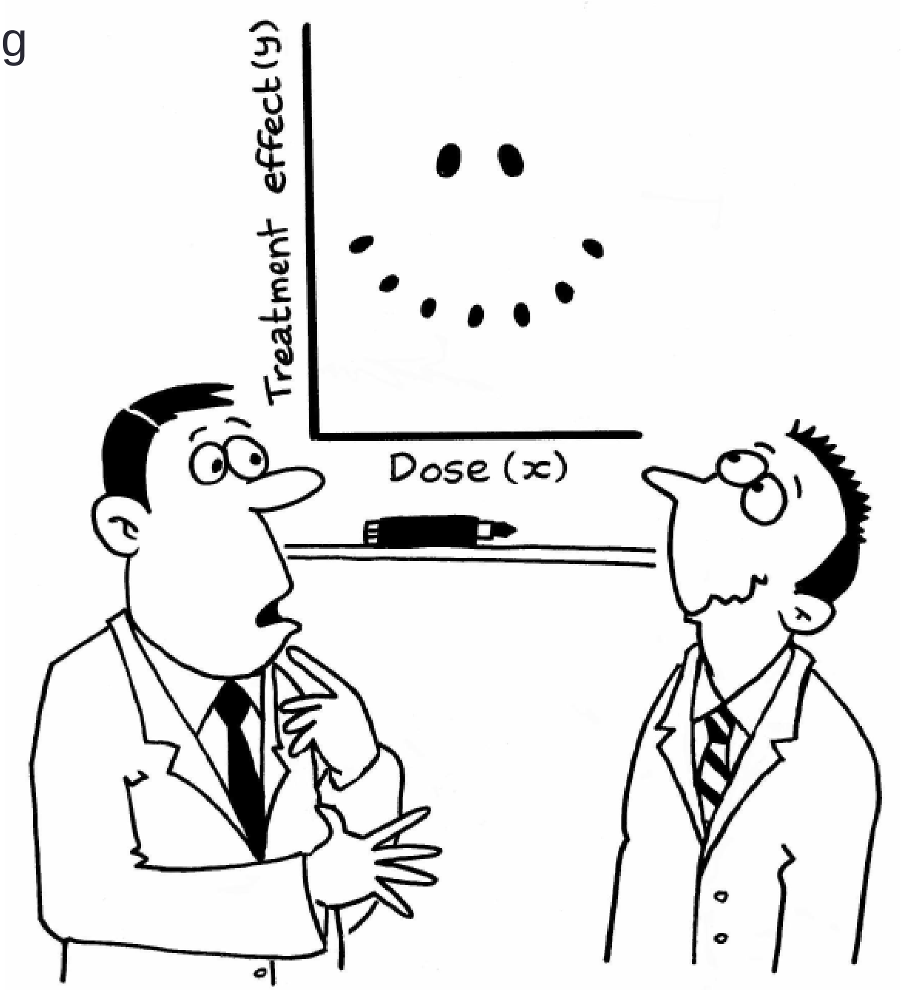
www.causeweb.org



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

VIRTUAL LAB 9

Multiple Regression

Example

- The measurements of 31 felled black cherry trees were measured for their diameter (in inches measured at 4 foot 6 inches above ground), height (in feet) and volume (cubic foot). We want to use diameter and height to predict the volume of the tree.

$\hat{y} = -57.9877 + 4.708_{\text{Diameter}} + 0.3393_{\text{Height}}$

$SSE = 421.92 \quad SSR = 7684.16 \quad TSS = 8106.08$

Questions

- Test the overall significance of the model.

Source	df	SS	MS	F	pvalue
Model	2	7684.16	3842.08	254.97	~0
Error	28	421.92	15.07		

H_0: $\beta_1=\beta_2=0$ versus H_A: At least one ≠ 0

F=254.97 with p-value close to 0. Reject the null hypothesis and conclude there is some significant information in the explanatory variables regarding the response

- Test the individual significance for the Diameter of the trees.

$$s_{Diameter} = 0.2643$$

H_0: $\beta_{diameter} = 0$ versus H_A: $\beta_{diameter} \neq 0$; t=17.813; pvalue ~0. Reject the null hypothesis and conclude there does appear to be a significant linear relationship between diameter and volume in black cherry trees.

- Now test the significance of Height. Should either be removed from the model if we use an α of 0.01?

$$s_{Height} = 0.1302$$

H_0: $\beta_{height} = 0$ versus H_A: $\beta_{height} \neq 0$; t=2.606; pvalue =0.0145. Fail to reject the null hypothesis and conclude there does NOT appear to be a significant linear relationship between diameter and volume in black cherry trees (at α=0.01).

If we use α=0.01, then we would consider removing height.

Categorical Predictor variable

- Develop both effects coding and dummy / reference coding for a categorical variable with 4 categories.

There are 3 dummy variables needed..X_1, X_2, X_3; let's call the 4 levels (categories) A, B, C, D

	X_1	X_2	X_3		X_1	X_2	X_3
A	1	0	0		1	0	0
B	0	1	0		0	1	0
C	0	0	1		0	0	1
D	0	0	0		-1	-1	-1

Another example

- Using the same data set, the land in which the tree grew can be classified as either dry or not dry (0 if dry and 1 if not dry). The regression equation for this data set is:

$\hat{y}$ = -61.63 + 5.33_{Diameter} + 0.31_{Height} – 4.50_{dry}

With s_{dry} = 2.31 and SSE = 369.6

Questions

- What is the interpretation of the dry coefficient?

Black cherry trees growing in not dry conditions are expected to have on average 4.5 cubic feet less in volume than trees growing in dry conditions for a given diameter and height.

- Calculate the test of significance for the dry variable.

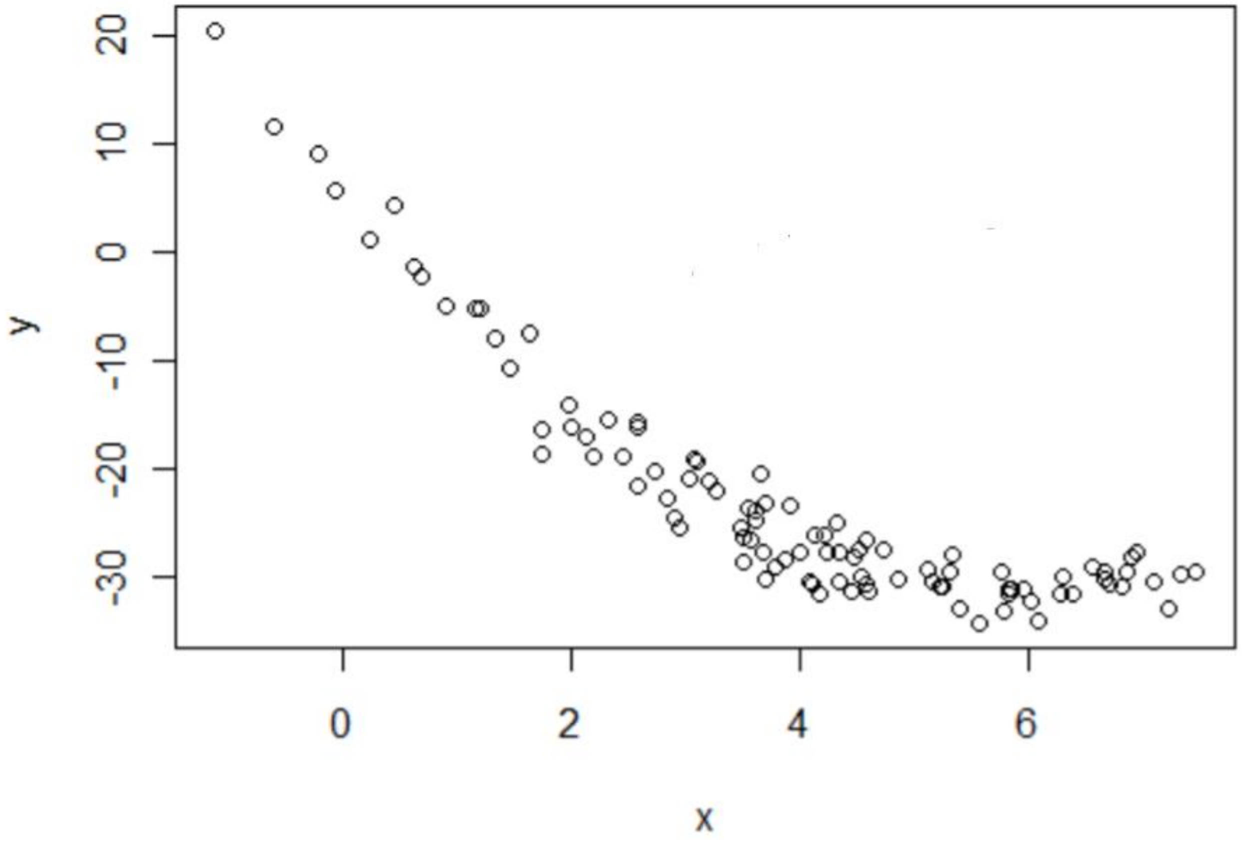
H_0: $\beta_{dry} = 0$ versus H_A: $\beta_{dry} \neq 0$; t=1.948; pvalue =0.062. Fail to reject the null hypothesis and conclude there does NOT appear to be a significant linear relationship between dry conditions of the land and volume in black cherry trees (at α=0.01).

- Calculate R^2 and R^2_{Adj}

$R^2 = 0.954$ and $R^2_{Adj} = 0.544$

And another….

The plot is fitted with a quadratic model for x predicting y. From the below plot, what can you determine about the sign of the coefficient estimate for the quadratic term of x?



Positive