Advanced Statistics Midterm

Group: Potato

November 11

1 Probability

Let X be a random variable. X has an MGF, M_X, defined as

$$M_X(t) := \mathbb{E}\exp(tX)$$

a) Show that if $X \sim \mathcal{N}(\mu, \sigma^2)$ its MGF exists and demonstrate its form.

Let g(X) = $\exp(tX)$. This allows me to rewrite M_X into the form $\mathbb{E}\, g(X)$. g(X) is a function on a random variable and thus is a random variable itself. Following we apply the information given in the lecture notes "2.2 Expectations"[1] to write out the expression as

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Where f is the pdf of X. Since X is normally distributed we know that the pdf is the normal density function. Therefore, the expression can be written as

$$\int_{-\infty}^{\infty} e^{tx}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

Continuing we define the standard normal random variable $z = \frac{x-\mu}{\sigma}$, which implies that $x = z\sigma + \mu$. We do so to utilize the change of variables technique. By definition of z it follows that $\frac{dz}{dx} = \frac{1}{\sigma}$ and $dx = \sigma dz$. With this information we can rewrite the expression as

$$\int_{-\infty}^{\infty} e^{t(z\sigma+\mu)}\frac{\sigma}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = e^{t\mu}\int_{-\infty}^{\infty} e^{zt\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$

Set a = $t\sigma$ and use the information that

$$e^{za}e^{-\frac{1}{2}z^2} = e^{za-\frac{1}{2}z^2} = e^{-\frac{1}{2}(z-a)^2+\frac{1}{2}a^2}$$

To rewrite the expression to

$$e^{t\mu}e^{\frac{1}{2}a^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-a)^2}$$

What remains in the integral is the normal density function $\mathcal{N}(z; \mu = t, \sigma^2 = 1)$, which is 1 when integrated over the entire range. Thus:

$$M_X(t) = e^{t\mu}e^{\frac{1}{2}\sigma^2 t^2}$$

b) Show that if $X \sim$ Poisson(λ) its MGF exists and demonstrate its form.

Following the same logic of 1a), i write out the expression as

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{e^{tx}\lambda^x}{x!}$$

From here we use the fact that $e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$ and that $e^{tx}\lambda^x = (e^t\lambda)^x$ to derive the formula:

$$M_X(t) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(e^t\lambda)^x}{x!} = e^{-\lambda}e^{e^t\lambda} = e^{\lambda(e^t-1)}$$

c) Show that if $X \sim$ t(1), the MGF does not exist.

t(1) is the Student's t-distribution with degrees of freedom 1. The existence of a MGF for a random variable entails that all the moments exist for that random variable. Therefore, to show that the MGF does not exist for $X \sim$ t(1), we will show that the first moment, the mean, does not exist. We will use the information from the lecture notes that if $\mathbb{E}\,|g(X)| = \infty$ then the expectation does not exist. Now we prove that $\mathbb{E}\,|X| = \infty$

$$\mathbb{E}\,|X| = \int_{-\infty}^{\infty} |x| f(x) dx = 2\pi^{-1} \int_{0}^{\infty} \frac{x}{x^2+1} dx$$

$$= 2\pi^{-1} \left[\frac{ln(x^2+1}{2} \right]_0^\infty = \pi^{-1} \left[ln(x^2+1) \right]_0^\infty$$

$$\pi^{-1} \left[ln(x^2+1) \right]_0^T = \pi^{-1} ln(T^2+1)$$

From this result it is clear that $\mathbb{E}\,|X| \to \infty$ as $T \to \infty$. It then follows that the first moment of random variable X does not exist, and therefore, neither does the MGF of X.

d) Prove Lemma 1.

Lemma 1. *If X has an MGF, then so does aX + b for any constant a, b $\in \mathbb{R}$ and $M_{aX+b}(t) = e^{bt} M_X(at)$.*

Through the definition of $M_X(t)$ given in this exercise and by the rule of exponents we can expand $M_{aX+b}(t)$:

$$M_{aX+b}(t) = \mathbb{E}\, e^{t(aX+b)} = \mathbb{E}\left[e^{aXt} e^{bt} \right]$$

By using the well known fact that constants can be dragged out of the expectation(also shown in lecture notes proposition 2.6 under 2.2 Expectations), it can be simplified to

$$e^{bt}\, \mathbb{E}\left[e^{aXt} \right]$$

Lastly, we set m = at to clearly illustrate the final result

$$e^{bt}\, \mathbb{E}\left[e^{aXt} \right] = e^{bt}\, \mathbb{E}\left[e^{mX} \right] = e^{bt} M_X(m) = e^{bt} M_X(at)$$

e) Prove Lemma 2.

Lemma 2. *If X, Y have MGFs and are independent, then so does Z=X+Y and $M_Z(t) = M_X(t) M_Y(t)$.*

We start by expanding the MGF of Z and substitute Z for X + Y.

$$M_Z(t) = \mathbb{E}\, e^{tZ} = \mathbb{E}\, e^{tX+tY} = \mathbb{E}\left[e^{tX} e^{tY} \right]$$

Let g(X) = $e^t X$. g is a function on a random variable and thus is itself a random variable. g(X) and g(Y) are independent due to the random variables X and Y being independent. Hence, i can use the fact that for two independent random variables, say V and W, $\mathbb{E}[VW] = \mathbb{E}[V]\,\mathbb{E}[W]$ to prove Lemma 2:

$$\mathbb{E}\left[e^{tX} e^{tY} \right] = \mathbb{E}[e^{tX}]\,\mathbb{E}[e^{tY}] = M_X(t) M_Y(t)$$

2 Statistical inference

Suppose we observe $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, I)$.

a) Show that $\hat{\mu} := X$ is (i) the MLE and (ii) UMVU for μ.

(i) The likelihood function of the random vector X is defined as

$$L(\mu; X) = (2\pi)^{-n/2} det(I)^{-1/2} \exp\left(-\frac{1}{2}(X-\mu)' I^{-1} (X-\mu)\right)$$

Where n denotes the n random variables in X. The determinant of the identity matrix I is just 1, therefore, the equation can be simplified to

$$L(\mu; X) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(X-\mu)' I^{-1} (X-\mu)\right)$$

To find the MLE of μ we have to calculate the log-likelihood ℓ_μ by taking the log of the likelihood function:

$$\ell_\mu = logL(\mu; X) = -\frac{n}{2} log(2\pi) - \frac{1}{2}(X-\mu)' I^{-1} (X-\mu)$$

Continuing, we need to take the derivative of ℓ_μ with respect to μ and equate to zero. We use the information that $\dfrac{\partial v' A v}{\partial v} = 2Av$ if A is not a function of v and A is symmetric. Matrix I is not dependent of $v = (X-\mu)$ and I is symmetric, hence I can use this information to solve the derivative:

$$\frac{\partial \ell_\mu}{\partial \mu} = -I^{-1}(X-\mu) = 0$$

This equation is solved when $X - \mu = 0 \to \mu = X$. As a result, $\hat{\mu} = X$ is the MLE.

(ii) To show that $\hat{\mu}$ is UMVU, we first find a complete and sufficient statistic T=T(X). To do so, we make use of the factorization theorem, which states that T(X) is sufficient iff there exist functions $g_\theta \geq 0$ and $h \geq 0$ such that $p_\theta = g_\theta(T(x))h(x)$[1], to reformulate the pdf of X into this form. The pdf of X is:

$$f(X) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(X-\mu)' I^{-1} (X-\mu)\right)$$

Let T(X) = X, h(X) = $(2\pi)^{-n/2}$ and $g_\mu(t) = \exp\left(-\frac{1}{2}(t-\mu)' I^{-1} (t-\mu)\right)$. Then the pdf can be rewritten as

$$f(X) = g_\mu(T(X))h(X)$$

By there factorization theorem, the statistic T(X)=X is a sufficient statistic. The multivariate normal distribution is part of the exponential family in canonical form, thus T(X) = X is also complete.

With the acquired complete and sufficient statistic T(X), I can use the Lehmann - Scheffe theorem to find the UMVU estimator η:

$$\eta(T) = \mathbb{E}\left[\hat{\mu} | T(X)\right] = \mathbb{E}[X|X] = X = \hat{\mu}$$

From this it is evident that $\hat{\mu}$ is UMVU for μ.

b) Compute the risk of $\hat{\mu}$ under quadratic loss

By section "3.3.1 Risk of an estimator" in the lecture notes we know that risk is defined as

$$R(\theta, T) := \mathbb{E}_{P_\theta} L(\theta, T(X))$$

In our case, T is the estimator $\hat{\mu}$ and θ is μ. The loss function to be used in the risk computation is the quadratic loss function, defined as

$$L(\theta, t) = ||g(\theta) - t||^2$$

Additionally, we know from the notes that the this risk function satisfies a decomposition into bias and variance terms. As proven in a), the estimator $\hat{\mu}$ is unbiased, hence the risk function can be broken down to

$$R(\mu, \hat{\mu}) = \mathbb{E}_{P_\mu} ||\mu - \hat{\mu}||^2 = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Continuing, we know that the variance of the random vector X is the identity matrix. This implies that the variance of each random variable X_i in random vector X has variance of 1. Knowing this, we can finally compute the risk of $\hat{\mu}$ under quadratic loss:

$$R(\mu, \hat{\mu}) = \mathbb{E}_{P_\mu} ||\mu - \hat{\mu}||^2 = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n$$

c) Consider the following estimator $\hat{\delta} = \left(1 - \frac{n-2}{||X||^2}\right) X$. Show that if $n \geq 3$, then $R(\mu, \hat{\delta}) < R(\mu, \hat{\mu})$ for all μ.

We start of by rewriting the estimator as

$$\hat{\delta}(X) = X - g(X), \qquad \text{Where} \qquad g(X) = \frac{(n-2)}{||X||^2} X$$

From this we can use Stein's unbiased risk estimate $\hat{R}$ from the theorem given to show that $R(\mu, \hat{\delta}) < R(\mu, \hat{\mu})$. To begin, we calculate $||g(X)||^2$:

$$||g(X)||^2 = ||\frac{(n-2)}{||X||^2} X||^2 = \frac{(n-2)^2}{||X||^2}$$

Following, we calculate $\nabla g(X)$:

$$\nabla g(X) = \nabla \left(\frac{(n-2)}{||X||^2} X \right) = \left(\nabla \frac{(n-2)}{||X||^2} \right) X + \frac{(n-2)}{||X||^2} \nabla X$$

$$= \left(\frac{-2(n-2)X}{||X||^4} \right) X + \frac{(n-2)}{||X||^2} I = \frac{-2(n-2)}{||X||^4} X X^T + \frac{(n-2)}{||X||^2} I$$

Based on these calculations we can find

$$tr(\nabla g(X)) = \frac{-2(n-2)}{||X||^4} tr(X X^T) + \frac{(n-2)}{||X||^2} tr(I) = -2 \frac{(n-2)}{||X||^4} ||X||^2 + \frac{(n-2)}{||X||^2} n$$

$$tr(\nabla g(X)) = \frac{(n-2)^2}{||X||^2}$$

Continuing, we can use what we have found to calculate Stein's unbiased risk estimate:

$$\hat{R} = n + ||g(X)||^2 - 2tr(\nabla g(X))$$

$$= n + \frac{(n-2)^2}{||X||^2} - 2 \frac{(n-2)^2}{||X||^2} = n - \frac{(n-2)^2}{||X||^2}$$

Lastly, since $\mathbb{E}_\mu(\hat{R}) = n - \frac{(n-2)^2}{||X||^2} = R(\mu, \hat{\delta})$, we can conclude that $R(\mu, \hat{\delta}) < R(\mu, \hat{\mu})$ for $n \geq 3$ because:

$$n - \frac{(n-2)^2}{||X||^2} < n \qquad \text{for } n \geq 3$$

d) Explain why (a) and (c) are not in contradiction.

For (a) we found that $\hat{\mu}(X) = X$ is the MLE and UMVU for μ, whereas in (c) we found another estimator for μ which has a lower risk than the UMVU. This is not a contradiction due to the James-Stein estimator being biased. Thus, although the James-Stein estimator has a lower risk, $\hat{\mu}$ still has the lowest variance among all unbiased estimators.

3 Regression

a) Prove Theorem 4.6.

The Wald test is defined as:

$$W = \frac{1}{\sigma^2}(R\hat{\beta} - r)'[R(XX')^{-1}R]^{-1}(R\hat{\beta} - r)$$

We know that $W \sim \chi^2(d)$ where d is the number of restrictions. Secondly, we know that

$$\frac{(n-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-K)$$

This can be rearranged to

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{(n-K)}\chi^2(n-K)$$

Using this and the information given before the exercise, we can see that:

$$F = \frac{(R\hat{\beta} - r)'[R(XX')^{-1}R]^{-1}(R\hat{\beta} - r)/d}{\hat{\sigma}^2} = \frac{(\sigma^2 W)/d}{\hat{\sigma}^2} \sim \frac{\sigma^2\chi^2(d)/d}{\frac{\sigma^2}{n-K}\chi^2(n-K)} = \frac{\chi^2(d)/d}{\frac{\chi^2(n-K)}{n-K}} \sim \mathcal{F}_{d,n-K}$$

Thus, $F \sim \mathcal{F}_{d,n-K}$ if $R\beta = r$. Consequently, the test $\phi(F) = 1\{F > k_\alpha\}$, where k_α is the $1-\alpha$ quantile of $\mathcal{F}_{d,n-K}$, has the level α. So if $R\beta = r$, i.e. the H_0 is true, then

$$P(\phi(F)) = P(F > k_\alpha) = \alpha$$

b) Simulate many datasets from a normal regression model such that the restriction Rβ = r holds (you should try a variety of different R, β, r). For each of your datasets perform the F test with α = 0.05. Record what proportion of the tests rejected. Comment on the results with reference to Theorem 4.6.

We chose to generate 4 different datasets through the formula

$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0,1)$$

To generate the data we created 4 sets of true beta values and used "np.random.random" to generate X matrices with 1000 rows and as many columns as there were true betas. This was used to generate the true y values. With the xs and ys we made beta estimates through the formula:

$$\hat{\beta} = (X'X)^{-1}Xy$$

For each dataset we defined different R matrices. The betas and R matrices for each test is as follows:

$$\text{Test 1: } \beta = [1,5,3,8]', \qquad R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad r = [5,8,16]'$$

$$\text{Test 2: } \beta = [2,4,1,9]', \qquad R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad r = [4,1,13]'$$

$$\text{Test 3: } \beta = [1,2,1,2]', \qquad R = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad r = [3,1]'$$

$$\text{Test 4: } \beta = [1,5,2,8,4]', \qquad R = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad r = [3,1]'$$

Continuing, use that Rank(R) = d and that n is the number of observations and K is the amount of predictors to calculate the F statistic. We compare this value up against the critical value we get from the $\mathcal{F}_{d,n-K}$ distribution at the $1-\alpha$ quantile, i.e. 0.95. If $F > k_\alpha$ we reject the H_0. By running each test 1000 times, we found test 1 rejected 5.8% of the time, test 2 3.7% of the time, test 3 5.7% and test 4 rejected 5.5% of tests.

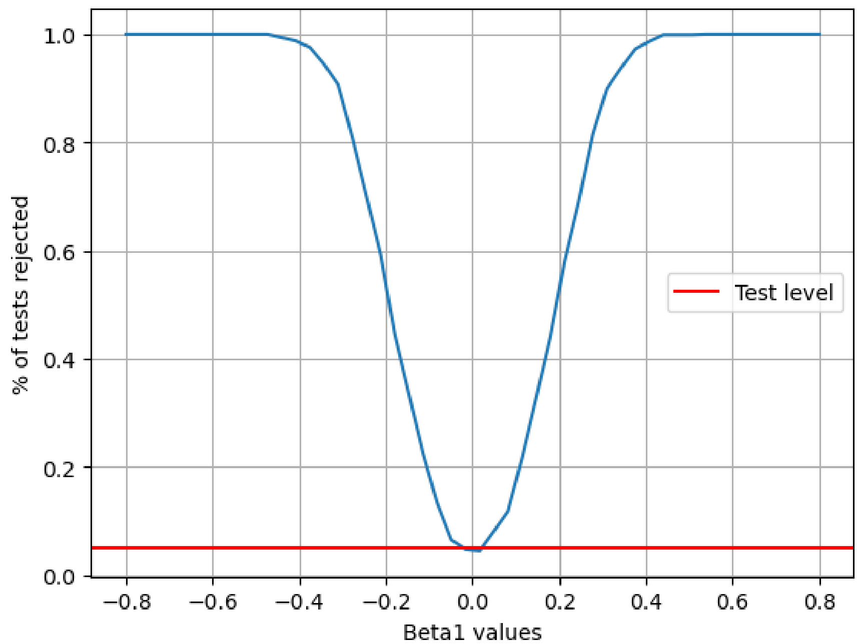
Our results are in line with theorem 4.6 in the notes. The theorem states that if $R\beta = r$, then $P(F > k_\alpha) = \alpha$, where k_α is the $1-\alpha$ quantile of $\mathcal{F}_{d,n-K}$. Our results are in line with this theorem because each test rejects the H_0 when its true about 5% of the time.

c) Now consider the special case where $R\beta = r$ is the restriction that $\beta_1 = 0$. What are R, r in this case? Simulate (many) datasets for a range of β_1 values and perform the F test that $\beta_1 = 0$. Record the proportion of the tests that reject and plot these proportions against β_1.

In this case R and r will be:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad r = [0]'$$

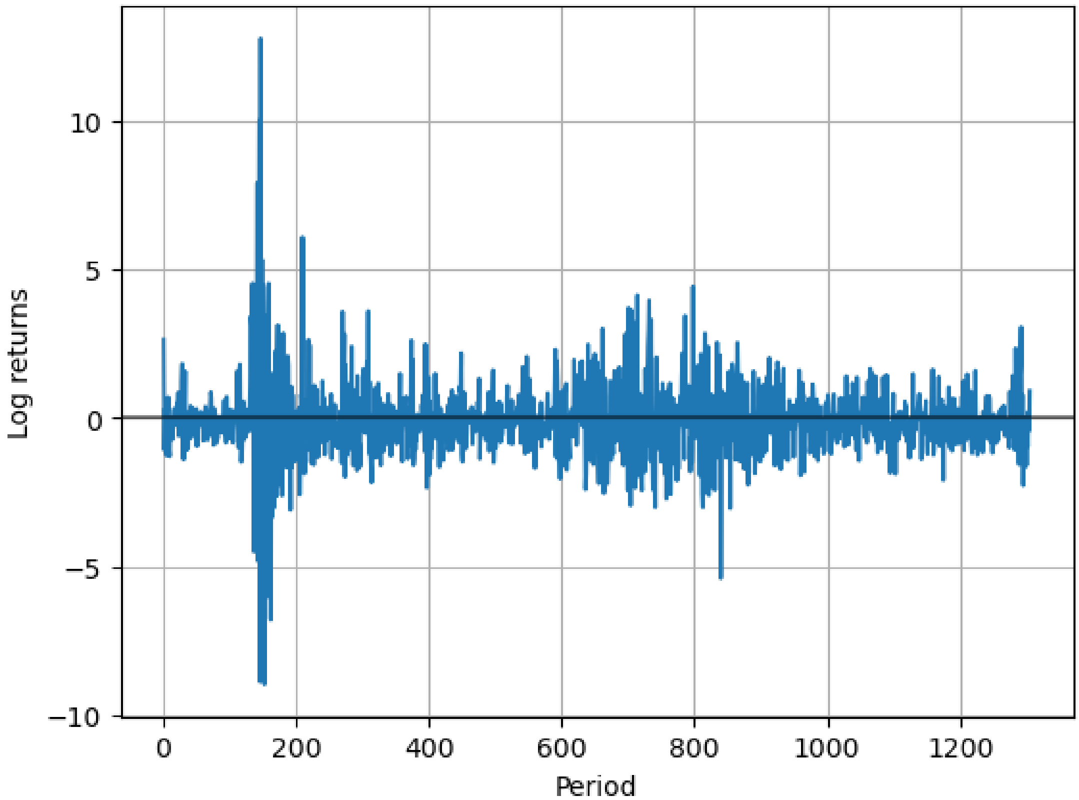
Where we X is a 1000X4 matrix. Plotting the percentage of tests rejected against different values of β_1 we get the plot:



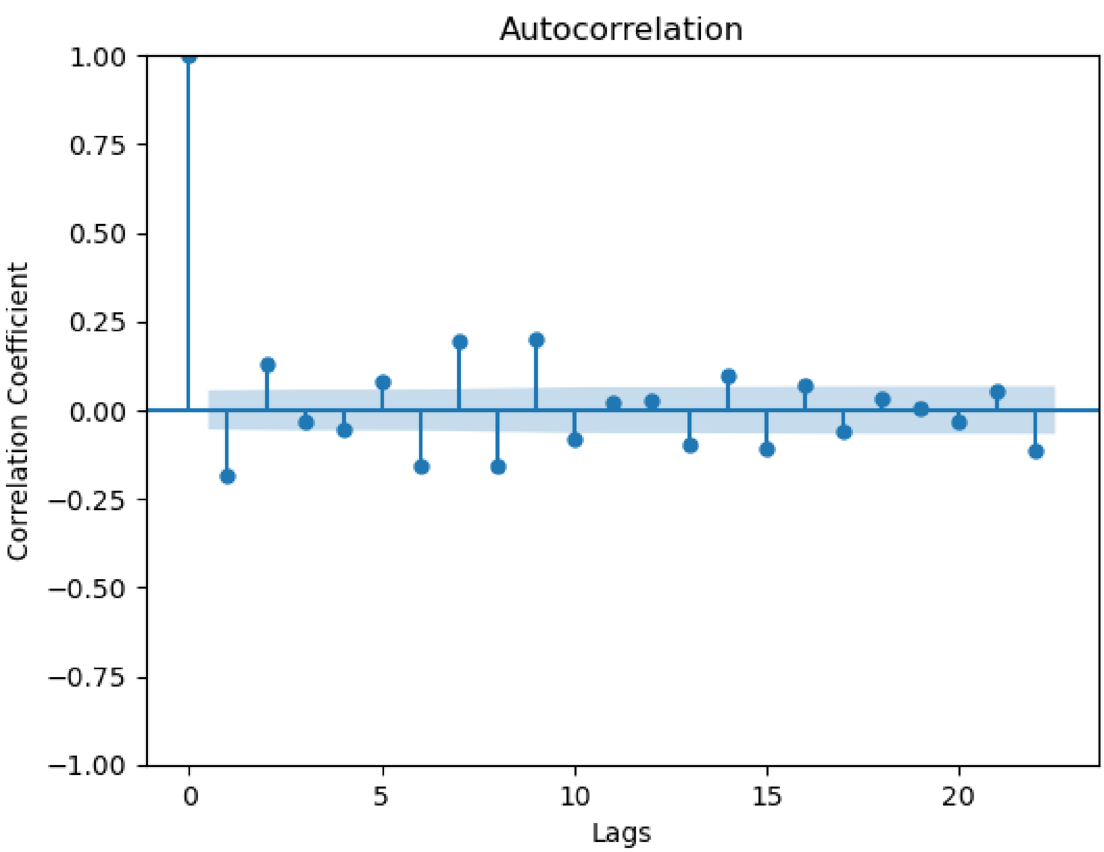
4 Time series

a) Convert the time series into returns and plot them. Plot also the ACF. Is there any evidence of serial correlation?

We converted the time series data in python by first removing the rows with missing values and used the definition of r_t defined in the exercise. This led to the following plot.



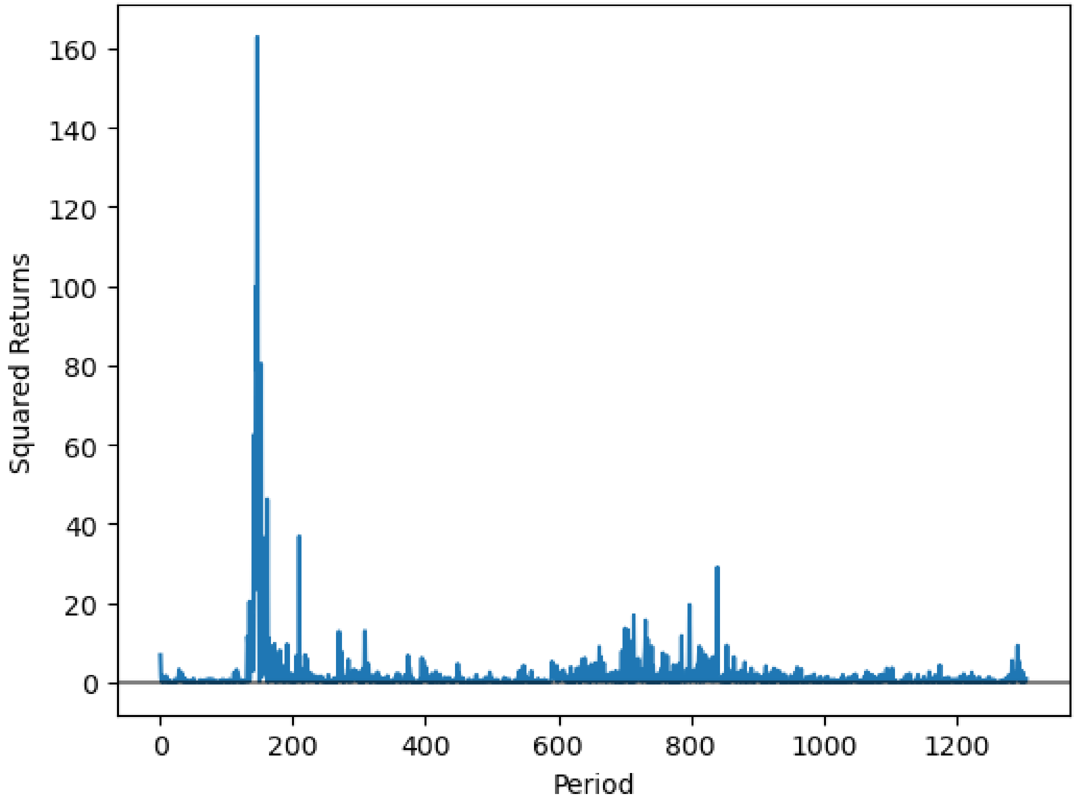
To plot the ACF we used the "plot_acf" function from statsmodels. This results in the ACF plot below.

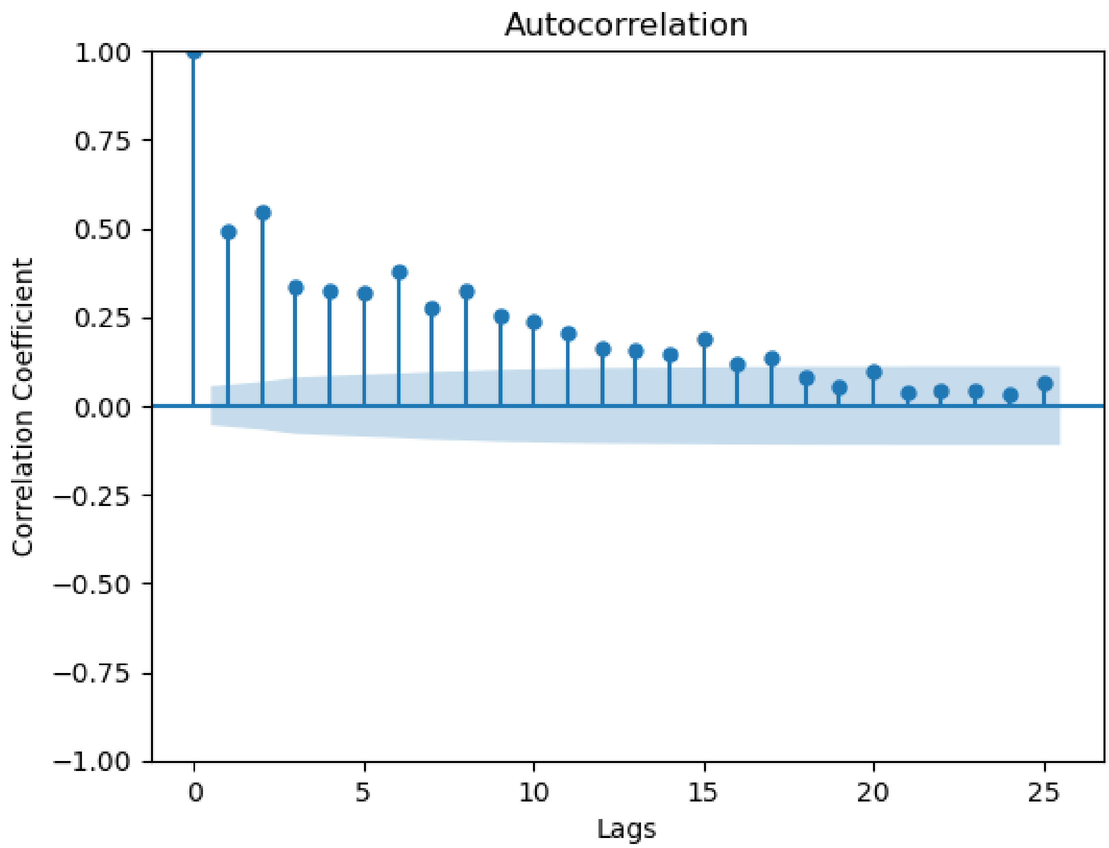


As we can see from the plot, there is evidence for serial correlation with lag 1 and 2. They both fall outside the 95 percent confidence interval and hence is statistically non-zero. The spikes in later lags might be due to some seasonality in the data. However, this can not be concluded without further analysis.

b) Do the same for the squared returns r_t^2. Is there any evidence of serial correlation in this series?

By following the same procedure, just with squared returns, we arrive at the plots:





Displays a much stronger serial correlation, suggesting lags all the way through 15 has an effect.

c) Show that if r_t follows a GARCH(1,1) with $\alpha + \beta < 1$ then r_t is a white noise sequence.

First, we prove that the sequence has a expected value of zero for all t:

$$\mathbb{E}(r_t) = \mathbb{E}(\sigma_t \epsilon_t) = \mathbb{E}(\sigma_t) * 0 = 0$$

This holds true for all t because $\epsilon \sim iid(0, 1)$. Continuing, we need to prove that it is a stationary process. To do so we need to prove that $\mathbb{E}(r_t^2) < \infty$:

$$\mathbb{E}(r_t^2) = \mathbb{E}(\sigma_t^2 \epsilon_t^2) = \mathbb{E}(\sigma_t^2)$$

To simplify, we define $A_{t-1} = \alpha\epsilon_{t-1}^2 + \beta$. Thus, we can rewrite the equation as following:

$$\sigma_t^2 = w + \alpha\sigma_{t-1}^2\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

$$\rightarrow \sigma_{t-1}^2(\alpha\epsilon_{t-1}^2 + \beta) + w$$

$$\rightarrow \sigma_t^2 = A_{t-1}\sigma_{t-1}^2 + w$$

It is possible to expand the equation by repeated use of σ^2 terms:

$$\sigma_t^2 = A_{t-1}\sigma_{t-1}^2 + w$$

$$= A_{t-1}A_{t-2}\sigma_{t-2}^2 + A_{t-1}w + w$$

$$= A_{t-1}A_{t-2}A_{t-3}\sigma_{t-3}^2 + A_{t-1A_{t-2}}w + A_{t-1}w + w$$

Continuing this expansion to infinity, we can write the equation as:

$$\sigma_t^2 = w(1 + \sum_{j=1}^{\infty}\prod_{i=1}^{j} A_{t-i})$$

From here, we can find the expectation of the variance at time t:

$$\mathbb{E}(\sigma_t^2) = w(1 + \sum_{j=1}^{\infty} \mathbb{E}\left(\prod_{i=1}^{j}(\alpha\epsilon_{t-1}^2 + \beta)\right))$$

$$= w(1 + \sum_{j=1}^{\infty}(\alpha + \beta)^j) = w(\sum_{j=0}^{\infty}(\alpha + \beta)^j)$$

This result converges, but only when $|\alpha + \beta| < 1$:

$$\mathbb{E}(r_t^2) = \mathbb{E}(\sigma_t^2) = \frac{w}{1 - \alpha - \beta}$$

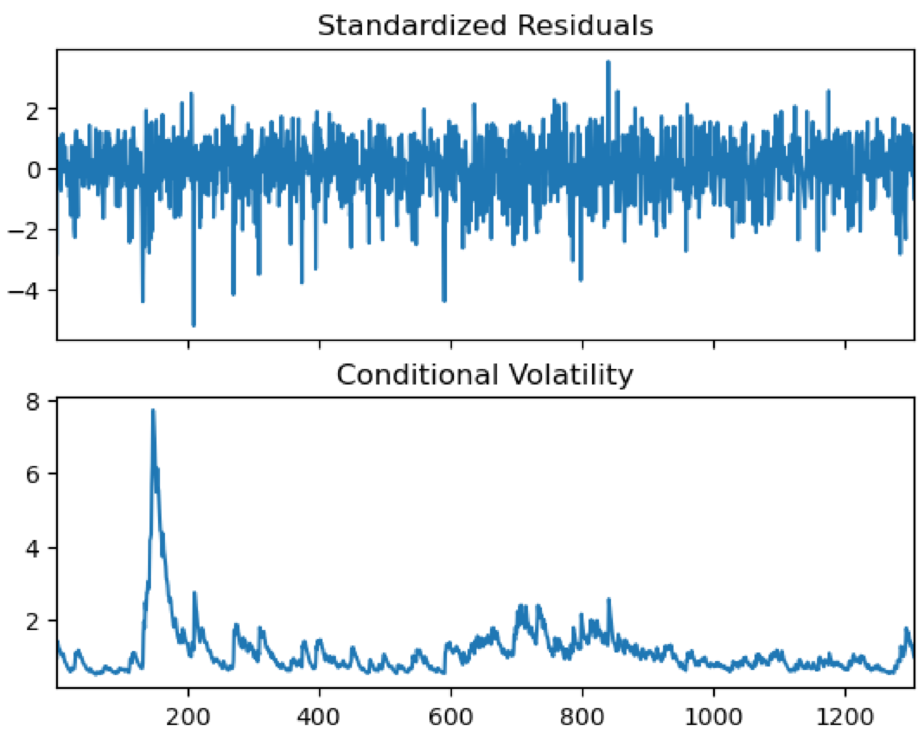
Thus, $\mathbb{E}(r_t^2) < \infty$. Lastly, we look at the covariance between the lags. The covariance between r_t and r_{t-h}, for some h, is:

$$Cov(r_{t+h}, r_t) = \mathbb{E}(\sigma_{t+h}\epsilon_{t+h}\sigma_t\epsilon_t) - \mathbb{E}(\sigma_{t+h}\epsilon_{t+h})\,\mathbb{E}(\sigma_t\epsilon_t)$$

$$Cov(r_{t+h}, r_t) = \begin{cases} \frac{w}{1-\alpha-\beta} & \text{if } h = 0 \\ 0 & \text{Otherwise} \end{cases}$$

As a result, we conclude that if a time series follows a GARCH(1,1) with $\alpha + \beta < 1$ then the times series is a white noise sequence.

d) Using arch, fit a GARCH(1, 1) model to r_t. Use the plot method of the resulting ARCHModelResult object to visualise the estimated volatility process.



References

[1] Adam Lee. Advanced statistics and alternative data types, 2024.