The *RAINDATA* dataset is a subset of Kaggle dataset which contains rainfall data of Australia which was collected by Bureau of Meteorology Australia. This subset includes 5 main cities (Sydney Airport, Melbourne Airport, Mildura, Watsonia & Brisbane), which are our groups for the month of January in 2016. In particular, the data include the following variables:

- **Date:** The date where the weather information was taken, in a YYYY-MM-DD format
- **Location:** The name of the city or location
- **MinTemp:** Minimum temperature in Celsius
- **MaxTemp:** Maximum temperature in Celsius
- **Rainfall:** Total rainfall in mm during the day
- **Evaporation:** The so-called Class A pan evaporation (mm) in the 24 hours to 9am
- **Sunshine:** Number of hours of sunshine during the day
- **WindGustDir:** The direction of the strongest wind gust in the 24 hours to midnight
- **WindGustSpeed:** The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- **WindDir9am:** Direction of the wind at 9am
- **WindDir3pm:** Direction of the wind at 3pm
- **WindSpeed9am:** The speed of the wind at 9am, in km/h
- **WindSpeed3pm:** The speed of the wind at 3pm, in km/h
- **Humidity9am:** The percentage of humidity in the air at 9 am
- **Humidity3pm:** The percentage of humidity in the air at 9 am
- **Pressure9am:** Atmospheric pressure (hpa) reduced to mean sea level at 9am
- **Pressure3pm:** Atmospheric pressure (hpa) reduced to mean sea level at 9am
- **Cloud9am:** Fraction of the sky that is obscured by clouds at 9 am. This is a categorical variable between 0 to 8, where 0 indicates no overcast, clear skies and 8 indicates a sky that is completely obscured
- **Cloud3pm:** Similarly, to Cloud9am, the fraction of the sky obscured at 3pm
- **Temp9am:** Temperature in Celsius at 9am
- **Temp3pm:** Temperature in Celsius at 9am
- **Day:** The date, used as a timestamp, from 1 to 31

This dataset is publicly available. For this project the dataset is downloaded from https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

Q3 a) From this SAS output, we ask you populate the covariance and correlation matrix on the errors for a group (city) for their first 5 measurements by hand, knowing a Compound Symmetry covariance structure was applied on the errors (no group effects or random effects were added).

Valeur estimée du paramètre de covariance					
Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
CS	Location	7.4695	6.8550	1.09	0.2759
Residual		39.5692	5.1683	7.66	<.0001

Solution –

$$Var(Y_{ij}) = \sigma^2 + \sigma_1$$

$$Cov(Y_{ij}, Y_{ik}) = \sigma_1$$

We can see that $\hat{\sigma}_1 = 7.4695$ & $\hat{\sigma}^2 = 39.5692$.

$$Var(Y_{ij}) = 39.5692 + 7.4695 = 47{,}0387$$

$$Cov(Y_{ij}, Y_{ik}) = 7.4695$$

Knowing that we are in presence of a Compound Symmetry structure on our errors,

$$\text{Cov(Ei)} = \begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$$

Thus, we find:

$$\widehat{\text{Cov(Ei)}} = \begin{bmatrix} 47{,}0387 & 7.4695 & 7.4695 & 7.4695 & 7.4695 \\ 7.4695 & 47{,}0387 & 7.4695 & 7.4695 & 7.4695 \\ 7.4695 & 7.4695 & 47{,}0387 & 7.4695 & 7.4695 \\ 7.4695 & 7.4695 & 7.4695 & 47{,}0387 & 7.4695 \\ 7.4695 & 7.4695 & 7.4695 & 7.4695 & 47{,}0387 \end{bmatrix}$$

As for the correlation matrix on the errors within one group we find,

$$\hat{\rho} = \frac{\hat{\sigma}_1}{\hat{\sigma}_1 + \hat{\sigma}^2} = \frac{7.4695}{7.4695+39.5692} = 0{,}15879$$

Knowing that the correlation matrix with a Compound Symmetry applied on our error term gives us,

$$\mathrm{Cor}(\mathrm{Ei}) = \begin{bmatrix} 1 & p & p & p & p \\ p & 1 & p & p & p \\ p & p & 1 & p & p \\ p & p & p & 1 & p \\ p & p & p & p & 1 \end{bmatrix}$$

thus, we find:

$$\widehat{\mathrm{Cor}(\mathrm{Ei})} = \begin{bmatrix} 1 & 0{,}15879 & 0{,}15879 & 0{,}15879 & 0{,}15879 \\ 0{,}15879 & 1 & 0{,}15879 & 0{,}15879 & 0{,}15879 \\ 0{,}15879 & 0{,}15879 & 1 & 0{,}15879 & 0{,}15879 \\ 0{,}15879 & 0{,}15879 & 0{,}15879 & 1 & 0{,}15879 \\ 0{,}15879 & 0{,}15879 & 0{,}15879 & 0{,}15879 & 1 \end{bmatrix}$$

b) Fit a basic linear regression using RainFall as a response variable and MinTemp, MaxTemp, Humidity3pm and WindDir3pm as explanatory variables without using a covariance structure on the errors. Provide the estimates.

Solution –

Solution pour effets fixes

Effet	WindDir3pm	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|	Alpha	Inférieur	Supérieur
Intercept		2.7825	7.4242	118	0.37	0.7085	0.05	-11.9195	17.4845
MinTemp		0.2008	0.2823	118	0.71	0.4782	0.05	-0.3582	0.7598
MaxTemp		-0.1954	0.2237	118	-0.87	0.3842	0.05	-0.6383	0.2476
Evaporation		-0.4825	0.2076	118	-2.32	0.0218	0.05	-0.8936	-0.07135
Humidity3pm		0.1068	0.06052	118	1.77	0.0801	0.05	-0.01301	0.2267
Sunshine		0.03639	0.2004	118	0.18	0.8562	0.05	-0.3605	0.4333
WindDir3pm	E	-0.5594	3.7800	118	-0.15	0.8826	0.05	-8.0448	6.9259
WindDir3pm	ENE	-3.1558	3.4282	118	-0.92	0.3592	0.05	-9.9446	3.6329
WindDir3pm	ESE	6.6340	4.8948	118	1.36	0.1779	0.05	-3.0591	16.3271
WindDir3pm	N	1.4004	4.4641	118	0.31	0.7543	0.05	-7.4397	10.2404
WindDir3pm	NE	-2.7977	3.5275	118	-0.79	0.4293	0.05	-9.7831	4.1876
WindDir3pm	NNE	-2.6506	4.0968	118	-0.65	0.5189	0.05	-10.7634	5.4621
WindDir3pm	NNW	3.0625	4.4932	118	0.68	0.4968	0.05	-5.8352	11.9602
WindDir3pm	NW	5.2369	5.2136	118	1.00	0.3172	0.05	-5.0875	15.5612
WindDir3pm	S	-1.5961	3.4219	118	-0.47	0.6418	0.05	-8.3724	5.1803
WindDir3pm	SE	3.4534	3.6255	118	0.95	0.3428	0.05	-3.7261	10.6329
WindDir3pm	SSE	2.0958	3.3687	118	0.62	0.5351	0.05	-4.5752	8.7668
WindDir3pm	SSW	12.1172	4.0796	118	2.97	0.0036	0.05	4.0385	20.1958
WindDir3pm	SW	-0.4701	3.5447	118	-0.13	0.8947	0.05	-7.4896	6.5495
WindDir3pm	W	3.9535	4.7571	118	0.83	0.4076	0.05	-5.4668	13.3738
WindDir3pm	WNW	8.1949	4.6314	118	1.77	0.0794	0.05	-0.9766	17.3664
WindDir3pm	WSW	0	.	.	.	.	.	.	.

c) Can we use model 3.B to do inference on rainfall in these Australian cities? State if you agree or disagree with this statement, explain why?

Solution –

First, we would have to determine if in fact there is correlation between the groups, which in this case are the cities. The data is indeed correlated because our measurements are taken over time for each city. Therefore, we disagree with this statement, we could not use this regression to estimate rainfall without accounting for correlation between observation. Because we cannot assume independence between our observations, we must alter our model to account for dependence between observations within one city.

d) Adapt model 3.B & apply an AR (1) covariance structure on the errors. Provide the estimated parameters.

Solution –

Solution pour effets fixes

Effet	WindDir3pm	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|	Alpha	Inférieur	Supérieur
Intercept		6.7606	7.1299	4	0.95	0.3967	0.05	-13.0353	26.5564
MinTemp		-0.2490	0.2553	115	-0.98	0.3315	0.05	-0.7547	0.2567
MaxTemp		-0.1570	0.2011	115	-0.78	0.4366	0.05	-0.5552	0.2413
Humidity3pm		0.08859	0.05098	115	1.74	0.0849	0.05	-0.01239	0.1896
WindDir3pm	E	1.4495	3.5345	27	0.41	0.6850	0.05	-5.8027	8.7017
WindDir3pm	ENE	-2.2605	3.2112	27	-0.70	0.4875	0.05	-8.8493	4.3284
WindDir3pm	ESE	6.3365	4.3636	27	1.45	0.1580	0.05	-2.6168	15.2899
WindDir3pm	N	0.2916	4.0550	27	0.07	0.9432	0.05	-8.0286	8.6118
WindDir3pm	NE	-2.8067	3.3532	27	-0.84	0.4099	0.05	-9.6869	4.0735
WindDir3pm	NNE	-2.7914	3.7803	27	-0.74	0.4666	0.05	-10.5480	4.9652
WindDir3pm	NNW	-0.1124	3.6375	27	-0.03	0.9756	0.05	-7.5760	7.3512
WindDir3pm	NW	6.6316	4.5072	27	1.47	0.1528	0.05	-2.6164	15.8795
WindDir3pm	S	-1.1754	3.3177	27	-0.35	0.7259	0.05	-7.9828	5.6320
WindDir3pm	SE	3.0003	3.3329	27	0.90	0.3760	0.05	-3.8382	9.8388
WindDir3pm	SSE	1.0015	3.1827	27	0.31	0.7554	0.05	-5.5288	7.5318
WindDir3pm	SSW	9.8850	3.8045	27	2.60	0.0150	0.05	2.0789	17.6911
WindDir3pm	SW	-0.8942	3.3281	27	-0.27	0.7902	0.05	-7.7229	5.9345
WindDir3pm	W	5.5007	4.0533	27	1.36	0.1860	0.05	-2.8160	13.8174
WindDir3pm	WNW	6.8624	4.2129	27	1.63	0.1150	0.05	-1.7819	15.5066
WindDir3pm	WSW	0	.	.	.	.	.	.	.
day		0.03237	0.1017	115	0.32	0.7508	0.05	-0.1691	0.2338

e) Adapt model 3.B & apply a compound symmetry structure on the errors. Compare models 3.D & 3.E using the AIC or BIC approach, is this method appropriate to compare these two models? Could we formally compare these models using a likelihood ratio test?

Solution -

3.D AR (1):

Tests d'ajustement	
-2 log-vraisemblance restreinte	835.1
AIC (préférer les petites valeurs)	839.1
AICC (préférer les petites valeurs)	839.2
BIC (préférer les petites valeurs)	838.4

3.E CS:

Tests d'ajustement	
-2 log-vraisemblance restreinte	845.5
AIC (préférer les petites valeurs)	849.5
AICC (préférer les petites valeurs)	849.6
BIC (préférer les petites valeurs)	848.7

We can use the AIC or BIC measures to compare these two models as they have the same mean model, they have the same explanatory variables in the mean part of the model. Even if we look at AIC or BIC we obtain the same result where the model we should chose is one with the AR (1) structure on the errors as both AIC (839.1<849.5) and BIC (838.4< 848.7) are smaller for this model then for the model with a Compound Symmetry structure on the errors.

We could not formally test these two models using a likelihood ratio test as neither of these models is the nested version of the other