

MTL 390 (Statistical Methods)
Major Examination Assignment 2 Report

Name: Arpit Saxena

Entry Number: 2018MT10742

1. Testing of Hypothesis

Let (X_1, X_2, \dots, X_m) be a random sample from a Binomial Distribution $B(n, p)$ with p unknown. Consider the hypothesis test of the null hypothesis $H_0 : p = p_0$ where $p_0 \in [0, 1]$ is fixed, against the alternative composite hypothesis $H_1 : p > p_0$. Find the uniformly most powerful test of size α .

Following are the number of heads obtained on tossing a coin 5 times and repeating it for 5 times. At 95% level of confidence, test the null hypothesis of the coin to be a fair coin against the alternative hypothesis of the coin to be heads-biased (i.e. it lands on heads with more probability than it does on tails).

Number of heads: 3 4 3 5 3

Answer

The joint pdf of (X_1, X_2, \dots, X_m) is

$$\begin{aligned} L(p; x_1, x_2, \dots, x_m) &= \prod_{i=1}^m \left\{ \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \right\} \\ &= \prod_{i=1}^m \binom{n}{x_i} p^{\sum x_i} (1-p)^{mn - \sum x_i} \end{aligned}$$

Let p' be a number greater than p_0 , then we take ratio of joint probability with parameter p_0 to that of p' and let C be the set of points such that the ratio is less than a constant k .

$$\left(\frac{p}{p'} \right)^{\sum x_i} \left(\frac{1-p}{1-p'} \right)^{mn - \sum x_i} \leq k$$

Now taking log on both sides, we get

$$\begin{aligned} \log \left\{ \frac{p}{p'} \right\} \cdot \sum_{i=1}^m x_i + \log \left\{ \frac{1-p}{1-p'} \right\} \cdot \left(mn - \sum_{i=1}^m x_i \right) &\leq \log k \\ \Rightarrow \sum_{i=1}^m x_i \cdot \left[\log \left\{ \frac{1-p}{1-p'} \right\} - \log \left\{ \frac{p}{p'} \right\} \right] &\geq mn \log \left\{ \frac{1-p}{1-p'} \right\} - \log k \\ \Rightarrow \sum_{i=1}^m x_i &\geq \frac{mn \log \left\{ \frac{1-p}{1-p'} \right\} - \log k}{\log \left\{ \frac{1-p}{1-p'} \right\} - \log \left\{ \frac{p}{p'} \right\}} = c \end{aligned}$$

Now we invoke Neyman-Pearson lemma to show that the region

$$C = \left\{ (x_1, x_2, \dots, x_m) : \sum_{i=1}^m x_i \geq c \right\}$$

is the best rejection region for testing the simple hypothesis $H_0 : p = p_0$ against the simple alternative hypothesis $H_1 : p = p'$.

Now we need to find c so that the rejection region C is of the desired size α . Under the null hypothesis, the random variable $\sum_{i=1}^m X_i$ follows the Binomial distribution $B(mn, p_0)$ because sum of binomial variables with the same probability p is a binomial variable with the first terms added.

Now we have to use the equation

$$\alpha \leq P \left(\sum_{i=1}^m X_i \geq c \right)$$

to find c . Note that for a binomial distribution like we have here, c has no closed form expression. Once we find c using numerical methods, the region

$$C = \left\{ (x_1, x_2, \dots, x_m) : \sum_{i=1}^m x_i \geq c \right\}$$

is the best critical region of size α for testing $H_0 : p = p_0$ against the alternative hypothesis $H_1 : p = p'$.

Observe that this region will be valid for all $p' > p_0$. Therefore, the region C is a uniformly most powerful critical region of size α for testing $H_0 : p = p_0$ against $H_1 : p > p_0$.

For the coin toss example, we have $p_0 = 0.5$ and we solve

$$\begin{aligned} \alpha &\leq P \left(\sum_{i=1}^m X_i \geq c \right) \\ \implies 0.05 &\leq (0.5)^{25} \left[\binom{25}{c} + \binom{25}{c+1} + \dots + \binom{25}{25} \right] \end{aligned}$$

Since we have a small number at hand, we can just use brute-force to find the value of c .

We find that $c \leq 17$ satisfies the inequality.

From the given data, $\sum x_i = 3 + 4 + 3 + 5 + 3 = 18 > 17$, and thus lies outside the critical region. Thus the null hypothesis of the coin to be a fair coin can be rejected at 95% confidence level.

2. Testing of Hypothesis

Number of claims filed at 3 insurance companies (in millions) over 5 years are shown in the following table. Using ANOVA method, test the appropriate hypothesis at 5% level of significance to decide if the mean number of claims filed in companies 1, 2 and 3 differ.

Company 1	Company 2	Company 3
15	13	23
17	21	16
16	15	14
14	16	14
13	18	18

Answer

We consider the following null and alternate hypotheses.

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

H_1 : At least one pair of sample means is significantly different

Computing group means,

$$\begin{aligned}\bar{x}_1 &= \frac{15 + 17 + 16 + 14 + 13}{5} = 15 \\ \bar{x}_2 &= \frac{13 + 21 + 15 + 16 + 18}{5} = 16.6 \\ \bar{x}_3 &= \frac{23 + 16 + 14 + 14 + 18}{5} = 17\end{aligned}$$

Computing grand mean

$$\bar{x}_G = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3}{3} = 16.2$$

Now we compute Sum of squares_{between}:

$$\begin{aligned}SS_b &= n \sum_{i=1}^k (\bar{x}_i - \bar{x}_G)^2 \\ &= 5 \times [(15 - 16.2)^2 + (16.6 - 16.2)^2 + (17 - 16.2)^2] \\ &= 11.2\end{aligned}$$

Then we need to compute Sum of squares_{within}

$$\begin{aligned}SS_w &= \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \\ &= [(15 - 15)^2 + (17 - 15)^2 + (16 - 15)^2 + (14 - 15)^2 + (13 - 15)^2 \\ &\quad + (13 - 16.6)^2 + (21 - 16.6)^2 + (15 - 16.6)^2 + (16 - 16.6)^2 + (18 - 16.6)^2 \\ &\quad + (23 - 17)^2 + (16 - 17)^2 + (14 - 17)^2 + (14 - 17)^2 + (18 - 17)^2] \\ &= 103.2\end{aligned}$$

Finally we have the Sum of squares_{total}

$$\begin{aligned}
 SS_{\text{total}} &= \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_G)^2 \\
 &= [(15 - 16.2)^2 + (17 - 16.2)^2 + (16 - 16.2)^2 + (14 - 16.2)^2 + (13 - 16.2)^2 \\
 &\quad + (13 - 16.2)^2 + (21 - 16.2)^2 + (15 - 16.2)^2 + (16 - 16.2)^2 + (18 - 16.2)^2 \\
 &\quad + (23 - 16.2)^2 + (16 - 16.2)^2 + (14 - 16.2)^2 + (14 - 16.2)^2 + (18 - 16.2)^2] \\
 &= 114.4
 \end{aligned}$$

Degrees of freedom for SS_b are $df_b = k - 1 = 2$.

Degrees of freedom for SS_w are $df_w = k(n - 1) = 3 \times 4 = 12$

We now compute the statistics

$$\begin{aligned}
 \text{mean square}_{\text{between}} : MS_{\text{between}} &= \frac{SS_b}{df_b} = \frac{11.2}{2} = 5.6 \\
 \text{mean square}_{\text{within}} : MS_{\text{within}} &= \frac{SS_w}{df_w} = \frac{103.2}{12} = 8.6 \\
 \text{F-statistic } F_0 &= \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{5.6}{8.6} = 0.651
 \end{aligned}$$

From the statistics tables, we find that

$$F_{k-1, N-k, \alpha} = F_{2, 12, 0.05} = 19.40$$

We note that the calculated F-statistic is less than the critical value, so the null hypothesis is not rejected.

Therefore, we can't reject the hypothesis that the number of claims submitted to the 3 insurance companies are same at 95% level of significance.

3. Analysis of correlation and regression

Give the general formula for Spearman's rank correlation coefficient. Using it, derive the simpler formula when distinct ranks are assumed. Emphasise where the distinctness is assumed.

Consider the following data. Find the Spearman's rank correlation coefficient using the general formula and the other formula with the distinct rank assumption. What is the difference in the results?

X	Y
106	7
100	27
86	2
101	50
99	29
103	29
99	20
113	12
112	6
110	17

Answer

Let us consider a sample of size n . The n raw scores X_i, Y_i are converted to ranks rg_{X_i}, rg_{Y_i} , and r_s is computed as

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}} \quad (1)$$

where,

- ρ denotes the Pearson correlation coefficient, but applied to rank variables,
- $cov(rg_X, rg_Y)$ is the covariance of the rank variables
- ρ_{rg_X} and ρ_{rg_Y} are the standard deviations of the rank variables

Let us directly work with ranks to derive the formula. We denote X_i and Y_i as the ranks. Then for each of these

$$\begin{aligned} \sum_{i=1}^n X_i &= 1 + 2 + \cdots + n \\ &= \frac{n(n+1)}{2} \end{aligned}$$

We note that the assumption of distinct ranks doesn't come into play here. This is because in case of ties an average of the tied ranks is given to both the variables and the sum becomes the same in that case as well

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ &= \frac{n+1}{2} \end{aligned}$$

Similarly, $\bar{Y} = \frac{n+1}{2}$

Now we calculate the variance.

$$\begin{aligned}
\sigma_X^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\
&= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2 \right] \\
&= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right]
\end{aligned}$$

We use the assumption of distinct ranks here to get

$$\sum_{i=1}^n X_i^2 = 1^2 + 2^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

Plugging in this value and the mean calculated earlier, we get

$$\begin{aligned}
\sigma_X^2 &= \frac{1}{n} \left[\frac{n(n+1)(2n+1)}{6} - n \left\{ \frac{n+1}{2} \right\}^2 \right] \\
&= \frac{1}{n} \left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \right] \\
&= \frac{1}{n} \frac{n(n+1)}{12} [2(2n+1) - 3(n+1)] \\
&= \frac{n+1}{12} (n-1) \\
&= \frac{n^2-1}{12}
\end{aligned}$$

Similarly, we get $\sigma_Y^2 = \frac{n^2-1}{12}$

Having calculated the variance of X_i, Y_i , we calculate their covariance.

$$\begin{aligned}
Cov(X, Y) &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \frac{1}{n} \sum_{i=1}^n [X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X} \bar{Y}] \\
&= \frac{1}{n} \left\{ \sum_{i=1}^n X_i Y_i - \bar{Y} \sum_{i=1}^n X_i - \bar{X} \sum_{i=1}^n Y_i + n \bar{X} \bar{Y} \right\} \\
&= \frac{1}{n} \left\{ \sum_{i=1}^n X_i Y_i - n \bar{Y} \bar{X} - n \bar{X} \bar{Y} + n \bar{X} \bar{Y} \right\} \\
&= \frac{1}{n} \left\{ \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right\} \\
&= \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X} \bar{Y} \\
&= -\frac{1}{2n} \sum_{i=1}^n \{X_i^2 + Y_i^2 - 2X_i Y_i - (X_i^2 + Y_i^2)\} - \bar{X} \bar{Y} \\
&= \frac{1}{2n} \sum_{i=1}^n \{X_i^2 + Y_i^2\} - \frac{1}{2n} \sum_{i=1}^n \{X_i^2 + Y_i^2 - 2X_i Y_i\} - \bar{X} \bar{Y}
\end{aligned}$$

We are once again invoking the assumption of distinct ranks to calculate the sum of X_i^2 to get

$$\begin{aligned}
Cov(X, Y) &= \frac{1}{2n} \times 2 \cdot \frac{n(n+1)(2n+1)}{6} - \frac{1}{2n} \sum_{i=1}^n \{X_i - Y_i\}^2 - \bar{X} \bar{Y} \\
&= \frac{(n+1)(2n+1)}{6} - \frac{1}{2n} \sum_{i=1}^n d_i^2 - \left\{ \frac{n+1}{2} \right\}^2 \\
&= \frac{n+1}{12} \{2(2n+1) - 3(n+1)\} - \frac{1}{2n} \sum_{i=1}^n d_i^2 \\
&= \frac{n+1}{12} \cdot (n-1) - \frac{1}{2n} \sum_{i=1}^n d_i^2 \\
&= \frac{n^2 - 1}{12} - \frac{1}{2n} \sum_{i=1}^n d_i^2
\end{aligned}$$

Where $d_i = X_i - Y_i$ is the difference of ranks.

Now we calculate the Spearman rank correlation coefficient using (1)

$$\begin{aligned}
 r_s &= \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \\
 &= \frac{\frac{n^2-1}{12} - \frac{1}{2n} \sum_{i=1}^n d_i^2}{\frac{n^2-1}{12}} \\
 \therefore r_s &= 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n d_i^2
 \end{aligned} \tag{2}$$

Thus we have obtained the formula for calculating the Spearman rank correlation coefficient in case of distinct ranks.

Now for the given data, we first convert the raw scores into ranks. These are tabulated below:

X	Y	rank X_i	rank Y_i	d_i	d_i^2
106	7	7	3	4	16
100	27	4	7	-3	9
86	2	1	1	0	0
101	50	5	10	-5	25
99	29	2.5	8.5	-6	36
103	29	6	8.5	-2.5	6.25
99	20	2.5	6	-3.5	12.25
113	12	10	4	6	36
112	6	9	2	7	49
110	17	8	5	3	9

Note that due to ties, there are some fractional ranks. Suppose there was a tie between values at ranks 2 and 3. Then we assign both of these rank 2.5 which is the mean of 2 and 3.

Now, we calculate the spearman rank correlation coefficient using the general formula (1).

$$\begin{aligned}
 Var(\text{rank } X) &= 8.2 \\
 Var(\text{rank } Y) &= 8.2 \\
 Cov(\text{rank } X, \text{rank } Y) &= -1.725 \\
 \therefore r_s &= \frac{Cov(\text{rank } X, \text{rank } Y)}{\sqrt{Var(\text{rank } X) Var(\text{rank } Y)}} \\
 &= \frac{-1.725}{\sqrt{8.2 \times 8.2}} \\
 &= -0.210
 \end{aligned}$$

Now, using the Spearman rank correlation coefficient distinct rank formula (2).

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n d_i^2$$

$$= -0.203$$

We note that there is only an absolute difference of 0.007 in these calculated values which is about 3.5% relative difference. We note that it's a small difference and the distinct rank formula may be useful even in case of ties for approximation purposes.

4. Analysis of correlation and regression

Consider the following data from the agricultural sector of Taiwan, 1958-1972. We are interested in seeing the effect of labor days (X_2) and the real capital input (X_3) on the real gross product (Y).

- Write the regression equation when regressing X_2 and X_3 on Y
- Derive the estimators for the coefficients using Ordinary Least Squares
- Find the estimates of Y using the estimators obtained. Also find the standard error.
- Find the R^2 and adjusted- R^2 values.

Year	Real gross product (millions of NT \$), Y	Labor days (millions of days), X_2	Real capital input (millions of NT \$), X_3
1958	16607.7	275.5	17803.7
1959	17511.3	274.4	18096.8
1960	20171.2	269.7	18271.8
1961	20932.9	267	19167.3
1962	20406	267.8	19647.6
1963	20831.6	275	20803.5
1964	24806.3	283	22076.6
1965	26465.8	300.7	23445.2
1966	27403	307.5	24939
1967	28628.7	303.7	26713.7
1968	29904.5	304.7	29957.8
1969	27508.2	298.6	31585.9
1970	29035.5	295.5	33474.5
1971	29281.5	299	34821.8
1972	31535.8	288.1	41794.3

Answer

The population regression equation is

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where ϵ is an error term.

The sample regression equation using the estimators is

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \epsilon_i$$

For a sample with n observations, we want to minimise

$$S = \sum_{i=1}^n \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\}^2$$

To do this, we use the first derivative tests and set the partial derivatives of S with respect to the estimators as 0. We get the equations

$$\begin{aligned} \frac{\partial S}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^n \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\} \\ \frac{\partial S}{\partial \hat{\beta}_2} &= - \sum_{i=1}^n X_{2i} \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\} \\ \frac{\partial S}{\partial \hat{\beta}_3} &= - \sum_{i=1}^n X_{3i} \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\} \end{aligned}$$

Setting these equations equal to zero, we obtain the normal equations

$$\begin{aligned} & \sum_{i=1}^n \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\} = 0 \\ \Rightarrow & \sum_{i=1}^n Y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^n X_{2i} - \hat{\beta}_3 \sum_{i=1}^n X_{3i} = 0 \\ & \sum_{i=1}^n X_{2i} \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\} = 0 \\ \Rightarrow & \sum_{i=1}^n X_{2i} Y_i - \hat{\beta}_1 \sum_{i=1}^n X_{2i} - \hat{\beta}_2 \sum_{i=1}^n X_{2i}^2 - \hat{\beta}_3 \sum_{i=1}^n X_{2i} X_{3i} = 0 \\ & \sum_{i=1}^n X_{3i} \left\{ Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \right\} = 0 \\ \Rightarrow & \sum_{i=1}^n X_{3i} Y_i - \hat{\beta}_1 \sum_{i=1}^n X_{3i} - \hat{\beta}_2 \sum_{i=1}^n X_{2i} X_{3i} - \hat{\beta}_3 \sum_{i=1}^n X_{3i}^2 = 0 \end{aligned}$$

Putting the values from the data, we obtain the equations as:

$$\begin{aligned} 371030 - 15\hat{\beta}_1 - 4310.2\hat{\beta}_2 - 382599.5\hat{\beta}_3 &= 0 \\ 107449076 - 4310.2\hat{\beta}_1 - 1241590.88\hat{\beta}_2 - 110882249\hat{\beta}_3 &= 0 \\ 9907328434 - 382599.5\hat{\beta}_1 - 110882249\hat{\beta}_2 - 10512155203.39\hat{\beta}_3 &= 0 \end{aligned}$$

Solving this system of equations, we get the values as

$$\hat{\beta}_1 = -28067.2, \hat{\beta}_2 = 147.94, \hat{\beta}_3 = 0.4036$$

Using these, we calculate the estimated Y values, which are written in the table below:

Year	Y	X_2	X_3	Estimated Y
1958	16607.7	275.5	17803.7	19875.84332
1959	17511.3	274.4	18096.8	19831.40448
1960	20171.2	269.7	18271.8	19206.71648
1961	20932.9	267	19167.3	19168.70228
1962	20406	267.8	19647.6	19480.90336
1963	20831.6	275	20803.5	21012.5926
1964	24806.3	283	22076.6	22709.93576
1965	26465.8	300.7	23445.2	25880.84072
1966	27403	307.5	24939	27489.7304
1967	28628.7	303.7	26713.7	27643.82732
1968	29904.5	304.7	29957.8	29101.08608
1969	27508.2	298.6	31585.9	28855.75324
1970	29035.5	295.5	33474.5	29159.3782
1971	29281.5	299	34821.8	30220.93848
1972	31535.8	288.1	41794.3	31422.49348

The standard error is given by

$$\sqrt{\frac{\sum(Y - Y_{est})^2}{n}} = 1416.13 \text{ NT \$}$$

Now we calculate the R^2 and adjusted- R^2 values.

$$\begin{aligned} R^2 &= 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}} \\ &= 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \\ &= 0.91 \\ \text{Adjusted-}R^2, \bar{R}^2 &= 1 - \frac{\frac{\text{Residual Sum of Squares}}{n-k}}{\frac{\text{Total Sum of Squares}}{n-1}} \end{aligned}$$

Here n is the number of variables in the sample ($=15$) and k is the number of parameters in the regression ($=3$)

$$\begin{aligned}\Rightarrow \bar{R}^2 &= 1 - \frac{\sum (Y_i - \hat{Y}_i)^2 / (n - k)}{\sum (Y_i - \bar{Y})^2 / (n - 1)} \\ &= 0.89\end{aligned}$$

Thus we see that 90% of the variation in Y is explained by the variations in X_2 and X_3 through this model.

5. Time Series Analysis

Show that every stationary MA(2) process without a unit root can be converted to an invertible process by suitably changing the coefficients and white noise random variables. Note that this can actually be done for all MA(q) processes. Is the following MA(2) process invertible?

$$X_t = e_t - 3.2e_{t-1} + 0.6e_{t-2}$$

Here e_t 's are white noise random variables normally distributed with mean 0 and variance 2. If not, convert it to an invertible MA(2) process.

Answer

A general MA(2) process looks like the following:

$$X_t = e_t + \lambda_1 e_{t-1} + \lambda_2 e_{t-2}$$

We can use the backshift operator B to write

$$\begin{aligned}X_t &= e_t + \lambda_1 B(e_t) + \lambda_2 B^2(e_t) \\ \Rightarrow X_t &= (1 + \lambda_1 B + \lambda_2 B^2)e_t\end{aligned}$$

We factorise the polynomial into linear factors to get

$$X_t = (1 - \mu_1 B)(1 - \mu_2 B)e_t$$

Thus we get the general characteristic polynomial of the MA(2) process as $(1 - \mu_1 B)(1 - \mu_2 B)$. The zeros of the polynomial are $\frac{1}{\mu_1}$ and $\frac{1}{\mu_2}$ respectively.

We note that the way to find if the process is invertible is that the absolute value of the zeros of the characteristic polynomial are greater than 1. If not, there would be atleast one zero which is less than 1. We note that the absence of unit roots means that no zero is exactly equal to 1.

We show that a process with the characteristic polynomial with one root μ taken as the reciprocal and variance of the white noise multiplied by μ^2 yields a process with the same autocovariance function.

Let us define $\lambda_0 = 1$. Then,

$$\begin{aligned}
Cov(X_t, X_{t+\tau}) &= E(X_t X_{t+\tau}) - E(X_t)E(X_{t+\tau}) \\
&= E(X_t X_{t+\tau}) \quad (\text{since } E(X_t) = 0 \forall t) \\
&= E \left[\left\{ \sum_{j=0}^q \lambda_j e_{t-j} \right\} \left\{ \sum_{k=0}^q \lambda_k e_{t-k+\tau} \right\} \right] \\
&= \sum_{j=0}^q \sum_{k=0}^q \lambda_j \lambda_k E(e_{t-j} e_{t-k+\tau})
\end{aligned}$$

Since e_t 's are independent and identically distributed with variance as σ^2

$$Cov(X_t, X_{t+\tau}) = \sum_{j=0}^q \sum_{k=0}^q \lambda_j \lambda_k \sigma^2 \delta_{t-j, t-k+\tau} \quad (\text{Where } \delta_{ij} \text{ is the Kronecker delta})$$

To remove the Kronecker delta, We make the following observations:

- When $\tau = 0$, the (j, k) pairs which will be included are $(0, 0), (1, 1), \dots, (q, q)$
- For $\tau = 1$, we'll have $(0, 1), (1, 2), \dots, (q-1, q)$
- For $\tau = 2$, we'll have $(0, 2), (1, 3), \dots, (q-2, q)$
- \vdots
- For $\tau = q-1$, we'll have $(0, q-1), (1, q)$
- For $\tau = q$, we'll have $(0, q)$
- For $\tau > q$, $t-j \neq t-k+\tau \forall j, k = 0, \dots, q$

So we'll have the following simplification:

$$Cov(X_t, X_{t+\tau}) = \begin{cases} \sigma^2 \sum_{j=0}^{q-\tau} \lambda_j \lambda_{j+\tau} & \text{if } \tau = 0, 1, \dots, q \\ 0 & \tau > q \end{cases} \quad (3)$$

Now for a process $X_t = (1 - \mu_1 B)(1 - \mu_2 B)e_t$, we have $\lambda_1 = -\mu_1 - \mu_2$ and $\lambda_2 = \mu_1 \mu_2$, then we get

$$\begin{aligned}
Cov(X_t, X_t) &= \sigma^2 [1 + (\mu_1 + \mu_2)^2 + \mu_1^2 \mu_2^2] \\
Cov(X_t, X_{t+1}) &= \sigma^2 [(-\mu_1 - \mu_2) + (-\mu_1 - \mu_2)(\mu_1 \mu_2)] \\
&= -\sigma^2 (\mu_1 + \mu_2)(1 + \mu_1 \mu_2) \\
Cov(X_t, X_{t+2}) &= \sigma^2 \mu_1 \mu_2 \\
Cov(X_t, X_{t+\tau}) &= 0 \forall \tau > 2
\end{aligned}$$

If we replace μ_1 by $\frac{1}{\mu_1}$ in the above equations, we'll get

$$\begin{aligned}
Cov(X_t, X_t) &= \sigma^2 \left[1 + \left(\frac{1}{\mu_1} + \mu_2 \right)^2 + \frac{1}{\mu_1^2} \mu_2^2 \right] \\
&= \frac{\sigma^2}{\mu_1^2} [\mu_1^2 + (1 + \mu_1 \mu_2)^2 + \mu_2^2] \\
&= \frac{\sigma^2}{\mu_1^2} [\mu_1^2 + 1 + 2\mu_1 \mu_2 + \mu_1^2 \mu_2^2 + \mu_2^2] \\
&= \frac{\sigma^2}{\mu_1^2} [1 + (\mu_1 + \mu_2)^2 + \mu_1^2 \mu_2^2] \\
Cov(X_t, X_{t+1}) &= -\sigma^2 \left(\frac{1}{\mu_1} + \mu_2 \right) \left(1 + \frac{1}{\mu_1} \mu_2 \right) \\
&= -\frac{\sigma^2}{\mu_1^2} (1 + \mu_1 \mu_2) (\mu_1 + \mu_2) \\
Cov(X_t, X_{t+2}) &= \sigma^2 \frac{1}{\mu_1} \mu_2 \\
&= \frac{\sigma^2}{\mu_1^2} \mu_1 \mu_2 \\
Cov(X_t, X_{t+\tau}) &= 0 \forall \tau > 2
\end{aligned}$$

We observe that we have the exact same autocovariance equations with the variance σ^2 divided by μ^2 .

Thus, we can change a root of the characteristic polynomial to its reciprocal by suitably changing the variance of the white noise variables to get the same autocovariance function. This way, we can change all roots which are less than 1 to become more than 1 and obtain an invertible process this way.

For the process $X_t = e_t - 3.2e_{t-1} + 0.6e_{t-2}$, the characteristic polynomial is $\theta(B) = 1 - 3.2B + 0.6B^2 = (1 - 3B)(1 - 0.2B)$. The zeros of this polynomial are $\frac{1}{3}, 5$. Since one of them is less than 1, this process is not invertible.

We can make this invertible by the procedure outlined above. We'll replace the root $\frac{1}{3}$ by 3 and change the variance of white noise which is currently 2 to $\frac{2}{\frac{1}{3}^2} = 18$. Thus, we obtain the process

$$\begin{aligned}
X_t &= (1 - \frac{1}{3}B)(1 - 0.2B)e_t \\
&= e_t - \frac{8}{15}e_{t-1} + \frac{1}{15}e_{t-2}
\end{aligned}$$

where e_t 's are white noise variables normally distributed with mean 0 and variance 18.

6. Time Series Analysis

Consider the stationary ARMA(p, q) process. Elaborate the general method of finding the variance and covariances. Use the method to find the variance and covariances & correlations of time difference upto 3 (i.e. ρ_1, ρ_2, ρ_3) of the stationary ARMA(1, 1) process.

Answer

Let $\{X_t\}_{t \in \mathbb{N}}$ be a time series following the stationary ARMA(p, q) process. Then we have the following recurrence relation:

$$X_t = \sum_{j=1}^p \alpha_j X_{t-j} + \sum_{j=1}^q \beta_j e_{t-j} + e_t$$

where e_t 's are independent and identically distributed random variables following a normal distribution with mean 0 and variance 1. We also take the boundary condition as $X_t = 0 \forall t < \max(p, q)$

Since this is a stationary process, it is implied that it will also satisfy wide sense stationarity, which means that

$$\begin{aligned} E(X_t) &= \mu \forall t \\ Cov(X_t, X_s) &= f(t - s) \end{aligned}$$

i.e. the mean of all the random variables of the process is constant and the covariance of the random variables at two time instances of the process depends only on the time difference between them.

We first find the mean of the random variables of the process:

$$\begin{aligned} X_t &= \sum_{j=1}^p \alpha_j X_{t-j} + \sum_{j=1}^q \beta_j e_{t-j} + e_t \\ \implies E(X_t) &= E\left(\sum_{j=1}^p \alpha_j X_{t-j} + \sum_{j=1}^q \beta_j e_{t-j} + e_t\right) \\ \implies E(X_t) &= \sum_{j=1}^p \alpha_j E(X_{t-j}) + \sum_{j=1}^q \beta_j E(e_{t-j}) + E(e_t) \end{aligned}$$

Now we use the fact that e_t 's have a mean of 0 to get

$$\begin{aligned}
E(X_t) &= \sum_{j=1}^p \alpha_j E(X_{t-j}) \\
\implies \mu &= \sum_{j=1}^p \alpha_j \mu && \text{(Since the process is stationary)} \\
\implies \left(\sum_{j=1}^p \alpha_j - 1 \right) \mu &= 0
\end{aligned}$$

Now assuming that $\sum_{j=1}^p \alpha_j \neq 0$, we get

$$\mu = 0$$

Therefore we have the following result:

$$E(X_t) = 0 \quad \forall t \tag{4}$$

We denote γ_τ as the covariance of random variables in this process at time τ apart. Then,

$$X_t = \sum_{j=1}^p \alpha_j X_{t-j} + \sum_{j=1}^q \beta_j e_{t-j} + e_t$$

Multiplying by $X_{t-\tau}$ on both sides, we get

$$\begin{aligned}
X_t X_{t-\tau} &= \sum_{j=1}^p \alpha_j X_{t-j} X_{t-\tau} + \sum_{j=1}^q \beta_j e_{t-j} X_{t-\tau} + e_t X_{t-\tau} \\
\implies E(X_t X_{t-\tau}) &= E \left(\sum_{j=1}^p \alpha_j X_{t-j} X_{t-\tau} + \sum_{j=1}^q \beta_j e_{t-j} X_{t-\tau} + e_t X_{t-\tau} \right) \\
\implies E(X_t X_{t-\tau}) &= \sum_{j=1}^p \alpha_j E(X_{t-j} X_{t-\tau}) + \sum_{j=1}^q \beta_j E(e_{t-j} X_{t-\tau}) + E(e_t X_{t-\tau}) \\
&&& \text{(Since expectation is a linear operator)}
\end{aligned}$$

Using (4), we have $Cov(X_t, X_s) = E(X_t X_s) - E(X_t)E(X_s) = E(X_t X_s)$

$$\begin{aligned}
\implies Cov(X_t, X_{t-\tau}) &= \sum_{j=1}^p \alpha_j Cov(X_{t-j}, X_{t-\tau}) + \sum_{j=1}^q \beta_j E(e_{t-j} X_{t-\tau}) + E(e_t X_{t-\tau}) \\
\implies \gamma_\tau &= \sum_{j=1}^p \alpha_j \gamma_{\tau-j} + \sum_{j=1}^q \beta_j E(e_{t-j} X_{t-\tau}) + E(e_t X_{t-\tau})
\end{aligned}$$

Now we note that $X_{t-\tau}$ is a function of the white noise variables $e_1, \dots, e_{t-\tau}$ and since they are independent from each other, e_t is independent from all of $e_1, \dots, e_{t-\tau}$ and thus $E(e_t X_{t-\tau}) = \text{Cov}(e_t, X_{t-\tau}) + \cancel{E(e_t)E(X_{t-\tau})}^0 = \text{Cov}(e_t, X_{t-\tau}) = 0$. Therefore, we get

$$\gamma_\tau = \sum_{j=1}^p \alpha_j \gamma_{\tau-j} + \sum_{j=1}^q \beta_j E(e_{t-j} X_{t-\tau}) \quad (5)$$

Using the previous logic, we can see that $E(e_t X_s) = 0$ whenever $t > s$ i.e. e_t and X_s are independent whenever $t > s$. In lieu of this observation, we split into the following two cases:

– **Case 1:** $\tau > q$

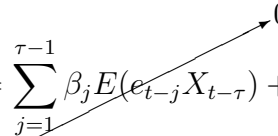
Here, $j < \tau \implies t - j > t - \tau$ for all $j = 1, \dots, q$. This implies, by our previous observation, that $E(e_{t-j} X_{t-\tau}) = 0$ for all $j = 1, \dots, q$

Thus the equation simplifies to:

$$\gamma_\tau = \sum_{j=1}^p \alpha_j \gamma_{\tau-j} \quad (6)$$

– **Case 2:** $0 < \tau \leq q$

In this case, we split up the summation into two parts as:

$$\begin{aligned} \sum_{j=1}^q \beta_j E(e_{t-j} X_{t-\tau}) &= \sum_{j=1}^{\tau-1} \beta_j E(e_{t-j} X_{t-\tau}) + \sum_{j=\tau}^q \beta_j E(e_{t-j} X_{t-\tau}) \\ &= \sum_{j=\tau}^q \beta_j E(e_{t-j} X_{t-\tau}) \end{aligned}$$


For each of the terms we'll need to expand $X_{t-\tau}$ to find the coefficient of e_{t-j} in it, which will give us the expectation.

Thus we get the equation as:

$$\gamma_\tau = \sum_{j=1}^p \alpha_j \gamma_{\tau-j} + \sum_{j=\tau}^q \beta_j E(e_{t-j} X_{t-\tau}) \quad (7)$$

Now, we find the variance of X_t .

$$X_t = \sum_{j=1}^p \alpha_j X_{t-j} + \sum_{j=1}^q \beta_j e_{t-j} + e_t$$

Multiplying by X_t on both sides, we get

$$\begin{aligned}
X_t^2 &= \sum_{j=1}^p \alpha_j X_{t-j} X_t + \sum_{j=1}^q \beta_j e_{t-j} X_t + e_t X_t \\
\Rightarrow E(X_t^2) &= \sum_{j=1}^p \alpha_j E(X_{t-j} X_t) + \sum_{j=1}^q \beta_j E(e_{t-j} X_t) + E(e_t X_t) \\
\Rightarrow Var(X_t) &= \sum_{j=1}^p \alpha_j \gamma_j + \sum_{j=1}^q \beta_j E(e_{t-j} X_t) + E \left(e_t \left\{ \sum_{j=1}^p \alpha_j X_{t-j} + \sum_{j=1}^q \beta_j e_{t-j} + e_t \right\} \right) \\
\Rightarrow Var(X_t) &= \sum_{j=1}^p \alpha_j \gamma_j + \sum_{j=1}^q \beta_j E(e_{t-j} X_t) + \sum_{j=1}^p \alpha_j \cancel{E(e_t X_{t-j})}^0 + \sum_{j=1}^q \beta_j \cancel{E(e_t e_{t-j})}^0 + E(e_t^2)
\end{aligned}$$

Therefore, we have the variance equation as:

$$Var(X_t) = \sum_{j=1}^p \alpha_j \gamma_j + \sum_{j=1}^q \beta_j E(e_{t-j} X_t) + 1 \quad (8)$$

Now we consider a stationary ARMA(1, 1) process with the equation

$$X_t = \alpha X_{t-1} + \beta e_{t-1} + e_t$$

To find the variance, we use (8).

$$\begin{aligned}
Var(X_t) &= \sum_{j=1}^p \alpha_j \gamma_j + \sum_{j=1}^q \beta_j E(e_{t-j} X_t) + 1 \\
&= \alpha \gamma_1 + \beta E(e_{t-1} X_t) + 1 \\
&= \alpha \gamma_1 + \beta E(e_{t-1} \{ \alpha X_{t-1} + \beta e_{t-1} + e_t \}) + 1 \\
&= \alpha \gamma_1 + \alpha \beta E(e_{t-1} X_{t-1}) + \beta^2 E(e_{t-1}^2) + \beta \cancel{E(e_{t-1} e_t)}^0 + 1
\end{aligned}$$

Now using $E(e_{t-1}^2) = Var(e_{t-1}) = 1$ and $E(e_t X_t) = 1 \forall t$, we get

$$Var(X_t) = \alpha \gamma_1 + \alpha \beta + \beta^2 + 1 \quad (9)$$

Note we have yet to find γ_1 which we'll do next. Note that since $0 < \tau = 1 \leq q = 1$, we'll use (7) to calculate.

$$\begin{aligned}
\gamma_\tau &= \sum_{j=1}^p \alpha_j \gamma_{\tau-j} + \sum_{j=\tau}^q \beta_j E(e_{t-j} X_{t-\tau}) \\
\Rightarrow \gamma_1 &= \alpha \gamma_0 + \beta E(e_{t-1} X_{t-1}) \\
\Rightarrow \gamma_1 &= \alpha \gamma_0 + \beta \\
\Rightarrow \gamma_1 &= \alpha Var(X_t) + \beta
\end{aligned} \quad (10)$$

Now using (9), we get

$$\begin{aligned}
& \gamma_1 = \alpha(\alpha\gamma_1 + \alpha\beta + \beta^2 + 1) + \beta \\
& \implies \gamma_1 = \alpha^2\gamma_1 + \alpha^2\beta + \alpha\beta^2 + \alpha + \beta \\
& \implies (1 - \alpha^2)\gamma_1 = \alpha^2\beta + \alpha\beta^2 + \alpha + \beta \\
& \implies \gamma_1 = \frac{\alpha^2\beta + \alpha\beta^2 + \alpha + \beta}{1 - \alpha^2}
\end{aligned} \tag{11}$$

Now that we have found γ_1 , we'll describe the other results using it since they get very messy otherwise. From (10), we have

$$\begin{aligned}
& \gamma_1 = \alpha \text{Var}(X_t) + \beta \\
& \implies \gamma_1 = \alpha\gamma_0 + \beta
\end{aligned}$$

Now dividing both sides by γ_0 and setting $\frac{\gamma_1}{\gamma_0}$ to ρ_1 , we get

$$\begin{aligned}
& \rho_1 = \alpha + \frac{\beta}{\gamma_0} \\
& \implies \rho_1 = \alpha + \frac{\beta}{\alpha\gamma_1 + \alpha\beta + \beta^2 + 1}
\end{aligned} \tag{Using (9)}$$

For $\tau > 1 = q$, we can use the equation for case 1 i.e. (6). We have

$$\begin{aligned}
& \gamma_\tau = \sum_{j=1}^p \alpha_j \gamma_{\tau-j} \\
& \implies \gamma_\tau = \alpha\gamma_{\tau-1}
\end{aligned}$$

Now dividing both sides by the variance to the correlations,

$$\begin{aligned}
& \rho_\tau = \alpha\rho_{\tau-1} \\
& \therefore \rho_2 = \alpha\rho_1 = \alpha^2 + \frac{\alpha\beta}{\alpha\gamma_1 + \alpha\beta + \beta^2 + 1} \\
& \rho_3 = \alpha\rho_2 = \alpha^3 + \frac{\alpha^2\beta}{\alpha\gamma_1 + \alpha\beta + \beta^2 + 1}
\end{aligned}$$

Therefore, we have found the values of ρ_1, ρ_2, ρ_3 as

$$\begin{aligned}
\rho_1 &= \alpha + \frac{\beta}{\alpha\gamma_1 + \alpha\beta + \beta^2 + 1} \\
\rho_2 &= \alpha^2 + \frac{\alpha\beta}{\alpha\gamma_1 + \alpha\beta + \beta^2 + 1} \\
\rho_3 &= \alpha^3 + \frac{\alpha^2\beta}{\alpha\gamma_1 + \alpha\beta + \beta^2 + 1}
\end{aligned}$$

where,

$$\gamma_1 = \frac{\alpha^2\beta + \alpha\beta^2 + \alpha + \beta}{1 - \alpha^2}$$