Chapter 15 – Correlation and Regression

Now, we measure both an X ('independent' or 'predictor') and a Y ('dependent') variable – both RV's – for each person or subject in a study. In *correlation analysis*, we view both RV's as varying together and wish to quantify the linear association between them. More specifically, we assume that the couples (x,y) have a Bivariate Normal distribution, which entails five population parameters: μ_x, σ_x, μ_y, σ_y, and ρ (the last one is the 'population correlation coefficient'). In *regression analysis*, on the other hand, conditional on X = x, we assess the linear association by fitting the expected line

$$E(Y/\,X = x) = \mu_{y/x} = \alpha + \beta x \qquad (*)$$

Here, we need to estimate the intercept (α) and the slope (β). We also need to estimate the standard error, $\sigma_{y/x}$, associated with the expected mean line (*); this linear model requires that $\sigma_{y/x}$ does not vary with x.

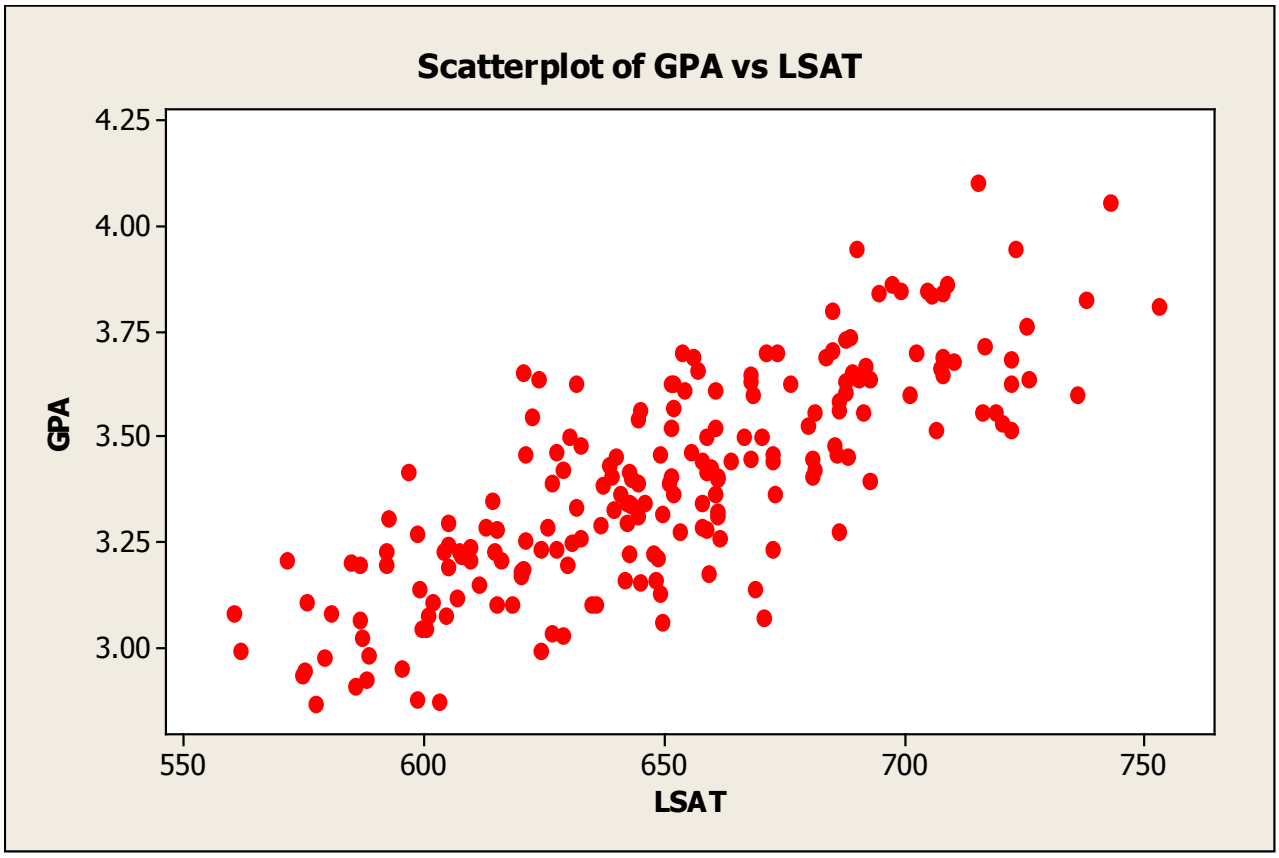
Scatter Plots – The first thing to do with bivariate data is to plot the data with the dependent variable on the y-axis and the independent variable on the x-axis. In the handout (p.1), is the scatter plot of the n = 8 points plotting y = height (cm) versus x = age (months); note here the up-sloping linear association. On p.2, we see the scatter plot for the n = 38 points for Example 15.2 (p.410), plotting y = mileage (mpg) versus x = car weight (pounds). In the latter plot, there is a down-sloping (somewhat linear) relationship. In Figure 15.3 on p.410, there does not seem to be any linear association.

The Correlation Coefficient – We estimate the above population correlation coefficient (ρ) with the sample counterpart (r) given by:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum xy - n\times\bar{x}\times\bar{y}}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

The calculation of r by hand is best done via a table such as Table 15.3 on p.412, and coincides with the plots in Figures 15.5 and 15.6. Note that the correlation coefficient (r) is positive when the trend is upward, negative when it's downward, and always between –1 and 1.

<u>LSAT – Law School GPA Example</u>. Here n = 200 (students), r = 0.796



Note the elliptical pattern in the data. From this plot, we can 'eyeball' that $\bar{x} \approx 660$, $\bar{y} \approx 3.4$, $s_x \approx 40$, $s_y \approx 0.25$, and $r \approx 0.80$. Indeed, we have:

Descriptive Statistics: LSAT, GPA

Variable	N	Mean	StDev	Minimum	Median	Maximum
LSAT	200	650.21	41.11	560.25	649.50	752.99
GPA	200	3.3967	0.2544	2.8659	3.3971	4.1082

Correlations: LSAT, GPA

Pearson correlation of LSAT and GPA = 0.796

Whereas *correlation analysis* treats X and Y equally, *regression* treats Y as the dependent variable – and depending on X.

<u>Simple Linear Regression</u> – Another way to express equation (*) is (conditional on observing x)

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{for } i = 1 \ldots n,$$

where the ε_i are called the error terms. By *assumption*, we have that for all values of i = 1 … n, $\varepsilon_i \sim N(0, \sigma_e^2)$; hence, $y_i \sim N(\alpha + \beta x_i, \sigma_e^2)$. From our discussion above, we realize that this is really a conditional statement (conditional upon $X = x_i$). Our next goal is to estimate and test hypotheses related to the parameter estimates of α, β and σ_e^2; we do so using the *principle of least squares*. Our objective function to minimize is the error sum of squares (SSE) function:

$$SSE(\alpha, \beta) = \Sigma \varepsilon_i^2 = \Sigma (y_i - \alpha - \beta x_i)^2 \qquad (**)$$

Differentiating (**) with respect to ('wrt') α and β, and then setting to zero, and solving will give us the LSE's (least squares estimates), 'a' and 'b'. Differentiating wrt α, setting to zero, and solving gives:

$$\bar{y} = a + b\bar{x} \qquad \textbf{(NE1)}$$

This expression, called the first '*normal equation*' (NE), implies that the regression line goes through the point of averages. The second NE is obtained by differentiating wrt β and setting to zero:

$$\Sigma x_i y_i = a \Sigma x_i + b \Sigma x_i^2 \qquad \textbf{(NE2)}$$

Simultaneous solution of NE1 and NE2 gives the slope LSE formula:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$$

Once this is obtained, we use (NE1) for find the intercept LSE (a).

The *predicted values* of the y_i are $\hat{y}_i = a + bx_i$, and the *residuals* are then the difference between the actual and the predicted: $e_i = y_i - \hat{y}_i$. The residual sum of squares (the text and Minitab *mistakenly* call this SSE) is SSE = Σe_i^2. It is very important, then, to note the difference between errors and residuals: residuals are the actual differences between the actual and predicted y's, whereas errors are never observed – in much the same way that the estimated slope (b) is observed and calculated and used to estimate the true slope (β), which is never observed (unless the whole population is sampled). Finally, our estimate of σ_e^2 is the mean square of the residuals,

$$s_e^2 = SSE/(n-2)$$

The square root of this term – which estimates σ_e - is called the RMS of the residuals on p.423.

<u>Example 15.1 continued</u> – we now calculate our estimates (LSE's for α and β and s_e^2 for σ_e^2) by hand. Note $\bar{x} = 54$, $\bar{y} = 116$, n = 8. Summing down columns, we get $\sum(x_i - \bar{x})(y_i - \bar{y}) = 3531$, $\sum(x_i - \bar{x})^2 = 3348$, so b = 3531/3348 = 1.05466, a = 116 – 1.05466*54 = 59.0484, and the fitted line is $\hat{y}_i = 59.04839 + 1.05466\,x_i$, giving us the last two columns.

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$\hat{y}$	e
1	24	87	-30	-29	870	900	84.36	2.64
2	48	101	-6	-15	90	36	109.67	-8.67
3	60	120	6	4	24	36	122.33	-2.33
4	96	159	42	43	1806	1764	160.30	-1.30
5	63	135	9	19	171	81	125.49	9.51
6	39	104	-15	-12	180	225	100.18	3.82
7	63	126	9	10	90	81	125.49	0.51
8	39	96	-15	-20	300	225	100.18	-4.18

Here, SSE = $\Sigma e_i^2 = 211.997$, so $s_e^2 = 211.997/6 = 35.333 = 5.944^2$.

Predictions – The text makes a good point on pp.428-9 (for this example) in that when prediction of y = height is made with no x = age information, our prediction is $\bar{y} = 116$ cm, with SE of $s_y = 23.71$ cm. On the other hand, with the information that x = 60 months, the (regression) predicted value is $\hat{y} = 122.33$ cm, with SE of $s_e = 5.944$. *Note the huge drop in the SE!*

As we can see from the handout, Minitab reports the SSE under 'Residual Error SS', s_e^2 (often called the MSE for 'mean square error') under 'Residual Error MS', s_e under 'S', it gives the LSE's (along with some t-tests and p-values), and it also gives and 'Analysis of Variance' or ANOVA table. At the bottom of this table is the 'Total SS' – more correctly called the *corrected* total SS – which is equal in general to SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$. The Lemma on p.429 shows that $SSE = (1 - r^2)SST$. Usually, 'r^2' is expressed as 'R^2', so that

$$R^2 = \frac{SS\,\mathrm{Re}\,g}{SST},$$

where SSReg (Regression SS) = SST – SSE. In the above example (and handout), note that $3724.0/3936.0 = 0.94614$, which corresponds with the reported value of R^2. The usual interpretation of R^2 is 'the proportion of the total variability that is explained by the regression line'. Finally, Theorem 15.3 establishes that

$$s_e = s_y\sqrt{1-r^2}\,\sqrt{\frac{n-1}{n-2}} \qquad \approx s_y\sqrt{1-r^2} \text{ (for large n).}$$

In the above example (where n = 8 is *not* large), we have

$$5.944 = 23.713\sqrt{1-0.9727^2}\,\sqrt{\frac{7}{6}},$$

which is correct.

Read 15.7 to understand 'the regression paradox'.

Hypothesis Testing (section 15.8) – Paramount is the test of whether the linear regression line is flat ($\beta = 0$) or that there is zero correlation between X and Y ($\rho = 0$), and these parameters are related by the equation:

$$\beta = \rho \frac{\sigma_y}{\sigma_x}$$

It follows that a hypothesis of zero *correlation* can be tested by testing for zero *slope*; remembering that the LSE for β is b, for this latter test, though, we need to find the SE associated with b, SE_b. On p.435 bottom, it is shown that b is a linear combination of the (independent) y_i's, from which Theorems 15.4 and 15.5 result. The first shows that b is an *unbiased estimate* of β. The latter theorem shows that

$$\text{Var(b)} = \frac{\sigma_e^{\,2}}{\sum (x_i - \bar{x})^2}$$

Since we don't know σ_e^2, we'll use s_e^2 in its place:

$$\text{SE}_\text{b} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{b}{r}\sqrt{\frac{1-r^2}{n-2}}$$

In the above example, $SE_b = (1.0547/0.9727) \times \sqrt{0.05385/6} = 0.10273$. Now, we have all the ingredients to find a CI for the true slope (β) or test hypotheses related to it. Since the degrees of freedom associated with it are $(n-2)$, if we want the 95% CI for β, the relevant t-value for this example is $t_{0.975} = 2.4469$, so the CI is:

$1.0547 \pm 2.4469 * 0.10273 = 1.0547 \pm 0.2514$ or (0.8033,1.3061)

We're 95% confident that the true slope lies between these values.

***Beware: correlation is not causation!* (Section 15.9)**