JÖNKÖPING UNIVERSITY
*Jönköping International Business School*

# Assignment 1

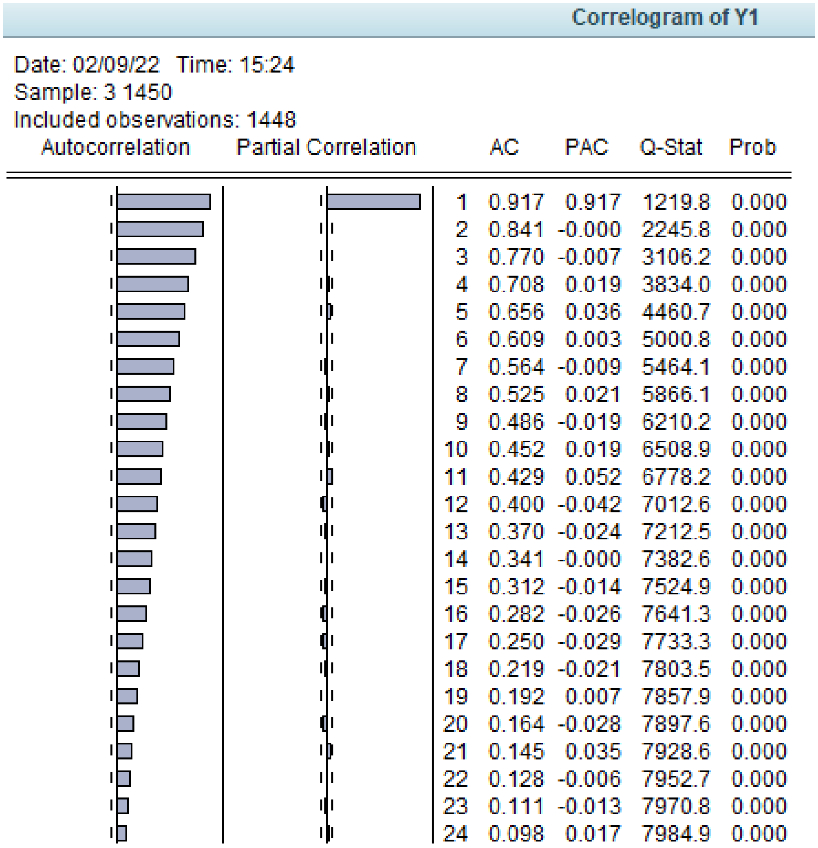
**COURSE**: *FSSS23 - Analytical Methods for Economic and Financial Analysis*
**PROGRAMME**: *International Financial Analysis*
**AUTHOR**: Sophie Dick (25.12.1999), Enrique Höner (03.11.1997)
**TUTOR**: *Pär Henrik Sjölander*

*All calculations were made with EViews Version 12 LITE*

# Q.1.1

## Y1:

### 1. Model identification



Correlogram of Y1

Date: 02/09/22 Time: 15:24
Sample: 3 1450
Included observations: 1448

| Lag | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|
| 1 | 0.917 | 0.917 | 1219.8 | 0.000 |
| 2 | 0.841 | -0.000 | 2245.8 | 0.000 |
| 3 | 0.770 | -0.007 | 3106.2 | 0.000 |
| 4 | 0.708 | 0.019 | 3834.0 | 0.000 |
| 5 | 0.656 | 0.036 | 4460.7 | 0.000 |
| 6 | 0.609 | 0.003 | 5000.8 | 0.000 |
| 7 | 0.564 | -0.009 | 5464.1 | 0.000 |
| 8 | 0.525 | 0.021 | 5866.1 | 0.000 |
| 9 | 0.486 | -0.019 | 6210.2 | 0.000 |
| 10 | 0.452 | 0.019 | 6508.9 | 0.000 |
| 11 | 0.429 | 0.052 | 6778.2 | 0.000 |
| 12 | 0.400 | -0.042 | 7012.6 | 0.000 |
| 13 | 0.370 | -0.024 | 7212.5 | 0.000 |
| 14 | 0.341 | -0.000 | 7382.6 | 0.000 |
| 15 | 0.312 | -0.014 | 7524.9 | 0.000 |
| 16 | 0.282 | -0.026 | 7641.3 | 0.000 |
| 17 | 0.250 | -0.029 | 7733.3 | 0.000 |
| 18 | 0.219 | -0.021 | 7803.5 | 0.000 |
| 19 | 0.192 | 0.007 | 7857.9 | 0.000 |
| 20 | 0.164 | -0.028 | 7897.6 | 0.000 |
| 21 | 0.145 | 0.035 | 7928.6 | 0.000 |
| 22 | 0.128 | -0.006 | 7952.7 | 0.000 |
| 23 | 0.111 | -0.013 | 7970.8 | 0.000 |
| 24 | 0.098 | 0.017 | 7984.9 | 0.000 |

The correlogram shows a spike at lag 1 for the SPAC and a decay for the SAC. This suggest that an autoregressive model with one lag (AR(1)) should be used.

### 2. Parameter estimation

Dependent Variable: Y1
Method: ARMA Maximum Likelihood (BFGS)
Date: 02/09/22 Time: 15:28
Sample: 3 1450
Included observations: 1448
Convergence achieved after 3 iterations
Coefficient covariance computed using outer product of gradients

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 0.409359 | 0.326445 | 1.253992 | 0.2100 |
| AR(1) | 0.919036 | 0.010528 | 87.29471 | 0.0000 |
| SIGMASQ | 1.009878 | 0.037521 | 26.91525 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.843231 | Mean dependent var | 0.455027 |
| Adjusted R-squared | 0.843014 | S.D. dependent var | 2.538949 |
| S.E. of regression | 1.005969 | Akaike info criterion | 2.853136 |
| Sum squared resid | 1462.303 | Schwarz criterion | 2.864071 |
| Log likelihood | -2062.670 | Hannan-Quinn criter. | 2.857217 |
| F-statistic | 3886.183 | Durbin-Watson stat | 2.005232 |
| Prob(F-statistic) | 0.000000 | | |

| | |
|---|---|
| Inverted AR Roots | .92 |

From this follows that estimated AR(1) model has the following form:

$$Y_t^1 = 0.409359 + 0.919036{*}Y_{t-1}^1$$

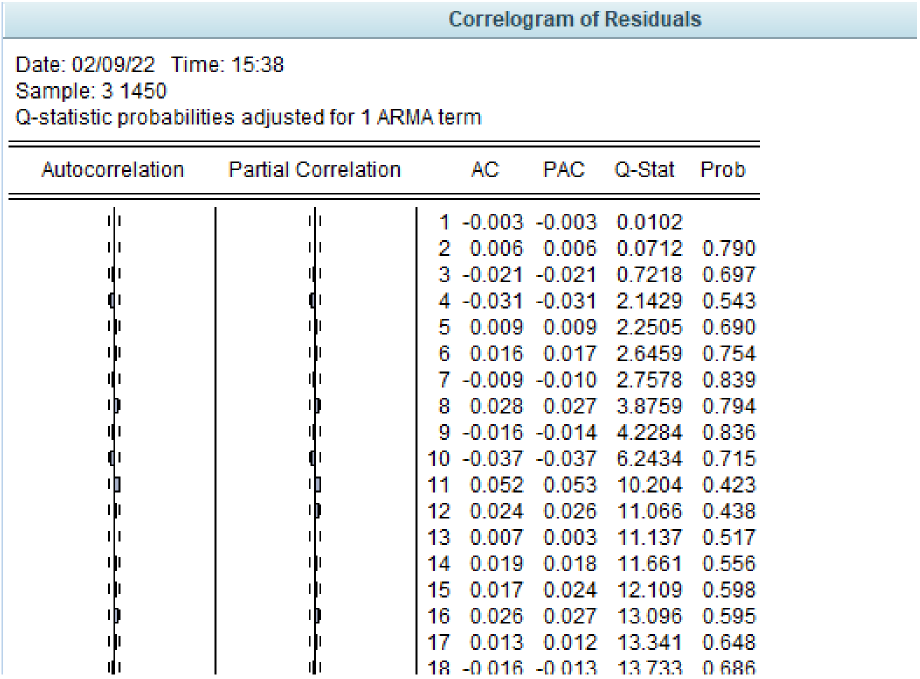
## 3. **Diagnostic testing**

i) Checking for autocorrelation in the residuals

The hypotheses for this first test are as follows:

$H_0$: $\rho_1 = \rho_2 = \ldots = \rho_K = 0$ *(no autocorrelation)*

$H_1$: $\rho_{(1,\ \ldots\ ,\ k)} \neq 0$ *(autocorrelation)*



Correlogram of Residuals

Date: 02/09/22 Time: 15:38
Sample: 3 1450
Q-statistic probabilities adjusted for 1 ARMA term

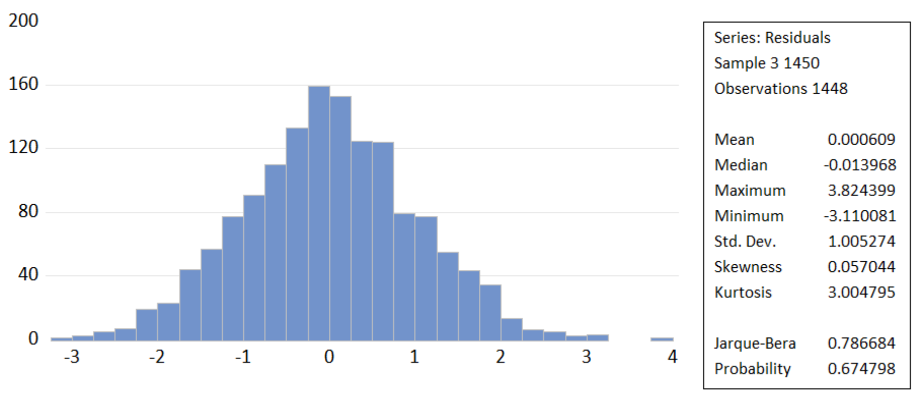
| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | -0.003 | -0.003 | 0.0102 | |
| | | 2 | 0.006 | 0.006 | 0.0712 | 0.790 |
| | | 3 | -0.021 | -0.021 | 0.7218 | 0.697 |
| | | 4 | -0.031 | -0.031 | 2.1429 | 0.543 |
| | | 5 | 0.009 | 0.009 | 2.2505 | 0.690 |
| | | 6 | 0.016 | 0.017 | 2.6459 | 0.754 |
| | | 7 | -0.009 | -0.010 | 2.7578 | 0.839 |
| | | 8 | 0.028 | 0.027 | 3.8759 | 0.794 |
| | | 9 | -0.016 | -0.014 | 4.2284 | 0.836 |
| | | 10 | -0.037 | -0.037 | 6.2434 | 0.715 |
| | | 11 | 0.052 | 0.053 | 10.204 | 0.423 |
| | | 12 | 0.024 | 0.026 | 11.066 | 0.438 |
| | | 13 | 0.007 | 0.003 | 11.137 | 0.517 |
| | | 14 | 0.019 | 0.018 | 11.661 | 0.556 |
| | | 15 | 0.017 | 0.024 | 12.109 | 0.598 |
| | | 16 | 0.026 | 0.027 | 13.096 | 0.595 |
| | | 17 | 0.013 | 0.012 | 13.341 | 0.648 |
| | | 18 | -0.016 | -0.013 | 13.733 | 0.686 |

It becomes apparent that none of the values exceed the confidence bound. Since all p-values are greater than 0.05 we can reject $H_0$ and assume that there is no autocorrelation in the residuals. Therefore, we can assume that the residuals are white noise.

ii) Jarque-Bera test

*$H_0$: Normal distribution*

*$H_1$: Non-normal distribution*



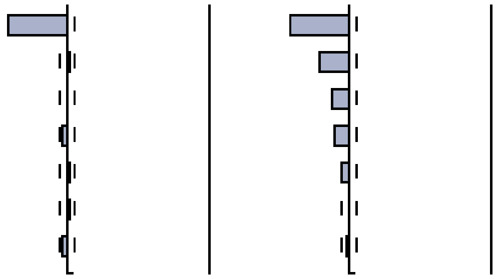

Since the Jarque-Bera statistic is 0.786 with a p-value of 0.67 > 0.05 we cannot reject $H_0$. There is no indication against normality, which leads us to assume a normal distribution.

**Y2:**

## 1. Model identification

**Correlogram of Y2**

Date: 02/09/22 Time: 15:49
Sample: 3 1450
Included observations: 1448




| | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|
| 1 | -0.442 | -0.442 | 283.21 | 0.000 |
| 2 | 0.020 | -0.218 | 283.78 | 0.000 |
| 3 | -0.007 | -0.121 | 283.86 | 0.000 |
| 4 | -0.025 | -0.102 | 284.75 | 0.000 |
| 5 | 0.018 | -0.053 | 285.24 | 0.000 |
| 6 | 0.018 | -0.002 | 285.71 | 0.000 |
| 7 | -0.026 | -0.021 | 286.71 | 0.000 |

The correlogram shows a spike at lag 1 for the SAC and a decay for the SPAC. This suggest that an moving average model with one lag (MA(1)) should be used.

## 2. Parameter Estimation

Dependent Variable: Y2
Method: ARMA Maximum Likelihood (BFGS)
Date: 02/09/22 Time: 15:51
Sample: 3 1450
Included observations: 1448
Convergence achieved after 4 iterations
Coefficient covariance computed using outer product of gradients

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 0.017256 | 0.011202 | 1.540441 | 0.1237 |
| MA(1) | -0.577106 | 0.021856 | -26.40486 | 0.0000 |
| SIGMASQ | 1.012539 | 0.037672 | 26.87749 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.253450 | Mean dependent var | 0.017762 |
| Adjusted R-squared | 0.252417 | S.D. dependent var | 1.165002 |
| S.E. of regression | 1.007294 | Akaike info criterion | 2.854761 |
| Sum squared resid | 1466.156 | Schwarz criterion | 2.865696 |
| Log likelihood | -2063.847 | Hannan-Quinn criter. | 2.858842 |
| F-statistic | 245.2857 | Durbin-Watson stat | 2.009779 |
| Prob(F-statistic) | 0.000000 | | |

| | |
|---|---|
| Inverted MA Roots | .58 |

From this follows that the estimated MA(1) model has the following form:

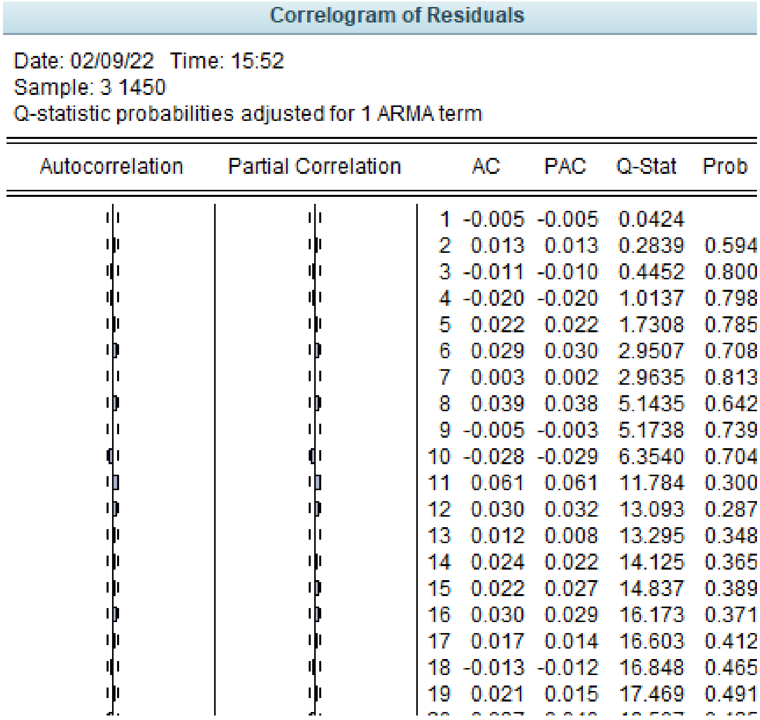
$$Y_t^2 = 0.017256 - 0.577106 * e_{t-1}$$

3. **Diagnostic testing**

i) Checking for autocorrelation in the residuals

*$H_0$: $\rho_1 = \rho_2 = \ldots = \rho_K = 0$ (no autocorrelation)*

*$H_1$: $\rho_{(1, \ldots, k)} \neq 0$ (autocorrelation)*



Correlogram of Residuals

Date: 02/09/22 Time: 15:52
Sample: 3 1450
Q-statistic probabilities adjusted for 1 ARMA term

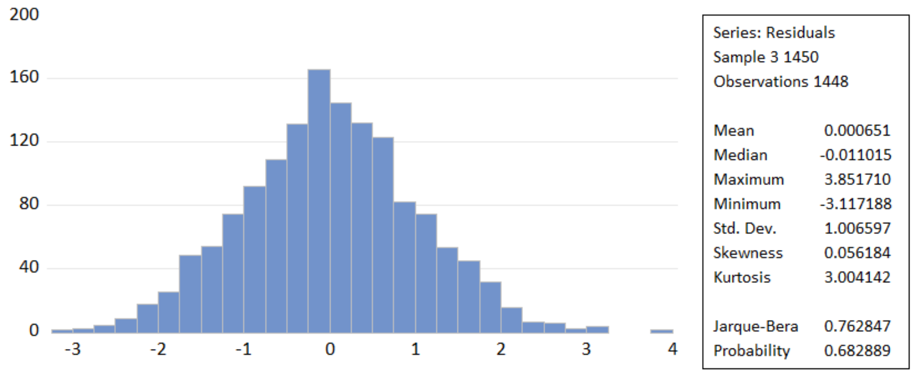
| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | -0.005 | -0.005 | 0.0424 | |
| | | 2 | 0.013 | 0.013 | 0.2839 | 0.594 |
| | | 3 | -0.011 | -0.010 | 0.4452 | 0.800 |
| | | 4 | -0.020 | -0.020 | 1.0137 | 0.798 |
| | | 5 | 0.022 | 0.022 | 1.7308 | 0.785 |
| | | 6 | 0.029 | 0.030 | 2.9507 | 0.708 |
| | | 7 | 0.003 | 0.002 | 2.9635 | 0.813 |
| | | 8 | 0.039 | 0.038 | 5.1435 | 0.642 |
| | | 9 | -0.005 | -0.003 | 5.1738 | 0.739 |
| | | 10 | -0.028 | -0.029 | 6.3540 | 0.704 |
| | | 11 | 0.061 | 0.061 | 11.784 | 0.300 |
| | | 12 | 0.030 | 0.032 | 13.093 | 0.287 |
| | | 13 | 0.012 | 0.008 | 13.295 | 0.348 |
| | | 14 | 0.024 | 0.022 | 14.125 | 0.365 |
| | | 15 | 0.022 | 0.027 | 14.837 | 0.389 |
| | | 16 | 0.030 | 0.029 | 16.173 | 0.371 |
| | | 17 | 0.017 | 0.014 | 16.603 | 0.412 |
| | | 18 | -0.013 | -0.012 | 16.848 | 0.465 |
| | | 19 | 0.021 | 0.015 | 17.469 | 0.491 |

It becomes apparent that none of the values exceed the confidence bound. Since all p-values are greater than 0.05 we can reject $H_0$ and assume that there is no autocorrelation in the residuals. Therefore, we can assume that the residuals are white noise.

ii) Jarque-Bera test

*$H_0$: Normal distribution*

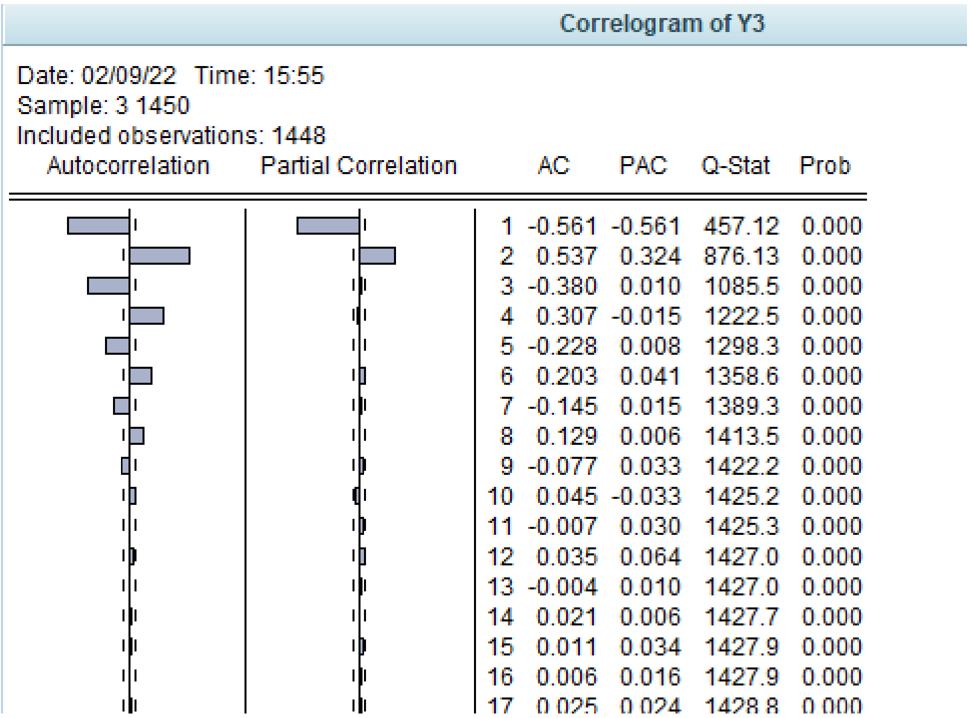
*$H_1$: non-normal distribution*




Since the Jarque-Bera statistic is 0.763 with a p-value of 0.68 > 0.05 we cannot reject $H_0$. There is no indication against normality, which leads us to assume a normal distribution.

## Y3:

### 1. Model identification



Correlogram of Y3

Date: 02/09/22 Time: 15:55
Sample: 3 1450
Included observations: 1448

| Lag | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|
| 1 | -0.561 | -0.561 | 457.12 | 0.000 |
| 2 | 0.537 | 0.324 | 876.13 | 0.000 |
| 3 | -0.380 | 0.010 | 1085.5 | 0.000 |
| 4 | 0.307 | -0.015 | 1222.5 | 0.000 |
| 5 | -0.228 | 0.008 | 1298.3 | 0.000 |
| 6 | 0.203 | 0.041 | 1358.6 | 0.000 |
| 7 | -0.145 | 0.015 | 1389.3 | 0.000 |
| 8 | 0.129 | 0.006 | 1413.5 | 0.000 |
| 9 | -0.077 | 0.033 | 1422.2 | 0.000 |
| 10 | 0.045 | -0.033 | 1425.2 | 0.000 |
| 11 | -0.007 | 0.030 | 1425.3 | 0.000 |
| 12 | 0.035 | 0.064 | 1427.0 | 0.000 |
| 13 | -0.004 | 0.010 | 1427.0 | 0.000 |
| 14 | 0.021 | 0.006 | 1427.7 | 0.000 |
| 15 | 0.011 | 0.034 | 1427.9 | 0.000 |
| 16 | 0.006 | 0.016 | 1427.9 | 0.000 |
| 17 | 0.025 | 0.024 | 1428.8 | 0.000 |

The correlogram shows a decay for the SAC and two spikes at lag 1 and 2 for the SPAC. This suggest that an autoregressive model with two lags (AR(2)) should be used.

### 2. Parameter Estimation

Dependent Variable: Y2
Method: ARMA Maximum Likelihood (BFGS)
Date: 02/09/22 Time: 15:51
Sample: 3 1450
Included observations: 1448
Convergence achieved after 4 iterations
Coefficient covariance computed using outer product of gradients

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 0.017256 | 0.011202 | 1.540441 | 0.1237 |
| MA(1) | -0.577106 | 0.021856 | -26.40486 | 0.0000 |
| SIGMASQ | 1.012539 | 0.037672 | 26.87749 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.253450 | Mean dependent var | 0.017762 |
| Adjusted R-squared | 0.252417 | S.D. dependent var | 1.165002 |
| S.E. of regression | 1.007294 | Akaike info criterion | 2.854761 |
| Sum squared resid | 1466.156 | Schwarz criterion | 2.865696 |
| Log likelihood | -2063.847 | Hannan-Quinn criter. | 2.858842 |
| F-statistic | 245.2857 | Durbin-Watson stat | 2.009779 |
| Prob(F-statistic) | 0.000000 | | |

| | |
|---|---|
| Inverted MA Roots | .58 |

From this follows that the estimated AR(2) model has the following form:

$$Y_t^3 = 0.038942 - 0.378933 * Y_{t-1}^3 + 0.324412 * Y_{t-2}^3$$

## 3. Diagnostic testing

i) Checking for autocorrelation in the residuals

$H_0$: $\rho_1 = \rho_2 = \ldots = \rho_K = 0$ *(no autocorrelation)*

$H_1$: $\rho_{(1, \ldots, k)} \neq 0$ *(autocorrelation)*

**Correlogram of Residuals**

Date: 02/09/22 Time: 15:59
Sample: 3 1450
Q-statistic probabilities adjusted for 2 ARMA terms

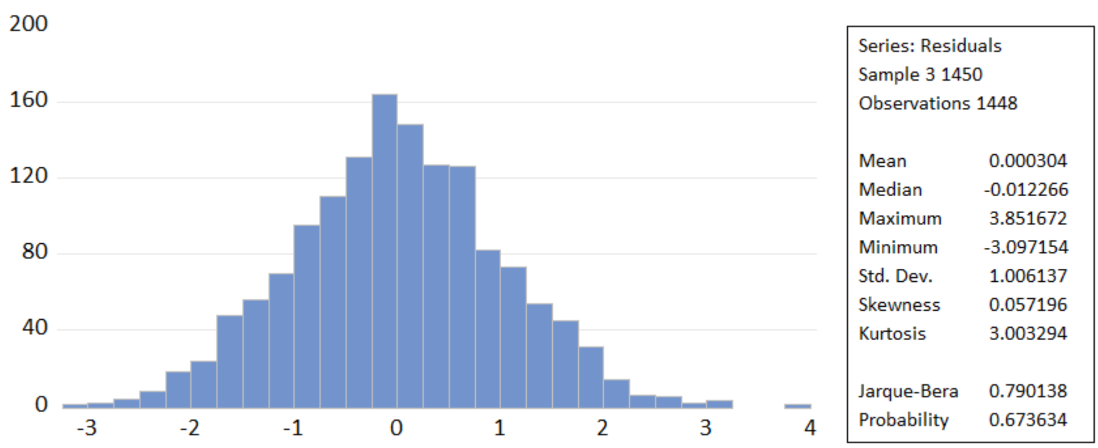
| | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|
| 1 | -0.003 | -0.003 | 0.0161 | |
| 2 | 0.010 | 0.010 | 0.1609 | |
| 3 | -0.003 | -0.003 | 0.1761 | 0.675 |
| 4 | -0.020 | -0.020 | 0.7700 | 0.680 |
| 5 | 0.026 | 0.026 | 1.7668 | 0.622 |
| 6 | 0.028 | 0.029 | 2.9012 | 0.574 |
| 7 | 0.005 | 0.004 | 2.9329 | 0.710 |
| 8 | 0.038 | 0.038 | 5.0773 | 0.534 |
| 9 | -0.004 | -0.002 | 5.0964 | 0.648 |
| 10 | -0.029 | -0.029 | 6.3047 | 0.613 |
| 11 | 0.062 | 0.061 | 11.890 | 0.220 |
| 12 | 0.030 | 0.032 | 13.242 | 0.210 |
| 13 | 0.013 | 0.009 | 13.474 | 0.263 |
| 14 | 0.024 | 0.021 | 14.342 | 0.279 |
| 15 | 0.022 | 0.027 | 15.078 | 0.302 |
| 16 | 0.031 | 0.029 | 16.458 | 0.286 |
| 17 | 0.017 | 0.014 | 16.895 | 0.325 |
| 18 | -0.012 | -0.012 | 17.122 | 0.378 |
| 19 | 0.021 | 0.015 | 17.762 | 0.404 |

It becomes apparent that none of the values exceed the confidence bound. Since all p-values are greater than 0.05 we can reject $H_0$ and assume that there is no autocorrelation in the residuals. Therefore, we can assume that the residuals are white noise.

ii) Jarque-Bera test

*$H_0$: Normal distribution*

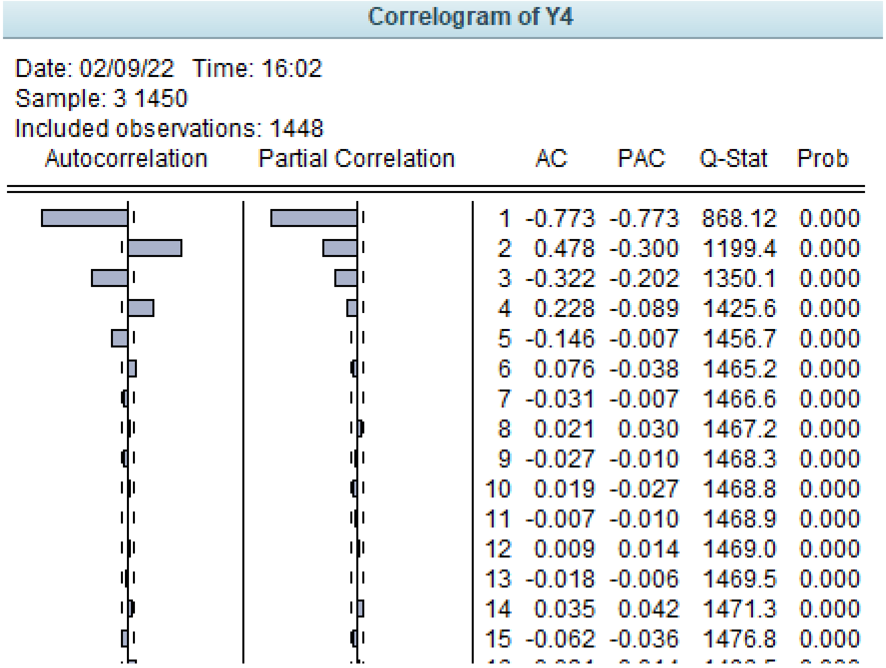
*$H_1$: non-normal distribution*




Since the Jarque-Bera statistic is 0.79 with a p-value of 0.67 > 0.05 we cannot reject $H_0$. There is no indication against normality, which leads us to assume a normal distribution.

**<u>Y4:</u>**

## 1. Model identification



Correlogram of Y4

Date: 02/09/22 Time: 16:02
Sample: 3 1450
Included observations: 1448

| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | -0.773 | -0.773 | 868.12 | 0.000 |
| | | 2 | 0.478 | -0.300 | 1199.4 | 0.000 |
| | | 3 | -0.322 | -0.202 | 1350.1 | 0.000 |
| | | 4 | 0.228 | -0.089 | 1425.6 | 0.000 |
| | | 5 | -0.146 | -0.007 | 1456.7 | 0.000 |
| | | 6 | 0.076 | -0.038 | 1465.2 | 0.000 |
| | | 7 | -0.031 | -0.007 | 1466.6 | 0.000 |
| | | 8 | 0.021 | 0.030 | 1467.2 | 0.000 |
| | | 9 | -0.027 | -0.010 | 1468.3 | 0.000 |
| | | 10 | 0.019 | -0.027 | 1468.8 | 0.000 |
| | | 11 | -0.007 | -0.010 | 1468.9 | 0.000 |
| | | 12 | 0.009 | 0.014 | 1469.0 | 0.000 |
| | | 13 | -0.018 | -0.006 | 1469.5 | 0.000 |
| | | 14 | 0.035 | 0.042 | 1471.3 | 0.000 |
| | | 15 | -0.062 | -0.036 | 1476.8 | 0.000 |

This suggest that an ARMA(1,1) model should be used. Since there is an exponential decay in both the SPAC and SAC.

## 2. Parameter Estimation

Dependent Variable: Y4
Method: ARMA Maximum Likelihood (BFGS)
Date: 02/09/22 Time: 16:03
Sample: 3 1450
Included observations: 1448
Convergence achieved after 6 iterations
Coefficient covariance computed using outer product of gradients

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| AR(1) | -0.597455 | 0.024793 | -24.09788 | 0.0000 |
| MA(1) | -0.501792 | 0.025793 | -19.45450 | 0.0000 |
| SIGMASQ | 0.995786 | 0.037350 | 26.66087 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.652537 | Mean dependent var | 0.003664 |
| Adjusted R-squared | 0.652056 | S.D. dependent var | 1.693474 |
| S.E. of regression | 0.998926 | Akaike info criterion | 2.838666 |
| Sum squared resid | 1441.899 | Schwarz criterion | 2.849601 |
| Log likelihood | -2052.194 | Hannan-Quinn criter. | 2.842746 |
| Durbin-Watson stat | 1.977742 | | |

| | |
|---|---|
| Inverted AR Roots | -.60 |
| Inverted MA Roots | .50 |

From this follows that the estimated ARMA(1,1) model has the following form:

$$Y_t^4 = 0.002971 - 0.597404 * Y_{t-1}^4 - 0.501962 * e_{t-1}$$

## 3. Diagnostic testing

i) Checking for autocorrelation in the residuals

The hypotheses for this first test are as follows:

$H_0$: $\rho_1 = \rho_2 = \ldots = \rho_K = 0$ *(no autocorrelation)*

$H_1$: $\rho_{(1, \ldots, k)} \neq 0$ *(autocorrelation)*

Date: 02/09/22 Time: 16:07
Sample: 3 1450
Q-statistic probabilities adjusted for 2 ARMA terms

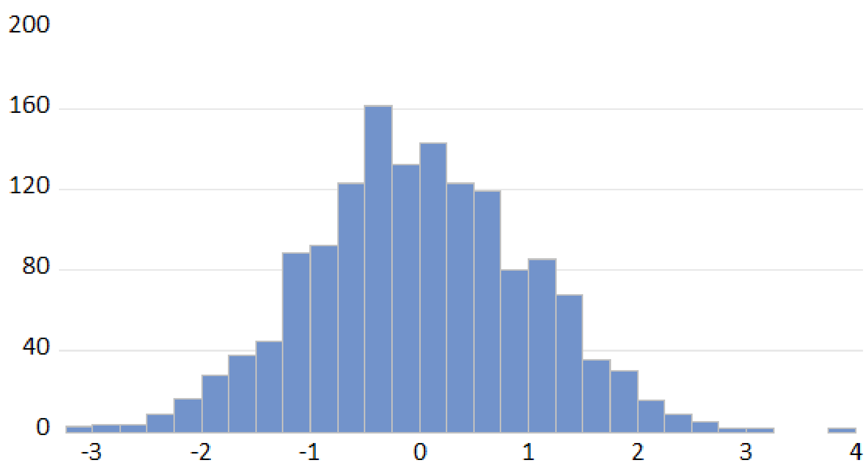
| Autocorrelation | Partial Correlation | | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|---|---|
| | | 1 | 0.010 | 0.010 | 0.1473 | |
| | | 2 | -0.019 | -0.019 | 0.6846 | |
| | | 3 | -0.023 | -0.022 | 1.4395 | 0.230 |
| | | 4 | 0.051 | 0.051 | 5.2529 | 0.072 |
| | | 5 | -0.020 | -0.022 | 5.8170 | 0.121 |
| | | 6 | -0.003 | -0.001 | 5.8271 | 0.212 |
| | | 7 | 0.029 | 0.031 | 7.0927 | 0.214 |
| | | 8 | -0.017 | -0.022 | 7.5157 | 0.276 |
| | | 9 | -0.034 | -0.031 | 9.2060 | 0.238 |
| | | 10 | 0.006 | 0.008 | 9.2658 | 0.320 |
| | | 11 | 0.022 | 0.017 | 9.9744 | 0.353 |
| | | 12 | 0.005 | 0.007 | 10.013 | 0.439 |
| | | 13 | 0.005 | 0.009 | 10.051 | 0.526 |
| | | 14 | -0.005 | -0.007 | 10.082 | 0.609 |
| | | 15 | -0.020 | -0.020 | 10.660 | 0.639 |
| | | 16 | 0.035 | 0.038 | 12.463 | 0.569 |

It becomes apparent that none of the values exceed the confidence bound. Since all p-values are greater than 0.05 we can reject $H_0$ and assume that there is no autocorrelation in the residuals. Therefore, we can assume that the residuals are white noise.

ii) Jarque-Bera test

$H_0$: *Normal distribution*

$H_1$: *non-normal distribution*




| | |
|---|---|
| Series: Residuals | |
| Sample 3 1450 | |
| Observations 1448 | |
| Mean | 0.000810 |
| Median | -0.025104 |
| Maximum | 3.823811 |
| Minimum | -3.099385 |
| Std. Dev. | 0.998190 |
| Skewness | 0.023353 |
| Kurtosis | 2.963322 |
| Jarque-Bera | 0.212781 |
| Probability | 0.899074 |

Since the Jarque-Bera statistic is 0.21 with a p-value of 0.899 > 0.05 we cannot reject $H_0$. There is no indication against normality, which leads us to assume a normal distribution.

# Q.1.2

As shown in Q.1.1 for Y1 the optimal model is an AR(1). However, when using a MA(1) model we get the following results:

Dependent Variable: Y1
Method: ARMA Maximum Likelihood (BFGS)
Date: 02/09/22 Time: 16:12
Sample: 3 1450
Included observations: 1448
Convergence achieved after 7 iterations
Coefficient covariance computed using outer product of gradients

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 0.453473 | 0.076207 | 5.950525 | 0.0000 |
| MA(1) | 0.756033 | 0.016991 | 44.49505 | 0.0000 |
| SIGMASQ | 2.723890 | 0.093864 | 29.01962 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.577154 | Mean dependent var | 0.455027 |
| Adjusted R-squared | 0.576569 | S.D. dependent var | 2.538949 |
| S.E. of regression | 1.652134 | Akaike info criterion | 3.844667 |
| Sum squared resid | 3944.193 | Schwarz criterion | 3.855602 |
| Log likelihood | -2780.539 | Hannan-Quinn criter. | 3.848748 |
| F-statistic | 986.1615 | Durbin-Watson stat | 0.947520 |
| Prob(F-statistic) | 0.000000 | | |
| Inverted MA Roots | -.76 | | |

Therefore, the estimated MA(1) model has the following form:

$Y_t^1 = 0.453473 + 0.756033 * e_{t-1}$

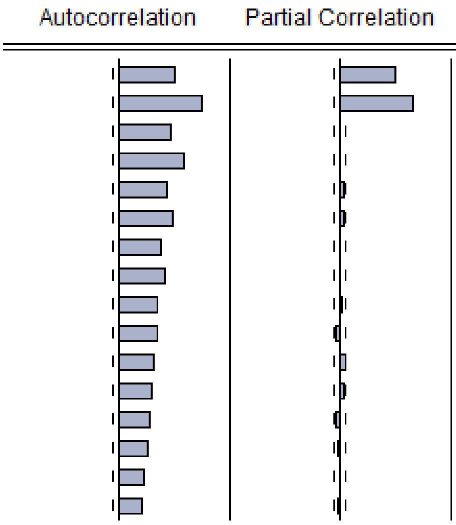
Diagnostic testing:

i) Testing for autocorrelation in the residuals

$H_0$: $\rho_1 = \rho_2 = \ldots = \rho_K = 0$ *(no autocorrelation)*

$H_1$: $\rho_{(1, \ldots, k)} \neq 0$ *(autocorrelation)*

Date: 02/09/22 Time: 16:17
Sample: 3 1450
Q-statistic probabilities adjusted for 1 ARMA term

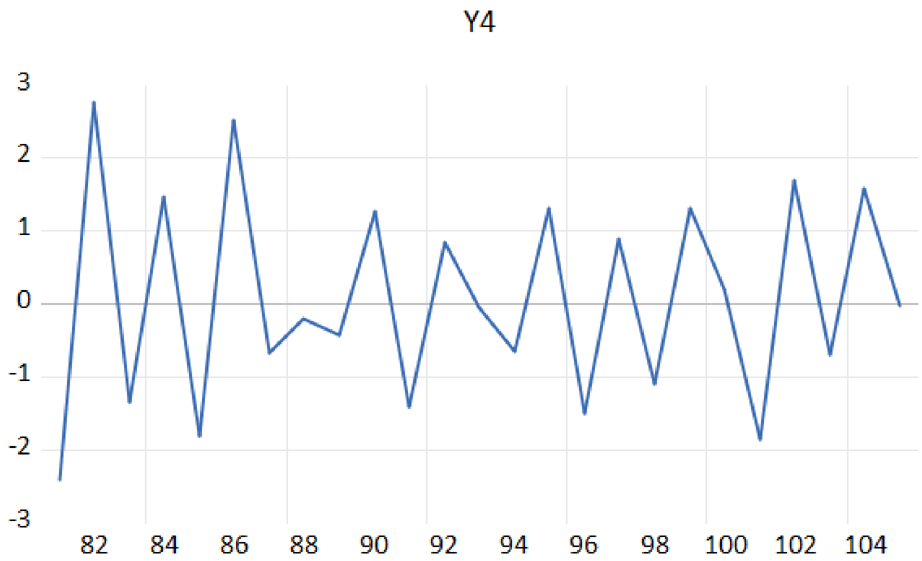



| | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|
| 1 | 0.525 | 0.525 | 400.42 | |
| 2 | 0.780 | 0.696 | 1282.8 | 0.000 |
| 3 | 0.488 | -0.007 | 1628.3 | 0.000 |
| 4 | 0.618 | -0.005 | 2183.5 | 0.000 |
| 5 | 0.445 | 0.026 | 2472.2 | 0.000 |
| 6 | 0.513 | 0.034 | 2855.8 | 0.000 |
| 7 | 0.397 | -0.007 | 3084.9 | 0.000 |
| 8 | 0.430 | 0.000 | 3354.2 | 0.000 |
| 9 | 0.355 | 0.018 | 3538.5 | 0.000 |
| 10 | 0.353 | -0.028 | 3720.7 | 0.000 |
| 11 | 0.328 | 0.046 | 3878.0 | 0.000 |
| 12 | 0.310 | 0.026 | 4018.5 | 0.000 |
| 13 | 0.283 | -0.039 | 4135.3 | 0.000 |
| 14 | 0.262 | -0.020 | 4236.2 | 0.000 |
| 15 | 0.241 | 0.002 | 4321.4 | 0.000 |
| 16 | 0.214 | -0.023 | 4388.5 | 0.000 |

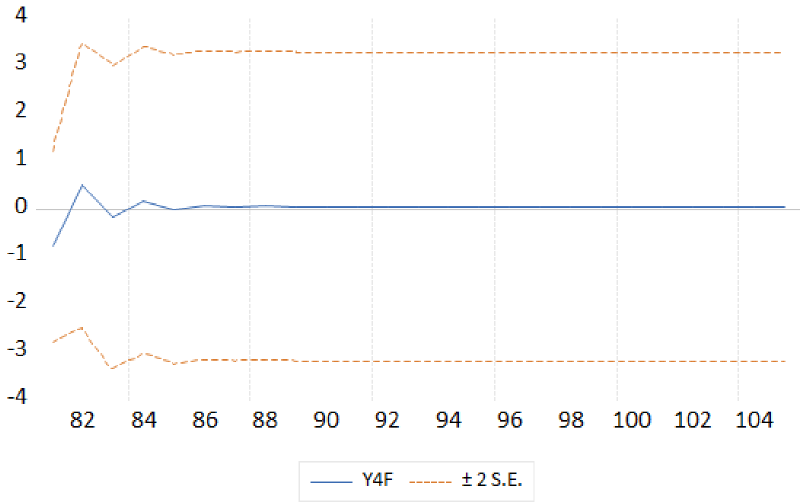
The output shows statistically relevant (because all p-values are <0.05) autocorrelation in the residuals. We must therefore reject $H_0$. This leads to biased results.

# <u>Q.1.3</u> (Dynamic out of sample forecasting for Y4)

Plot of the real outcomes of Y4 for the period 81 to 105:

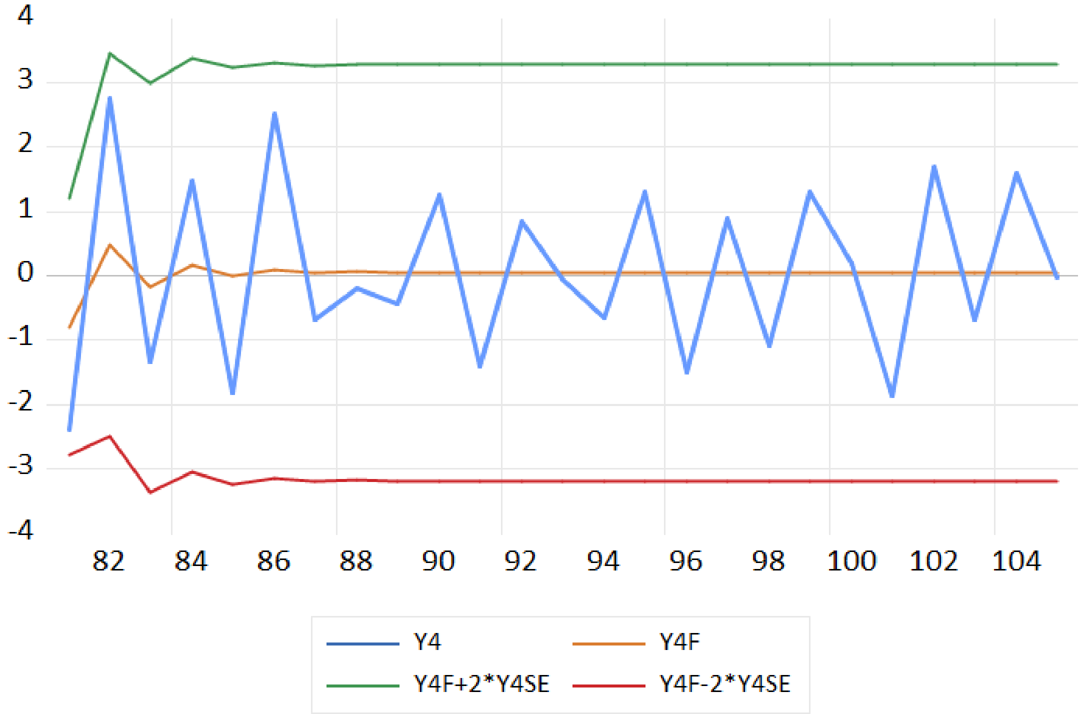



Forecast of Y4 for the period 81 to 105 using only the data from period 2 to 80:




| | |
|---|---|
| Forecast: Y4F | |
| Actual: Y4 | |
| Forecast sample: 81 105 | |
| Included observations: 25 | |
| Root Mean Squared Error | 1.299900 |
| Mean Absolute Error | 1.134750 |
| Mean Abs. Percent Error | 112.3799 |
| Theil Inequality Coef. | 0.810811 |
| Bias Proportion | 0.000937 |
| Variance Proportion | 0.860384 |
| Covariance Proportion | 0.138680 |
| Theil U2 Coefficient | 0.976934 |
| Symmetric MAPE | 181.2762 |

Comparison between the dynamic out of sample forecast and reality:



4
3
2
1
0
-1
-2
-3
-4
82
84
86
88
90
92
94
96
98
100
102
104
Y4
Y4F
Y4F+2*Y4SE
Y4F-2*Y4SE

# Q.1.4.1

Dependent Variable: SMOKER
Method: Least Squares
Date: 02/09/22 Time: 17:01
Sample: 1 1196
Included observations: 1196

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.123089 | 0.188356 | 5.962575 | 0.0000 |
| AGE | -0.004726 | 0.000829 | -5.700952 | 0.0000 |
| EDUC | -0.020613 | 0.004616 | -4.465272 | 0.0000 |
| INCOME | 1.03E-06 | 1.63E-06 | 0.628522 | 0.5298 |
| PCIGS79 | -0.005132 | 0.002852 | -1.799076 | 0.0723 |

| | | | |
|---|---|---|---|
| R-squared | 0.038770 | Mean dependent var | 0.380435 |
| Adjusted R-squared | 0.035541 | S.D. dependent var | 0.485697 |
| S.E. of regression | 0.476988 | Akaike info criterion | 1.361519 |
| Sum squared resid | 270.9729 | Schwarz criterion | 1.382785 |
| Log likelihood | -809.1885 | Hannan-Quinn criter. | 1.369531 |
| F-statistic | 12.00927 | Durbin-Watson stat | 1.943548 |
| Prob(F-statistic) | 0.000000 | | |

## Q.1.4.2

If the years of education of the average person increase by 1, the probability of being a smoker decreases by 0.020613 = 2.06%

# Q.1.4.3

**Logit Model:**

Since in the linear probability model very high cigarette prices could lead to a negative probability of smoking the Logit model uses odds ratios:

$Odds = \frac{P_i}{1-P_i}$ where $\frac{P_i}{1-P_i}$ $\in [0, \infty]$ with a midpoint at 1 (which can be seen by inserting 0.5 for $P_i$.

However, a distribution over a range of $[0, \infty]$ is heavily skewed. Furthermore, a lower bound of 0 does not exclude negative probabilities.

To work around this problem the natural logarithm of the odds ratio is taken:

$\ln(\frac{P_i}{1-P_i})$ where $\ln(\frac{P_i}{1-P_i})$ $\in [-\infty, \infty]$ with a midpoint at 0 (which can be seen by inserting 0.5 for $P_i$.

Therefore, a simple Logit model takes the following form:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 * X_i = Z_i$$

Solving for $P_i$:

$$\frac{P_i}{1-P_i} = e^{Z_i}$$

$$\rightarrow P_i = (1 - P_i) * e^{Z_i} = e^{Z_i} - e^{Z_i} * P_i$$

$$\rightarrow P_i + e^{Z_i} * P_i = P_i * (1 + e^{Z_i}) = e^{Z_i}$$

$$\rightarrow P_i = \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{e^{Z_i}}{(1+e^{-Z_i}) * e^{Z_i}} = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\beta_1+\beta_2*X_i)}}$$

When solving for $1 - P_i$ we get through the same transformation the following:

$$1 - P_i = \frac{1}{1+e^{Z_i}} = \frac{1}{1+e^{\beta_1+\beta_2*X_i}}$$

Going back to the odds ratio we now get:

$$\frac{P_i}{1-P_i} = \frac{\frac{e^{Z_i}}{1+e^{Z_i}}}{\frac{1}{1+e^{Z_i}}} = e^{Z_i}$$

When taking the natural logarithm, we get:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \ln(e^{Z_i}) = Z_i = \beta_1 + \beta_2 * X_i$$

The estimated Logit model from the assignment has the following form:

Dependent Variable: SMOKER
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 02/09/22 Time: 17:23
Sample: 1 1196
Included observations: 1196
Convergence achieved after 2 iterations
Coefficient covariance computed using the Huber-White method

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | 2.745082 | 0.821766 | 3.340466 | 0.0008 |
| AGE | -0.020853 | 0.003613 | -5.772377 | 0.0000 |
| EDUC | -0.090973 | 0.020548 | -4.427426 | 0.0000 |
| INCOME | 4.72E-06 | 7.27E-06 | 0.649033 | 0.5163 |
| PCIGS79 | -0.022319 | 0.012388 | -1.801631 | 0.0716 |

| | | | |
|---|---|---|---|
| McFadden R-squared | 0.029748 | Mean dependent var | 0.380435 |
| S.D. dependent var | 0.485697 | S.E. of regression | 0.477407 |
| Akaike info criterion | 1.297393 | Sum squared resid | 271.4495 |
| Schwarz criterion | 1.318658 | Log likelihood | -770.8409 |
| Hannan-Quinn criter. | 1.305405 | Deviance | 1541.682 |
| Restr. deviance | 1588.950 | Restr. log likelihood | -794.4748 |
| LR statistic | 47.26785 | Avg. log likelihood | -0.644516 |
| Prob(LR statistic) | 0.000000 | | |

| | | | |
|---|---|---|---|
| Obs with Dep=0 | 741 | Total obs | 1196 |
| Obs with Dep=1 | 455 | | |

With these estimated coefficients we can now calculate the probability that an individual is a smoker, considering their specific data.

$$P_i = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\beta_1+\beta_2 * AGE_i+\beta_3 * Education_i+\beta_4 * Income_i+\beta_5 * Price_i)}}$$

If we assume the following information about an individual: Age= 30 , Education= 20, Income= 15000 and Price (of cigarettes)=80 we get the following probability:

$$P_i = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(2.745082-0.020853*30-0.090973*20+0.00000472*15000-0.022319*80)}}$$

$$= 0.195496 \sim 19.5\%$$

Therefore, we can assume that a person with these characteristics has a probability of about 19.5% to be a smoker.

# Q.1.4.4

The marginal effect on the probability of smoking on age is given by the following formula:

$\hat{\beta}_4 * \bar{Y} * (1 - \bar{Y}) = -0.020853 * 0.38 * 0.62 = -0.0049129668$

(Where $\bar{Y}$ gives the percentage of smokers in the considered sample)

This could be interpreted in the following way: An increase in age by 1 unit (years) causes on average a decrease of probability of being a smoker by 0.5%.

# Q.1.4.5

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SMOKER
Method: ML - Binary Probit (Newton-Raphson / Marquardt steps)
Date: 02/10/22 Time: 11:14
Sample: 1 1196
Included observations: 1196
Convergence achieved after 2 iterations
Coefficient covariance computed using the Huber-White method

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.701906 | 0.506635 | 3.359234 | 0.0008 |
| AGE | -0.012965 | 0.002226 | -5.824174 | 0.0000 |
| EDUC | -0.056230 | 0.012554 | -4.479040 | 0.0000 |
| INCOME | 2.72E-06 | 4.45E-06 | 0.611796 | 0.5407 |
| PCIGS79 | -0.013794 | 0.007653 | -1.802412 | 0.0715 |

| | | | |
|---|---|---|---|
| McFadden R-squared | 0.030066 | Mean dependent var | 0.380435 |
| S.D. dependent var | 0.485697 | S.E. of regression | 0.477328 |
| Akaike info criterion | 1.296970 | Sum squared resid | 271.3598 |
| Schwarz criterion | 1.318236 | Log likelihood | -770.5881 |
| Hannan-Quinn criter. | 1.304982 | Deviance | 1541.176 |
| Restr. deviance | 1588.950 | Restr. log likelihood | -794.4748 |
| LR statistic | 47.77335 | Avg. log likelihood | -0.644304 |
| Prob(LR statistic) | 0.000000 | | |

| | | | |
|---|---|---|---|
| Obs with Dep=0 | 741 | Total obs | 1196 |
| Obs with Dep=1 | 455 | | |

In the Probit model all variables, except the income variable, are significant at the 10% level.

To calculate the marginal effect, the coefficient of the variable is multiplied with the value of the normal density function evaluated for all the X values for that individual.

If we multiply the Probit coefficient by approximately 1.81 we will get the Logit coefficient.

Age: -0.012965*1.81= -0.0235

This is about the same result we get from the Logit age coefficient.

# Q.1.4.6

Marginal effect of education:

$$\bar{Y} = \frac{455}{1196} = 0.38$$

$$\hat{\beta}_2 * \bar{Y} * (1 - \bar{Y}) = -0.056230 * 0.38 * 0.62 = -0.013247788$$

Marginal effect on probability of smoking of education: An increase of 1 unit (year) in education decreases the probability on the average person by ~0.1%.

# Q.1.4.7

**Uncensored data:**

Dependent Variable: HOURS
Method: Least Squares
Date: 02/10/22 Time: 12:00
Sample: 1 753
Included observations: 753

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1298.293 | 231.9451 | 5.597413 | 0.0000 |
| AGE | -29.55452 | 3.864413 | -7.647869 | 0.0000 |
| EDUC | 5.064135 | 12.55700 | 0.403292 | 0.6868 |
| EXPER | 68.52186 | 9.398942 | 7.290380 | 0.0000 |
| EXPERSQ | -0.779211 | 0.308540 | -2.525480 | 0.0118 |
| FAMINC | 0.028993 | 0.003201 | 9.056627 | 0.0000 |
| KIDSL6 | -395.5547 | 55.63591 | -7.109701 | 0.0000 |
| HWAGE | -70.51493 | 9.024624 | -7.813615 | 0.0000 |

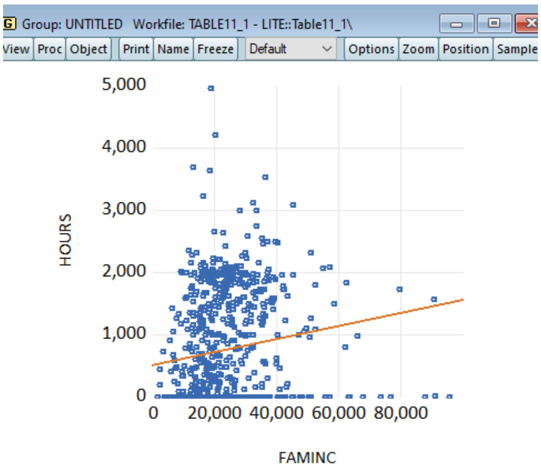
| | | | |
|---|---|---|---|
| R-squared | 0.338537 | Mean dependent var | 740.5764 |
| Adjusted R-squared | 0.332322 | S.D. dependent var | 871.3142 |
| S.E. of regression | 711.9647 | Akaike info criterion | 15.98450 |
| Sum squared resid | 3.78E+08 | Schwarz criterion | 16.03363 |
| Log likelihood | -6010.165 | Hannan-Quinn criter. | 16.00343 |
| F-statistic | 54.47011 | Durbin-Watson stat | 1.482101 |
| Prob(F-statistic) | 0.000000 | | |

The results out of this output are interpreted in the framework of the standard linear regression model. That means each of the slope coefficients above gives the marginal effect of that variable on the mean of the dependent variable, ceteris paribus.

For instance, if husband's wages go up by one-dollar, average hours worked by a married women decline by about 71 hours, ceteris paribus.

All coefficients, except education are statistically significant.

Now a plot, that shows two variables, namely FAMINC and HOURS simultaneously.




In this case, caution should be given towards the fact, that in the sample 325 women had 0 hours of work. In the plot above it is visible that the regression line is pressed down by all the zeros (that are the women with 0 working hours), leading to biased and inconsistent results.

Now the output of data with only women that have working hours:

**Censored data:**

Equation: UNTITLED Workfile: TABLE11_1 - LITE::Table11_1\

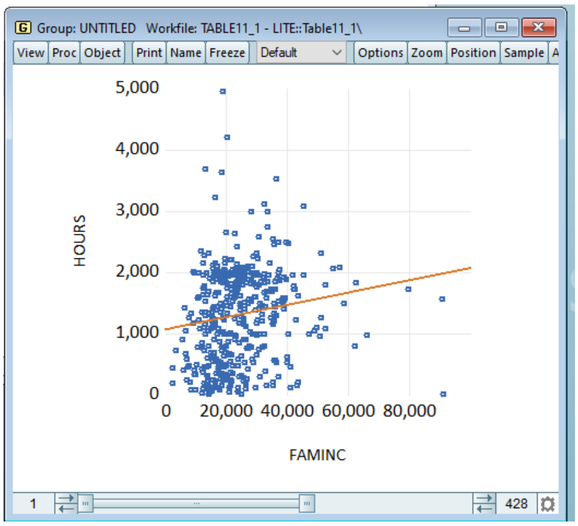
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: HOURS
Method: Least Squares
Date: 02/10/22 Time: 12:07
Sample: 1 753 IF HOURS>0
Included observations: 428

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1817.334 | 296.4489 | 6.130345 | 0.0000 |
| AGE | -16.45594 | 5.365311 | -3.067100 | 0.0023 |
| EDUC | -38.36287 | 16.06725 | -2.387644 | 0.0174 |
| EXPER | 49.48693 | 13.73426 | 3.603174 | 0.0004 |
| EXPERSQ | -0.551013 | 0.416918 | -1.321634 | 0.1870 |
| FAMINC | 0.027386 | 0.003995 | 6.855281 | 0.0000 |
| KIDSL6 | -243.8313 | 92.15717 | -2.645821 | 0.0085 |
| HWAGE | -66.50515 | 12.84196 | -5.178739 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.218815 | Mean dependent var | 1302.930 |
| Adjusted R-squared | 0.205795 | S.D. dependent var | 776.2744 |
| S.E. of regression | 691.8015 | Akaike info criterion | 15.93499 |
| Sum squared resid | 2.01E+08 | Schwarz criterion | 16.01086 |
| Log likelihood | -3402.088 | Hannan-Quinn criter. | 15.96495 |
| F-statistic | 16.80640 | Durbin-Watson stat | 2.107803 |
| Prob(F-statistic) | 0.000000 | | |

We will again plot a relationship between the two variables FAMINC and HOURS simultaneously, but for this case where women have working hours.




If we now compare the results, it is noticeable that the variable education in the regression of women that work now is statistically significant with a negative sign.

Still the results from the first and the second regression are biased and inconsistent, since we have an OLS estimate of censored regression models. In censored regression models the conditional mean of the error term is nonzero and the error is correlated with the regressors, which leads to the fact that the OLS estimators are biased and inconsistent.

If we now compare the plots, we can see that in the figure with the whole sample size, the zeroes (women that don't work) bias the regression down to the zeroes. In the other plot (where we included only women that work) no observation lies on the horizontal axis, since we excluded that data. But the results will again be misleading since we don't use information about the entire population. Therefore, the slope coefficients of the regression lines will be different.

## Q.1.4.8

Equation: UNTITLED Workfile: TABLE11_1 - LITE::Table11_1\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: HOURS
Method: ML - Censored Normal (TOBIT) (Newton-Raphson / Marquardt steps)
Date: 02/10/22 Time: 12:15
Sample: 1 753
Included observations: 753
Left censoring (value) at zero
Convergence achieved after 6 iterations
Coefficient covariance computed using observed Hessian

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | 1126.335 | 379.5851 | 2.967279 | 0.0030 |
| AGE | -54.10977 | 6.621301 | -8.172074 | 0.0000 |
| EDUC | 38.64634 | 20.68458 | 1.868365 | 0.0617 |
| EXPER | 129.8273 | 16.22972 | 7.999356 | 0.0000 |
| EXPERSQ | -1.844762 | 0.509684 | -3.619422 | 0.0003 |
| FAMINC | 0.040769 | 0.005258 | 7.754009 | 0.0000 |
| KIDSL6 | -782.3734 | 103.7509 | -7.540886 | 0.0000 |
| HWAGE | -105.5097 | 15.62926 | -6.750783 | 0.0000 |
| Error Distribution | | | | |
| SCALE:C(9) | 1057.598 | 39.06064 | 27.07579 | 0.0000 |

| | | | |
|---|---|---|---|
| Mean dependent var | 740.5764 | S.D. dependent var | 871.3142 |
| S.E. of regression | 707.2850 | Akaike info criterion | 10.08993 |
| Sum squared resid | 3.72E+08 | Schwarz criterion | 10.14520 |
| Log likelihood | -3789.858 | Hannan-Quinn criter. | 10.11122 |
| Avg. log likelihood | -5.033012 | | |
| Left censored obs | 325 | Right censored obs | 0 |
| Uncensored obs | 428 | Total obs | 753 |

This is a Tobit regression estimated by maximum likelihood. It uses the maximum likelihood method to estimate a model where some observations on the regressand are censored.

# Q.1.4.9

The signs of the different regressors are the same for both OLS-regressions above.

The education variable is only significant in the OLS sample where we disregard all the zeroes (only working women in the sample), but with a negative sign. Instead in the Tobit regression the education variable has a positive sign because the model is suitable for this type of censored data.

The issue is that we cannot interpret the Tobit coefficient of a regressor as giving the marginal impact of that regressor on the mean value of the observed regressand.

The reason for that is, that the Tobit regression models a unit change in the value of a regressor has two effects. First, the effect on the mean value of the observed regressand and second, the effect on the probability that Yi* is observed.

We have for example the coefficient for age, which is about -54 (Tobit regression), ceteris paribus. If age increases by one unit, the hours of work per year will decrease by about 54 hours per year and additionally the probability of a married woman entering the labour force also decreases. We are not able to compute the aggregate impact of an increase in age on the hours worked unless we know the probability to multiply with -54.

The probability Y* must lie between zero and one, meaning the product of each slope coefficient will be smaller than the coefficient itself. Therefore, the sign of the marginal impact will depend on the sign of the slope coefficient, for the probability of observing Yi* is always positive.

Important to note is, that all coefficients are statistically significant at the 10% level of significance.

# Q.1.4.10

We want to determine the influence of R&D, industry category and the two countries on the mean or average number of patents received by the 181 firms.

Dependent Variable: P90
Method: Least Squares
Date: 02/10/22 Time: 12:50
Sample: 1 181
Included observations: 181

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -250.8386 | 55.43486 | -4.524925 | 0.0000 |
| LR90 | 73.17202 | 7.970758 | 9.180058 | 0.0000 |
| AEROSP | -44.16199 | 35.64544 | -1.238924 | 0.2171 |
| CHEMIST | 47.08123 | 26.54182 | 1.773851 | 0.0779 |
| COMPUTER | 33.85645 | 27.76933 | 1.219203 | 0.2244 |
| MACHINES | 34.37942 | 27.81328 | 1.236079 | 0.2181 |
| VEHICLES | -191.7903 | 36.70362 | -5.225378 | 0.0000 |
| JAPAN | 26.23853 | 40.91987 | 0.641217 | 0.5222 |
| US | -76.85387 | 28.64897 | -2.682605 | 0.0080 |

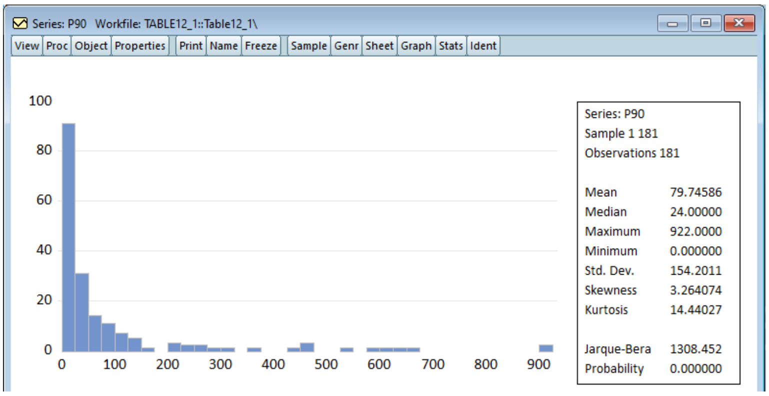
| | | | |
|---|---|---|---|
| R-squared | 0.472911 | Mean dependent var | 79.74586 |
| Adjusted R-squared | 0.448396 | S.D. dependent var | 154.2011 |
| S.E. of regression | 114.5253 | Akaike info criterion | 12.36791 |
| Sum squared resid | 2255959. | Schwarz criterion | 12.52695 |
| Log likelihood | -1110.296 | Hannan-Quinn criter. | 12.43239 |
| F-statistic | 19.29011 | Durbin-Watson stat | 1.946344 |
| Prob(F-statistic) | 0.000000 | | |

The p-value is zero for the variable LR90, indicating a positive relationship between the number of patents received and R&D expenditure.

The R&D variable is in the logarithmic form and the patent variable is in the linear form, so the interpretation would be, if you increase R&D expenditure by 1 %, the average number of patents received will increase by about 0.73, ceteris paribus.

From the other variables only chemistry, vehicles and US are statistically significant.

From interpretation, it can be said that average level of patents granted in the chemistry industry is higher by 47 patents and the average level of patents granted in the vehicle industry is lower by 192. On average US firms received 77 fewer patents than the base group.



Series: P90 Workfile: TABLE12_1::Table12_1\

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr

Descriptive Statistics for P90
Categorized by values of P90
Date: 02/10/22 Time: 12:55
Sample: 1 181
Included observations: 181

| P90 | Mean | Std. Dev. | Obs. |
|---|---|---|---|
| [0, 10) | 2.809524 | 2.728920 | 63 |
| [10, 20) | 13.76190 | 3.048028 | 21 |
| [20, 30) | 24.69231 | 2.897833 | 13 |
| [30, 40) | 35.86667 | 2.559762 | 15 |
| [40, 50) | 43.30000 | 2.406011 | 10 |
| [50, 60) | 52.75000 | 1.281740 | 8 |
| [60, 70) | 65.50000 | 1.914854 | 4 |
| [70, 80) | 75.33333 | 3.669696 | 6 |
| [80, 90) | 85.50000 | 4.041452 | 4 |
| [90, 100) | 95.33333 | 4.041452 | 3 |
| [100, 110) | 105.7500 | 3.774917 | 4 |
| [110, 120) | 111.0000 | NA | 1 |
| [120, 130) | 123.0000 | 0.000000 | 2 |
| [130, 140) | 137.3333 | 0.577350 | 3 |
| [140, 150) | 146.0000 | 1.414214 | 2 |
| [160, 170) | 165.0000 | NA | 1 |
| [200, 210) | 207.0000 | NA | 1 |
| [210, 220) | 213.5000 | 2.121320 | 2 |
| [230, 240) | 235.0000 | NA | 1 |
| [240, 250) | 246.0000 | NA | 1 |
| [250, 260) | 257.0000 | NA | 1 |
| [260, 270) | 260.0000 | NA | 1 |
| [270, 280) | 279.0000 | NA | 1 |
| [310, 320) | 313.0000 | NA | 1 |
| [350, 360) | 353.0000 | NA | 1 |
| [420, 430) | 428.0000 | NA | 1 |
| [450, 460) | 458.0000 | NA | 1 |
| [470, 480) | 470.0000 | 0.000000 | 2 |
| [530, 540) | 533.0000 | NA | 1 |
| [570, 580) | 577.0000 | NA | 1 |
| [620, 630) | 621.0000 | NA | 1 |
| [640, 650) | 640.0000 | NA | 1 |
| [660, 670) | 667.0000 | NA | 1 |
| [900, 910) | 900.0000 | NA | 1 |
| [920, 930) | 922.0000 | NA | 1 |
| All | 79.74586 | 154.2011 | 181 |

Most patents are in the interval of 0 to 30 patents.

The Jarque-Bera test confirms that the null hypothesis of normality can be rejected. But the assumption that the error term is normally distributed cannot hold for a simple LPM when the dependent variable is asymmetric. This indicates that LPM is not a suitable model for this data set since the dependent variable is Poisson distributed. A Poisson regression is more suitable.

Now a Poisson model will be estimated:

Equation: UNTITLED Workfile: TABLE12_1::Table12_1\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: P90
Method: ML/QML - Poisson Count (Newton-Raphson / Marquardt steps)
Date: 02/10/22 Time: 15:42
Sample: 1 181
Included observations: 181
Convergence achieved after 7 iterations
Coefficient covariance computed using the Huber-White method

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.745849 | 0.669183 | -1.114567 | 0.2650 |
| LR90 | 0.865149 | 0.084725 | 10.21132 | 0.0000 |
| AEROSP | -0.796538 | 0.328668 | -2.423531 | 0.0154 |
| CHEMIST | 0.774752 | 0.213134 | 3.635037 | 0.0003 |
| COMPUTER | 0.468894 | 0.263556 | 1.779103 | 0.0752 |
| MACHINES | 0.646383 | 0.390135 | 1.656821 | 0.0976 |
| VEHICLES | -1.505641 | 0.295253 | -5.099486 | 0.0000 |
| JAPAN | -0.003893 | 0.325974 | -0.011944 | 0.9905 |
| US | -0.418938 | 0.241899 | -1.731873 | 0.0833 |

| | | | |
|---|---|---|---|
| R-squared | 0.675516 | Mean dependent var | 79.74586 |
| Adjusted R-squared | 0.660424 | S.D. dependent var | 154.2011 |
| S.E. of regression | 89.85789 | Akaike info criterion | 56.24675 |
| Sum squared resid | 1388804. | Schwarz criterion | 56.40579 |
| Log likelihood | -5081.331 | Hannan-Quinn criter. | 56.31123 |
| Restr. log likelihood | -15822.38 | LR statistic | 21482.10 |
| Avg. log likelihood | -28.07365 | Prob(LR statistic) | 0.000000 |

In nonlinear models the R2 is not meaningful. Instead, the LR (likelihood ratio) is important. Here we have a significant LR statistic of 21282, which suggests that the explanatory variables are collectively important in explaining the conditional mean of patents, which is $\hat{\lambda}_i$ .

This can also be stated by comparing the restricted log-likelihood with the unrestricted log-likelihood function.

Interpretation: The LR90 coefficient (which is expressed in logarithmic form) is about 0.86, indicating that an increase in R&D expenditure by 1%, will increase the average number of patents given a firm by about 0.86%, ceteris paribus.

The dummy variable of the machines is in a semi-log model. The coefficient is 0.6464 and the average number of patents in the machines industry is higher by 90.86% compared to the comparison category. The coefficient of US variable is -0.4189 and the average number of patents is lower by -34.23%.

The Japan variable is statistically not significant.

# Q.1.4.11

The Poisson distribution has the feature that the mean and the variance of a Poisson-distributed variable are the same. This is called equidispersion.

If the variance is higher than the mean there is overdispersion. In this case the solution would be to use the negative binomial regression method instead of the Poisson method.

Equation: UNTITLED Workfile: TABLE12_1::Table12_1\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: (P90-P90F)^2-P90
Method: Least Squares
Date: 02/10/22 Time: 16:41
Sample: 1 181
Included observations: 181

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| P90F^2 | 0.185270 | 0.023545 | 7.868747 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.185812 | Mean dependent var | 7593.203 |
| Adjusted R-squared | 0.185812 | S.D. dependent var | 24801.26 |
| S.E. of regression | 22378.76 | Akaike info criterion | 22.87512 |
| Sum squared resid | 9.01E+10 | Schwarz criterion | 22.89279 |
| Log likelihood | -2069.199 | Hannan-Quinn criter. | 22.88229 |
| Durbin-Watson stat | 1.865256 | | |

In six steps we estimated the dispersion test to see if there is overdispersion or not.

Attention should be paid to the regression coefficient. If this one is positive and statistically significant, there is overdispersion and if it is negative, then there is under-dispersion.

We reject the Poisson model if the coefficient is statistically significant and don't reject it when its statistically insignificant.

As seen above the p-value is 0 and the coefficient is positive we can reject the assumption of equidispersion and conclude that there is overdispersion.

In this case the Poisson regression cannot be used since the standard errors are misleading. Therefore, the negative binomial regression model is more suitable for this type of data set.

# Q.1.4.12

Dependent Variable: P90
Method: QML - Negative Binomial Count (Newton-Raphson / Marquardt steps)
Date: 02/10/22 Time: 16:43
Sample: 1 181
Included observations: 181
QML parameter used in estimation: 1
Convergence achieved after 5 iterations
Coefficient covariance computed using the Huber-White method

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.406826 | 0.641560 | -0.634119 | 0.5260 |
| LR90 | 0.866808 | 0.072283 | 11.99182 | 0.0000 |
| AEROSP | -0.873849 | 0.411650 | -2.122797 | 0.0338 |
| CHEMIST | 0.665741 | 0.204927 | 3.248675 | 0.0012 |
| COMPUTER | -0.129847 | 0.263131 | -0.493470 | 0.6217 |
| MACHINES | 0.011432 | 0.272389 | 0.041970 | 0.9665 |
| VEHICLES | -1.516676 | 0.364282 | -4.163466 | 0.0000 |
| JAPAN | 0.122250 | 0.383324 | 0.318922 | 0.7498 |
| US | -0.689748 | 0.365911 | -1.885018 | 0.0594 |

| | | | |
|---|---|---|---|
| R-squared | 0.442121 | Mean dependent var | 79.74586 |
| Adjusted R-squared | 0.416173 | S.D. dependent var | 154.2011 |
| S.E. of regression | 117.8229 | Akaike info criterion | 9.363858 |
| Sum squared resid | 2387745. | Schwarz criterion | 9.522899 |
| Log likelihood | -838.4291 | Hannan-Quinn criter. | 9.428336 |
| Restr. log likelihood | -974.7010 | LR statistic | 272.5439 |
| Avg. log likelihood | -4.632205 | Prob(LR statistic) | 0.000000 |

The results out of the negative binomial regression model are interpreted in the same way as above for the Poisson model.

## Q.1.4.13

If LR90 increases by 1% the average number of patents increases by ~0.867%