Multiple Regression Worksheet

David Gerard

2018-12-07

Kentucky Derby

The Kentucky Derby is an annual horse race held every year at Churchill Downs. The data frame `ex0920` contains information on the winners for every year from 1896 to 2011. Researchers are interested in what variables are associated with the winning average speed. Variables include

- `Year`: year of Kentucky Derby.
- `Winner`: a character vector with the name of the winning horse.
- `Starters`: number of horses that started the race.
- `NetToWinner`: the net winnings of the winner, in U.S. dollars.
- `Time`: the winning time in seconds.
- `Speed`: the winning average speed, in miles per hour.
- `Track`: a factor indicating track condition with levels "Fast", "Good", "Dusty", "Slow", "Heavy", "Muddy", and "Sloppy".
- `Conditions`: a factor with with 2 levels of track condition, with levels "Fast" and "Slow".

You can load these data into R using:

```
library(Sleuth3)
data("ex0920")
head(ex0920)
```

1. What variable is the response? What are the explanatory variables? What variables are quantitative? What variables are categorical? What are the observational units?
2. Use the `ggpairs()` function in the `GGally` library to make a matrix plot of all the quantitative variables. Comment on any trends you notice.
3. Make some color coded scatterplots using `Conditions`. Comment on what you notice.
4. We'll model the effect of `Starters`, `Year`, and `Conditions` on `Speed`. Using just these four variables, fit a linear model based on your exploratory analysis from parts 2 and 3. Check the assumptions using residual plots. Adjust the model if needed. Iterate until you come up with a final model
5. Look at the regression summaries. What appears to be the effect of track condition on speed? Provide 95% confidence intervals in your statement.