STATISTICAL MODELLING Part IIC

Practical 4: More on linear models IAC/Lent 2011

Comments and corrections to ioana@statslab.cam.ac.uk

Open R and return to the data from the Welding Institute that we studied last time. Remember that we fit two linear models and used an ANOVA test to conclude a preference for the one that included a quadratic term. Now suppose that we want to predict the minimum diameter of the weld at 8 Amps and at 9 Amps. We have to supply new vectors for all explanatory variables used by the model; this can be done using a data frame. A data frame is a list of variables of the same length, but possibly of different types (numeric, character or logical); the rows and columns of a data frame can be named.

```
> NewPoints <- data.frame(w = c(8, 9) - mean(x))
> predict(LinMod2, NewPoints, se.fit = TRUE, interval = "prediction",
+     level = 0.95)
```

```
$fit
       fit      lwr      upr
1 3.725277 3.356960 4.093593
2 5.582471 5.218793 5.946150

$se.fit
         1          2
0.06093938 0.05426342

$df
[1] 18

$residual.scale
[1] 0.1643794
```

It is not a good idea to extrapolate models far outside the range of the data, e.g. to predict the minimum diameter of the resulting weld at 15 Amps.

Exercise: See if you can verify the calculations leading to the predicted values and prediction intervals. (You can skip the standard error of the predicted means.) Use the function qt for quantiles of the t-distribution.

The next data set is `mammals`, which is contained in the library of data sets that accompany the book *Modern Applied Statistics with S-PLUS* (MASS) by Venables and Ripley. It gives the body and brain weights (in kilograms and grams respectively) of 62 different species of land mammals. For details, try `?mammals`. Use

```
> library(MASS)
> dim(mammals)
```

```
[1] 62  2
```

```
> names(mammals)
```

```
[1] "body"  "brain"
```

```
> Species <- row.names(mammals)
> Species
```

```
 [1] "Arctic fox"                "Owl monkey"
 [3] "Mountain beaver"           "Cow"
 [5] "Grey wolf"                 "Goat"
 [7] "Roe deer"                  "Guinea pig"
 [9] "Verbet"                    "Chinchilla"
[11] "Ground squirrel"           "Arctic ground squirrel"
[13] "African giant pouched rat" "Lesser short-tailed shrew"
[15] "Star-nosed mole"           "Nine-banded armadillo"
[17] "Tree hyrax"                "N.A. opossum"
[19] "Asian elephant"            "Big brown bat"
[21] "Donkey"                    "Horse"
[23] "European hedgehog"         "Patas monkey"
[25] "Cat"                       "Galago"
[27] "Genet"                     "Giraffe"
[29] "Gorilla"                   "Grey seal"
[31] "Rock hyrax-a"              "Human"
[33] "African elephant"          "Water opossum"
[35] "Rhesus monkey"             "Kangaroo"
[37] "Yellow-bellied marmot"     "Golden hamster"
[39] "Mouse"                     "Little brown bat"
```

```
[41] "Slow loris"                "Okapi"
[43] "Rabbit"                    "Sheep"
[45] "Jaguar"                    "Chimpanzee"
[47] "Baboon"                    "Desert hedgehog"
[49] "Giant armadillo"           "Rock hyrax-b"
[51] "Raccoon"                   "Rat"
[53] "E. American mole"          "Mole rat"
[55] "Musk shrew"                "Pig"
[57] "Echidna"                   "Brazilian tapir"
[59] "Tenrec"                    "Phalanger"
[61] "Tree shrew"                "Red fox"
```

You can see the other datasets available by

```
> data()
```

Notice that `mammals` is a data frame and the column numbers do not count as a dimension. “Attaching” the column headings of the data frame to the R search path means that the corresponding columns can be accessed by simply giving their names. The opposite of `attach()` is `detach()`.

```
> attach(mammals, warn.conflicts = FALSE)
> x <- body
> y <- brain
```

Figure 1 plots brain weight as a function of body weight (*left*). Can you see any potential problems with fitting a regression line through this data? Would it be easier to fit a line though the log–log version on the *right*?

Exercise: Re-plot the original pairs and use the `identify()` function to identify each species. Try

```
> plot(x,y)
> identify(x,y,Species)
```

and left–click on the points to identify; right–click to quit. Do the same for the log–log version. Terminating the process returns the indeces of the observations identified.

Let’s have a look at two different linear models. Try the following in your R session.

```
> par(mfrow = c(1, 2))
> plot(x, y, main = "brain = f(body)", xlab = "body weight (kg)",
+     ylab = "brain weight (g)")
> plot(log(x), log(y), main = "log(brain) = f(log(body))",
+     xlab = "log(body weight (kg))", ylab = "log(brain weight (g))")
```

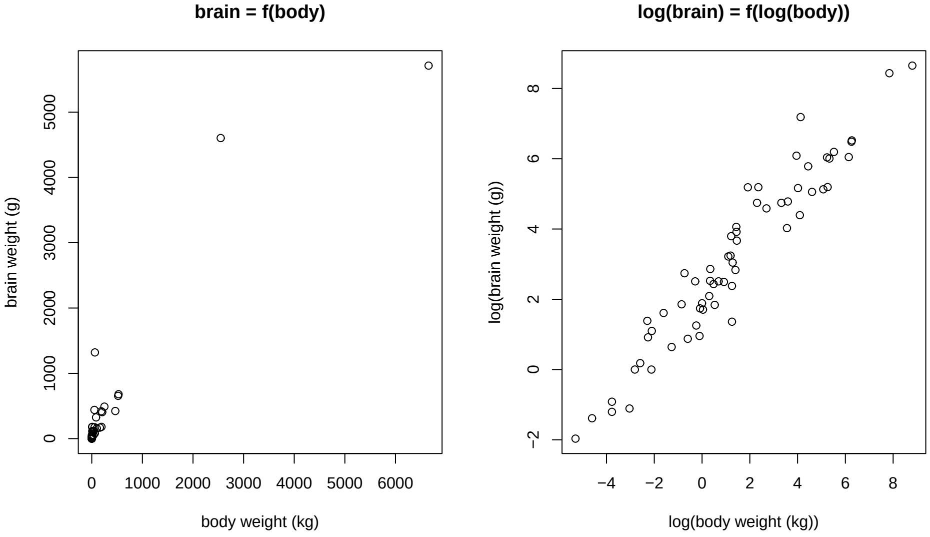

Figure 1: Brain weight plotted as a function of body weight for the mammals data; and log–log plot.

```
> Species.lm <- lm(y ~ x)
> summary(Species.lm)
> Species.log.lm <- lm(log(y) ~ log(x))
> summary(Species.log.lm)
```

Examine the residual plots with the following R code.

```
> plot(Species.lm)
> plot(Species.log.lm)
```

How can you use the `par()` function to view all 8 of the plots produced by the above two R calls?

Exercise: Explain the two models being fit in the above R code. Verify the calculations which give the multiple R^2 and adjusted multiple R^2 values for the second model. Might the influence of any of the points be of concern to us? Use the function `cooks.distance` and compare against the value $8/(n-2p)$.

Here is a more direct way to compare the two models graphically: (Could we compare these models with an ANOVA table?)

```
> par(mfrow = c(2, 2))
> plot(x, y, main = "brain = f(body)", xlab = "body weight (kg)",
+     ylab = "brain weight (g)")
> abline(Species.lm)
> qqnorm(rstudent(Species.lm))
> qqline(rstudent(Species.lm))
> plot(log(x), log(y), main = "log(brain) = f(log(body))",
+     xlab = "log(body weight (kg)", ylab = "log(brain weight (g)")
> abline(Species.log.lm)
> qqnorm(rstudent(Species.log.lm))
> qqline(rstudent(Species.log.lm))
```

Figure 2 shows the resulting plots of the data, fits, and residuals.

Exercise: Re-make the `qqnorm()` plot above, add the `qqline()` and then use the `identify()` function to identify any possible outliers. Do this for the linear model fit to the original data, and to the log–log transformed data. *Hint: save the output of the* `qqnorm()` *function call, and provide this as input to the* `identify()` *function.*

Exercise: Use your preferred model to compute 95% confidence sets for the slope term and for the pair consisting of the intercept and slope terms. Construct an estimate $\hat{Y}$ of the brain weight Y^* of a new mammal whose body weight is 30kg, and give a 95% prediction interval. Is it the case that $\mathbb{E}(\hat{Y}) = \mathbb{E}(Y^*)$? Comment.

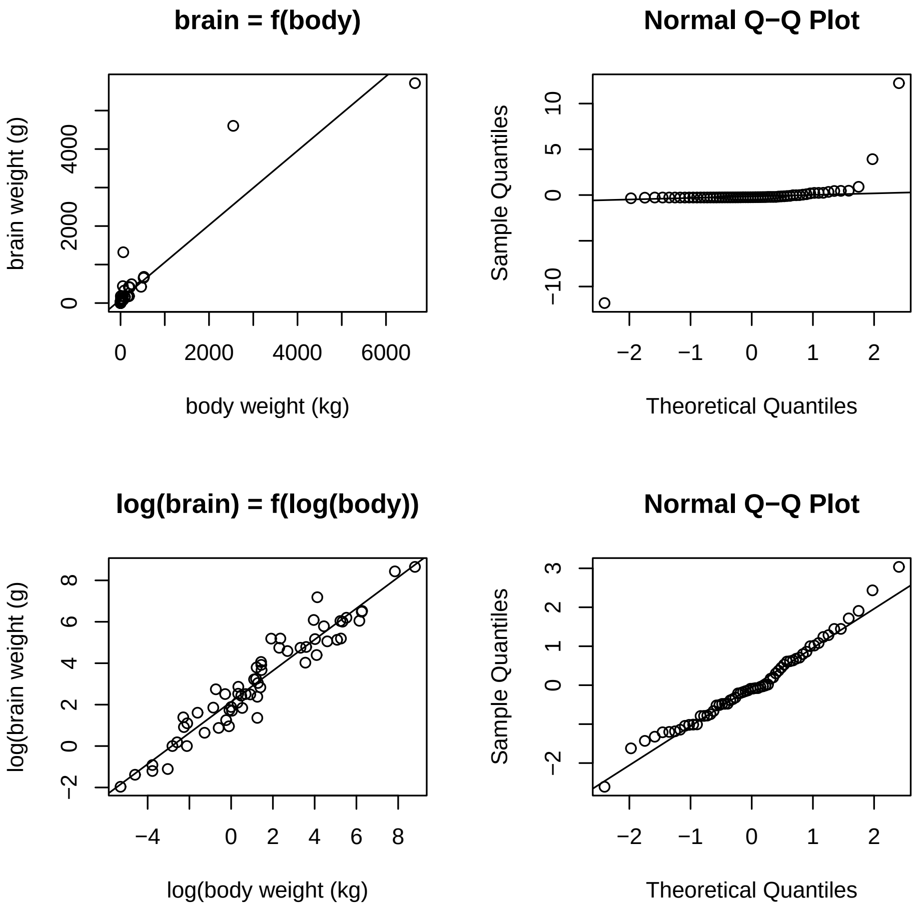

Figure 2: Linear model fit to mammal data, and log–log transformed data.