STATISTICAL MODELLING Part IIC
R revision sheet RDS/Lent 2015
NB this sheet is not a substitute for doing the practical sheets!

Basics of vectors and matrices

- Vectors: how to create them and extract components and sub-vectors. Operations such as `^`, `*`, `/`, `mean`, `length`, `sum`, `exp`, `log`.

```
> x <- 1:5
> x
[1] 1 2 3 4 5
> x*x
[1]  1  4  9 16 25
> x^3
[1]   1   8  27  64 125
> sum(x)
[1] 15
> mean(x)
[1] 3
> x <- c(1, 4, 2.8, 3)
> x
[1] 1.0 4.0 2.8 3.0
> x >= 3
[1] FALSE  TRUE FALSE  TRUE
> x[x >= 3]
[1] 4 3
```

- Similar knowledge of matrices in R.

```
> A <- matrix(1:6, 3, 2)
> A
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> A[1, ]
[1] 1 4
> A*A   # pointwise product
> t(A)   # transpose
> t(A) %*% A   # matrix multiplication
> solve(t(A) %*% A)   # matrix inverse
```

Random number generation

`rnorm`, `rchisq`, `rt`, `rf` (normal, χ^2, t, F). Distribution and quantile function are given by `pnorm`, `qnorm` etc.

```
> qchisq(0.05, df = 4, lower.tail = FALSE)
```

—the *upper* 5%-point of χ^2_4 distribution;

```
> pt(1.96, df = 10)
```

—$\mathbb{P}(T \leq 1.96)$ when $T \sim t_{10}$.

Factors

See practical 5 for more details, particularly the grouping of levels of factors. See also the `gl` ("generate levels") function.

Fitting models

This is the most important section. You should know...

- How to supply a model formula to `lm` or `glm` (possibly with interaction terms) (see `?formula`). Recall that an intercept term is included by default and corner point constraints are enforced when factors are present.
- How to interpret the diagnostic plots produced by `plot` applied to an `lm` object (see `plot.lm`).
- Understand the output of `summary` applied to fitted models (see `summary.lm`, `summary.glm`), and interpret the deviance of a GLM obtained by `mod$dev` where `mod` is a fitted `glm` object
- Understand the output of `anova` applied to a pair of nested models (see `anova.lm`, `anova.glm`).
- Understand `predict` applied to fitted models. See `predict.lm`—pay attention to the `object`, `newdata`, `se.fit` and `interval` arguments; `predict.glm`—pay attention to the `object`, `newdata`, `se.fit` and `type` arguments (for the `link` and `response` options).
- What `model.matrix` applied to a fitted model returns.
- Similarly the roles of `coef`, `residuals`, `fitted.values`, `hatvalues` (leverage), `cooks.distance` applied to `lm` objects.

Things you don't need to know for the exam

The `paste` function, `sapply`, the intricacies of producing plots in R, how to write your own function, `ellipse`, `rstudent`, `stepAIC`, ...

Past papers

For doing some of the past papers (depending on how far back you go), you'll need to know that `dbinom` and `dpois` give the probability mass functions of the binomial and Poisson distributions respectively. The `log = TRUE` option for these gives the log of these pmf's. Also, if `fit1` and `fit2` are `glm` objects, then

```
anova(fit, fit, test = "Chisq")
```

and

```
anova(fit, fit, test = "LR")
```

are the same.