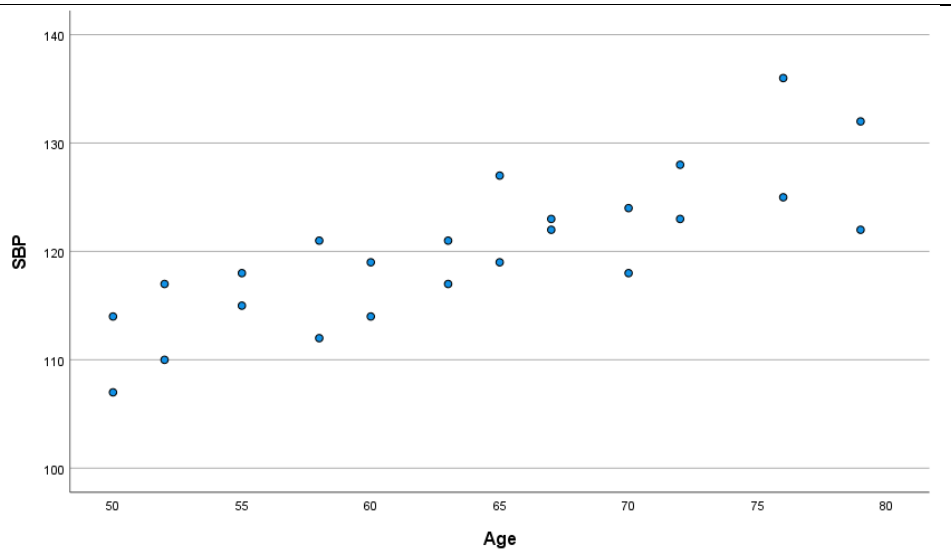
Practice Problem Topic 4: Simple Linear Regression

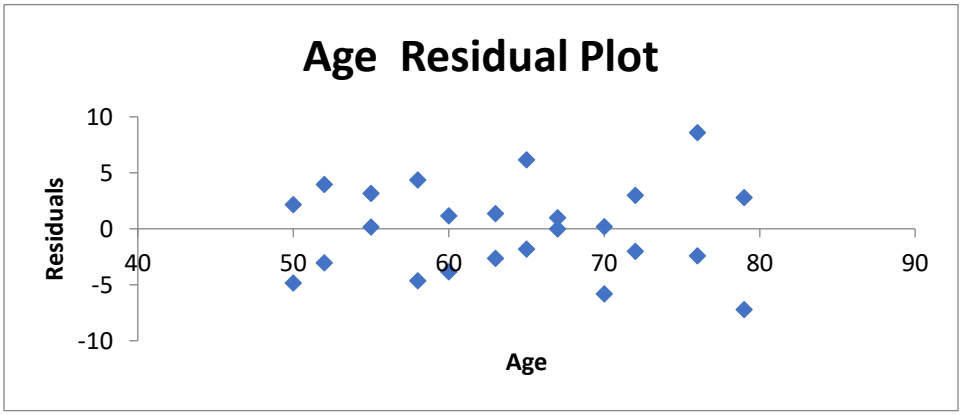
Continuing on the Theme of Different Factors/Variables Affecting Blood Pressure

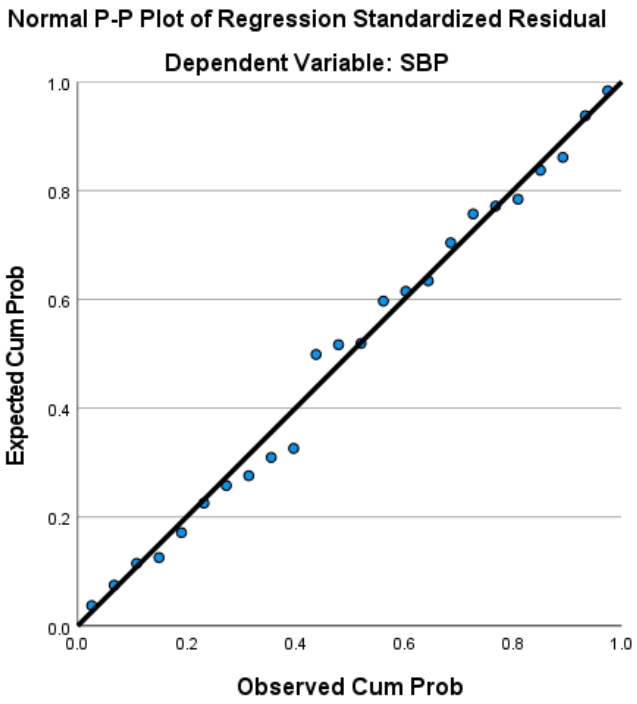
Does age (explanatory variables) have an effect on blood pressure (response variable)?

A sample of 24 people between 50 and 79 years of age was randomly selected from a general population who are not defined by any health issues. The age (in years) and systolic blood pressure (SBP) (in mmHg) of each person was recorded obtaining data as shown below. Use the analysis shown in the SPSS output below (with some values missing), to answer the questions. All assumptions are met for the required analysis. [This is a hypothetical data set, but it is patterned after real data collected by Canadian Health Measures Survey, 2009 to 2011.]

Age (Years)	SBP (mmHg)
50	107
52	110
55	115
58	112
60	114
63	121
65	127
67	122
70	124
72	128
76	136
79	132
50	114
52	117
55	118
58	121
60	119
63	117
65	119
67	123
70	118
72	123
76	125
79	122







Descriptive Statistics			
	Mean	Std. Deviation	N
SBP	120.17	6.722	24
Age	63.92	9.103	24

Model Summary[b]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.810[a]	.656	.641	4.028
a. Predictors: (Constant), Age				
b. Dependent Variable: SBP				

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	682.305	1	682.305	42.044	1.6E-06
	Residual	357.028	22	16.229		
	Total	1039.333	23			
a. Dependent Variable: SBP						
b. Predictors: (Constant), Age						

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	81.923	5.955		13.757	<.001	69.573	94.273
	Age	.598	.092	.810	6.484	1.6E-06	.407	.790

a. Dependent Variable: SBP

(a) Although the start of the question says, "assume that all the assumptions are met for the required analysis" and SPSS SLR output is provided, explain in detail why simple linear regression (SLR) would be the analysis for relating age to SBP and verify that all the assumptions are met.

- There are two variables, and both are quantitative. Logically, the explanatory variable would be age and the purpose of the research problem is to see if age affects (or is related to) SBP.
- The response variable is a continuous, quantitative variable, that is, SBP.
- **Purpose** of the study: To test whether there is a relationship between age and SBP, that is, does age affect SBP.

- **Assumptions:**
  1. The relationship between age and SBP is linear as shown in the XY graph.
  2. The standard deviation deviations of the SBP responses are approximately equal for each value of the predictor variable (age) as indicated by the residual plot.
  3. The Normal P-P plot forms approximately a straight line indicating that the SBP responses are approximately normal for each value of the predictor variable.
  4. The XY graph and the residual plot do not show any serious outliers.
  5. We can assume the observations of the response variable (SBP) are independent of each other and are random.

>>>>>>>>>>

(b) According to the regression model, what would you predict to be the SBP of a person who is 74 years old?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{y} = 81.923 + 0.598x$$
$$\hat{y} = 81.923 + 0.598(74) = 126.175$$

The predicted SBP of a person who is 74 years old is 126.175.

(c) According to the regression model, what would you predict to be the SBP of a person who is 25 years old? Would this be a reliable prediction?

$$\hat{y} = 81.923 + 0.598(25) = 96.873$$

No, because that is beyond the observed range of x, which is from 50 to 79 years of age.

.

(d) The data set above shows that a person who was 79 years old had a SBP reading of 122 mmHg. According to this regression model, what was the residual (error) of that observation?

$$\hat{y} = 81.923 + 0.598(79) = 129.165$$

Residual (error) = (Actual – Predicted)

$$= (y_i - \hat{y}_i) = 122 - 129.165 = -7.165 \text{ mmHg}$$

(e) What percentage of variability in SBP is explained by age in this regression model?

$$R^2 = \frac{SS_{REGR}}{SS_{TOTAL}} = \frac{682.305}{1039.333} = 0.656484$$

Therefore, 65.65% of the variability in SBP is explained by age in this regression model.

(f) Calculate the linear correlation coefficient for the relationship between age and SBP. What would be the exact P-value of the correlation coefficient if you were doing a two-tailed test? If you were doing a one-tailed test.

Since the slope of the regression line is positive, calculate the positive square root of the coefficient of determination.

$$r = +\sqrt{R^2} = +\sqrt{0.656484} = +0.8102$$

For a two-tailed test for the correlation coefficient, the exact P-value **equals** the exact P-value for the ANOVA F-test = 1.6×10^{-6}

For a one-tailed test for the correlation coefficient, the exact P-value is **half** of the exact P-value for the ANOVA F-test = $(1.6 \times 10^{-6})/2 = 8.0 \times 10^{-7}$

(g) What is the standard error of the model?

$$SS_{Error} = SS_{TOTAL} - SS_{REGR} = 1039.333 - 682.305 = 357.028$$

$$\hat{\sigma} = \sqrt{MS_{Error}} = \sqrt{\frac{SS_{Error}}{n-2}} = \sqrt{\frac{357.028}{24-2}} = \sqrt{16.228545} = 4.02847$$

(h) At the 5% significance level, carry out the most appropriate test, using the F-distribution, to determine whether the overall model is significant, that is, whether there is a relationship between age and SBP.

H_0: $\beta_1 = 0$ (There is no relationship between age and SBP.)
H_a: $\beta_1 \neq 0$ (There is a relationship between age and SBP.)

$$SS_{Error} = SS_{TOTAL} - SS_{REGR} = 1039.333 - 682.305 = 357.028$$

$$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{MS_{REGR}}{MS_{ERROR}}$$

$$= \frac{682.305/1}{357.028/(24-2)} = \frac{682.305}{16.228545} = 42.0435$$

$df = (1, n-2) = (1, 24-2) = (1, 22)$ P-value is: P < 0.001

There is extremely strong evidence against H_0. Since P-value ≤ α, reject H_0

At the 5% significance level, the overall model is significant, that is, whether there is sufficient evidence that there is a relationship between age and SBP.

(i) At the 5% significance level, carry out the most appropriate test to test if there is a positive linear relationship between age and SBP. Give the P-value based on the appropriate statistical table and the exact P-value.

H_0: $\beta_1 = 0$ (There is no relationship between age and SBP.)
H_a: $\beta_1 > 0$ (There is a positive relationship between age and SBP.)

Estimate: $\hat{\beta}_1 = 0.598$ SE(Estimate): $SE(\hat{\beta}_1) = 0.092$

$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.598}{0.092} = 6.500$ Note that there is just a rounding error.

$df = n - 2 = 24 - 2 = 22$

Based on the t-table: P-value: P < 0.0005

For a one-tailed t-test, the exact P-value is **half** of the exact P-value for the ANOVA F-test
= $(1.6 \times 10^{-6})/2 = 8.0 \times 10^{-7}$

There is extremely strong evidence against H_0
Since P-value < α, reject H_0

At the 5% significance level, the data provide sufficient evidence that there is a positive relationship between age and SBP.

(j) Compare your answers in parts (h) and (i) with respect to the value of the test statistic, the df and the exact P-value.

The *t*-statistic $t = +\sqrt{F} = \sqrt{42.0435} = 6.484$
[Any difference is due to rounding.]

The denominator df of the F-test = the df of the t-test = 22

For a one-tailed t-test, the exact P-value is **half** of the exact P-value for the ANOVA F-test
= $(1.6 \times 10^{-6})/2 = 8.0 \times 10^{-7}$

(k) Calculate a 95% confidence interval for the slope of the regression line for the relationship between age and SBP. Then, based on this confidence interval, can you conclude that the slope is significant. Explain your answer.

For a 95% confidence level, α = 1 – 0.95 = 0.05
At df = 24 – 2 = 22, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.074$

Estimate: $\hat{\beta}_1 = 0.598$

SE(Estimate): $SE(\hat{\beta}_1) = 0.092$

Note: if it was not given, you can calculate:

$$S_{xx} = (n-1)s_x^2 = (24-1)(9.103)^2 = 1905.886007$$

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{4.02847}{\sqrt{1905.886007}} = 0.092277$$

$$\hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$$
$$0.598 \pm 2.074 \times 0.092277$$
$$0.598 \pm 0.19138$$
$$(0.407, 0.789)$$

Interpretation: We can be 95% confident that the slope of the regression line for the relationship between age and SBP is between 0.407 and 0.789 mmHg/year.

Based on this confidence interval, the slope for SBP against age is significant since this confidence interval for the slope does <u>not</u> contain 0.

(l) Calculate a 99% confidence interval for the slope of the regression line for the relationship between age and SBP.

For a 99% confidence level, α = 1 – 0.99 = 0.01
At df = 24 – 2 = 22, $t_{\alpha/2} = t_{0.01/2} = t_{0.005} = 2.819$

$$\hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$$
$$0.598 \pm 2.819 \times 0.092277$$
$$0.598 \pm 0.26013$$
$$(0.338, 0.858)$$

Interpretation: We can be 99% confident that the slope of the regression line for the relationship between age and SBP is between 0.338 and 0.858 mmHg/year.

(m) Calculate a 99% confidence interval for the mean SBP for people who are 74 years of age.

Parameter: $\mu(Y \mid x = 74)$

Estimate: $\mu(Y \mid x = 74) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 81.923 + 0.598(74) = 126.175$

[Note this is from part (b)]

For a 99% confidence level, α = 1 – 0.99 = 0.01

At df = 24 – 2 = 22, $t_{\alpha/2} = t_{0.01/2} = t_{0.005} = 2.819$

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$$

$$S_{xx} = (n-1)s_x^2 = (24-1)(9.103)^2 = 1905.886007$$

$$126.175 \pm 2.819 \times 4.02847 \sqrt{\frac{1}{24} + \frac{(74-63.92)^2}{1905.886007}}$$

$$126.175 \pm 2.819 \times 4.02847 \times 0.308186$$

$$126.175 \pm 3.4998$$

$$(122.675, 129.675)$$

Conclusion: It is estimated with 99% confidence that the mean SBP for people who are 74 years is between 122.675 and 129.675 mmHg.

(n) Calculate a 99% prediction interval for the SBP for a person who is 74 years of age (or a 99% confidence interval for all single observations of a person who is 74 years of age).

Parameter: $\mu(Y \mid x = 74)$

Estimate: $\mu(Y \mid x = 74) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 81.923 + 0.598(74) = 126.175$

For a 99% confidence level, α = 1 – 0.99 = 0.01

At df = 24 – 2 = 22, $t_{\alpha/2} = t_{0.01/2} = t_{0.005} = 2.819$

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$$

$$S_{xx} = (n-1)s_x^2 = (24-1)(9.103)^2 = 1905.886007$$

$$126.175 \pm 2.819 \times 4.02847 \sqrt{1 + \frac{1}{24} + \frac{(74-63.92)^2}{1905.886007}}$$

$$126.175 \pm 2.819 \times 4.02847 \times 1.046412$$

$$126.175 \pm 11.8833$$

$$(114.292, 138.058)$$

Conclusion: It is estimated with 99% confidence that the SBP for a person who is 74 years of age is between 114.292 and 138.058 mmHg.

>>>>>>>>>>