HW 1

STA 2201S Assignment 1. *due Friday, February 10 at the beginning of class*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. DO NOT include in this summary printouts of computer code with the relevant selections highlighted. All computer code used to obtain the results summarized in the response should be provided as an appendix. In this appendix you may highlight the relevant results.

1. Exercise 9.1.1, Davison (p.425)
2. Exercise 9.2.6, Davison (p. 438)
3. Exercise 9.4.3, Davison (p.462)
4. Problem 9.6.7, Davison (p. 466)
5. Exercises 10.3.1 and 10.2.2, Davison (p. 486 and 479)
6. The data file for the "incentives to publish" paper by Franzoni et al.[1] is posted on the course web page. You can read this into R by using the command `read.table(url("http://www.utstat.utoronto.ca/reid/sta2101s/incentives.data"))`; the file `incentives.csv`, as well as Franzoni et al.'s original spreadsheet, are also posted, for those of you who are using SAS or SPSS or ... The original article and supplementary material are also on the web page.

   (a) One of the authors' analyses is a linear regression using the logarithm of the acceptance rate as the response variable, defined as:

   $$accrate_{it} = \log\left(\frac{totpub_{it}+1}{submitted_{it}}\right),$$

   where i indexes countries ($i = 1, \ldots, 30$), t indexes time (either year, or year-1, depending on the variable). Write the mathematical model implied by their Specifications I, II, and III as described in the supporting online material, carefully

Mixed linear model

- $y = X\beta + Zb + \epsilon, \quad b \sim N(0, \Omega_b), \quad \epsilon \sim N(0, \Omega)$
- Example 9.17:
  $y_{ij} = \mu + b_i + \epsilon_{ij}, \quad j = 1, \ldots \qquad ; i = 1, \ldots, q$
- $\Omega = \sigma^2 I_n \qquad \Omega_b = \sigma_b^2 I_q$
- $$X = 1_n, \qquad Z = \begin{pmatrix} 1_{n_1} & 0 & \ldots & 0 \\ 0 & 1_{n_2} & \ldots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1_{n_q} \end{pmatrix}$$
- $\tilde{b}_i = \dfrac{\bar{y}_{i.} - \bar{y}_{..}}{1 + \hat{\sigma}^2/(n_i \hat{\sigma}_b^2)} \qquad \text{var}(\tilde{b}_i) = \dfrac{1}{1/\hat{\sigma}_b^2 + n_i/\hat{\sigma}^2}$

Example 9.18

```
> summary(rat.mixed)
Linear mixed model fit by REML
Formula: y ~ week + (week | rat)
   Data: rat.growth
  AIC  BIC logLik deviance REMLdev
 1097 1115 -542.3     1089    1085
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 rat      (Intercept) 119.532  10.9331
          week         12.495   3.5348  0.184
 Residual              33.842   5.8174
Number of obs: 150, groups: rat, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept) 156.0533     2.1590   72.28
week         43.2667     0.7275   59.47

Correlation of Fixed Effects:
     (Intr)
week 0.007
```

... example 9.18

```
> summary(separate.lm)

Call:
lm(formula = y ~ week + factor(rat) + week:factor(rat), data =

Residuals:
   Min     1Q Median     3Q    Max
-16.80  -3.35   1.00   2.80  13.20

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        155.400      4.506  34.486  < 2e-16 ***
week                42.200      1.840  22.939  < 2e-16 ***
factor(rat)2        -9.800      6.373  -1.538 0.127601
factor(rat)3         5.400      6.373   0.847 0.399036
factor(rat)4         5.600      6.373   0.879 0.381874
factor(rat)5       -17.800      6.373  -2.793 0.006376 **
factor(rat)6         8.200      6.373   1.287 0.201482
factor(rat)7       -10.000      6.373  -1.569 0.120109
```

... example 9.18

```
> options(contrasts = c("contr.sum", "contr.poly"))
> summary(lm(y ~ week + factor(rat)+ week:factor(rat), data =

Call:
lm(formula = y ~ week + factor(rat) + week:factor(rat), data =

Residuals:
   Min     1Q Median     3Q    Max
-16.80  -3.35   1.00   2.80  13.20

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       156.05333    0.82271 189.683  < 2e-16 ***
week               43.26667    0.33587 128.820  < 2e-16 ***
factor(rat)1       -0.65333    4.43041  -0.147 0.883094
factor(rat)2      -10.45333    4.43041  -2.359 0.020464 *
factor(rat)3        4.74667    4.43041   1.071 0.286861
factor(rat)4        4.94667    4.43041   1.117 0.267169
factor(rat)5      -18.45333    4.43041  -4.165 7.12e-05 ***
factor(rat)6        7.54667    4.43041   1.703 0.091948 .
factor(rat)7      -10.65333    4.43041  -2.405 0.018239 *
```

... example 9.18

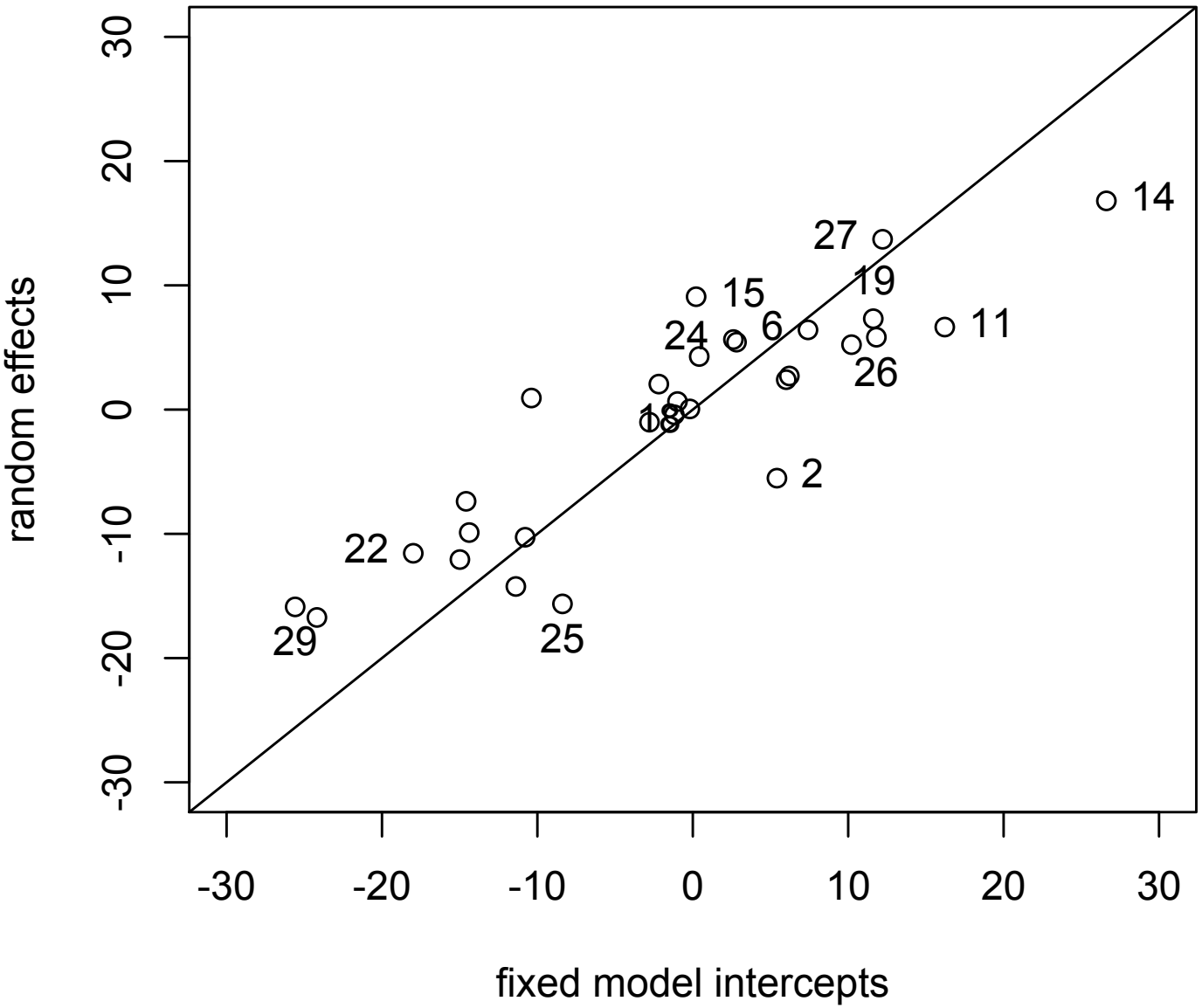
```
>separate.lm2 = lm(y ~ I(week - mean(week)) + factor(rat) + I(

Coefficients:
                                  Estimate Std. Error
(Intercept)                     242.58667    0.47499
I(week - mean(week))             43.26667    0.33587
factor(rat)1                     -2.78667    2.55790
factor(rat)2                      5.41333    2.55790
factor(rat)3                     10.21333    2.55790
factor(rat)4                    -10.38667    2.55790

> sum(coef(separate.lm2)[3:31])
[1] 1.186667
> fixef = c(coef(separate.lm2)[3:31], -1.186667)

> plot(fixef, ranef(rat.mixed)$rat[,1], xlim = c(-30, 30),
+ ylim = c(-30, 30) , xlab = "fixed model intercepts",
+ ylab = "random effects")
> identify()
```

... example 9.18


random effects
fixed model intercepts
30
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30
14
27
19
15
11
24
6
26
2
22
29
25

Example 9.15

```
> anova(lm(log(angle) ~ replicate*recipe*temperature,
+ data = cake))
Analysis of Variance Table

Response: log(angle)
                              Df Sum Sq Mean Sq F value
replicate                     14 8.1594 0.58282
recipe                         2 0.1862 0.09308
temperature                    5 2.0509 0.41018
replicate:recipe              28 1.3427 0.04796
replicate:temperature         70 1.3448 0.01921
recipe:temperature            10 0.1757 0.01757
replicate:recipe:temperature 140 2.6949 0.01925
Residuals                      0 0.0000
```

... example 9.15

```
> anova(lm(log(angle) ~ replicate + recipe + replicate:recipe
+ + temperature + recipe:temperature, data = cake))
Analysis of Variance Table

Response: log(angle)
                    Df Sum Sq Mean Sq F value    Pr(>F)
replicate           14 8.1594 0.58282 30.2975 < 2.2e-16 ***
recipe               2 0.1862 0.09308  4.8388 0.0088214 **
temperature          5 2.0509 0.41018 21.3233 < 2.2e-16 ***
replicate:recipe    28 1.3427 0.04796  2.4929 0.0001278 ***
recipe:temperature  10 0.1757 0.01757  0.9132 0.5218715
Residuals          210 4.0397 0.01924
```

... example 9.15

```
> cake.aov = aov(log(angle) ~ temperature*recipe+
+ Error(replicate/recipe), data = cake)
> summary(cake.aov)

Error: replicate
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 14 8.1594 0.58282

Error: replicate:recipe
          Df  Sum Sq  Mean Sq F value Pr(>F)
recipe     2 0.18616 0.093081   1.941 0.1624
Residuals 28 1.34274 0.047955

Error: Within
                    Df Sum Sq Mean Sq F value Pr(>F)
temperature          5 2.0509 0.41018 21.3233 <2e-16 ***
temperature:recipe  10 0.1757 0.01757  0.9132 0.5219
Residuals          210 4.0397 0.01924
```

... example 9.15

```
> summary(cake.aov,
+ split = list(temperature = list(L=1, Q=2, dev=3:5)))

  ...

Error: Within
                              Df Sum Sq Mean Sq  F value Pr(>F)
temperature                    5 2.0509 0.41018  21.3233 <2e-16
  temperature: L               1 1.9253 1.92527 100.0845 <2e-16
  temperature: Q               1 0.0207 0.02074   1.0783 0.3003
  temperature: dev             3 0.1049 0.03497   1.8178 0.1450
temperature:recipe            10 0.1757 0.01757   0.9132 0.5219
  temperature:recipe: L        2 0.0038 0.00192   0.0998 0.9051
  temperature:recipe: Q        2 0.0082 0.00409   0.2128 0.8085
  temperature:recipe: dev      6 0.1636 0.02727   1.4179 0.2090
Residuals                    210 4.0397 0.01924
```

Reference: Modern Applied Statistics with S, §10.2

Design and Analysis of Experiments (Montgomery), §12.2

Example 9.15

```
> model.tables(cake.aov, type="means", se = T)
temperature:recipe
           recipe
temperature A     B     C
        175 3.350 3.270 3.293
        185 3.433 3.355 3.331
        195 3.409 3.428 3.428
        205 3.493 3.443 3.405
        215 3.638 3.505 3.516
        225 3.535 3.537 3.538
Warning message:
In model.tables.aovlist(cake.aov, type = "means", se = T) :
  SEs for type 'means' are not yet implemented
```

See Table 9.26

$$y_{rmt} = \mu + \beta_r + \gamma_m + u_{rm} + \xi_t + \tau_{rt} + \epsilon_{rmt}, \quad u_{rm} \sim N(0, \sigma_u^2), \epsilon_{rmt} \sim N(0, \sigma^2)$$

Principles (C&D, §7.2 "Non-specific effects")

- "aspects of the system under study that may well correspond to systematic differences in the variables being studies, but which are of no, or limited, direct concern"
- examples: clinical trial carried out at several centres; agricultural field trials at a number of different farms; sociological study in a number of different countries; laboratory experiments with different sets of apparatus
- "it may be necessary to take account of such features in one of two different ways..."

C&D, §7.2.2 "Stable treatment effect"

- model:

$$E(Y_{tci}) = \alpha_c + x_{ci}^T \beta + \delta_t$$

- no treatment / centre interaction
- should α_c be ?fixed? or ?random?
- "effective use of a random-effects representation will require estimation of the variance component corresponding to the centre effects"
- "even under the most favourable conditions the precision achieved in that estimate will be at best that from estimating a single variance from a sample of a size equal to the number of centres"
- "... very fragile unless there are at least, say, 10 centres and preferably considerably more"

... C&D, §7.2.2 "Stable treatment effect"

- "if centres are chosen by an effectively random procedure from a large population of candidates, ... the random-effects representation has an attractive tangible interpretation. This would not apply, for example, to the countries of the EU in a social survey."
- some general considerations in linear mixed models:
  - in balanced factorial designs, the analysis of treatment means is unchanged
  - in other cases, estimated effects will typically be 'shrunk', and precision improved
  - "representation of the nonspecific effects as random effects involves independence assumptions which certainly need consideration and may need some empirical check"

... C& D, §7.2.3 “Unstable treatment effect”

- “ if there is an interaction between an explanatory variable [e.g. treatment] and a nonspecific variable”
- i.e. the effects of the explanatory variable change with different levels of the nonspecific factor
- “the first step should be to explain this interaction, for example by transforming the scale on which the response variable is measure or by introducing a new explanatory variable”
- example: two medical treatments compared at a number of centres show different treatment effects, as measured by an ratio of event rates
- possible explanation: the difference of the event rates might be stable across centres
- possible explanation: the ratio depends on some characteristic of the patient population, e.g. socio-economic status
- “an important special application of random-effect models for interactions is in connection with overviews, that is, assembling of information from different studies of essentially the same effect”

Design of Studies (C&D, Ch.2)

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

... design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed
- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)
- example: recent *Science* article, and letters, on the benefits (or not) of single sex schools
- in some areas new investigations can be set up and completed relatively quickly; design of individual studies may then be less important

... design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context
- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives
- latter will require confirmatory studies

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...
- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible
- the unit of analysis may not be the unit of interpretation – ecological bias
- on the whole, limited detail is needed in examining the variation within the unit of study

Types of investigations

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual
- experiment: investigator has complete control over treatment assignment
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic