LECTURE - Regression

Correlation Model

and

SLR without intercept

- Part 4

Correlation Model: X and Y Jointly Normally Distributed:

In regression set up, we consider X is known and Y is a random variable and conditionally given $X = x_0$, Y is distributed as normal. Now suppose that y and x are jointly distributed according to the bivariate normal distribution. Without getting into the density function it can be said that ρ is the correlation between X and Y.

The conditional distribution of y for a given value of x is normal with mean $E(Y|X) = \beta_0 + \beta_1 X$ and Variance σ^2 where β_0, β_1 and σ^2 are all function of mean, variance and correlation of the joint bivariate normal distribution. Maximum likelihood estimates under normality provides the same formula as least squares. Here we are trying make inferences regarding population correlation ρ.

The point estimate of ρ is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

It is often necessary and useful to test ρ against 0. i.e

$H_0 : \rho = 0 \ vs \ H_1 : \rho \neq 0$, test statistic $= t - stat = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Under H_0, the above statistic follows t-distribution with (n-2) degrees of freedom. Hence reject H_0 if $|t - stat| > t_{\alpha/2,n-2}$

One sided test against 0 can be performed similarly (look at the testing table for ρ in Lecture 3-2). This test is equivalent to the test for $\beta_1 = 0$. Both have the same test-statistic value though formula appears to be different.

The general inference procedure for ρ is little more complicated. It is based on the fact that for a reasonably large sample size ($n \geq 30$),

$$Z = arctanh(r) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \approx N\left(\frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right), (\sqrt{n-3})^{-2}\right)$$

Using the above fact, the tw0-sided test is performed as follows:

$H_0 : \rho = \rho_0 (\neq 0)\ vs\ H_1 : \rho \neq \rho_0$

test statistic $= Z - stat = \frac{arctanh\,(r) - arctanh\,(\rho_0)}{(\sqrt{n-3})^{-1}}$

Reject H_0 if $|Z - stat| > z_{\alpha/2}$

One sided hypothesis against a non-zero value can be performed similarly.

Using the same above transformation, $100(1-\alpha)\%$ confidence interval for ρ can be constructed as

$$tanh\Big(arctanh\,(r) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\Big) \le \rho \le tanh\Big(arctanh\,(r) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\Big)$$

where $\quad tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}.$

<u>Problem 2.10:</u> The weight and systolic blood pressure of 26 randomly selected males in the age group 25-30 are shown below (Summary only) Assume that weight and blood pressure (BP) are jointly normally distributed.

a. Find a regression line relating systolic blood pressure to weight.

b. Estimate the correlation coefficient.

c. Test the hypothesis that $\rho = 0$.

d. Test the hypothesis that $\rho = 0.6$.

e. Find a 95% CI for ρ.

$$\sum x = 4743, \sum y = 3786, \sum x^2 = 880545, \sum y^2 = 555802, \sum xy = 697076$$

To find the least square line (or estimated regression line) we start with the slope, (i.e $\hat{\beta}_1 = b_1$)

First, we need to find the average of X and Y data points as

$\bar{X} = \frac{\sum X}{n} = \frac{4743}{26} = 182.42, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{3786}{26} = 145.62,$

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{697076 - 26 \times 182.42 \times 145.62}{880545 - 26 \times 182.42^2} = \frac{6422.23}{15312.35} = 0.4194$$

and

$$\hat{\beta}_0 = b_o = \bar{Y} - b_1\bar{X} = 145.62 - 0.4194 \times 182.42 = 69.1044$$

Hence, the estimated regression line is $\hat{Y} = 69.1044 + 0.4194 \times X$

(b) The sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2).(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}} = \frac{6422.23}{\sqrt{(15312.35).(555802 - 26 \times 145.62^2)}} = 0.7735$$

(c) Testing against 0 using $\alpha = 0.05$ is

$H_0 : \rho = 0 \; vs \; H_1 : \rho \neq 0,$

test statistic $= t - stat = \frac{0.7735\sqrt{26-2}}{\sqrt{1-0.7735^2}} = 5.9786$

Reject H_0 if $|t - stat| > t_{0.05/2,26-2} = 2.064$

Rejection (Critical) Region for the test

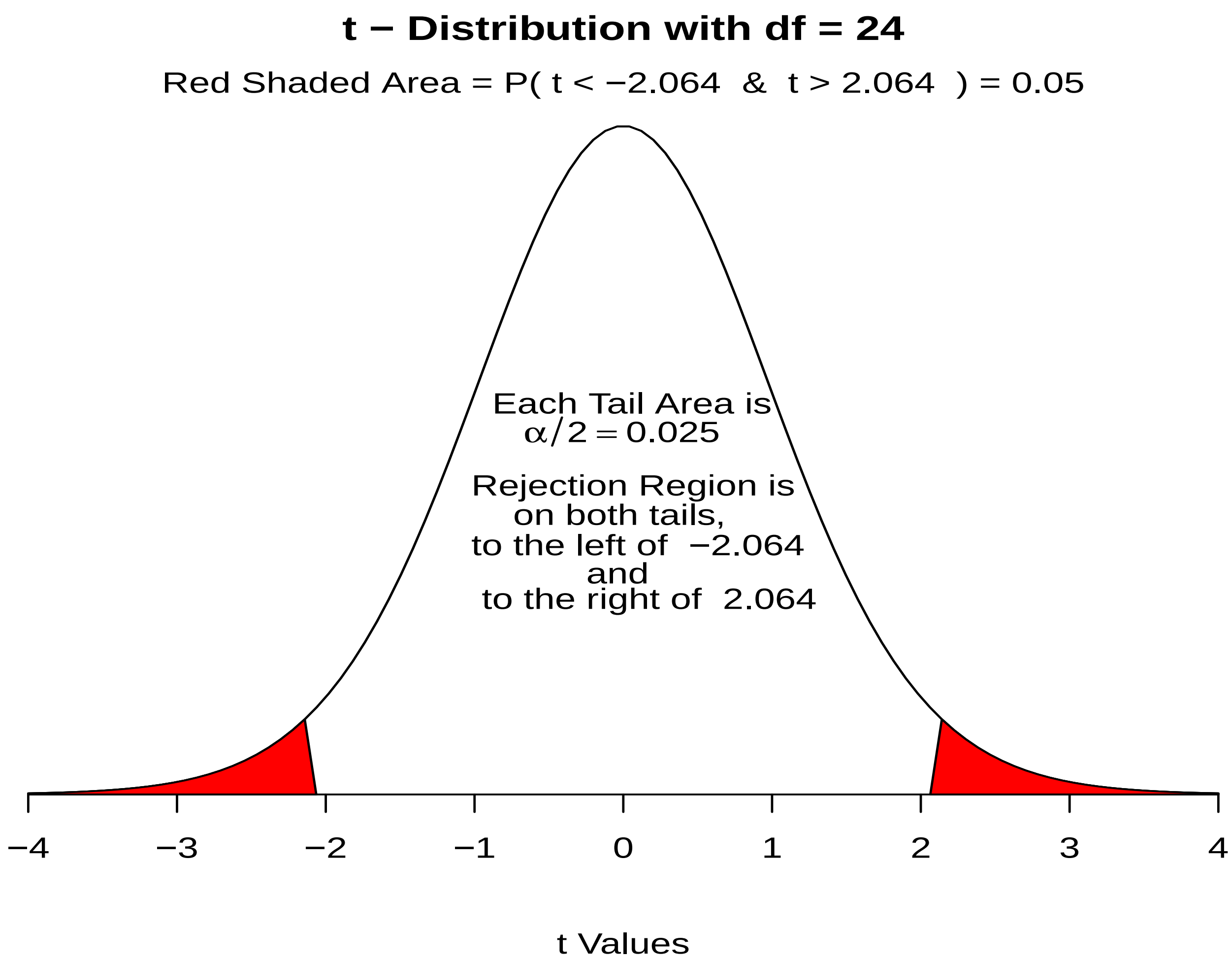

Figure 1: t Distribution with df=24 - Both Tail Area $\alpha = 0.05$

(d) To test against 0.6, we see that

$H_0 : \rho = 0.6 (\neq 0) \ vs \ H_1 : \rho \neq 0.6$

$$\text{test statistic} = Z - stat = \frac{arctanh\,(0.7735) - arctanh\,(0.6)}{(\sqrt{26-3})^{-1}}$$

$$= \frac{\frac{1}{2}\,ln\left(\frac{1+0.7735}{1-0.7735}\right) - \frac{1}{2}\,ln\left(\frac{1+0.6}{1-0.6}\right)}{(\sqrt{26-3})^{-1}}$$

$$= \frac{1.0290 - 0.6932}{(\sqrt{23})^{-1}} = 1.60$$

Reject H_0 if $\quad |Z - stat| > z_{0.05/2} = 1.96$

But $|Z - stat| = 1.60 \not> z_{0.05/2} = 1.96$, hence we do not reject null hypothesis. That is to say that there is not sufficient evidence to say that ρ is different from 0.6.

(e) To find a 95% confidence interval for ρ is

$$tanh\Big(arctanh\,(r) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\Big) \;\le\; \rho \;\le\; tanh\Big(arctanh\,(r) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\Big)$$

$$tanh\Big(arctanh\,(0.7735) - \frac{1.96}{\sqrt{26-3}}\Big) \;\le\; \rho \;\le\; tanh\Big(arctanh\,(0.7735) + \frac{1.96}{\sqrt{26-3}}\Big)$$

$$tanh\Big(1.0290 - 0.4087)\Big) \;\le\; \rho \;\le\; tanh\Big(1.0290 + 0.4087)\Big)$$

$$tanh\Big(0.6203\Big) \;\le\; \rho \;\le\; tanh\Big(1.4377\Big)$$

$$\frac{e^{0.6203} - e^{-0.6203}}{e^{0.6203} + e^{-0.6203}} \;\le\; \rho \;\le\; \frac{e^{1.4377} - e^{-1.1.4377}}{e^{1.4377} + e^{-1.4377}}$$

$$0.5513 \;\le\; \rho \;\le\; 0.8932$$

where $\quad tanh(u) = \frac{e^{u}-e^{-u}}{e^{u}+e^{-u}} \quad and \quad arctanh(u) = \frac{1}{2}\; ln\Big(\frac{1+u}{1-u}\Big).$

REGRESSION THROUGH THE ORIGIN

Some regression situations seem to imply that a straight line passing through the origin should be fit to the data. A no-intercept regression model often seems appropriate in analyzing data from chemical and other manufacturing processes. For example, the yield of a chemical process is zero when the process operating temperature is zero.

The "no-intercept" model assumes that $\beta_0 = 0$ which implies that the SLR model is

$$Y = \beta_1 \,.X + e$$

After going through the same least-square theory the estimate of β_1 (the slope) is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i \, y_i}{\sum_{i=1}^{n} x_i^2}$$

Hence the estimated regression line is $\hat{y} = \hat{\beta}_1 \,.x$

Estimate of σ (i.e standard error of estimate) is

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i.y_i}{n-1}}$$

Remark: Note that the divisor is (n-1) as there is only one β in the model.

Making the same assumption about the errors i.e $e_i \sim N(0, \sigma^2)$ we can obtain all the confidence and prediction intervals and perform the inferences as in the SLR model with intercept.

For no intercept model the $100(1-\alpha)\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \;\pm\; t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \quad \text{where} \quad S_{xx} = \sum_{i=1}^{n} x_i^2$$

.

Similarly $100(1-\alpha)\%$ confidence interval for $E(Y|X = x_0)$ (i.e True average value of Y for given value of X $= x_0$) is

$$\hat{\beta}_1 x_0 \ \pm \ t_{\alpha/2,n-1}\,\hat{\sigma}\sqrt{\frac{x_0^2}{S_{xx}}}$$

Remark: Note that the above interval is x_0 times the interval for β_1. As a result length of the interval for $E(Y|X = x_0)$ when $x_0 = 0$ is 0 which is vary different from SLR model with intercept.

By the same theory, $100(1-\alpha)\%$ prediction interval for a future value of Y at X $= x_0$ is

$$\hat{\beta}_1 x_0 \ \pm \ t_{\alpha/2,n-1}\,\hat{\sigma}\sqrt{1+\frac{x_0^2}{S_{xx}}}$$

In the no-intercept case the fundamental analysis-of-variance identity becomes

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i{}^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{where} \quad SST = \sum_{i=1}^{n} y_i^2, \quad SSR = \sum_{i=1}^{n} \hat{y_i}^2, \text{ and } \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

So the no-intercept model analogue for R^2 would be,

$$R_0^2 = \frac{\sum_{i=1}^{n} \hat{y_i}^2}{\sum_{i=1}^{n} y_i^2} = \frac{SSR}{SST}$$

Note that in the intercept model R^2 indicates the proportion of variability around $\bar{y}$ explained by regression. But in the model w.o intercept R_0^2 statistic indicates the proportion of variability around the origin (zero) accounted for by regression. We occasionally find that R_0^2 is larger than R^2 even though the residual mean square (which is a reasonable measure of the overall quality of the fit) for the intercept model is smaller than the residual mean square for the no-intercept model. This arises because is computed using uncorrected sums of squares.

Problem - 2.11: Consider the weight and blood pressure data in Problem 2.10.(above). Fit a no-intercept model to the data and compare it to the model obtained in Problem 1. Which model would you conclude is superior?

Solution: The previous formulas can be used to find out the $\hat{\beta}_1$ and other terms for the model without intercept (Solve it on your own). Plots and SAS outputs are attached at the end.

With Intercept Model: $R^2 = 59.83\%$ MSE $= 75.36$ $\hat{\sigma} = 8.68$

Without Intercept Model: $R^2 = 99.29\%$ MSE $= 158.71$ $\hat{\sigma} = 12.6$

Though R^2 for the intercept model (which is 59.83%) is much lower that the R^2 for the without intercept model (which is 99.29%), but comparing R^2 for these two models is not a justified process as denominators of R^2 for formulas are not same. But looking at the mean square error we see that MSE and then the standard error is much better for the "with intercept" model. Hence "with intercept" model is preferred.

GRAPH of LEAST SQUARE LINES WITH AND WITHOUT INTERCEPT

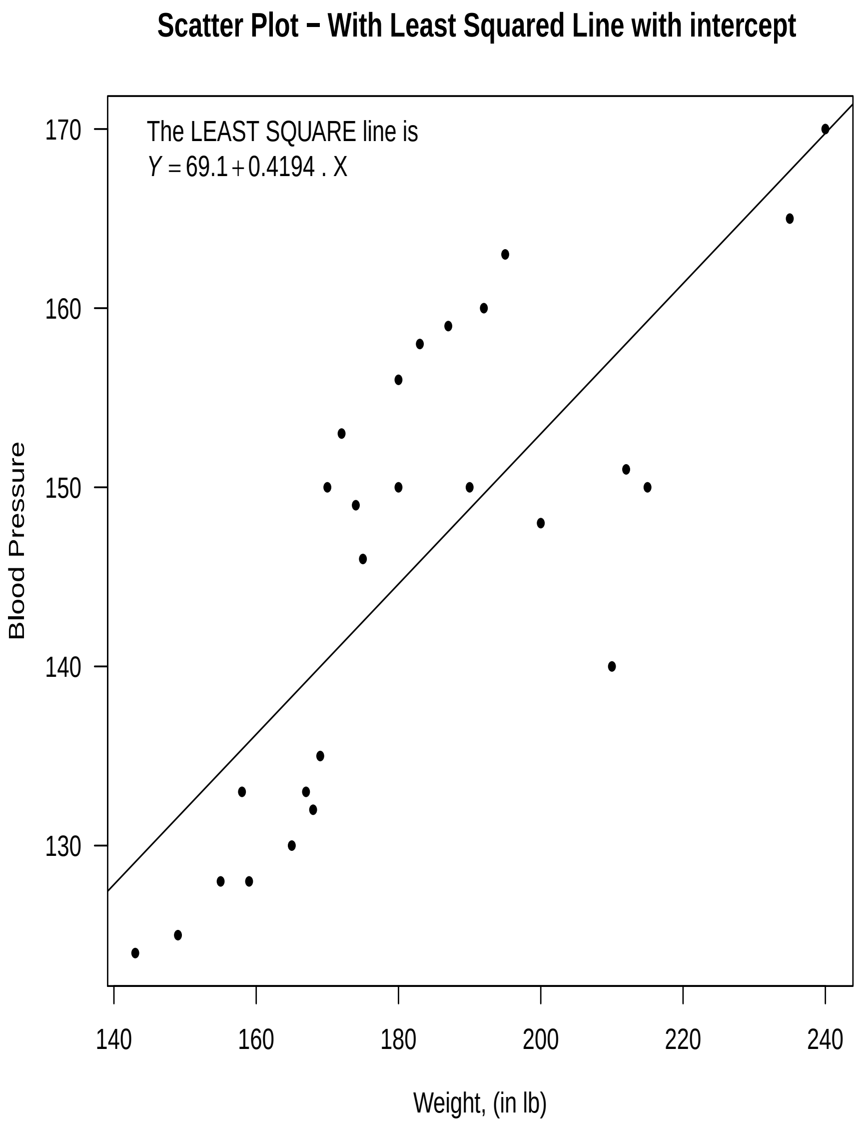

Figure 2: LS-Line with intercept

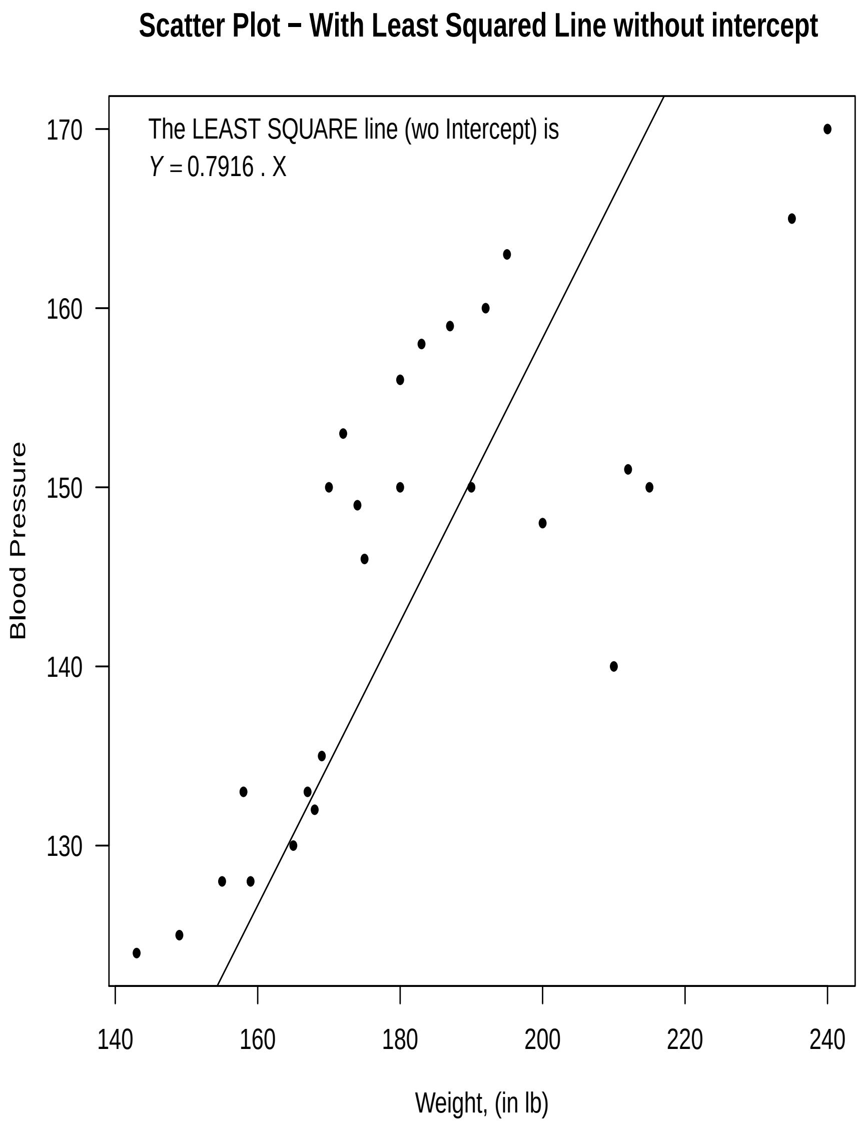

Figure 3: LS-Line without intercept

REGRESSION WITH INTERCEPT

The REG Procedure
Model: MODEL1
Dependent Variable: sysbp sysbp

Number of Observations Read	26
Number of Observations Used	26

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	2693.58122	2693.58122	35.74	<.0001
Error	24	1808.57262	75.35719		
Corrected Total	25	4502.15385			

Root MSE	8.68085	**R-Square**	0.5983
Dependent Mean	145.61538	**Adj R-Sq**	0.5815
Coeff Var	5.96149		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	69.10437	12.91013	5.35	<.0001
weight	weight	1	0.41942	0.07015	5.98	<.0001

REGRESSION WITHOUT INTERCEPT

The REG Procedure
Model: MODEL1
Dependent Variable: sysbp sysbp

Number of Observations Read	26
Number of Observations Used	26

Note: No intercept in model. R-Square is redefined.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	551834	551834	3477.06	<.0001
Error	25	3967.68174	158.70727		
Uncorrected Total	26	555802			

Root MSE	12.59791	**R-Square**	0.9929
Dependent Mean	145.61538	**Adj R-Sq**	0.9926
Coeff Var	8.65149		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
weight	weight	1	0.79164	0.01343	58.97	<.0001