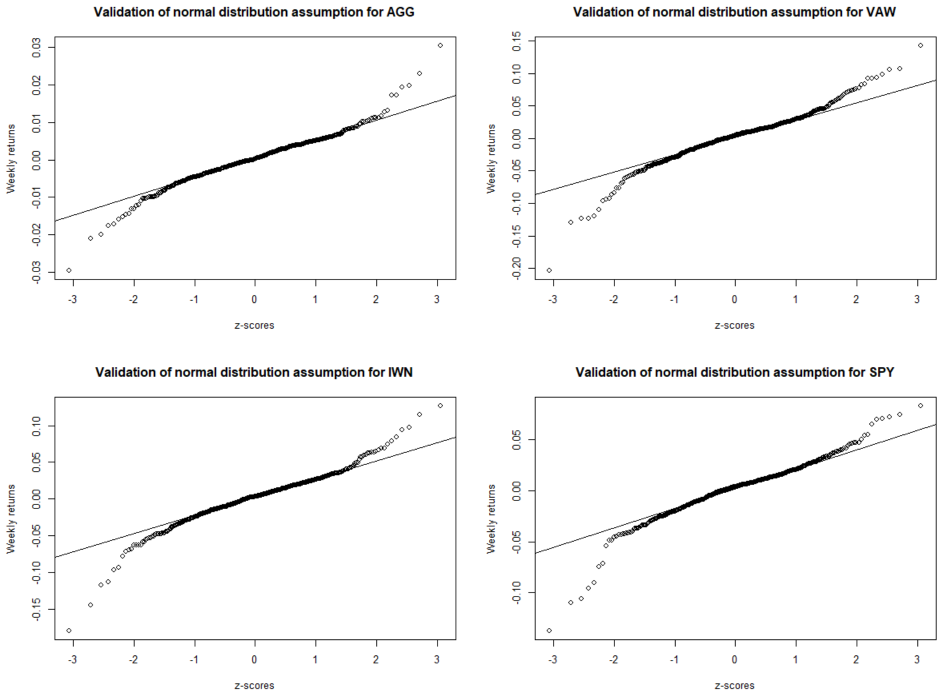


Figure 5: *Quantile quantile plots of ETFs*

g)

To determine the average returns confidence intervals, t-test with (n-1) degree of freedom is used and for variance chi-square test.

The average returns confidence intervals

ETF	$P(\bar{X} \pm t_{1-\alpha/2} * \frac{S}{\sqrt{n}})$	$\Leftrightarrow$	confidence interval
AGG	$P(0.000265757 \pm 1.965215 * \frac{0.005975841}{21.30728}$	$\Leftrightarrow$	(-0.0002854073, 0.0008169213)
VAW	$0.00179379 \pm 1.965215 * 0.03608286 * \frac{0.03608286}{21.30728}$	$\Leftrightarrow$	(-0.001534208 , 0.005121788)
IWN	$0.001187679 \pm 1.965215 * 0.03608286 * \frac{0.03201547}{21.30728}$	$\Leftrightarrow$	(-0.001765174, 0.004140533)
SPY	$0.001360105 \pm 1.965215 * 0.03608286 * \frac{0.02478601}{21.30728}$	$\Leftrightarrow$	(-0.000925960, 0.003646171)

Table 4: Table of 95% confidence intervals

Variance confidence intervals

ETF	$\left[\frac{(n-1)*s^2}{\chi^2_{97.5}}, \frac{(n-1)*s^2}{\chi^2_{2.5}}\right]$	$=$	confidence interval
AGG	$\left[\frac{(453)*0.005975841^2}{513.8655}, \frac{(453)*0.005975841^2}{395.9219}\right]$	$=$	3.148088e-05, 4.085891e-05
VAW	$\left[\frac{(453)*0.03608286^2}{513.8655}, \frac{(453)*0.03608286^2}{395.9219}\right]$	$=$	0.001147759, 0.001489672
IWN	$\left[\frac{(453)*0.0.03201547^2}{513.8655}, \frac{(453)*0.03201547^2}{395.9219}\right]$	$=$	0.0009035838, 0.001172758
SPY	$\left[\frac{(453)*0.02478601^2}{513.8655}, \frac{(453)*0.02478601^2}{395.9219}\right]$	$=$	0.0005415792, 0.0007029136

Table 5: Table of 95% confidence intervals for the variance parameter

From the resulst above it obvious that lenghts of the intervals differ. For average returns, The lowest range between upper and lower limit can be found at AGG. From that we can conclude that estimated average of weekly returns is closest to sample mean. The range of variance confidence interval is also the lowest. Overall it corresponds with the fact that for AGG, normal model has best fit, with most of the densities covered. Confidence intervals of VAW and IWN, have the least differences between each other. It can be assumed, that their models have similar level of accuracy (or inaccuracy) which can be also observed on their almost identically shaped histograms.

h)

Finding the mentioned confidence intervals with non-parametric bootstrap follows the methodology described in the Pregnant Women's cigarette consumption example in the textbook. First, the weekly returns are sampled 10000 times, then, the mean of each simulated sample is calculated. It seems that the sampled means are asymptotically normally distributed which confirms the effects of central limit theorem. Densities of simulated means are illustrated bellow.

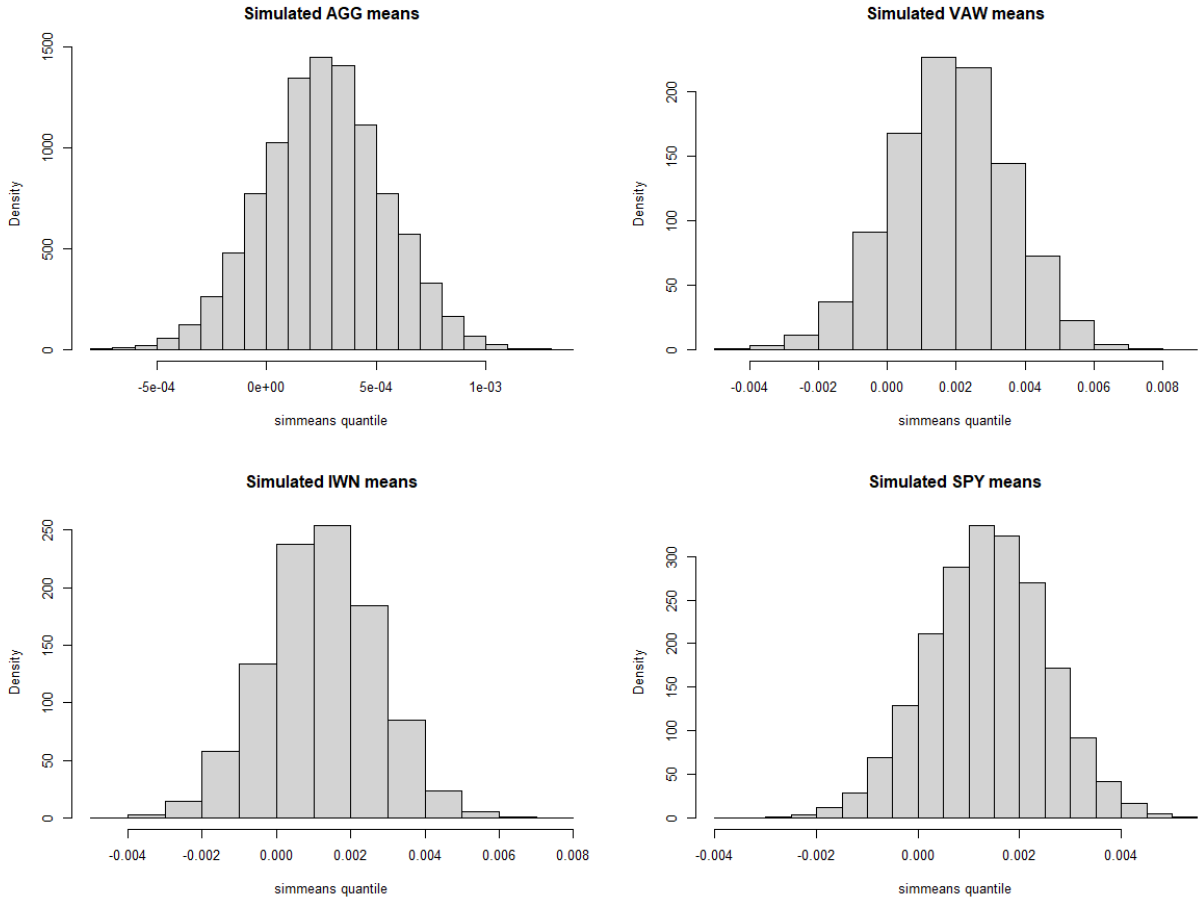


Figure 6: *Histograms of simulated means*

Confidence intervals

By using non parametric bootstrap method, confidence intervals are calculated as 2.5% and 97.5% quantiles of simulated parameters.

ETF	α	confidence intervals of simulated means	confidence interval of simulated variances
AGG	0.05	-0.000282520, 0.000809965	2.846097e-05, 4.382414e-05
VAW	0.05	-0.001598103, 0.005077533	0.001034601, 0.001603738
IWN	0.05	-0.001801865, 0.004117005	0.0008019725, 0.0012732107
SPY	0.05	-0.0009344601, 0.0036252122	0.0004841044, 0.0007654180

Table 6:

AGG,VAW,IWN intervals of means don't differ much from interval values obtained with t-tests, basically, only on fiths and sixths decimal places.

ETF	α	difference of mean conf. ints. (ttest - simulated)	diference of variances conf ints. (ttest-simulated)
AGG	0.05	2.887390e-06, -6.956265e-06	-3.019905e-06, 2.965230e-06
VAW	0.05	-6.389524e-05, -4.425504e-05	-0.0001131574, 0.0001140659
IWN	0.05	-3.669099e-05, -2.352792e-05	-0.0001016113, 0.0001004528
SPY	0.05	-8.500048e-06, -2.095875e-05	-5.747480e-05, 6.250447e-05

Table 7:

The most of the differences are negative, it seems that estimates from t and chi square distributions tends to be lower than simulated parameters. The values gained from simulation have significantly more degrees of freedom, that can provide better accuracy.

i)

For comparing mean value, whether it differs from zero, the one-sample t-test is used. The outputs from R are following

ETF	t-statistic	df	$P-value$	H_{alt}	est. mean	95 % confidence interval	
AGG	0.94757	453	0.3439	true mean is not equal to 0	0.000265757	-0.0002854073, 0.0008169213	$pvalue > \alpha$
VAW	1.0593	453	0.2901	true mean is not equal to 0	0.00179379	-0.001534208 , 0.005121788	$pvalue > \alpha$
IWN	0.79044	453	0.4297	true mean is not equal to 0	0.001187679	-0.001765174, 0.004140533	$pvalue > \alpha$
SPY	1.1692	453	0.2429	true mean is not equal to 0	0.001360105	-0.000925960, 0.003646171	$pvalue > \alpha$

Table 8: Table of t-test results

In all 4 cases, the pvalue ≥ 0.1 on the 95% confidence interval, which can be interpret as there is not enough evidence against H_0, concluding that the average weekly return do not differ significantly from saving the money under the pillow. It may be interpreted as the ETFs aren't suitable for low scale personal savings. I think that these indexes may be suitable for banks,national banks and large funds, which priority is to "rotate" their money more than savings.

j)

The lowest average returns provides AGG, the highest VAW. Assuming indepence, we can perform two sample (Welch) t-test without pooled variance. First we find t-test statistic

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} = \frac{0.00179379 - 0.000265757}{(0.03608286^2/454) + (0.005975841^2/454)} = 0.86313$$

$$\nu = \frac{((s_1^2/n_1) + (s_2^2/n_2))^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} = \frac{((0.03608286^2/454) + (0.005975841^2/454))^2}{\frac{(0.03608286^2/454)^2}{453} + \frac{(0.005975841^2/454)^2}{453}}$$

$$H_0 : \mu_{max} - \mu_{min} = 0 \; H_1 : \mu_{max} - \mu_{min} \neq 0$$

$$p - value = 2 * P(T > |t_{obs}|) = 0.3885 > 0.05$$

On the 95% confidence interval, we cannot reject the null hypothesis of the two ETFs providing equal average returns.

Data2

k)

Scatter plots

Observing Relation between Geometric mean of returns and maximum Time under Water,illustrated on figure 7, we can see, that the relation tends to have negative correlation. MaxTuW indicates the time for regaining historical peak so the returns will be lower when more time is needed.

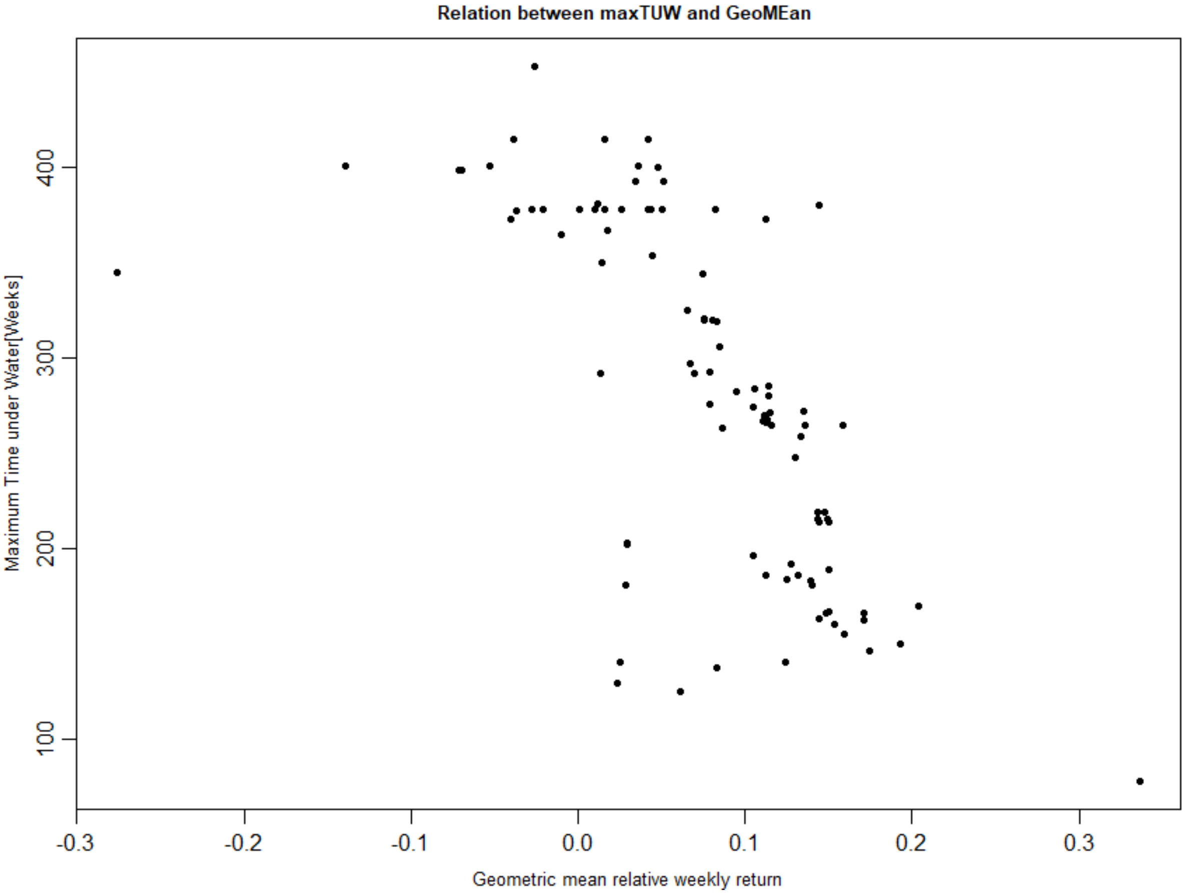


Figure 7: *Relation between GeoMEan and maxTUW*

The relation between Volatility and Conditional value at Risk is illustrated on figure 8. The points are very close to each other signalizing strong relationship with negative correlation. If the Volatility is high, which can signal bad condition or some troubles of the activum, the expected loss of the 5 percent worst situations will be larger (the loss grows in negative scale as an opposite of returns).

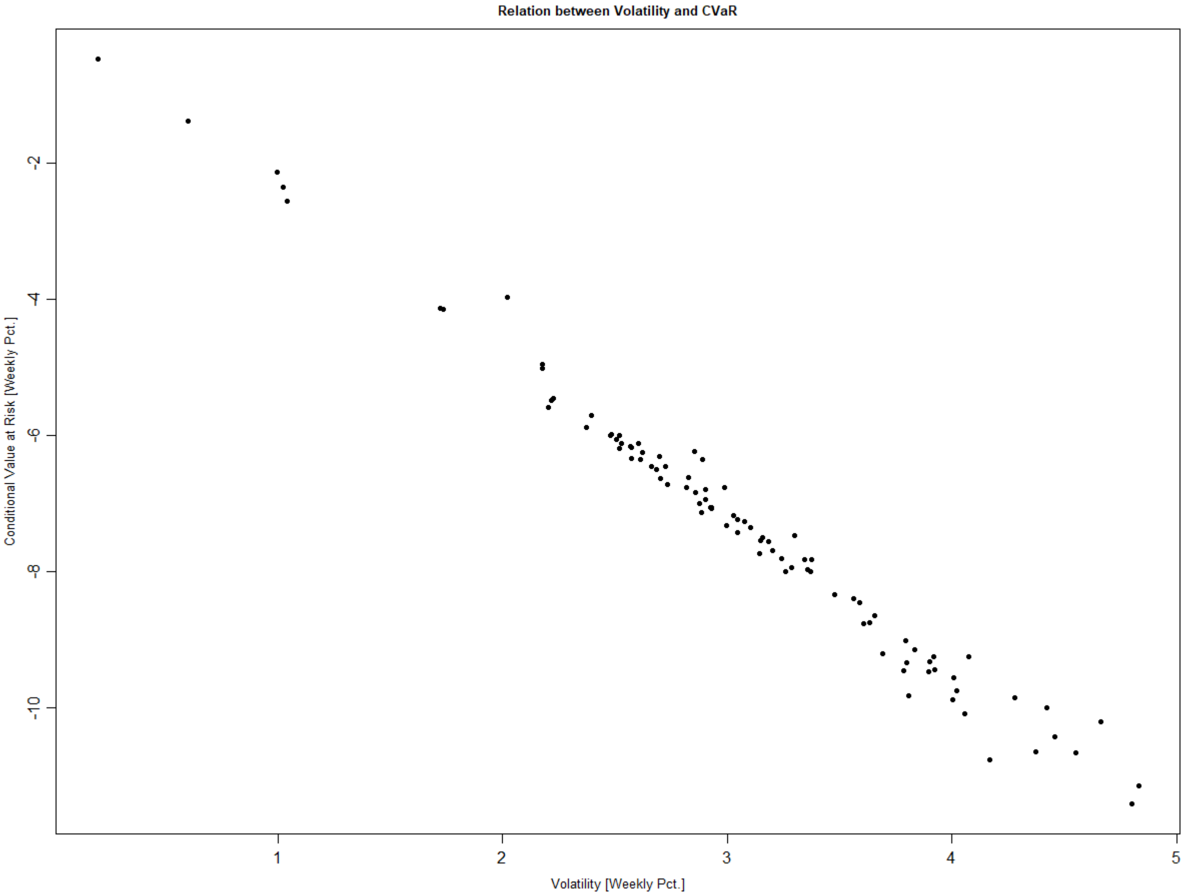


Figure 8: *Relation between Volatility and CVaR*

Although the points are scattered as it can be seen on figure 9, it seems there is a positive corelation between Maximum time under water and Maximum Drown Down. It make sense that the largest possible losses in a given period are greater when more time is needed for regaining historical peak.

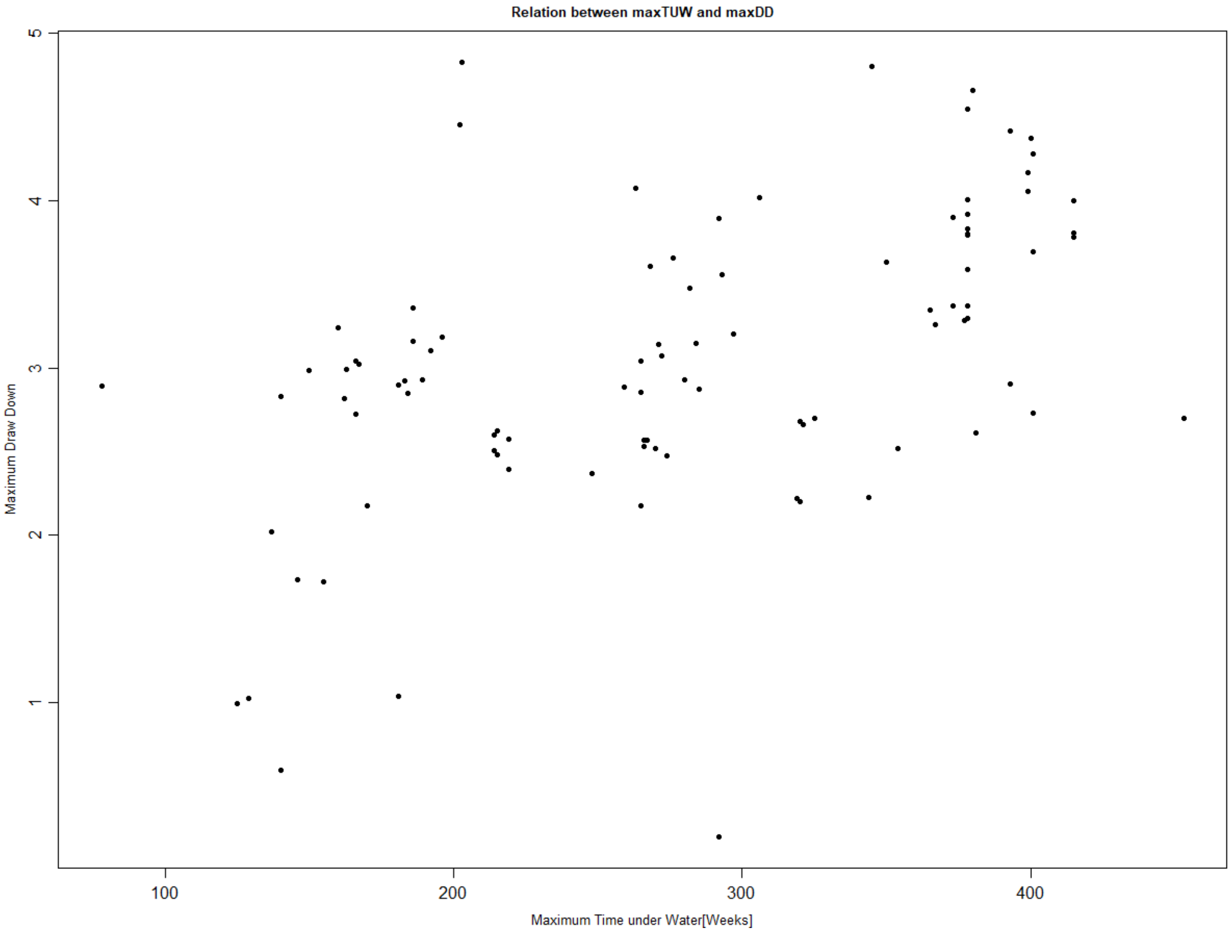


Figure 9: *Relation between maxTUW and maxDD*

The relation between Volatility and Maximum Drown down is illustrated on figure 10. The correlation is positive, meaning that the higher is volatility then the largest possible losses in a given period grows which can be interpreted as that the possible losses are greater when the activum is unstable.

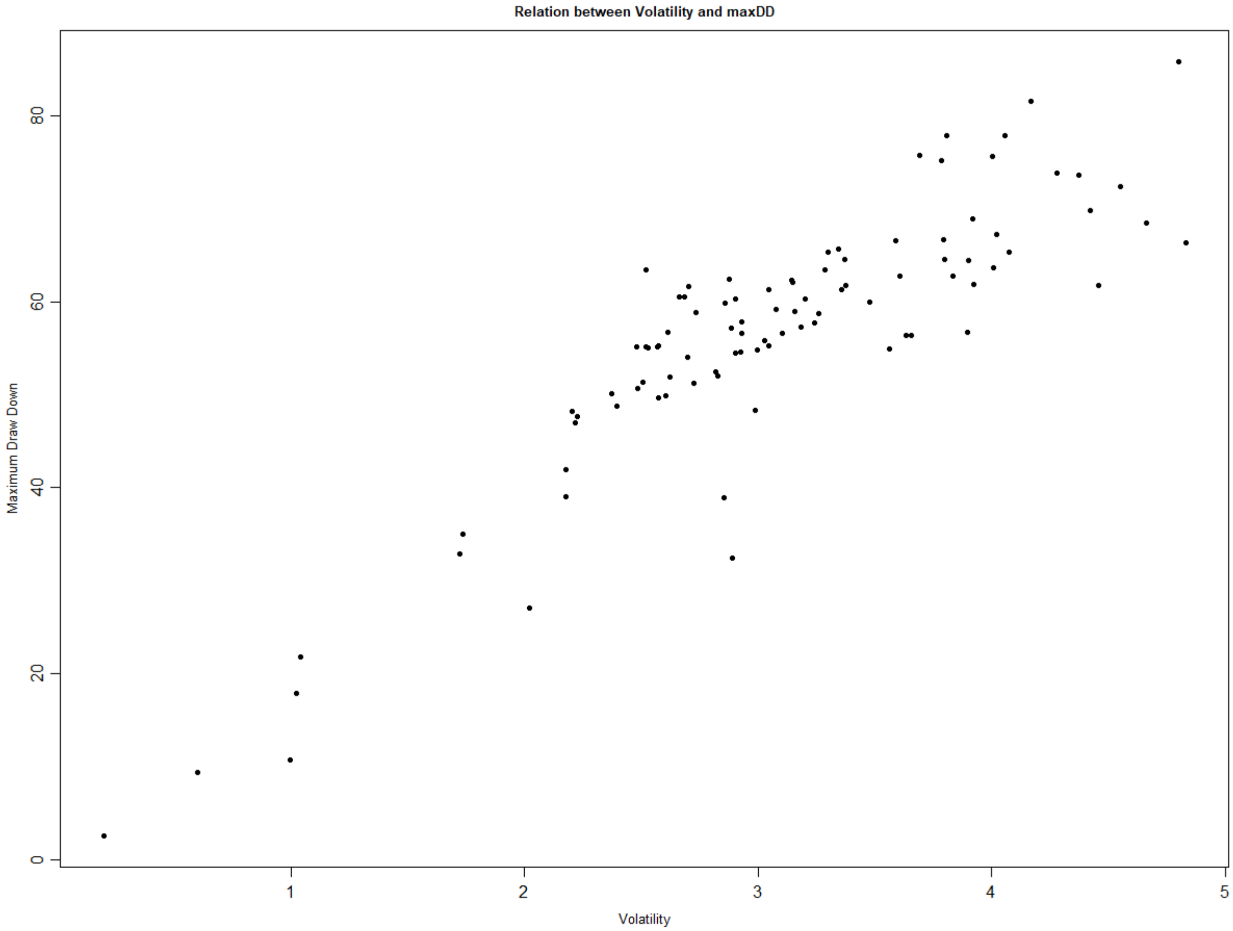


Figure 10: *Relation between Volatility and maxDD*

Correlation table The output from R with calculated correlation is following:

```
> #correlations in r
> cor(etfSum_analyse$Geo.mean,etfSum_analyse$maxTuW)
[1] -0.668348
> cor(etfSum_analyse$Volatility, etfSum_analyse$CVaR)
[1] -0.9922225
> cor(etfSum_analyse$maxTuW, etfSum_analyse$maxDD)
[1] 0.6468686
> cor(etfSum_analyse$maxDD, etfSum_analyse$Volatility)
[1] 0.8797983
```

Correlations by hand are calculated in R as a ratio between covariance of two random variables and their standard deviaton. The value starts to differ on fiths or sixths decimal places, therefore they are almost identical as a correlations given by r function cor. wti the value of -0.9922225, the strongest correlation is between VOlatility and CVAR, the weekest correlation is between Maximum time under water and Maximum Dorn down with the value 0.6468686.

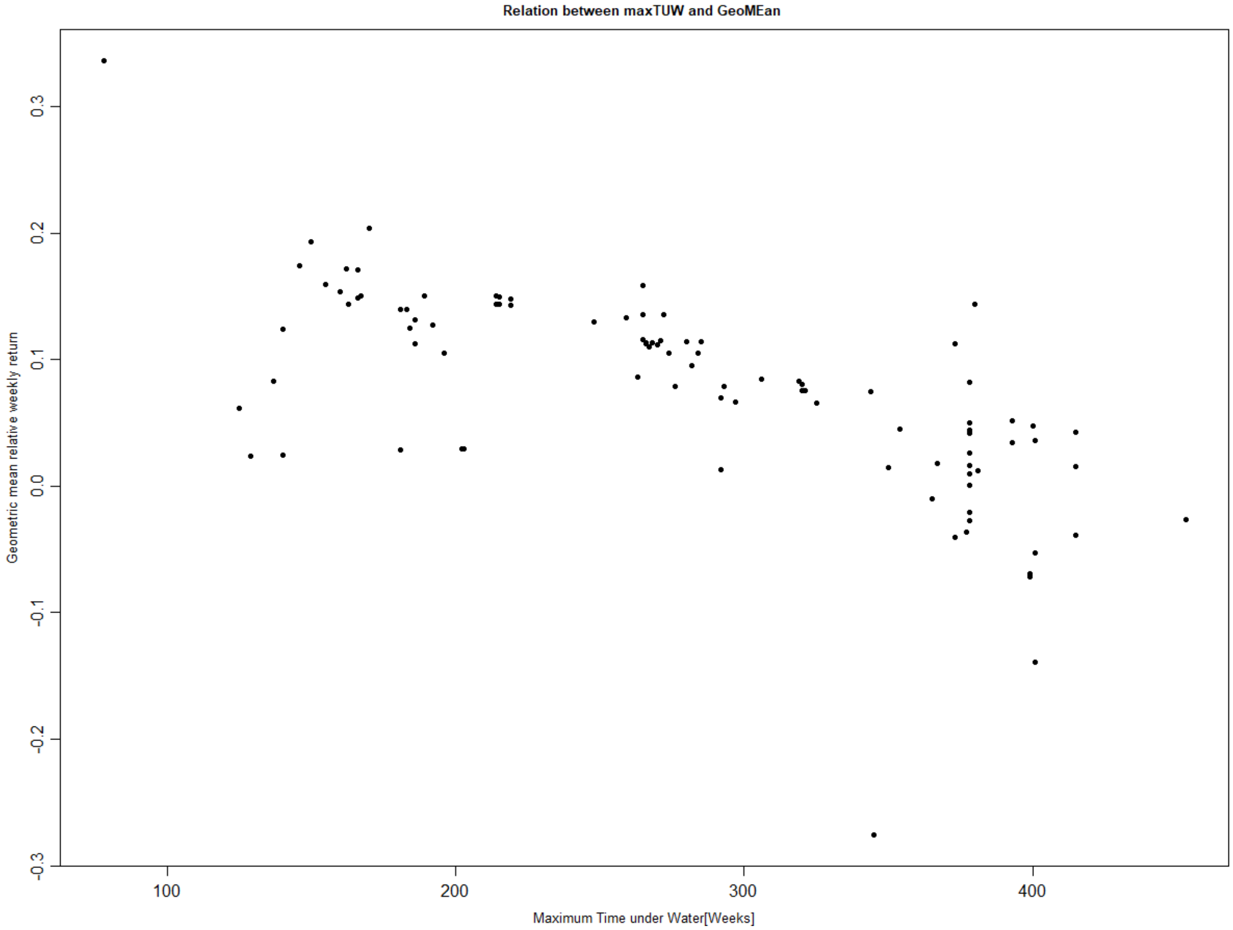
l)

As a explanatory variable, I have chosen Maximum Time Under water, mainly because its a time variable so it makes sence for me to model how geometric mean change with time needed for regaining historical peak.

Statement: Geometric mean of Weekly returns decreases linearly with the Maximum Time under water in Weeks increasing.

Assumption: Errors ϵ_i are random variables, independently and identically distributed with ϵ_i $N(0, \sigma^2$

Scatter plot:



Linear regression model calculated by hand in R

```
> ## A simple linear regression manual calculation
> ## Read data
> y <- etfSum_analyse$Geo.mean
> x <- etfSum_analyse$maxTuW
> ## Calculate averages
> xbar <- mean(x)
> ybar <- mean(y)
> ## Parameters estimates
> Sxx <- sum((x - xbar)^2)
> beta1hat <- sum((x - xbar)*(y - ybar)) / Sxx
> beta0hat <- ybar - beta1hat * xbar
> ##summary
> #average of Geometric means of relative weekly returns
> print(xbar)
[1] 280.1895
> #average Maximum Time under water
> print(ybar)
[1] 0.07690416
> ## Parameters estimates
> print(Sxx)
[1] 792778.6
> #the least squares estimates
> print(beta1hat)
[1] -0.0005885228
> print(beta0hat)
[1] 0.2418021
```

Plot of the model

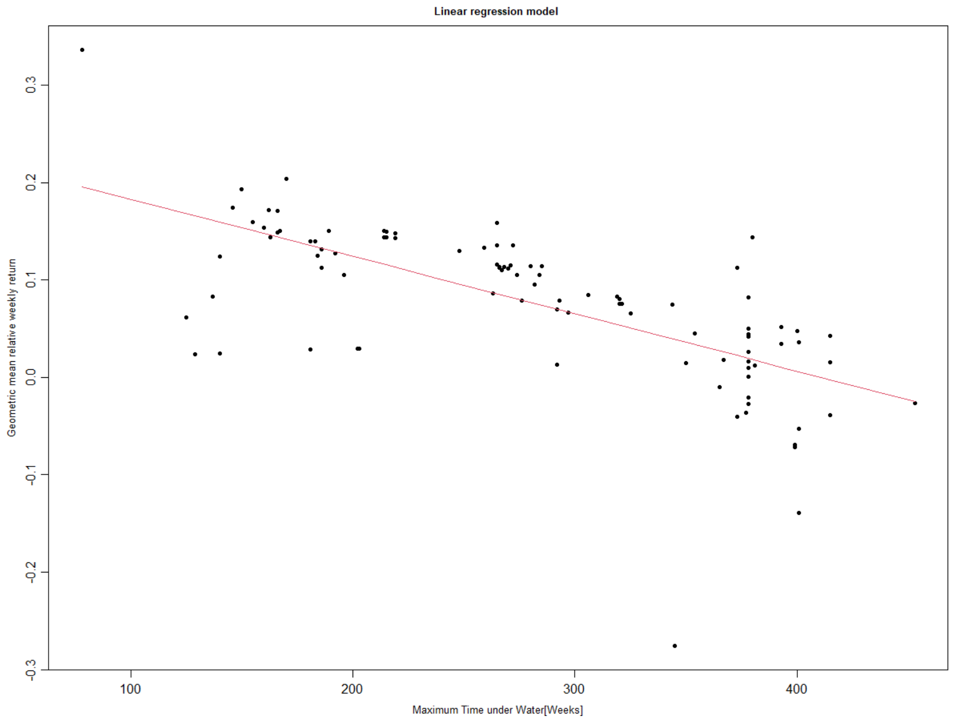


Figure 11: *Linear regression model*

Summary output from R

```
> summary(fit)

Call:
lm(formula = y ˜ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31441 -0.01396  0.01816  0.03021  0.14049

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.418e-01  2.002e-02  12.080  < 2e-16 ***
x           -5.885e-04  6.792e-05  -8.665 1.36e-13 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 0.06048 on 93 degrees of freedom
Multiple R-squared:  0.4467,  Adjusted R-squared:  0.4407
F-statistic: 75.08 on 1 and 93 DF,  p-value: 1.362e-13
```

The coefficients are same as the ones calculated ”by hand”

$$\S_{xx} = -792778.6$$

$$\hat{\beta}_1 = -0.0005885228$$

$$\hat{\beta}_0 = 0.2418021$$

n)

To estimate in the linear regression model how much of the variation in the outcome (Y) is explained by the input (x) we can use Coefficient of determination R^2 which is equal to the squared sample correlation coefficient (ρ^2) (proof in the textbook, eq 5.83).

$$r^2 = \frac{\sum_i (y_i - \hat{y})^2}{(y_i - \bar{y})^2} = 0.4466891 \tag{11}$$

44.7 percent of the variation is explained.