Level M Regression Models Examples

Scottish Hills

The hills dataset (loaded below) gives the record times in 1984 for 35 Scottish hill races. The variables are

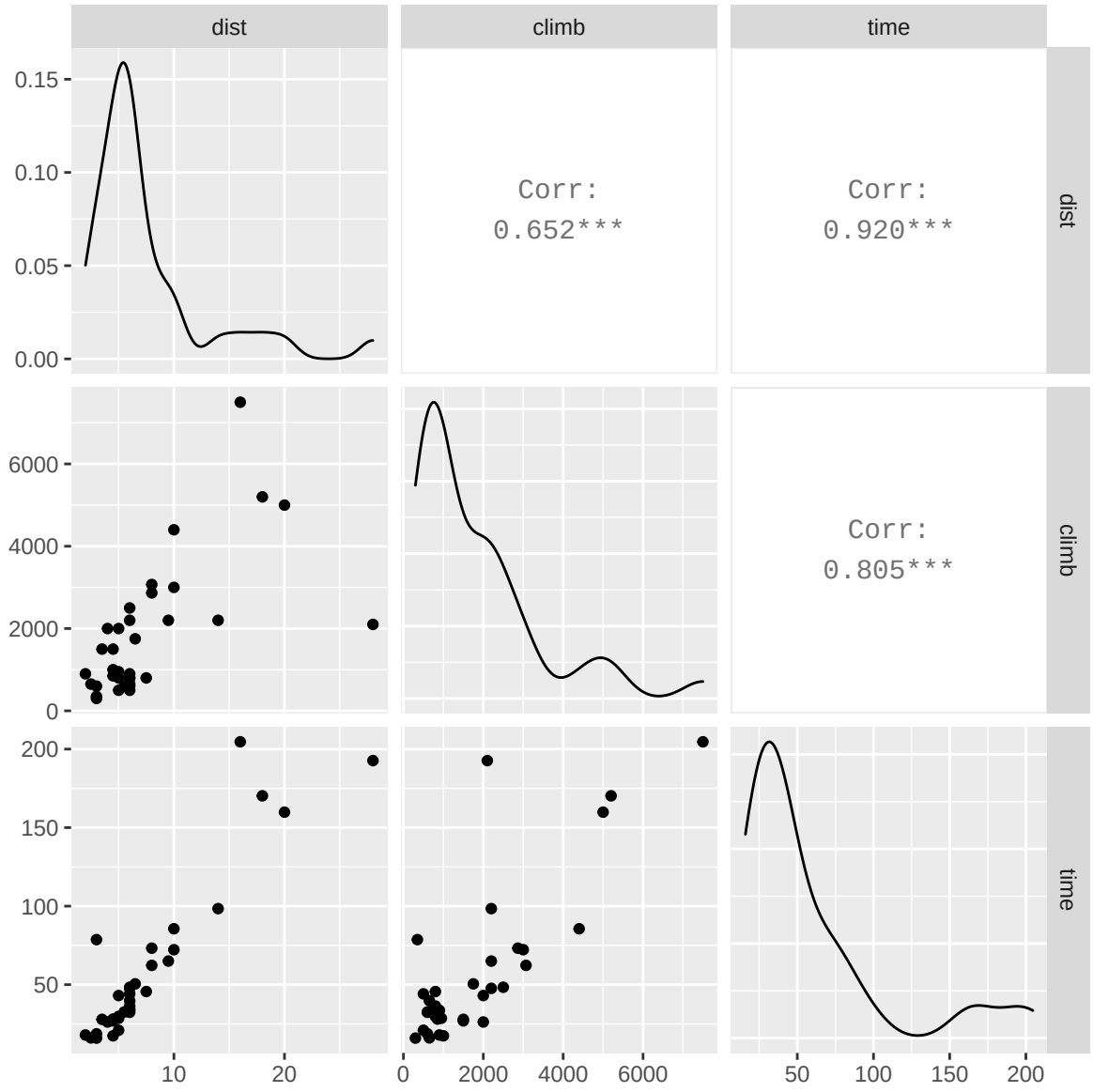
- `dist` distance in miles (on the map),
- `climb` total height gained during the route, in feet,
- `time` record time in minutes.

The goal is to predict `time` from the variables `dist` and `climb`.

```
library(MASS)
library(GGally)
data(hills)
head(hills)
```

```
##              dist climb   time
## Greenmantle   2.5   650 16.083
## Carnethy      6.0  2500 48.350
## Craig Dunain  6.0   900 33.650
## Ben Rha       7.5   800 45.600
## Ben Lomond    8.0  3070 62.267
## Goatfell      8.0  2866 73.217
```

```
ggpairs(hills)+ theme( plot.margin = margin(0,0,0,0, "cm"),
     plot.background = element_rect(
          fill = "transparent",
          colour = "transparent",
          size = 1))
```



Perform model diagnostics for the multiple linear regression model described, with time as the response variable.

In particular you can try doing the following:

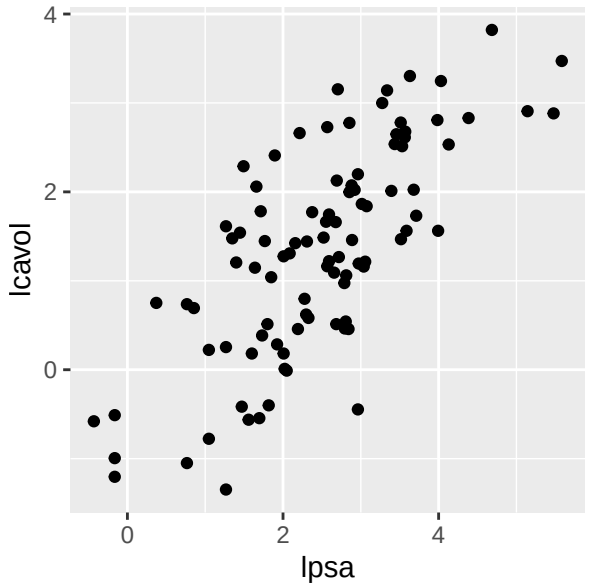
- Check whether all assumption are satisfied.
- Check whether a `log` transformation provides better a prediction,
- Check whether there are any outliers,

Prostate Cancer

A study was conducted on 97 men with prostate cancer who were due to receive a radical prostatectomy (a surgical procedure). A doctor was interested in the relationship between the logarithm of the size of the cancer (`lcavol`) and a potential continuous predictor variable, the logarithm of the prostate specific antigen (`lpsa`).

1. What may be the reason for using logarithmic transformations of the data rather than the original measurements?

2. Comment on the scatterplot below with respect to the research interests behind the study.



3. The following linear model, Model 1, was fitted to these data, with the log cancer volume, `lcavol`, as the response (Y) and the log prostate specific antigen, `lpsa`, as the explanatory variable (x),

   $$\text{Model 1}: \quad Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, 97, \quad \epsilon_i \sim N(0, \sigma^2), \quad \epsilon_i\text{'s independent.}$$

   For the two plots provided in Figure 3, explain for each plot which assumption of the normal linear model it is useful for assessing and comment specifically on whether or not the assumptions appear valid in this context.