Ryerson University
Department of Mathematics

MTH404 Probability and Statistics II Assignment II

By: Sophia Rybnik

Student Number: 501015789

Section Number: 2

Date: April 14th, 2022

I, Sophia Rybnik, am solely responsible for the content of the following report.

S.R.

Introduction

The following report contains the derivation of the Fisher Information matrix and of the Gamma distribution and the maximum likelihood estimation of its parameters given a set of data.

Additionally, an in-depth analysis of the relationship between oil prices and the S&P 500 and Nasdaq, respectively, was conducted. In the report, we build a simple linear regression model for both cases. The model for oil returns vs the Nasdaq returns over the period from January 1997 to October 2006 yields a beta of 0.012960%. Over the same time period, the model for oil returns vs the S&P 500 yields a beta of 0.010944. The interpretation of the value of beta is the change in the dependent variable when the independent variables experiences a one-unit change. That is, for each 1% return observed from the S&P 500, according to the model we can expect to see a 0.045330% return over the mean return in oil. Similarly, for each 1% return observed from the Nasdaq, we can expect to see a 0.021258% return over the mean return in oil.

Beyond the financial data part of the report, the main topic covered was hypothesis testing. The test statistics and their distributions are built under the assumption of the null hypothesis. Additionally, in cases where the sample size is large, the CLT is applied.

Please note that some important findings and conclusions have been included at the end of the report. A bibliography can be found on the last page of the report.

Assignment 2

Sophia Rybnik

14/04/2022

Questions

Q1)

a.

The Maximum likelihood estimation is a method that find the values of $\hat{\alpha}$ and $\hat{\beta}$that result in the curve that best fits the data. These estimators are given by the 'fitdistr' function in R. With $\hat{\alpha} \approx 4.8623$ and $\hat{\beta} \approx 0.098910$, we are most likely to observe the given sample.

Recall that the standard error of an estimate of a parameter is the standard deviation of its sampling distribution, i.e. how accurate the estimator is in comparison with the population parameter. In this case, $SE(\hat{\alpha}) \approx 0.21035$ and $SE(\hat{\beta}) \approx 0.0045075$.

Shape	Rate
4.862254857 (0.210345254)	0.098909597 (0.004507526)

b.

As n = 100, the CLT can be applied. Given that $\alpha = 0.05 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.975$

The 95% confidence interval for the parameters is built as follows:

$$\hat{\theta} \pm Z_{1-\frac{\alpha}{2}} S.E(\hat{\theta})$$

95% confidence interval for α:

$$4.862254857 \pm (1.96)(0.210345254) \Leftrightarrow (4.4450, 5.2745)$$

95% confidence interval for β:

$$0.098909597 \pm (1.96)(0.004507526) \Leftrightarrow (0.090075, 0.10774)$$

We can conclude from the given sample that 95% of the time, the population distribution shape will be between 4.4450 and 5.2745. Moreover, 95% of the time, the population distribution rate will be between 0.090075 and 0.10774.

c.

The Fisher Information Matrix for the Normal Distribution is derived as follows: The parameters of the Gamma distribution, β and α, can be estimated by the Method of Moments as follows:

1) Moment Generating Function of the Gamma distribution: Recall that the MGF of a distribution is given by:

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= \int_0^\infty e^{tx} f(x)dx \end{aligned}$$

Take $t < \beta$. It can be proven that the MGF of the Gamma distribution is not defined for $t \geq \beta$.

$$\begin{aligned} M_X(t) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x} dx & & let\ u = (\beta - t)x \Leftrightarrow du = (\beta - t)dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{u}{\beta - t}\right)^{\alpha-1} e^{-u} \frac{du}{\beta - t} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)(\beta - t)^\alpha} \int_0^\infty u^{\alpha-1} e^{-u} du & & Note : \Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \\ \Leftrightarrow M_X(t) &= \frac{\Gamma(\alpha)\beta^\alpha}{\Gamma(\alpha)(\beta - t)^\alpha} \\ &= \left(\frac{\beta}{\beta - t}\right)^\alpha \\ &= \left(1 - \frac{t}{\beta}\right)^{-\alpha} \end{aligned}$$

2) First and Second Moments of the MGF:

$$\begin{aligned} M'_X(t) &= -\alpha \left(1 - \frac{t}{\beta}\right)^{-\alpha-1} \left(\frac{-1}{\beta}\right) \\ &= \frac{\alpha}{\beta} \left(\frac{\beta - t}{\beta}\right)^{-\alpha-1} \\ &= \frac{\alpha\beta^\alpha}{(\beta - t)^{\alpha+1}} \end{aligned}$$

$$M'_X(0) = E[X] = \frac{\alpha}{\beta}$$

$$\begin{aligned} M''_X(t) &= \alpha\beta^\alpha(-\alpha - 1)(\beta - t)^{-\alpha-2}(-1) \\ &= \frac{\alpha\beta^\alpha(\alpha + 1)}{(\beta - t)^{\alpha+2}} \end{aligned}$$

$$M''_X(0) = \frac{\alpha(\alpha + 1)}{\beta^2}$$

3) Estimation of the Parameters:

$$
\begin{aligned}
E[X] &= \bar{X} \\
&\Leftrightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \\
&\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\alpha}{\beta}
\end{aligned}
$$

$$
\begin{aligned}
E[X^2] &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\
&\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{\alpha(\alpha+1)}{\beta^2} \\
&\Leftrightarrow \alpha = \beta\bar{X}
\end{aligned}
$$

$$
\begin{aligned}
\mathit{Then}, \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2 &= \frac{\beta\bar{X}(\beta\bar{X}+1)}{\beta^2} \\
&\Leftrightarrow \beta(\bar{X^2} - \frac{1}{n}\sum_{i=1}^{n} x_i^2) = -\bar{X} \\
&\Leftrightarrow \beta = \frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}}
\end{aligned}
$$

$$
\mathit{Hence}, \qquad \alpha = \beta\bar{X} \Leftrightarrow \alpha = \frac{\bar{X^2}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}}
$$

$$
\therefore \hat{\beta} = \frac{\bar{X}}{\hat{\sigma^2}}, \quad \hat{\alpha} = \frac{\bar{X^2}}{\hat{\sigma^2}}
$$

Now, using the above derived estimators of the two parameters, the corresponding estimates based on the sample data are $\hat{\alpha} \approx 4.7269$ and $\hat{\beta} \approx 0.09616$. From part a) it is clear that the estimate is very close to the values given by the Maximum-likelihood fitting function in R.

Q2)

Let $X = (X_1, X_2, ..., X_n)$ be a normally distributed random sample with $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$. $X \sim f(x|\theta)$, with $\theta = (\mu, \sigma^2)$.

Recall that the PDF of the normal distribution is given by $f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}{\frac{(x-\mu)}{\sigma^2}}^2}$.

The Fisher Information Matrix for the Normal Distribution is derived as follows:

1) Maximum Likelihood Function:

$$
\begin{aligned}
L(\mu, \sigma^2, x) &= \Pi_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2} \frac{(x-\mu)}{\sigma^2}^2} \\
&= (2\pi\sigma^2)^{\frac{-n}{2}} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2}
\end{aligned}
$$

2) Log Likelihood Function:

$$
l(\mu, \sigma^2, x) = \frac{-n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2
$$

3) For a multi-parameter distribution, $\theta = (\theta_1, \theta_2, ..., \theta_k)^T$.

The 1st order derivative of the Log Likelihood function with respect to θ is a k dimensional vector:

$\frac{\partial l(\theta,x)}{\partial \theta} = \left(\frac{\partial l(\theta)}{\partial \theta_1}, \cdots \frac{\partial l(\theta)}{\partial \theta_k} \right)^T$

The 2nd order derivative of the Log Likelihood function with respect to θ is a $k \cdot k$ matrix:

$\frac{\partial^2 l(\theta,x)}{\partial \theta^2} = \left[\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right]_{i=1,..,k; j=1,...,k}$

For the Normal distribution, $\theta = (\mu, \sigma^2)^T$, $\frac{\partial l(\theta,x)}{\partial \theta}$ is a 2x1 vector, and $\frac{\partial^2 l(\theta,x)}{\partial \theta^2}$ is a 2x2 matrix.

The Fisher Information Matrix for a multi-parameter distribution is defined as:

$$
\begin{aligned}
I(\theta) &= E \left[\frac{\partial l(\theta)}{\partial \theta} \left(\frac{\partial l(\theta)}{\partial \theta} \right)^T \right] \\
&= cov \left(\frac{\partial l(\theta)}{\partial \theta} \right) \\
&= -E \left(\frac{\partial^2 l(\theta)}{\partial \theta^2} \right)
\end{aligned}
$$

4) The 1st and 2nd derivatives with respect to μ and σ can be found.

1st and 2nd Derivatives with respect to μ:

$$
\begin{aligned}
\frac{\partial l(\mu, \sigma^2, x)}{\partial \mu} &= \sum_{k=1}^{n} \frac{x_k - \mu}{\sigma^2} \\
&= \sum_{k=1}^{n} \left(\frac{x_k}{\sigma^2} - \frac{\mu}{\sigma^2} \right) \\
\frac{\partial^2 l(\mu, \sigma^2, x)}{\partial \mu^2} &= \frac{-n}{\sigma^2}
\end{aligned}
$$

1st and 2nd Derivatives with respect to σ:

For simplicity, let $\tau = \sigma^2$. Then, $l(\mu, \tau, x) = \frac{-n}{2} ln(2\pi) - \frac{n}{2} ln(\tau) - \frac{1}{2\tau} \sum_{k=1}^{n} (x_k - \mu)^2$.

$$
\begin{aligned}
\frac{\partial l(\mu, \tau, x)}{\partial \tau} &= \frac{-n}{2\tau} + \frac{1}{2\tau^2} \sum_{k=1}^{n} (x_k - \mu)^2 \\
\frac{\partial^2 l(\mu, \tau, x)}{\partial \tau^2} &= \frac{n}{2\tau^2} - \frac{1}{\tau^3} \sum_{k=1}^{n} (x_k - \mu)^2
\end{aligned}
$$

Mixed 2nd Partial Derivatives:

$$\begin{aligned}\frac{d^2l(\theta)}{d\mu\tau} &= \frac{\partial l}{\partial\mu}\left[\frac{\partial l}{\partial\tau}\right]\\ &= \frac{\partial l}{\partial\mu}\left[\frac{-n}{2\tau}+\frac{1}{2\tau^2}\sum_{k=1}^{n}(x_k-\mu)^2\right]\\ &= \frac{-1}{\tau^2}\sum_{k=1}^{n}(x_k-\mu)\end{aligned}$$

$$\begin{aligned}\frac{d^2l(\theta)}{d\tau\mu} &= \frac{\partial l}{\partial\tau}\left[\frac{\partial l}{\partial\mu}\right]\\ &= \frac{\partial l}{\partial\tau}\left[\sum_{k=1}^{n}\frac{x_k-\mu}{\tau}\right]\\ &= \frac{-1}{\tau^2}\sum_{k=1}^{n}(x_k-\mu)\end{aligned}$$

5) The components of the Fisher Information Matrix can be combined as follows:

$$\begin{aligned}I(\theta) &= -E\left[\frac{\partial^2 l(\theta)}{\partial\theta^2}\right]\\ &= -E\begin{bmatrix}\frac{\partial^2 l(\theta)}{\partial\mu^2} & \frac{\partial^2 l(\theta)}{\partial\mu\partial\tau}\\ \frac{\partial^2 l(\theta)}{\partial\tau\partial\mu} & \frac{\partial^2 l(\theta)}{\partial\tau^2}\end{bmatrix}\\ &= -E\begin{bmatrix}\frac{-n}{\tau} & \frac{-1}{\tau^2}\sum_{k=1}^{n}(x_k-\mu)\\ \frac{-1}{\tau^2}\sum_{k=1}^{n}(x_k-\mu) & \frac{n}{2\tau^2}-\frac{1}{\tau^3}\sum_{k=1}^{n}(x_k-\mu)^2\end{bmatrix}\end{aligned}$$

Note that:

i) The expectation of the sum of the deviations from the mean is zero, i.e. $E\left[\sum_{k=1}^{n}(x_k-\mu)\right]=0$

ii) The expected value of $\sum_{k=1}^{n}(x_k-\mu)^2$ is $n\tau$.

$$\begin{aligned}E\left[\frac{n}{2\tau^2}-\frac{1}{\tau^3}\sum_{k=1}^{n}(x_k-\mu)^2\right] &= \frac{n}{2\tau^2}-\frac{n}{\tau^2}\\ &= \frac{-n}{2\tau^2}\end{aligned}$$

Finally, subbing back σ^2 for τ, the information matrix of the Normal distribution is:

$$I(\theta)=\begin{bmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}\end{bmatrix}$$

Q3)

$H_0 : p_1 = p_2$ $H_1 : p_1 > p_2$

The hypotheses constitute a one-tailed test for the difference of sample proportions. The null hypothesis will be rejected if the proportion of acceptable electronic components of the foreign supplier (p_1) is significantly greater than the proportion of that of the domestic supplied (p_2).

Note that $n_1 + n_2 = 80 + 100 = 180$ and $(n_1 + n_2)p_1 = (100 + 80)0.9 = 162$. Hence, $n_1 + n_2 > 30$ and $(n_1 + n_2)p_1 > 5$

Since the null hypothesis states that $p_1 = p_2$, then then the pooled sample proportion is an estimator of the common population proportion:

$$\begin{aligned}\hat{p} &= \frac{\hat{p}_1 * n_1 + \hat{p}_2 * n_2}{n_1 + n_2} \\ &= \frac{(0.9)(100) + (0.7)(80)}{100 + 80} \\ &\approx 0.8111\end{aligned}$$

The test statistic is the z-score:

$$\begin{aligned}Z_{p,2} &= \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \\ &= \frac{0.9 - 0.7}{\sqrt{\frac{0.8111(1-0.8111)}{100} + \frac{0.8111(1-0.8111)}{80}}} \\ &\approx 3.406\end{aligned}$$

For the given problem, the rejection region is $Z_2 > Z_{1-\alpha}$, where $\alpha = 0.05 \Leftrightarrow 1 - \alpha = 0.95$. Given that $Z_{0.95} = 1.645$, $3.406 > 1.645$, hence the sample falls into the rejection region, and H_0 is rejected. It can be concluded that p_1 is statistically greater than p_2. Thus, the foreign supplier has a greater proportion of acceptable electronic components than the domestic supplier.

Q4)

a.

A	B	C	D
33	32	31	29
38	40	37	34
36	42	35	32
40	38	33	30
31	30	34	33
35	34	30	31

The given data was transformed into a dataframe, and inputted into the R ANOVA command. The table has been summarized below:

Source of Variation	SS	df	Mean Squares	F-test
Treatments	77.50	3	25.833	2.3883
Error	216.33	20	10.817	
Total	293.83	23		

b.

To make a conclusion based on the 4 programs, the test hypotheses are set up as follows:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 :$ At least one μ_i is different, i = 1, 2, 3, or 4

Recall that $F = \frac{MST}{MSE} \sim F(k-1, n-k)$. Given that $\alpha = 0.05$,$k = 4$, and $n = 24$, the null hypothesis can be rejected if $F > F_{0.95}(4, 20)$ or $p - value < 0.05$.

At a significance level of 5%, $F = 2.3883 \ngtr F_{0.95} = 3.0984$. Alternatively, $p - value \approx 0.085282 \nless 0.05$.

As shown, both methods fail to reject the null hypothesis. Therefore, it cannot be concluded whether there are or are not significant differences between the effectiveness of the programs.

c.

Note that $n_b = n_c = 6$. Since the sample size is small, the difference of means follows a t-student distribution. Under the assumption of normality, independence of samples B and C, and equal population variances, the confidence interval for the difference of means is built as follows:

$$(\bar{X_b} - \bar{X_c}) \pm t_{1-\frac{\alpha}{2}}(n-k)S\sqrt{\frac{1}{n_b} + \frac{1}{n_c}}$$

where $S = \sqrt{MSE} = \sqrt{10.817}$ from the ANOVA table.

Since $1 - \alpha = 0.95$,
$\Leftrightarrow \alpha = 0.05$
$\Leftrightarrow 1 - \frac{\alpha}{2} = 0.975$

Then, the 95% confidence interval is given by:

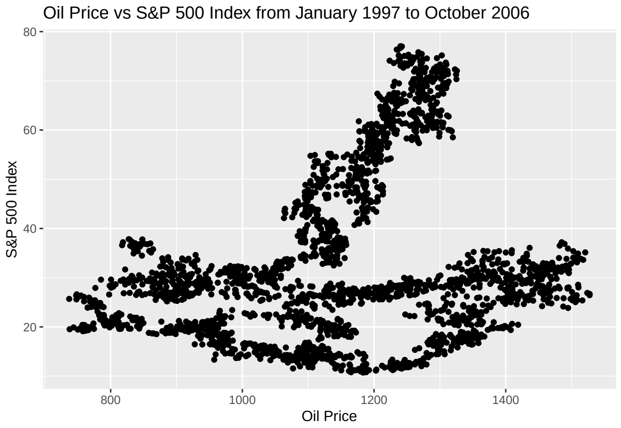
$$(36 - 33.333) \pm t_{0.975}(20)\sqrt{10.817\left(\frac{1}{6} + \frac{1}{6}\right)}$$
$$\approx 2.6667 \pm 3.9610$$

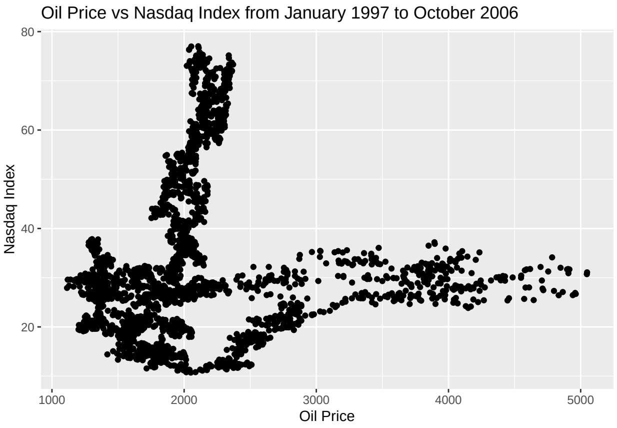
Therefore, the 95% confidence interval for the difference of means of the programs B and C is $(-1.2943, 6.6276)$.

Since the confidence interval is a range of likely values for the difference in means, if it contains zero then no difference between the sample means is a likely possibility. Based on this interval, we conclude that there is no statistically significant difference in the mean number of treads between the programs because the the 95% confidence interval includes the null value, zero.

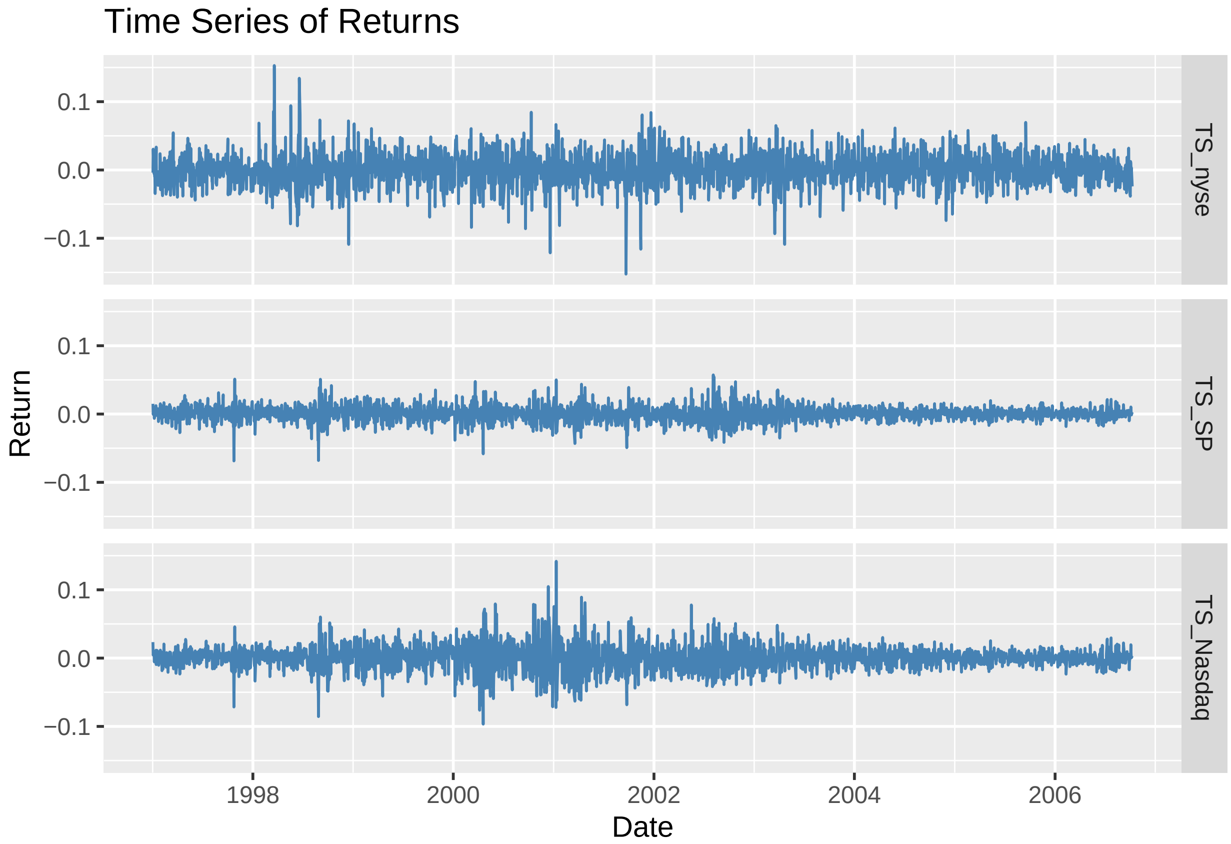
Q5)

a.





b.



c.

A simple linear regression model is given as $Y_i = \alpha + \beta X_i + \epsilon_i, \;\; i = 1, 2, ..., n.$

The estimators of α and β, $\hat{\alpha}$ and $\hat{\beta}$ respectively, are the least square estimators if they minimize the sum of the squares of the difference between the dependent values of the data and the model, i.e.

$$\begin{aligned} SSE(\hat{\alpha}, \hat{\beta}) &= min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - (\alpha + \beta X_i))^2 \end{aligned}$$

After differentiating the SSE with respect to the parameters and setting them equal to zero, we can define the least square estimators as follows:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\alpha} = \bar{Y} - \beta\bar{X}$$

Linear regression model for Oil Returns vs Nasdaq Returns:

Using the data from part b), using R we can find:

$$n = 2442$$
$$\sum X_i \approx 0.97290$$
$$\sum X_i^2 \approx 0.83342$$
$$\sum Y_i \approx 1.50871$$
$$\sum Y_i^2 \approx 1.3674$$
$$\sum X_i Y_i \approx 0.018310$$

Then, it follows that:

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n} \\ \Leftrightarrow S_{xy} &\approx 0.017709 \end{aligned}$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \\ \Leftrightarrow S_{xx} &\approx 0.83303 \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} \\ \Leftrightarrow S_{yy} &\approx 1.3665 \end{aligned}$$

Then, $\hat{\beta} \approx 0.021258$ and $\hat{\alpha} \approx 0.00060934$. Thus, the estimated regression line is $\hat{y} = 0.00060934 + 0.021258x$.

Linear regression model for Oil Returns vs S&P 500 Returns:

Using the data from part b), using R we can find:

$$n = 2442$$
$$\sum X_i \approx 0.74754$$
$$\sum X_i^2 \approx 0.33014$$
$$\sum Y_i \approx 1.50871$$
$$\sum Y_i^2 \approx 1.3674$$
$$\sum X_i Y_i \approx 0.015417$$

Then, it follows that:

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} X_iY_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n} \\ \Leftrightarrow S_{xy} &\approx 0.014955 \end{aligned}$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \\ \Leftrightarrow S_{xx} &\approx 0.32991 \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} \\ \Leftrightarrow S_{yy} &\approx 1.3665 \end{aligned}$$

Then, $\hat{\beta} \approx 0.045330$ and $\hat{\alpha} \approx 0.00060394$. Thus, the estimated regression line is $\hat{y} = 0.00060394 + 0.045330x$.

The interpretation of the value of beta is the change in the dependent variable when the independent variable experiences a one-unit change. In this case, the value of the beta means that for each 1% return observed from the S&P 500, we can expect to see a 0.045330% return in oil over the mean return in oil.

d.

The Analysis of Variance can be done using the conclusions of the least square estimator method from above.

The total Sum of Squares is: $TotalSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \Leftrightarrow TotalSS = S_{yy}$

Then, the Sum of Squares for regression:

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(\alpha + \beta Xi - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(\bar{Y} - \beta\bar{X} + \beta Xi - \bar{Y})^2 \\ &= \beta^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} \\ \Leftrightarrow SSR &= \frac{(S_{xy})^2}{S_{xx}} \end{aligned}$$

Combining the above, the Sum of Squares for Error is given by $SSE = TotalSS - SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$.

The ANOVA table for the linear regression model with Nasdaq returns as independent variable:

Source of Variation	SS	df	Mean Squares	F-test
Regression	0.00037646	1	0.00037646	0.67239
Error	1.3661	2440	0.00055988	
Total	1.366481	2441		

The ANOVA table for the linear regression model with S&P returns as independent variable:

Source of Variation	SS	df	Mean Squares	F-test
Regression	0.00067790	1	0.00067790	1.2111
Error	1.3658	2440	0.00055976	
Total	1.3665	2441		

To test whether S&P returns have an impact on Oil returns, we need to test about the validity of the model, i.e. whether β is significantly different from zero. The hypotheses are set up as follows:

$$H_0 : \beta = \beta_0 \qquad H_1 : \beta \neq \beta_0$$

From the ANOVA table for regression we know that $F = 1.2111$. The null hypothesis can be rejected if $F > F_{1-\alpha}(1, n-2)$.

Using R, $F = 1.2111 \not> F_{0.95}(1, 2440) \approx 3.8456$. Since the observed value of F does not fall into the rejection region, we fail to reject H_0. There is not enough evidence to conclude that the daily returns on the S&P index have an impact on daily oil returns.

e.

The linear correlation coefficient for a collection of n pairs of coordinates in a sample is given as follows:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Using the values found above, $r = \frac{0.014955}{\sqrt{1.3665 \cdot 0.32991}} \approx 0.022273$.

The positive value of r indicates that when daily returns on the S&P are positive, daily returns on oil prices also tend to be positive. Since |r| is much closer to 0 than it is to 1, the linear relationship between oil returns and S&P returns is not strong. This confirms our findings from part d).

The hypothesis for a positive linear relationship between oil returns and S&P 500 returns, a test against zero correlation is set up as follows:

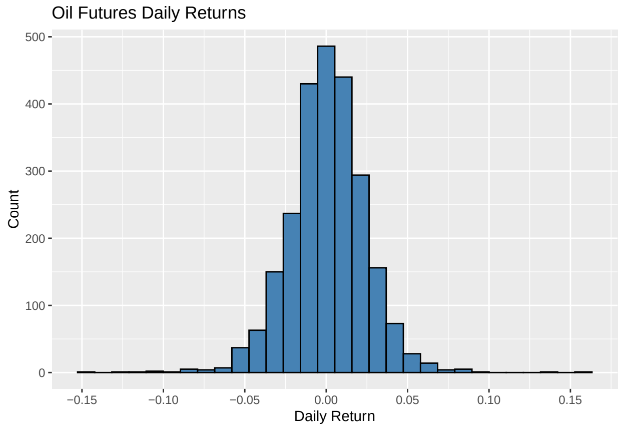
$$H_0 : \rho = 0 \qquad H_1 : \rho > 0$$

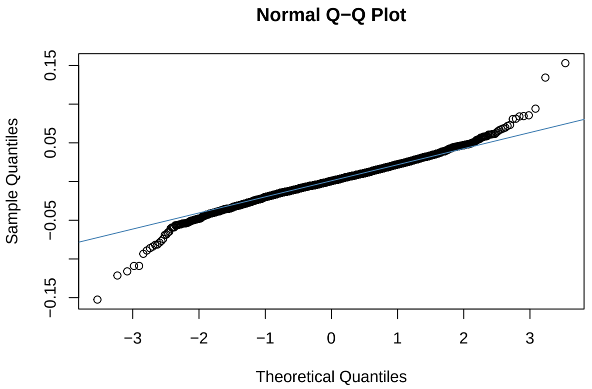
The test statistic is $t_r = r\sqrt{\frac{n-2}{1-r^2}}$, whose distribution under the null hypothesis is $t_r \sim t(n-2)$. Using the above calculated correlation coefficient, $r \approx 0.022273$, we find that $t_r \approx 1.1005$. The null hypothesis can be rejected if $t_r > t_{1-\alpha}(n-2)$.

Using R, $t_r \approx 1.1005 \ngtr t_{0.95}(2440) \approx 1.6455$. Since the observed value of t does not fall into the rejection region, we fail to reject H_0. There is not enough evidence to conclude that the daily returns on the S&P index and oil are linearly positively related.

Q6)

a.





b.

The hypotheses for the chi-square goodness of fit test are set up as follows:

$$H_0 : X \sim N(\mu, \sigma^2) \qquad H_1 : X \nsim N(\mu, \sigma^2)$$

The decision rule to reject the null hypothesis is achieved by computing the test statistic, $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. If $\chi^2 > \chi^2_{1-\alpha}(k - p - 1)$, the null hypothesis may be rejected.

Using R, the intervals for data were set up as follows, in order to have at least 5 expected occurrences in each bin:

```
      Interval Observed Frequency Expected Frequency
 [-0.08,-0.06)                  8           11.90539
 [-0.06,-0.05)                 26           26.86118
 [-0.05,-0.04)                 50           65.47620
 [-0.04,-0.03)                116          133.83605
 [-0.03,-0.02)                178          229.40777
 [-0.02,-0.01)                344          329.76100
     [-0.01,0)                445          397.51545
      [0,0.01)                465          401.86122
   [0.01,0.02)                354          340.69502
   [0.02,0.03)                224          242.22540
   [0.03,0.04)                120          144.42116
   [0.04,0.05)                 60           72.20859
   [0.05,0.06)                 19           30.27471
      [0.06,8]                 22           14.74984
```

Then, we compute $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx 50.918$. After merging, k = 14 and p = 2 remains as previously. So, $\chi^2_{0.95}(14 - 2 - 1) = \chi^2_{0.95}(11) \approx 19.675$. Since $\chi^2 = 50.918 > \chi^2_{0.95}(14 - 2 - 1) = 19.67$, we can reject the null hypothesis and claim that the oil returns are not normally distributed.

c.

The test hypotheses for the Kolmogorov-Smirnov test are set up as follows:

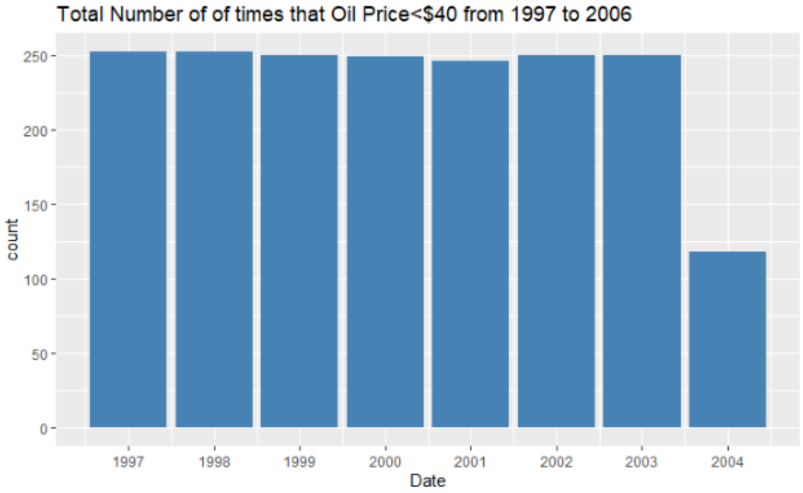
$$H_0 : X \sim N(\mu, \sigma^2) \qquad H_1 : X \nsim N(\mu, \sigma^2)$$

To test the distribution of the oil returns with a normal distribution, 2442 data points (the number of days of return data) were simulated from a normal distribution. The simulated data has the same mean and variance as the distribution of the oil returns. Using the Kolmogorov-Smirnov test in R, since the outputted $p - value = 2.2e - 16 < \alpha$, the null hypothesis can be rejected. Hence, the data does not follow a normal distribution.

Conclusion

In completing the assignment, I was able to take a closer look at validating my hypothesis in order to arrive at statistically significant conclusions. A part of the assignment I would like to take a closer look at is deriving a relationship between NYSE Oil prices and Index values over the period from 1997 to 2006.

The shape of the scatter plots in Q5) were interesting to note. At around $40 USD, the oil price compared to both indexes maps out a relatively straight line. At prices less than $40, the plot of oil prices versus index values is extremely volatile in comparison. Just from observation, we cannot confidently say that the price of oil and the value of the index is linearly correlated if the oil price is greater than $40. After some inspection, we can see that the price of oil was below $40 only before 2004. In fact, prices increased in 29 of the 40 months between September 2003 and December 2006, which explains the significant drop in the frequency of the bars.



The rise in oil prices during fiscal year 2004-05 was due to various factors, including increasing demand from China and other emerging economies. On the supply side, promises from Saudi Arabia and other members of the Organization of Petroleum Exporting Countries (OPEC) to pump more oil fell below projections with limited spare capacity for production. Additionally, Iraq, whose economy relied almost solely on revenue from oil exports, was coming into its 2nd year of the war, raising geopolitical tensions in the middle east. Although in part d) and e) we proceed to see that there is a positive linear relationship between returns on Oil and the S&P 500, without further analysis, the observations from the scatter plot do not help arrive to a conclusion.

I would also like to bring attention to Q6). A requirement to be able to apply the Chi-Square distribution is that the data needs to be “binned” into distinct categories, and that the expected frequency has at least 5 observations in every bin. If you take a look at my code, you will see that in order to bin the data adequately, I simulated data from a normal distribution with same sample size, mean and variance as oil returns. This helped give me an idea of which bins needed to be merged, in order to meet the requirement of 5 observations per bin. From there, I followed the steps of the goodness of fit test, by comparing the observed frequencies with the expected theoretical frequencies. After some research, I realize that an implication of simulating and then artificially binning data is that some important information is lost. Since the chi-squared test requires binned data, there is room for error in this method.

Using two different methods, we were able to reject the notion that the returns of oil futures follow a normal distribution. In fact, although the assumption of normality is widely used in financial modeling, distribution curves of asset returns tend to have fatter tails than that of a normal distribution. An example of that is the average monthly returns of the S&P 500 since 1950. The high kurtosis of certain asset return’s distributions means that extreme events occur more frequently in reality than what a normal distribution predicts.

Bibliography

“Canadian Energy Pricing Trends 2000-2010 - Energy Facts.” *CER*, Government of Canada, 15 Nov. 2020, https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/archive/canadian-energy-pricing-trends-2011/canadian-energy-pricing-trends-2000-2010-energy-facts.html.

Olivares, P. (2022, January). Data Analysis in Finance and Environment.

“Simple Linear Regression.” *Carnegie Mellon University Department of Statistics and Data Science*, https://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pdf.

Taylor, Courtney. “Compare Two Population Proportions with This Hypothesis Test.” *ThoughtCo*, ThoughtCo, 3 Apr. 2019, https://www.thoughtco.com/two-population-proportions-hypothesis-test-4075530.

Wolpert, Robert L. “Fisher Information & Efficiency - Duke University.” *Duke University Department of Statistical Science*, https://www2.stat.duke.edu/courses/Spring16/sta532/lec/fish.pdf.

Yiu, Tony. “Are Stock Returns Normally Distributed?” *Medium- Towards Data Science*, 29 Mar. 2020, https://towardsdatascience.com/are-stock-returns-normally-distributed-e0388d71267e#:~:text=For%20example%2C%20the%20return%20of,(a.k.a.%20the%20investment's%20risk).

Zheng, Songfeng. “Fisher Information and Cramer-Rao Bound.” *Missouri State Math Department*, https://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Fisher_info.pdf.