Build a model predicting Value ($000’s).

1. Comment on R2 and adjusted R2. Interpret adjusted R2.

> bonus <- read.csv(choose.files(), header=TRUE)

> attach(bonus)

> model1 = lm(Value ~ Size + Age + Garage + Rooms + Baths)

> summary(model1)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 110.1807 154.2452 0.714 0.5266

Size 0.2162 0.0800 2.703 0.0736 .

Age -2.1893 1.0160 -2.155 0.1201

Garage -38.4491 57.7353 -0.666 0.5531

Rooms -0.7231 33.4941 -0.022 0.9841

Baths 1.1132 45.3529 0.025 0.9820

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.57 on 3 degrees of freedom

Multiple R-squared: 0.9599, Adjusted R-squared: 0.8931

F-statistic: 14.36 on 5 and 3 DF, p-value: 0.02628

R² = 0.9599 and R²a = 0.8931. They are both high, which is good. But, not that close together, which is bad. They are a bit far apart because we have several insignificant variables in the model which are not significant. So, we are taken the penalty for including them and they are not providing us with explanatory benefit in predicting Value.

R²a: 89.31% of the variation in Value is being explained by Size, Age, Garage, Rooms, and Baths, once we have taken a penalty for the number of X’s we are using in our model.

1. Build a scatterplot matrix and comment appropriately.

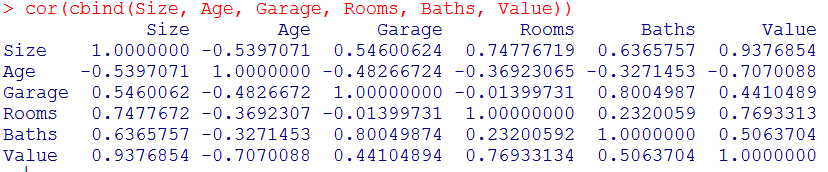
> Vars = data.frame(Size, Age, Garage, Rooms, Baths, Value)

> pairs(Vars, upper.panel=NULL)



We are mostly concerned with the bottom row here with every variable as the X and Value as the Y variable. We see that all slopes are positive except for Age. (Although note that Garage has a positive slope in the plot but a negative one in the coefficient output.) Size has the strongest linear relationship with Value as most points fall close to a line.

1. Build a correlation matrix and comment.



As they go with the scatterplots above, we see that only Age has a negative correlation while the rest have a positive correlation. It is surprising that the correlations are all of moderate or high strength when the variables are not significant in the output.

1. Interpret the slope for Garage.

For every 1 car added to the Garage, the Value of the home decreases by $38,449.10, holding all other variables constant.

1. Which variables are significant? How do you know?

If we use α = 0.10, only Size is significant as its p-value = 0.0736.

If we use α = 0.05 or 0.01, none of them are significant since all of their p-values are greater than 0.05.