Question 1

The data is obtained from a clinical trial, where the peak expiratory flow rate (PEFR) was measured repeatedly on children with asthma. This is an example of a repeated measurement study. The aim is to compare PEFR measurements over time between groups receiving placebo and the active treatment, salmeterol xinafoate (SM).

As the dataset is in long format, there are 4 rows of data for every patient, each corresponding to a PEFR measurement every three months for a year. The baseline PEFR of every patient is recorded in a separate column.

Table 1: Baseline characteristics

	Randomised group	
	Placebo (n=73)	**SM (n=75)**
Baseline PEFR (l/min) Mean (SD) to 2 d.p.	289.75 (69.85)	296.62 (70.04)

There were 157 children randomised to either SM or placebo; for reasons that are unknown to us there are only a total of 148 children included in the dataset. Table 1 shows that overall, the number of patients in each randomised group is reasonably balanced. However, PEFR is not very balanced at baseline; it is higher in the active treatment group by an average of 6.87 l/min. There are no missing data for baseline PEFR.

Figure 1: Spaghetti plots showing PEFR over time for 15 randomly sampled patients

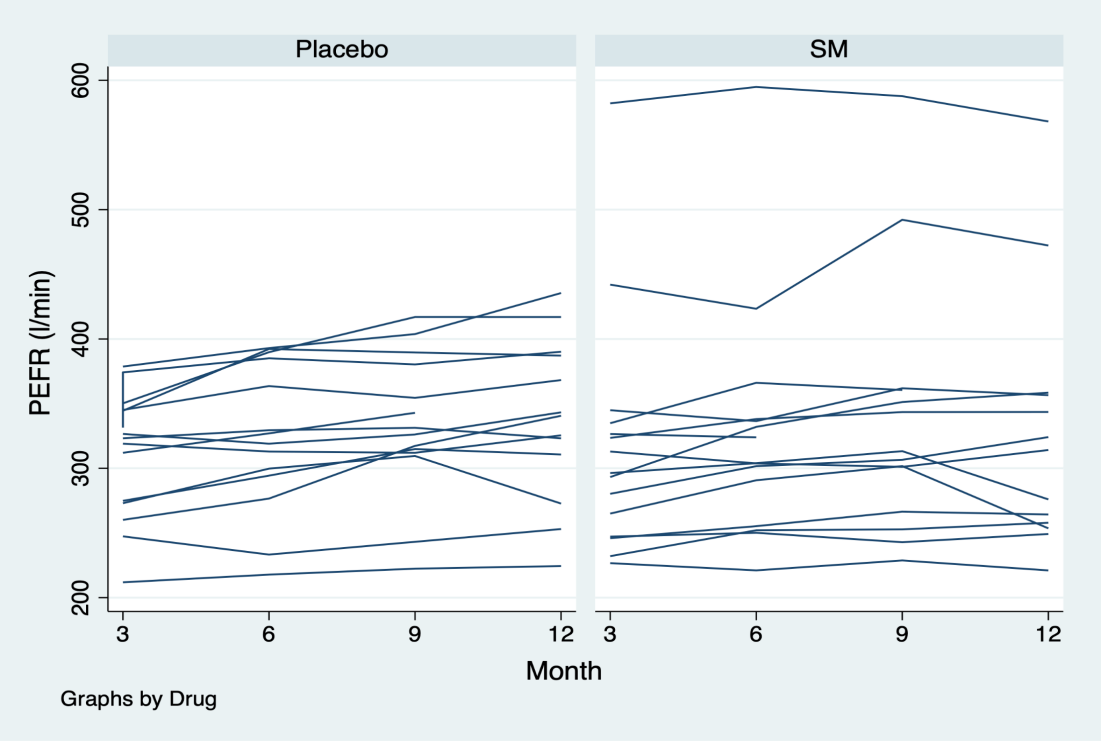


Figure 1 shows a spaghetti plot for 15 randomly sampled patients in each group. Overall, there seems to be a similar trend in PEFR in both groups, with seemingly more variability in the active treatment group. However, it is important to note that this just shows a subset of the full dataset, and cannot infer anything definitively.

Figure 2: Mean PEFR for each treatment arm over time

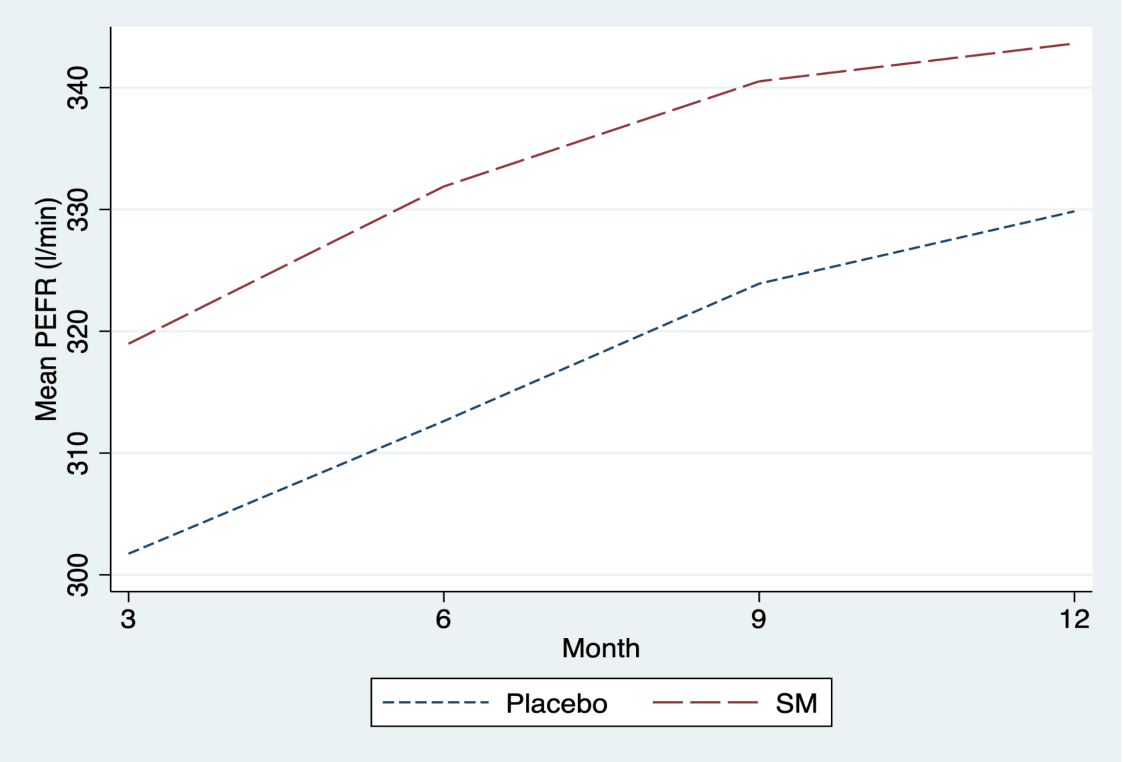


Figure 2 shows the mean PEFR every 3 months for 12 months. This gives us a better idea of the overall trend over time within each group. The lines almost look parallel; the trend is very similar, but PEFR in the treatment group is higher throughout.

Table 2: Number of patients with observed data throughout the duration of the year

Follow-up time (number of months)	**Randomised group**	
	Placebo n (%)	**SM** n (%)
3	73 (100)	75 (100)
6	60 (82.19)	70 (93.33)
9	56 (76.71)	67 (89.33)
12	53 (72.60)	64 (85.33)

Table 2 shows that there are many more people lost to follow-up in the placebo group over the duration of the 12 months, with only 72.6% retained at the end of the 12 months, 12.73% lower than the active treatment group. It is important to be aware of the missing data and differences in the number of measurements obtained per patient, as this may cause problems later during analysis.

Question 2

To account for the correlation between the responses from the same patient, we summarise each patient's multiple measurements by the mean.

Table 3: Summary statistics of mean PEFR

	Mean PEFR over the 12 months (l/min)	
Randomised group	**Mean (SD)**	**Sample size (n)**
Placebo	311.04 (71.42)	73
SM	330.54 (71.24)	75

Table 3 shows that on average, mean PEFR is higher in the active treatment group by 19.5 l/min, a huge difference. The standard deviation is very similar.

Figure 3: Histogram of mean PEFR in each treatment arm

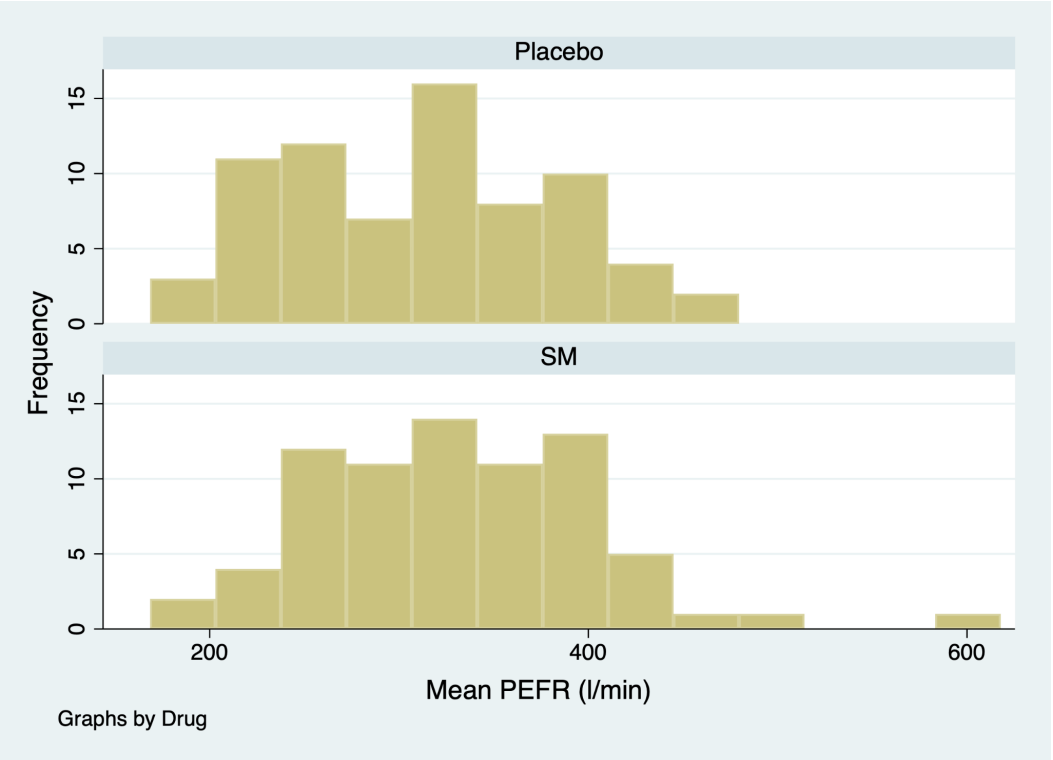


Figure 3 summarises the mean PEFR by randomised group in a histogram. The responses in each treatment arm look *approximately* normally distributed, with a clear outlier in the active treatment group.

The mean PEFR of each patient over time is used as the response variable in regression, adjusting for baseline to account for the imbalance seen in Table 1. A weighted regression is used, weighting by the number of observations per patient, giving more weight to the patients with more observations.

Table 4: Treatment effect estimates (ANCOVA)

Mean PEFR over 12 months (l/min), adjusting for baseline		
Treatment effect	**95% confidence interval**	**p-value**
11.30	1.70 to 20.90	0.02

The ANCOVA analysis suggests that mean PEFR is higher by 11.30 l/min on average for the SM group than those receiving the placebo (adjusting for baseline). The confidence interval does not include 0 and the p-value is significant at the 5% level, providing strong evidence that there is an association between the treatment and mean PEFR over the 12 months.

Figure 4: Model checking plots (ANCOVA model)

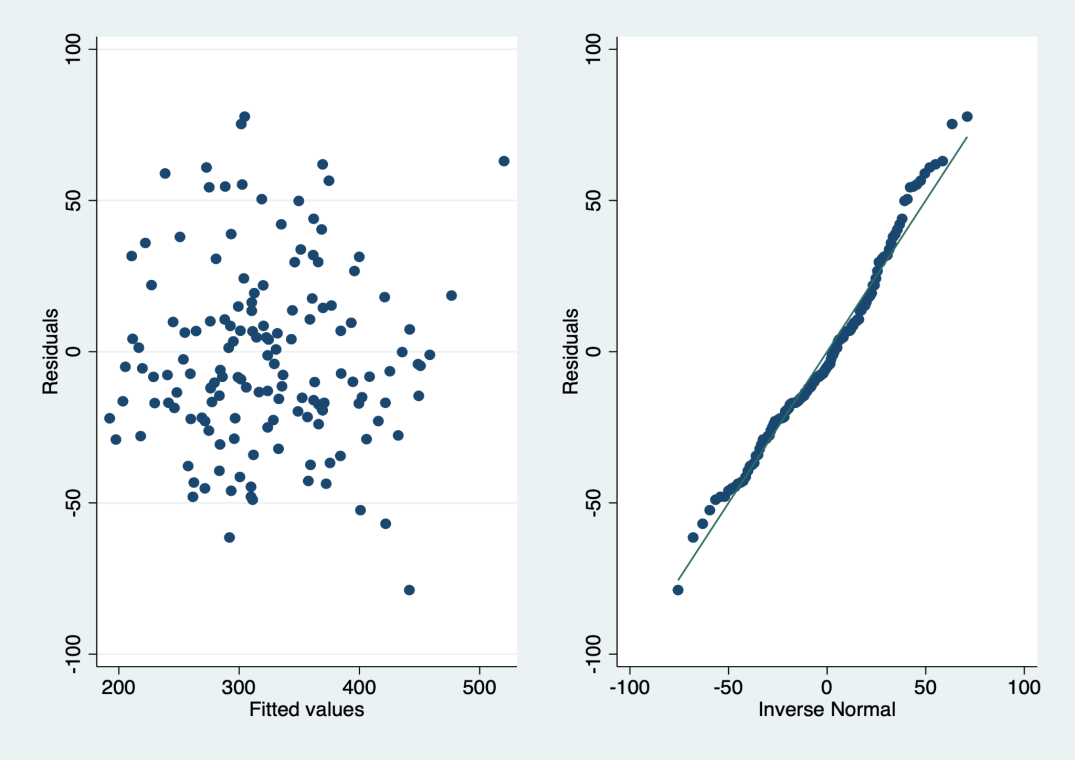


Figure 4 shows that model assumptions seem to hold; variances of the error terms are approximately equal, and the error terms are normally distributed (with very slight deviations towards the tails).

Question 3

We now fit a hierarchical model (more specifically a random intercepts model), including a single covariate for intervention and a simple variance structure, allowing the intercept to vary between patients. This preserves the information given, in contrast to the summary statistics analysis where each patient's measurements were collapsed into a single measure (the mean).

In this case, the random effects are:

- u_j: level 2 residuals (patient-level residuals), with $u_j \sim N(0,\ \sigma_u^2)$
- e_{ij}: level 1 residuals (the error term), with $e_{ij} \sim N(0,\ \sigma_e^2)$

Table 5: Treatment effect estimates (Unadjusted hierarchical model)

PEFR (l/min)		
Treatment effect	**95% confidence interval**	**p-value**
19.41	-3.39 to 42.22	0.10

Table 6: Variance component estimates (Unadjusted hierarchical model)

Variance component	**Estimate**	**95% confidence interval**
σ_u^2	4886.28	3867.77 to 6173.01
σ_e^2	347.68	301.01 to 401.59

As shown in Table 5, the random intercepts model suggests PEFR is higher by 19.41 l/min for a patient randomised to SM compared to a patient with the same random effect u_j on the placebo. Although the estimate is quite high, the confidence interval is wide and includes 0, and the p-value is not significant at the 5% level. The estimate does not seem very precise, bringing some doubt into whether the active treatment actually increases PEFR.

The patient-level variance is bigger than the variance between individual measurements, as shown in Table 6. The ICC is 0.93; a massive 93% of the variability is due to variation between patients.

Figure 5: Model checking plots (Unadjusted hierarchical model)

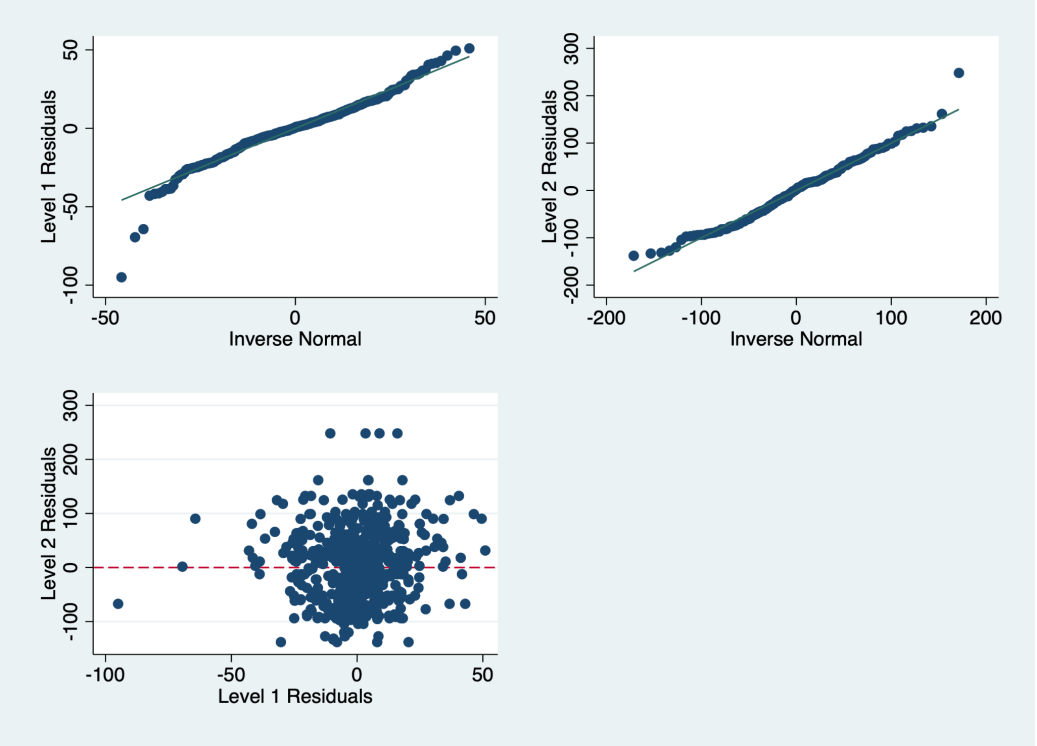


Figure 5 shows that the model assumptions seem to hold. The level 1 and level 2 residuals are approximately normally distributed (although there are *slight* heavier tails for the level 1 residuals), and they are uncorrelated.

The same model is now refitted to adjust for month and baseline PEFR.

Table 7: Treatment effect estimates (Adjusted hierarchical model)

PEFR (l/min)		
Treatment effect	**95% confidence interval**	**p-value**
11.72	2.41 to 21.04	0.01

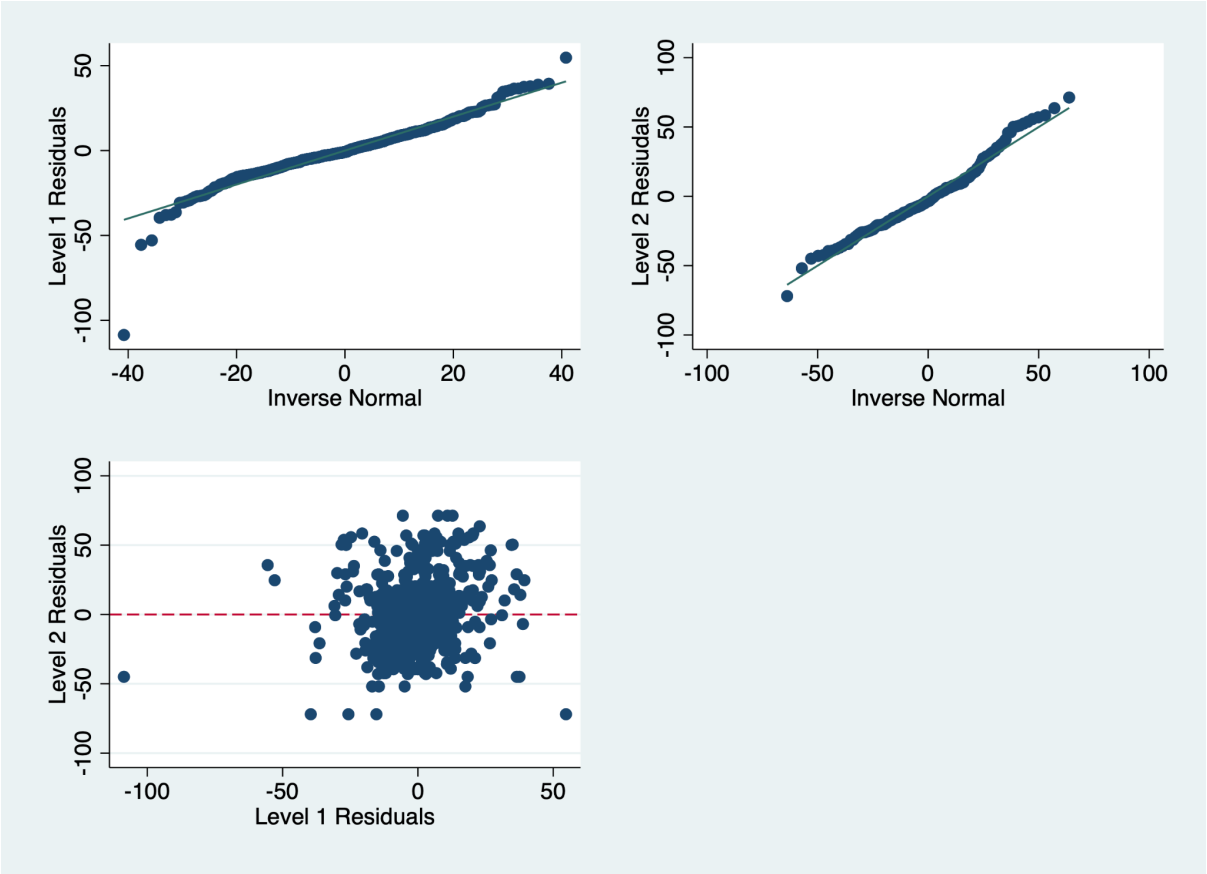
Table 8: Variance component estimates (Adjusted hierarchical model)

Variance component	**Estimate**	**95% confidence interval**
σ_u^2	742.19	575.26 to 957.56
σ_e^2	266.68	230.97 to 307.92

The adjusted hierarchical model suggests average PEFR is higher by 11.72 l/min for a patient randomised to SM compared to a patient with the same random effect u_j on the placebo (Table 7). This estimate is similar to that of the summary statistics analysis (Table 4). However, it is lower than the estimate from the unadjusted hierarchical model (Table 5), which now accounts for baseline PEFR, shown to be higher initially for the SM group in Table 1. Table 7 also shows that the confidence interval is now much narrower and excludes 0, and the p-value is now significant at the 5% level, providing strong evidence that there is a positive association between the treatment and PEFR.

The patient-level variance has also decreased by a massive amount (compare Tables 6 and 8). Some of the unexplained variance has now been explained by including month as a covariate, which was shown to have a positive effect on PEFR over time as shown in Figure 2. The ICC is 0.74; 74% of the variability is due to variation between patients. Although this is still quite high, a strong correlation of about 0.7 within a cluster (i.e. the patient) is expected for a repeated measurement study.

Figure 6: Model checking plots (Adjusted hierarchical model)



Again, model assumptions seem to hold from Figure 6.

Question 4

When investigating the correlation of the residuals within each patient, there was high correlation between measurements that were taken successively, which decreased as the observations were spaced further apart. Therefore, the correlation structure chosen for the marginal model is an autoregressive structure. We also use robust standard errors to correct for possible misspecification of correlation structure. However, robust standard errors are not as effective in the presence of missing data.

The marginal model is adjusted for month and baseline PEFR.

Table 9: Treatment effect estimates (Marginal model with autoregressive corr. structure)

PEFR (l/min)		
Treatment effect	**95% confidence interval**	**p-value**
10.04	0.61 to 19.46	0.04

The marginal model suggests that PEFR is higher by 10.04 l/min on average for the SM group than those receiving the placebo (adjusting for month and baseline PEFR). Comparing the results with the adjusted hierarchical model (Table 8), the estimate is now lower, the confidence interval has shifted towards 0, and the p-value has slightly increased. Nevertheless, the confidence interval does not include 0 and the p-value is significant at the 5% level, providing strong evidence of an association.

The autoregressive correlation structure is the most similar to that found in the residuals. However, the hierarchical models fitted earlier assumes an underlying exchangeable correlation structure, where the correlation between all pairwise observations within a patient is equal. If a marginal model using an exhangable correlation structure is used, the estimates should be very similar.

Table 10: Treatment effect estimates (Marginal model with exchangeable corr. structure)

PEFR (l/min)		
Treatment effect	**95% confidence interval**	**p-value**
11.73	2.55 to 20.92	0.01

This is shown in Table 10, which presents nearly identical results to Table 8.

```
* Set working directory
cd "/Users/rhaecandice/Desktop/Medical Stats 2/ICA"

* start a log file
log using ica.log

* download data
use icadata, clear

*** Question 1

browse // overview of the dataset, already ordered by patient number and month

tabdisp id month, cell(pefr)                // clearly some missing data
bysort drug: tabdisp id month, cell(pefr)   // split by intervention

recode month 1=3 2=6 3=9 4=12       // recode to months 3,6,9 and 12 for graphical purposes
reshape wide pefr, i(id) j(month)   // reshape to wide format
sample 15 if drug== 0, count        // sampling without replacement from each drug group
sample 15 if drug== 1, count
reshape long
sort id month

line pefr month, by(drug) c(L) ytitle("PEFR (l/min)")
    xlabel(3(3)12)  // spaghetti plots of 15 randomly sampled from each group
                    // x axis labels starts from 3, increasing by 3 until 12

* baseline characteristics table
use icadata, clear
reshape wide pefr, i(id) j(month)   // reshape to wide format
tab drug                            // quantity in each randomised group
bysort drug: sum(baseline)          // mean and sd for baseline

reshape long
recode month 1=3 2=6 3=9 4=12       // recode to months 3,6,9 and 12 for graphical purposes
collapse (mean) pefr, by(drug month)// collapse pefr to mean at each time point by group
sort month drug
reshape wide pefr, i(month) j(drug)
lab var pefr0 "Placebo"
lab var pefr1 "SM"
line pefr0 pefr1 month, sort lp(- _) ytitle("Mean PEFR (l/min)") xlabel(3(3)12)
    legend(ring(0) pos(10))     // line graph showing mean pefr for each group

* compare 'obs' with quantity randomised at the start for missing values
use icadata, clear
reshape wide pefr, i(id) j(month)
bysort drug: sum(pefr1)
bysort drug: sum(pefr2)
bysort drug: sum(pefr3)
bysort drug: sum(pefr4)
bysort drug: sum(baseline)


*** Question 2
use icadata, clear

* collapse measurements of each patient to the mean
* form a new column 'n' with how many observations for each patient
collapse (mean) pefr baseline (count) n = month, by(id drug)

sort drug id
browse                              // re-check the data
bysort drug: sum(pefr)              // mean and sd of mean pefr
hist pefr, by(drug, col(1)) freq
    xtitle(Mean PEFR (l/min))       // histogram looks fairly normal

* fit ANCOVA model to summary data weighted by number of observations per patient
regress pefr drug baseline [aw=n]

predict xb                          // put fitted values in variable xb
predict res, res                    // put residuals in variable res
scatter res xb, yline(0, lp("-")) sav(grph1)    // model check: plot residuals vs fitted
qnorm res, sav(grph2)                           // model check: normal plot of residuals
gr combine gph1.gph gra2.gph                    // combine model checking graphs
```

```
75
76
77  *** Question 3
78
79  use icadata, clear
80
81  * Unadjusted hierarchical model
82  mixed pefr drug || id:                          // random intercepts model
83  estat icc                                       // calculate iic
84
85  * Model checking (unadjusted)
86  predict eij, res                                // level 1 residuals
87  predict ui, reffects                            // level 2 residuals
88  list id month pefr eij ui if id < 4, sep(4)     // uj are constant within a patient
89  qnorm eij, ytitle("Level 1 Residuals") sav(ggr1)                // model check: level 1 residuals
90  qnorm ui if month == 1, ytitle("Level 2 Resiudals") sav(ggr2)   // model check: level 2 residuals
91  scatter ui eij, xtitle("Level 1 Residuals") ytitle("Level 2 Residuals")
92      yline(0, lp("-")) sav(ggr3)         // model check: uncorrelated level 1 and 2 residuals
93  gr combine ggr1.gph ggr2.gph ggr3.gph   // combine model checking graphs
94
95  * check auto-correlation (level 1 residuals)
96  keep id month eij
97  reshape wide eij, i(id) j(month)
98  cor eij*
99  gr matrix e* // seems to be fine
100
101 * Adjusted hierarchical model
102 use icadata, clear
103 mixed pefr baseline month drug || id:   // adjusting for month and baseline pefr
104 estat icc                               // calculate icc
105
106 * Model checking (adjusted)
107 predict eij2, res
108 predict ui2, reffects
109 qnorm eij2, ytitle("Level 1 Residuals") sav(g21)
110 qnorm ui2 if month == 1, ytitle("Level 2 Resiudals") sav(g22)
111 scatter ui2 eij2, xtitle("Level 1 Residuals") ytitle("Level 2 Residuals") yline(0, lp("-")) sav(
    g23)
112 gr combine g21.gph g22.gph g23.gph
113
114 * mixed pefr baseline month drug || id: , res (ar 1, t(month))
115 // mixed model with autoregressive correlation structure
116
117
118 *** Question 4
119 use icadata, clear
120
121 * checking correlation between these residuals within each patient
122 regress pefr drug baseline month
123 predict res, res
124 reshape wide res pefr, i(id) j(month)
125 correlate res*      // most resembles autoregressive structure
126
127 use icadata, clear
128 xtset id month                                      // set patient id and time variable (month)
129 xtgee pefr drug month baseline, cor(ar 1) robust    // GEE autoregressive 1, robust SEs
130
131 xtgee pefr drug month baseline, robust // very similar to adjusted hierarchical model
132
133 * close log file
134 log close
135
```