This is a function fitting problem as it arose in marine engineering. The Hydrodynamics Lab in Delft Netherlands made 22 hull models similar to the commercial Standfast 43 yacht, and, by measured them in their towing tank, determined each hull's drag resistance at 14 different speeds. The published results record 6 measures of hull shape and the resulting residual resistance.

- `CoB`: longitudinal position of the center of buoyancy (i.e., the center of gravity of the displaced water) as a percentage of the (water line) length of the hull. Ranges from 0% to –5%.
- `P.C`: prismatic coefficient—the ratio of the displaced water volume to a maximal volume using the maximal cross-sectional of the displaced water times the (water line) length. Ranges form .53 to .6
- `L.D`: length-displacement ratio—the ratio of the length to the cube root of the displaced water volume. Ranges form 4.34 to 5.14
- `B.D`: beam-draught ratio—beam is the maximal width of hull at the water line; draught (a.k.a. draft) is the depth of the displaced water. Draught does not include the fin keel or rudder. Ranges 2.73 to 5.35
- `L.B`: length-beam ratio. Ranges 2.73 to 3.64
- `Fr`: Froude number—ratio of boat speed to $\sqrt{gL}$. $Fr^2 \propto$ the ratio of wavelength of the boat wave to the boat length. If wavelength is about twice boat length (which corresponds to $Fr \sim 0.5$) the boat is always pushing up from the wave trough traveling with it. The result is large drag. If wavelength is much shorter than boat length ($Fr \sim 0$), the wave effect is averaged out. Resistance is a strong function of $Fr$. Each model was tested at 14 different Froude numbers ranging from .125 to .450.
- `RR`: residuary resistance per unit weight of displacement— the skin friction has been subtracted, hence 'residuary'.

1. `read.csv` the file `yacht_hydrodynamics.csv`; make a summary of the data.frame. One of the above ranges is incorrect; report the incorrect and correct value.
2. The most brain-dead way to analyze such data is to use a model where the dependent variable `RR` is a linear function of all the independent variables. This is easily done in R:

   ```
   lin1=lm(RR~.,data=df)
   summary(lin1)
   AIC(lin1)
   ```

   Now and in all the following models record the 'Residual standard error', 'Adjusted R-squared', and AIC (Akaike's Information Criterion). All of the $p$-values you'll see are quite small (which says nothing about the model being useful or correct). Report any/all significant variables. Examine the `str` of this model; see all the stuff that is crammed into 'columns' of `lin1`. Calculate `sqrt(sum(lin1$residuals^2)/301)`; see that it is the same as 'Residual standard error'. (I hope this clearly tells you what 'Residual standard error' is.)
3. Since `RR` has a range of variation greater than 10, the next nearly brain-dead way to analyze such data is to use a model where the log of the dependent variable (`RR`) is a linear function of all the independent variables. This is easily done in R:

   ```
   log1=lm(log(RR)~.,data=df)
   summary(log1)
   AIC(log1)
   ```

Note 'Residual standard error' is much reduced but that is largely because $\log(RR)$ is smaller than $RR$. We can make a fair comparison: the `log1` model gives $\log(RR)$; undo that with exp, and form the 'Residual standard error' directly. Or you could try to do the reverse and take the log of the `lin1$fitted.values` can calculate the difference from `log(df$RR)`:

```
sqrt(sum((exp(log1$fitted.values)-df$RR)^2)/301)
sqrt(sum((log(lin1$fitted.values/df$RR)^2))/301)
```

You will find that the second fails; take a look at `lin1$fitted.values[1:4]` and explain the cause of the failure.

The log model does beat the linear model.

4. `plot(df$Fr,df$RR)` and see highly non-linear behavior. A log-log plot: `plot(df$Fr,df$RR,log="xy")` looks almost linear. The physicists should know that a linear log-log plot suggests a power law ($y = Ax^B$) relationship between the variables. While the log-log plot is not exactly linear, I think it's of interest to see what value of $B$ would be a good fit.

   ```
   log2=lm(log(RR)~log(Fr),data=df)
   summary(log2)
   AIC(log2)
   plot(log(df$Fr),log(df$RR))
   abline(a=7.25812,b=4.69046)
   ```

5. On the log-log plot the effect of the hull-shape variation seems larger at small $Fr$, whereas on a normal (linear-linear) plot the variations seem large at large $Fr$. This is essentially saying that the percent errors are large at small $Fr$ whereas the absolute errors are large at large $Fr$.

   ```
   plot(df$Fr,df$RR)
   curve(exp(7.25812+4.69046*log(x)),.125,.45,add=T)
   ```

   Which type of variations are important to the consumer of yachts may depend on the use: racers in low wind conditions may well be more concerned about the absolutely small variations in $RR$ at small $Fr$. The 'best' model may not be the one with the smallest 'residual standard error' (and which error: from log or linear cases?).

6. There is a bit of a curve in the log-log plot; let's let there be a quadratic term and do the brain-dead thing of adding everything else. (The range of variation of the other quantities is small, hence the choice not to log the other variables. Note that trying to log `CoB` would result in errors as `CoB` includes the value zero.)

   ```
   log3=lm(log(RR)~poly(log(Fr),2)+.-Fr,data=df)
   ```

   `-Fr` removes `Fr` which is already in the `poly` and hence need not be re-included by the `+.`. No harm is done if you omit this detail.

7. `plot(log(df$Fr),log3$residuals)` shows a bit of a curve, so let's add another term. (Worried yet?)

   ```
   log4=lm(log(RR)~poly(log(Fr),3)+.-Fr,data=df)
   ```

   AIC and residual error have been reduced, but we have a bunch of non-significant terms.

8. `log5=lm(log(RR)~poly(log(Fr),3)+CoB,data=df)`

   Worse without non-significant terms. Let's bring some back.

9. `log6=lm(log(RR)~poly(log(Fr),3)+CoB+L.D,data=df)`

Marginal improvement with all almost significant terms. Try more:

10. `log7=lm(log(RR)~poly(log(Fr),3)+CoB+L.D+L.B,data=df)`

    Wow! a big change in $Pr(>|t|)$-values, nearly the same error as the all-in case.

11. `log8=lm(log(RR)~poly(log(Fr),3)+CoB+L.D+L.B+P.C,data=df)`

    `log9=lm(log(RR)~poly(log(Fr),3)+CoB+L.D+L.B+B.D,data=df)`

    Much the same result including either `B.D` or `P.C`, but adding both reduces the significance of all.

12. Hardcopy the below plots and comment on the results.

    ```
    plot(df$Fr,df$RR-exp(log8$fitted.values))
    plot(df$Fr,log8$residuals)
    sqrt(sum((exp(log8$fitted.values)-df$RR)^2)/299)
    ```

13. But maybe this is better

    ```
    lin2=lm(RR ~ poly(Fr,4) + CoB*Fr + P.C*Fr, data =df)
    plot(df$Fr,lin2$residuals)
    ```

    The terms like `CoB*Fr` are called interaction terms and are discussed ch. 3, p. 87. They don't work exactly as you might expect: a look at the `Coefficients:` table shows that such a term actually includes three things `CoB+Fr+I(CoB*Fr)` (but the `Fr` term is already in `poly(Fr,4)` hence the `NA`). Hardcopy the above plot and comment on which/why is the best according to you.

14. We'll learn in chapter 3 that if you, for example, `plot(log8)`, R will display a series of diagnostic plots (hit return to advance to the next plot). Take a look at `log8` and `lin2` diagnostic plots.

15. Finally note that we found *linear* relationships between hull geometry and drag. If you believe the formulas you can make the drag arbitrarily small (or even negative) by decreasing `CoB` or `L.D` or by increasing `L.B`, `B.D` or `P.C`. Why do you suppose this was not done?

16. In this project we have 'by hand' explored finding better models. It should come as no surprised that R has various methods of automatic 'model selection'. `leaps` is one R package I've used for this, but its not a topic we'll hit in this course.