Lecture 32

Continuing regression from Lecture 31

Example

Research was conducted to measure the stretchability of Mozzarella cheese. In particular, the researchers were interested in how temperature related to elongation (measured as percentage increase) at breakage. They gathered the following data

x (temperature)	y (elongation)
59	118
60	143
63	182
65	210
68	247
70	243

Find the equation of the regression line.

Note that

$\sum x_i = 385 \quad \sum x_i^2 = 24799 \quad \bar{x} = 385/6 = 64.16$

$\sum y_i = 1143 \quad \sum y_i^2 = 231655 \quad \bar{y} = 1143/6 = 190.5$

$\sum x_i y_i = 74464 \quad n = 6$

So $s_x = \sqrt{\dfrac{24799 - 6(64.16)^2}{5}} = 4.3550$

$s_y = \sqrt{\dfrac{231655 - 6(190.5)^2}{5}} = 52.7513$

so $r = \dfrac{1}{5} \dfrac{1}{4.355} \dfrac{1}{52.7513} (74464 - 6(64.16)(190.5))$

$= 0.9763$ recall $b_1 = r \dfrac{s_y}{s_x}$ $b_0 = \bar{y} - b_1 \bar{x}$

So slope and intercept estimates are

$b_1 = (0.9763)\left(\dfrac{\cancel{4.3550}}{52.7513}\right) = 11.83$

$b_0 = 190.5 - 11.83(64.16) = -568.34$

In otherwords the fitted regression line is

$$\text{Elongation} = -568.34 + 11.83\,\text{Temperature}.$$

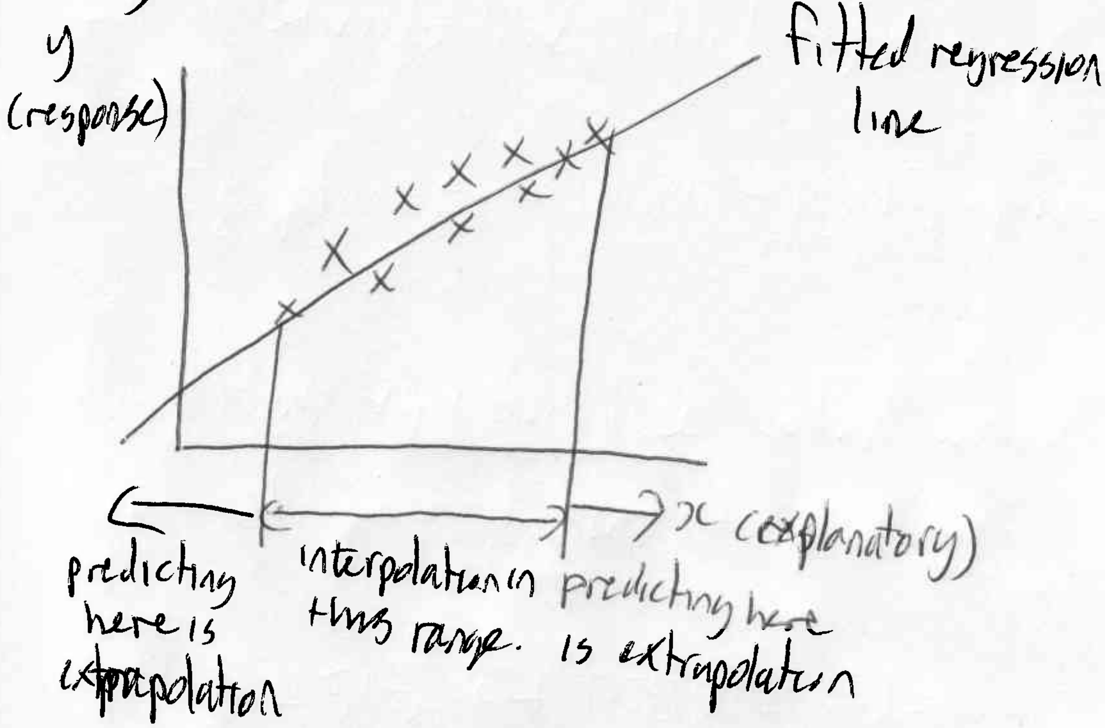
Using this regression line what would we predict the elongation to be at temperature 66.5?

$$-568.34 + 11.83(66.5) = \underline{218.355}$$

<u>Can we always use regression line to predict response given explanatory?</u>

In general NO!! It is usually only safe to predict at values of the explanatory within the range of the data observed. We call this <u>interpolation</u>. Predicting at values outside the range of the observed explanatory variables is called <u>extrapolation</u> it is dangerous.

eg

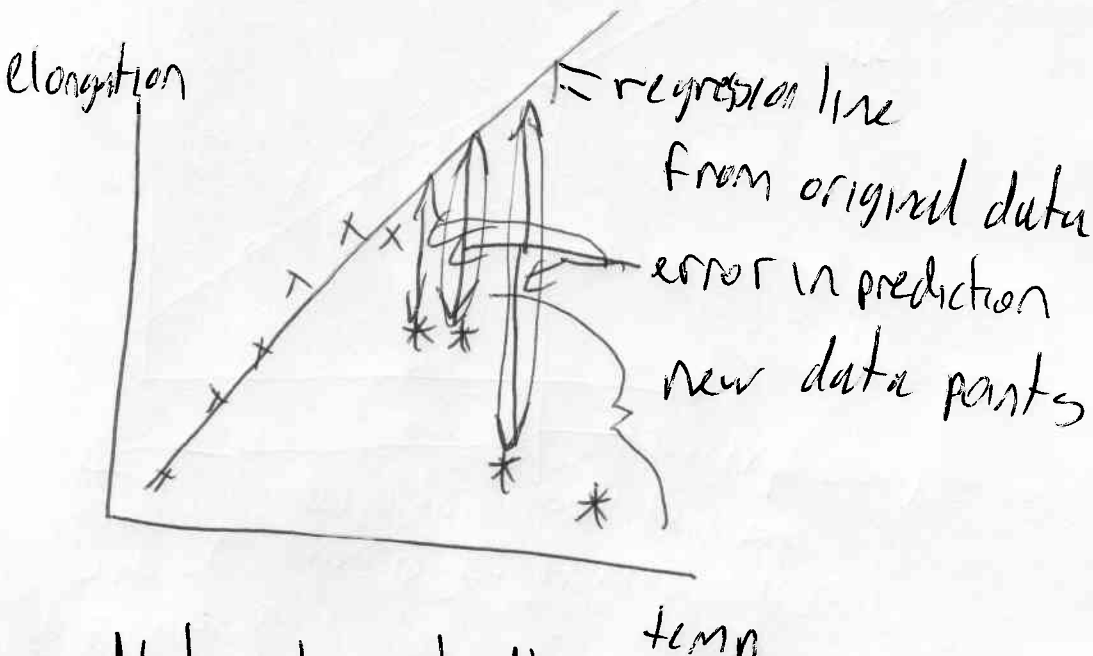


Why is it dangerous? Because we cannot be sure that the linear pattern we saw over the observed range continues outside this range

For instance further data on the mozzerella cheese study came in and was

x (temp)	y (elongation)
72	203
74	197
78	135
83	132

Rough sketch of situation

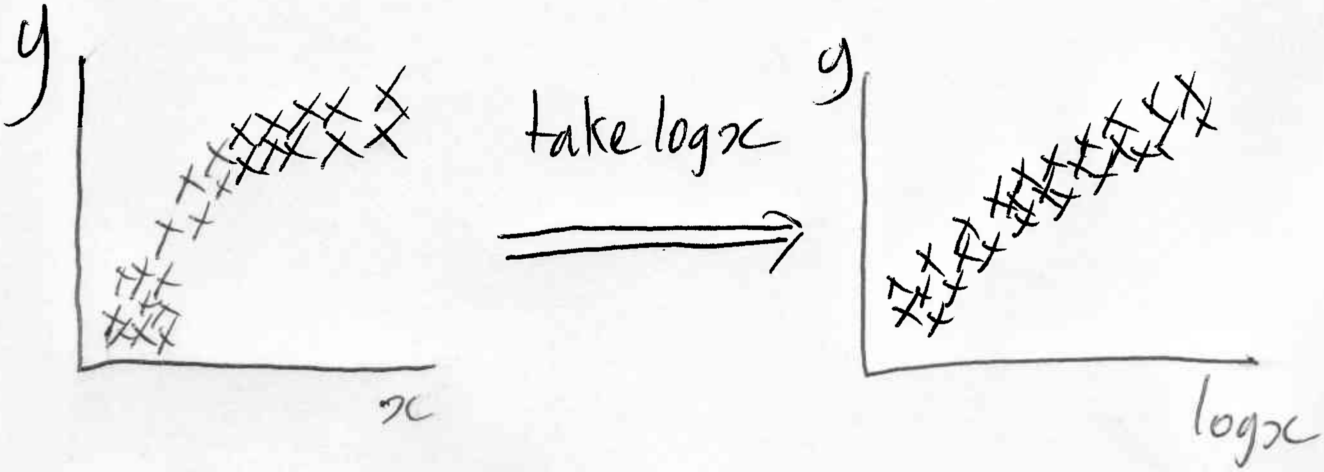


Notice how badly the regression line would have done at predicting the new points.

Transformation

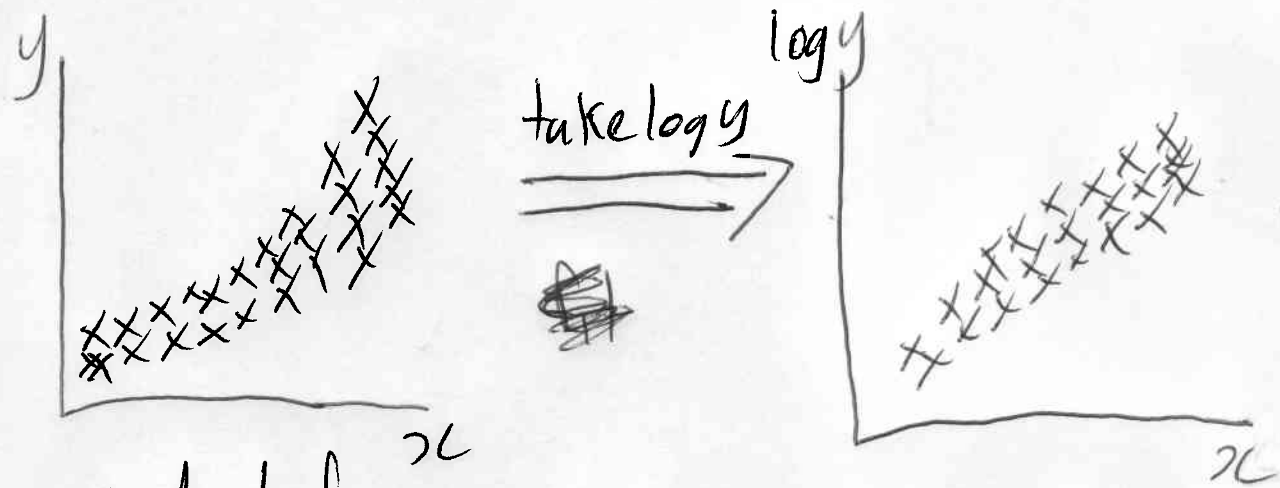
Sometimes a scatterplot will not initially show what looks like a reasonable linear relationship between the two variables. This might lead you to think that a linear regression line would be inappropriate. Some of the time it is possible to transform the data and produce a more linear relationship between the two variables. We discuss 3 cases here (your textbook discusses somemore)

1.



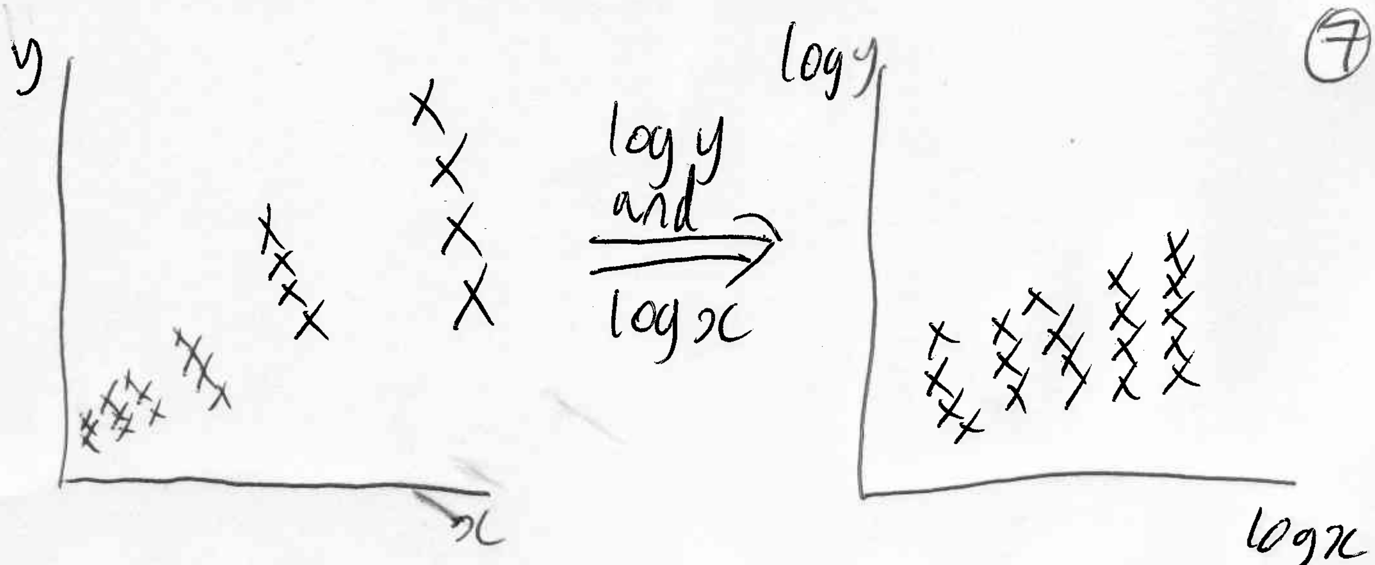
Instead of $y = b_1 x + b_0$ fit $y = b_1 \log x + b_0$

2



instead of $y = b_1 x + b_0$ fit $\log y = b_1 x + b_0$

3.



instead of

$y = b_1 x + b_0$

fit $\log y = b_1 \log x + b_0$

<u>R-sq (aka R^2)</u>

R^2 is the square of the correlation. It can be used to judge how well the linear regression line fit the observed data.

Note

$0 \leq R^2 \leq 1$

no linear relationship

linear relationship models data perfectly.