

1. Consider the data set `davis.data` which, among other things gives the reported weight (`repwt`) and actual weight (`weight`) of some experimental subjects (reported weight is what they said they weighed). The idea here is that we would like to just ask people “How much do you weigh?” instead of actually having to weigh them. Thus in the model $weight_i = \beta_0 + \beta_1 repwt_i + \epsilon_i$ we would hope that the slope β_1 would be equal to 1.

- (a) Fit the model with `weight` as the y -variable and `repwt` as the x -variable.

We find the fitted model to be:

$$weight_i = 5.46 + .926 repwt_i + \epsilon_i$$

with an r^2 value of .6979, indicating that about 70% of the variability in actual weight is explained by the subjects' reported weight.

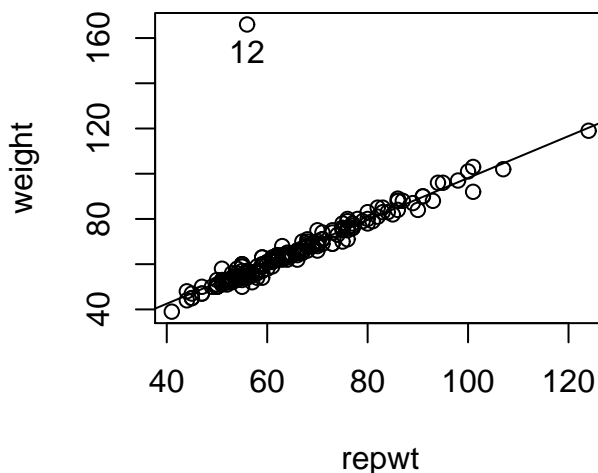
- (b) Compute a 95% confidence interval for the `repwt` slope. Is 1 in the confidence interval?

Computing a 95% confidence interval for the slope we obtain

$$0.83 < \beta_1 < 1.01$$

which does in fact contain 1 as we hoped.

- (c) Now plot the data. Notice anything odd? Which observation number is that? Look through the data by hand (with your eyeballs?) if you like. Assuming you typed `plot(repwt,weight)` you can type `identify(repwt,weight)` and click the point. Click some more points if you like. You'll have to right-click the plot and choose `Quit` to get the R prompt back. Turn in your plot and tell me what looks funny.



Looking at the plot above, observation 12 is nothing like the others and does not fit the trend.

- (d) Re-compute the slope and intercept without the offending point Then re-compute the confidence interval. What happens?

The model without the offending point is

$$weight_i = 2.83 + .96reput_i + \epsilon_i$$

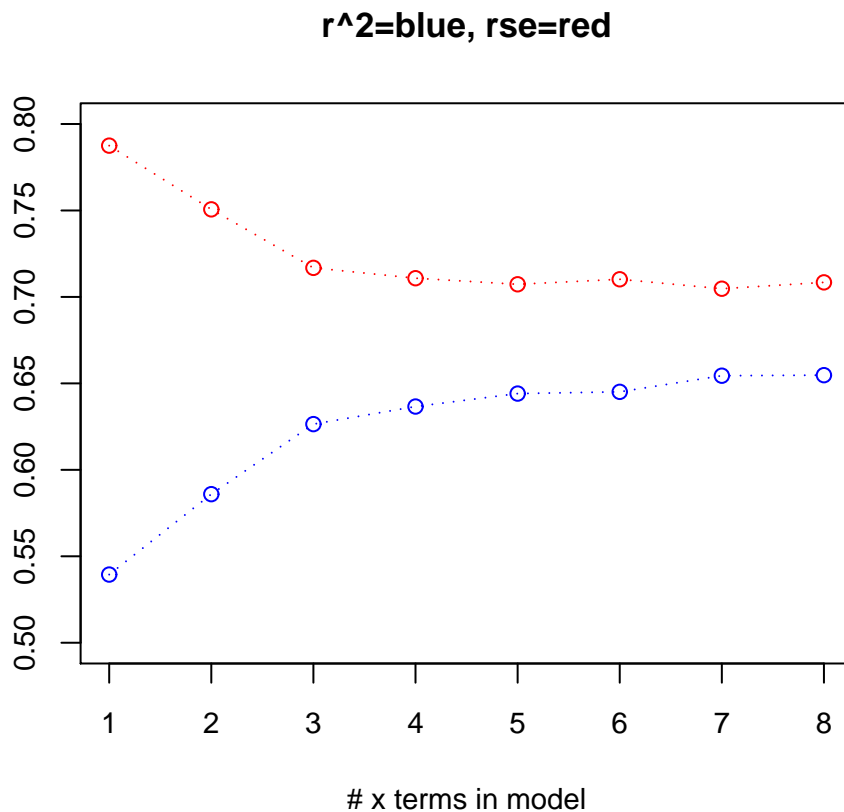
and the confidence interval is (0.93328930, 0.981006), which does not contain 1.

2. The dataset prostate.data comes from a study of 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with *lpsa* as the response and *lcavol* as the predictor. Record the residual standard error and r^2 . Now add *lweight*, *svi*, *lpph*, *age*, *lcp*, *pgg45* and *gleason* to the model one at a time. For each model record the residual standard error and the r^2 .

We tabulate our results as follows:

# <i>x</i> -terms in model	1	2	3	4	5	6	7	8
r^2	0.539	0.586	0.626	0.637	0.644	0.645	0.654	0.655
std.error	0.787	0.751	0.717	0.711	0.707	0.710	0.705	0.708

For a graphic, see the plot below.



Note that the r^2 in blue is always an increasing line (we showed in class that this will always happen). On the other hand the residual standard errors sometimes decrease and sometimes increase.

- Using the model with `lpsa` as the response (i.e. y) and the other variables as predictors, compute 90 and 95 percent confidence intervals for the slope corresponding to `age`. Using just these intervals, what can be said about the p -value for `age` in the regression summary?

The 90 percent confidence interval is $(-0.038, -0.001)$, which we note does not contain zero. Thus we would reject $H_0 : \beta_{age} = 0$ at the 10% level, and the p -value is necessarily less than .10.

The 95 percent confidence interval is $(-0.042, 0.003)$, which we note does contain zero. Thus we would fail to reject $H_0 : \beta_{age} = 0$ at the 5% level, and the p -value is necessarily greater than .05.

From this we can conclude that the p -value for the `age` slope is somewhere between .05 and .10.