Q17
P.488

We shall use the data to conduct a complete regression analysis.

First of all we shall investigate whether there is any relationship between the bivariates - x and y, at all. We will use Spearmans method.

If we find that a relationship exists we shall then investigate whether this relationship can be characterized by a linear regression model. We will use Pearson's formula.

Spearman Method

We ranked x_s and y_s seperately from lowest to highest. If there were tied values, we would apply an average rank for those tied values. However, too many ties compromises the accuracy of the Spearman Rank Correlation Method.

x	R_i	y	S_i	$R_i * S_i$
99.0	1	28.8	15	15
101.1	2	27.9	14	28
112.1	9	17.1	6	54
112.4	10	18.9	7	70
102.7	3	27.0	13	39
103.0	4	25.2	12	48
113.6	11	16.0	4	44
113.8	12	16.7	5	60
107.0	6	21.5	10	60
105.4	5	22.8	11	55
115.1	13	13.0	2	26
110.8	8	19.6	8	64
115.4	14	13.6	3	42
120.0	15	10.8	1	15
108.7	7	20.9	9	63
Σ	120		Σ 120	Σ 683

Spearmen's Rank correlation Coefficient

$$r_s = \frac{\sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}}$$

where n is the sample size, or number of bivariate pairs

$$r_s = \frac{683 - \frac{15(15+1)^2}{4}}{\frac{15(15^2-1)}{12}}$$

$$= \frac{683 - 960}{280}$$

$$= -0.9893$$

The closer $|r_s|$ is to 1 the stronger the relationship ~~to~~ between x_s and y_s. However r_s does not tell the functional form of the relationship. It just says there is some relationship $|r_s|$ value must be at least 50%.

Note that in this example we had negative value of r_s. This means that whatever the relationship is, as x increases, y decreases, or vice versa. If we had a positive value, then it would mean that as x increases y increases. But again, the functional form of the relationship cannot be determined by Spearman's method.

So now that we have established that there is some relationship, we want to find what functional form can describe the relationship. Lets use the simplest functional form first. The Pearson method enables us to test whether a linear model can describe the relationship between x and y. Pearson's Coefficient of Linear Correlation (r) measures the strength of a linear relationship between x and y. To pass we need $|r|$ at least 50%. The closer to 1 the stronger the linear correlation

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

We must note with ALL correlation methods that correlation does not imply causality. It simply means there is a relationship between x and y, but it does not say x is causing y or vice versa.

Note also that Pearson's method measures strength of LINEAR relationship. If it turns out that the linear model does not fit the data, then the Pearson r value is not relevant.

x	y	x^2	y^2	xy
99.0	28.8	9801	829.44	2851.2
101.1	27.9	10221.2	778.41	2820.69
112.1	17.1	12566.41	292.41	1916.91
112.4	18.9	12633.76	357.21	2124.36
102.7	27.0	10547.9	729	2772.9
103.0	25.2	10609	635.04	2595.6
113.6	16.0	12904.96	256	1817.6
113.8	16.7	12950.44	278.89	1900.46
107.0	21.5	11449	462.25	2300.5
105.4	22.8	11109.16	519.84	2403.12
115.1	13.0	13248.01	169	1496.3
110.8	19.6	12276.64	384.16	2171.68
115.4	13.6	13317.16	184.96	1569.44
120.0	10.8	14400	116.64	1296
108.7	20.9	11815.69	436.81	2271.83
1640.1	299.8	179850.33	6430.06	32298.59

$\bar{x} = 109.34$ $\bar{y} = 19.99$

[Very laborious and error-prone process. In practice we use computers for this analysis]

$$r = \frac{15(32298.59) - (1640.1)(299.8)}{\sqrt{15(179850.33) - (1640.1)^2}\sqrt{15(6430.06) - (299.8)^2}}$$

$$r = \frac{484478.5 - 491701.98}{(88.47)(81.06)}$$

$$= 0.99927$$

[Please check my arithmetic, I may have miscalculated somewhere.]

So we have a very strong linear correlation

So now that we have established that

1) there is a relationship (Spearman)

2) that relationship is a linear model (Pearson)

We now need to determine the parameters of the linear model. We know that a linear function is of the form

$$y = a + bx$$

Lets use $\hat{y}$ to denote y-values along the linear model and y to denote y-values in the data.

So,

$$\hat{y} = a + bx$$

It turns out that,

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$= \frac{15(32298.59) - (1640.1)(299.8)}{15(179850.33) - (1640.1)^2}$$

$$= -0.923$$

and

$$a = \bar{y} - b\bar{x}$$

$$= 19.99 - (-0.923)(109.34)$$

$$= 120.91$$

So we have the linear model

$$\hat{y} = 120.91 - 0.923x$$

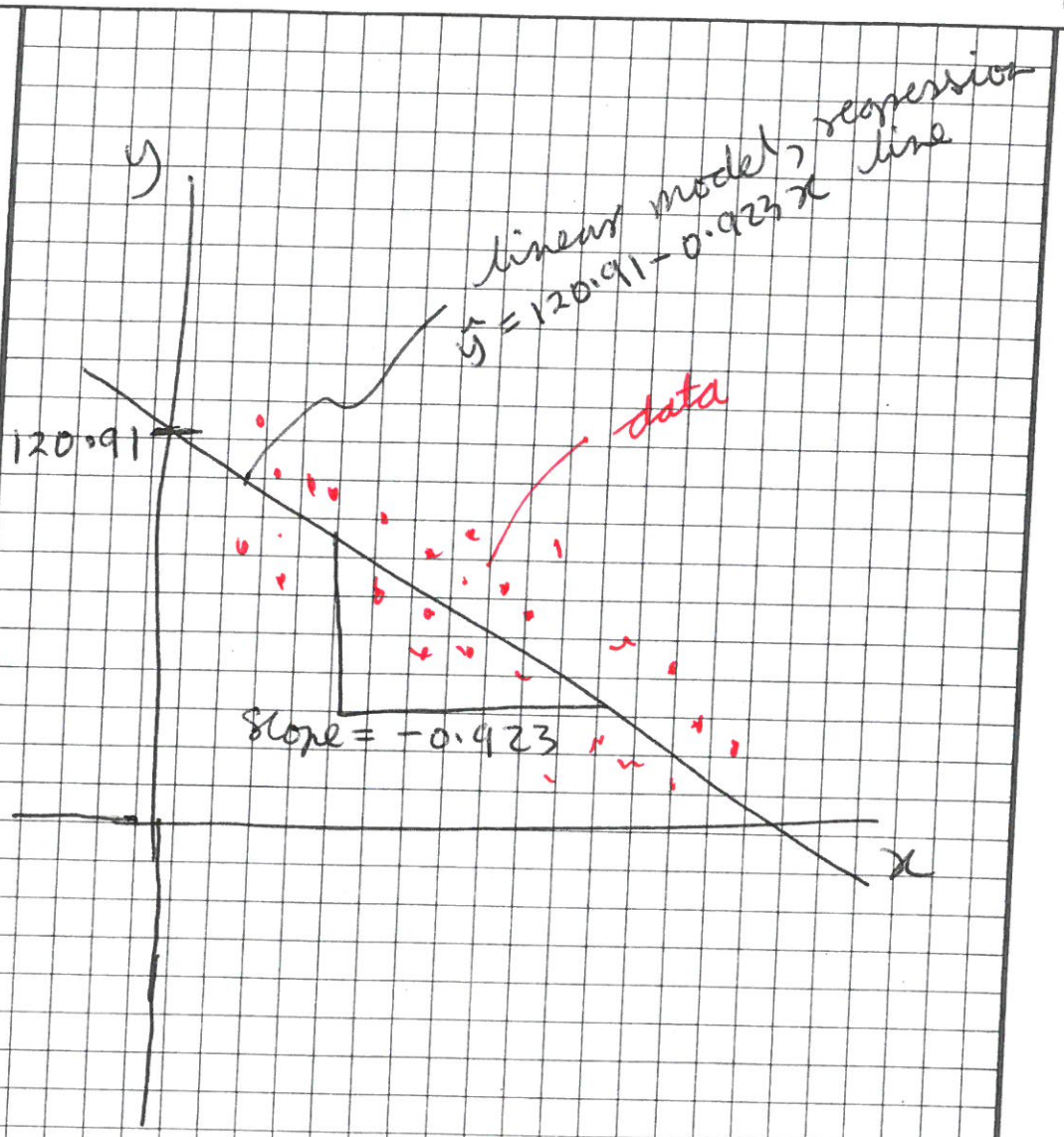
Now how well does this model describe (or account for) the variability in our data?

r^2 can measure this.

$$r^2 = (0.9927)^2$$

$$= 0.9854$$

In my profession, any linear model you develop must have r^2 ≥ 75% otherwise it will be rejected. Theoretically, any value 0.5 and above is acceptable.



Another check you can perform before running the Pearson calculations, is to draw a scatterplot of your data. If the points appear to be following some linear trend, then Yes, pursue the linear model.

Checking the Adequacy of the Linear Model

We plot the residuals versus fitted values (or theoretical values), $\hat{y}$.

For each x-value, the residual

$$e_i = y_i - \hat{y}_i$$

$\hat{y}_i$ — from regression line

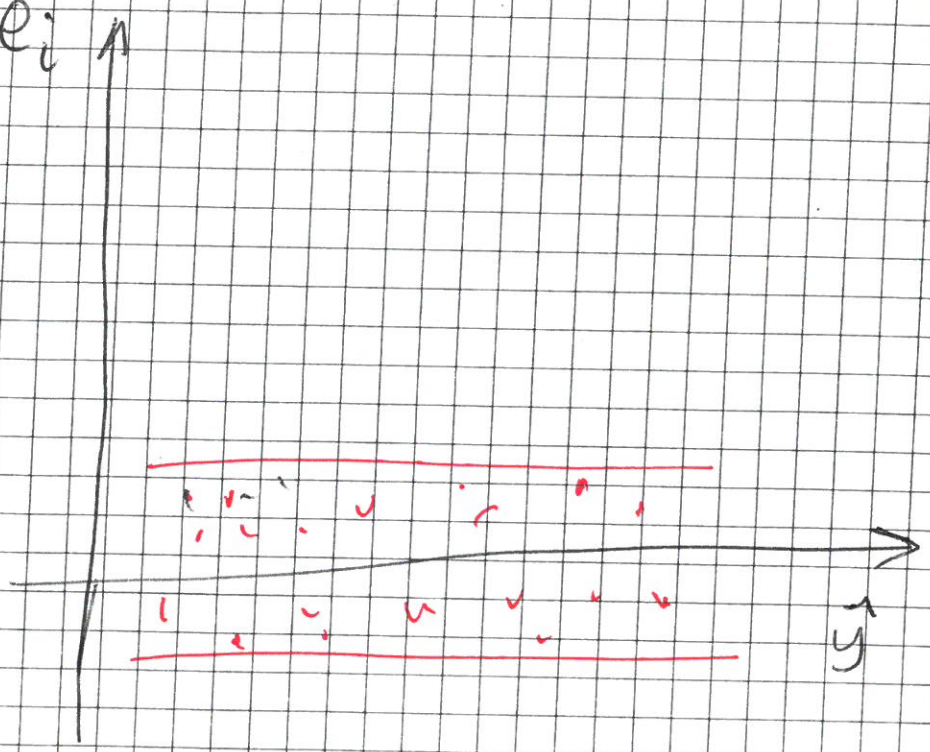
y_i — from data

We shall make a scatter plot of the residuals

x	y	fitted values $\hat{y} = 120.91 - 0.923x$	residuals $e_i = y - \hat{y}$
99.0	28.8	29.53	-0.733
101.1	27.9	27.59	0.3053
112.1	17.1	17.44	-0.3417
112.4	18.9	17.16	1.7352
102.7	27.0	26.12	0.8821
103.0	25.2	25.84	-0.641
113.6	16.0	16.06	-0.0572
113.8	16.7	15.87	0.8274
107.0	21.5	22.15	-0.649
105.4	22.8	23.63	-0.8258
115.1	13.0	14.67	-1.6727
110.8	19.6	18.64	0.9584
115.4	13.6	14.39	-0.7958
120.0	10.8	10.15	0.65
108.7	20.9	20.58	0.3201

[I used Excel to do the calcs, so my numbers may look a little different than doing it manually]

So we plot



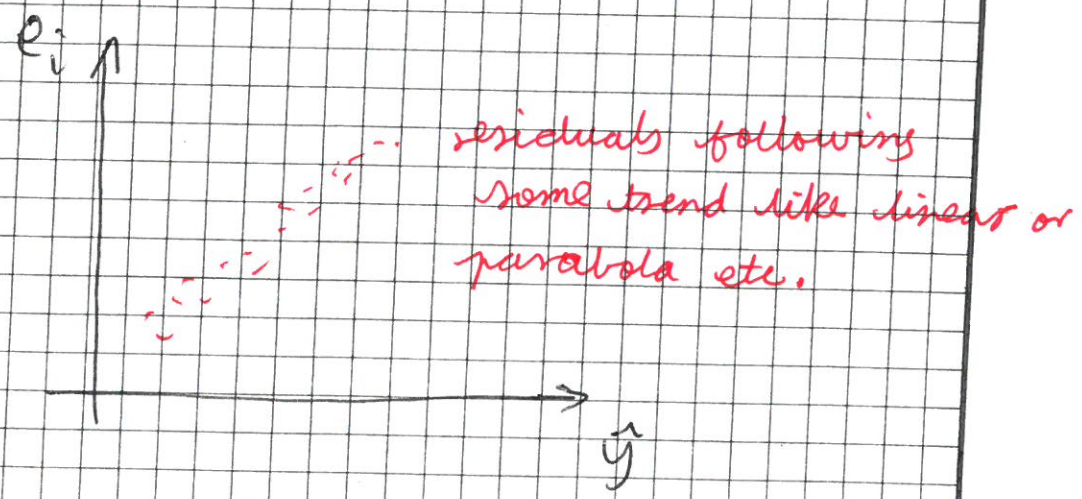
You must see a random scatter of the points in a band across the graph.

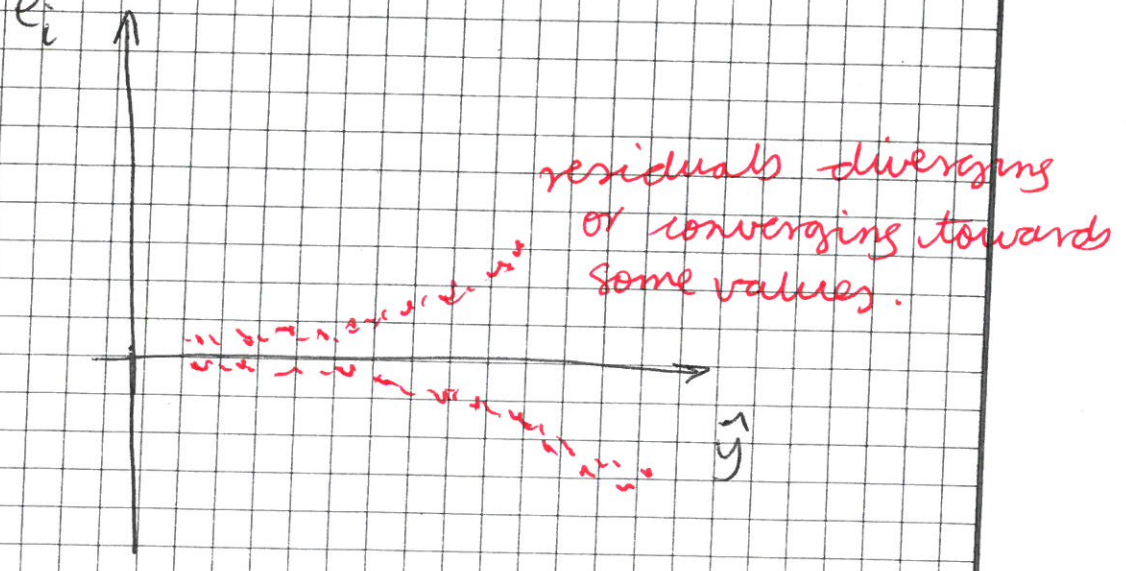
If so then CONGRATULATIONS!

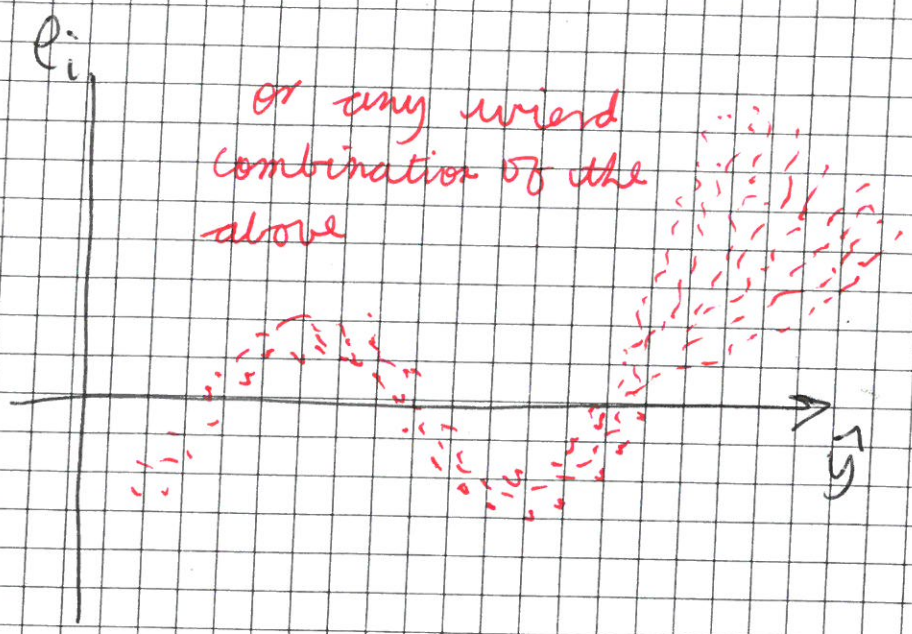
You developed a dependable linear regression model.

For small data sets it may be difficult to identify the band.

If you see any of the following, your linear model has ~~failed~~ FAILED!







This one is actually REAL. This is what I got on my first Regression model as a rookie engineer working on the Florida bridge maintenance program. It was an unmitigated disaster. I will never forget it. Ha!

But it turns out the correct regression model for my bridge data was a 3rd order polynomial (a cubic function)

So now that we confirmed the adequacy of our model, we can use it to make predictions.

But the predictor values MUST ALWAYS be within the range of the dated that we used to build the regression model.

So what y value can we predict for $x = 100$?

$$\hat{y} = 120.91 - 0.923(100) = 28.061$$

$x = 117$?

$$\hat{y} = 120.91 - 0.923(117) = 12.919$$

$x = 85$?

NOT POSSIBLE!!

Our data starts at $x = 99$ through $x = 120$. We ~~can bar~~ cannot predict for x-value outside of this range. So not possible

$x = 130$? NOT POSSIBLE

If we have $y = 27$ can we predict an x-value? YES.

If we have ~~y~~ $y = 58.2$ can we predict an x-value?

ABSOLUTELY NOT!

58.2 is beyond the y-values we have in the data.