Chapter VI

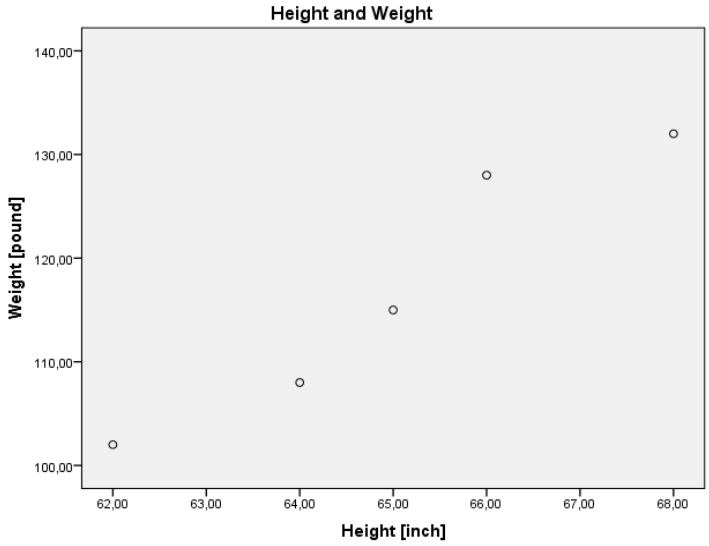
Correlation
and
Simple Linear Regression Analysis

Solutions

Part I:

1.

1.



2.
There appears to be a linear relation between weight and height.

3.
Many different straight lines can be drawn to provide a linear approximation of the relationship between height and weight. In part 4 we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

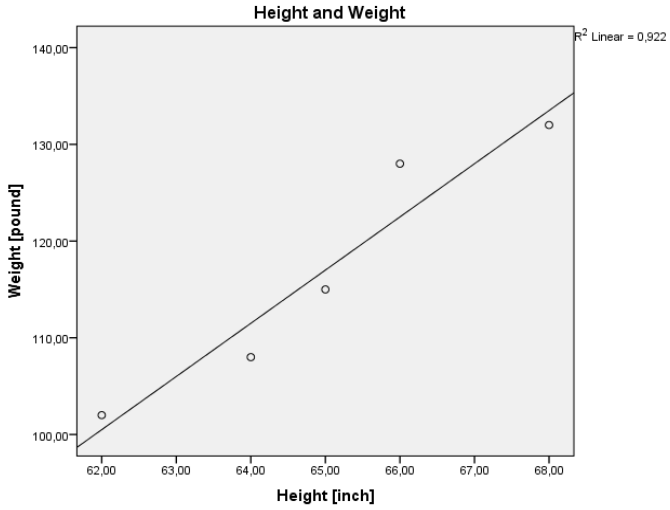
4.

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$
68	132	4624	8976
64	108	4096	6912
62	102	3844	6324
65	115	4225	7475
66	128	4356	8448
325	585	21145	38135

$$\begin{cases} 5b_0 + 325b_1 = 585 \\ 325b_0 + 21145b_1 = 38135 \end{cases} \Rightarrow \quad b_0 = -240.5, \quad b_1 = 5.5$$

$$y^* = -240.5 + 5.5x$$



5.

$y^*(63) = 106$ pounds.

2.

1.

Working Table

x_i	y_i	y_i^*	$(y_i^* - \bar{y})^2$	$(y_i - y_i^*)^2$
2.6	3300	3301.36	121549.85	1.84960000
3.4	3600	3766.24	13511.7376	27635.7376
3.6	4000	3882.46	54037.6516	13815.6516
3.2	3500	3650.02	0.00040000	22506.0004
3.5	3900	3824.35	30397.9225	5722.9225
2.9	3600	3475.69	30383.9761	15452.9761
	21900		249881.138	85135.1378

$$\bar{y} = \frac{21900}{6} = 3650$$

$$SSR = 249881.138, \quad SSE = 85135.1378, \quad SST = SSR + SSE = 335016.2758.$$

2.

$$r^2 = \frac{249881.138}{335016.2758} = 0.7458776067 \approx 0.746$$

About 74.6% of the monthly salary is explained by the GPA. It is a relatively good fit.

3.

$$r = +\sqrt{0.7458776067} = 0.8636420594 \approx 0.8636.$$

There is a direct relationship between the GPA and the monthly salary.

4.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{85135.1378}{6-2}} = 145.8896311, \qquad s_{b_1} = \frac{145.8896311}{\sqrt{0.74}} = 169.5932513$$

$$t_{stat} = \frac{581.1}{169.5932513} = 3.426433514$$

Step 3:
Because of

$$t_{stat} = 3.426433514 > 2.776 = t_{4;0.05},$$

we reject H_0.

(Or because of $p - value = 0.0266 \le 0.05 = \alpha$, we reject H_0.)

5.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$MSR = \frac{249881.138}{1} = 249881.138$$

$$MSE = s^2 = \frac{85135.1378}{6-2} = 21283.75$$

$$F = \frac{249881.138}{21283.75} = 11.74046575$$

Step 3:

Because of

$$F = 11.74046575 \ge 7.71 = F_{0.05;1;4},$$

H_0 will be rejected.

6.

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	29846.86	1	249881.138	11.74
Error	85135.14	4	21283.79	
Total	335000	5		

7.

$$s = 145.89, \quad \bar{x} = 3.2, \quad \sum_{i=1}^{6} (x_i - \bar{x})^2 = 0.74 ,$$

$$y^*(3) = 3533.8, \quad s_{y_p^*} = 145.89 \cdot \sqrt{\frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 68.54$$

$$\beta_1 \in \left[3533.8 - 2.776 \cdot 68.54,\ 3533.8 + 2.776 \cdot 68.54\right] = \left[3343.53,\ 3724.07\right].$$

8.

$$s_{ind} = 145.89 \cdot \sqrt{1 + \frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 161.19$$

$$y_p \in \left[3533.8 - 2.776 \cdot 161.19,\ 3533.8 + 2.776 \cdot 161.19\right] = \left[3086.34,\ 3981.26\right]$$

Part II: *SPSS*

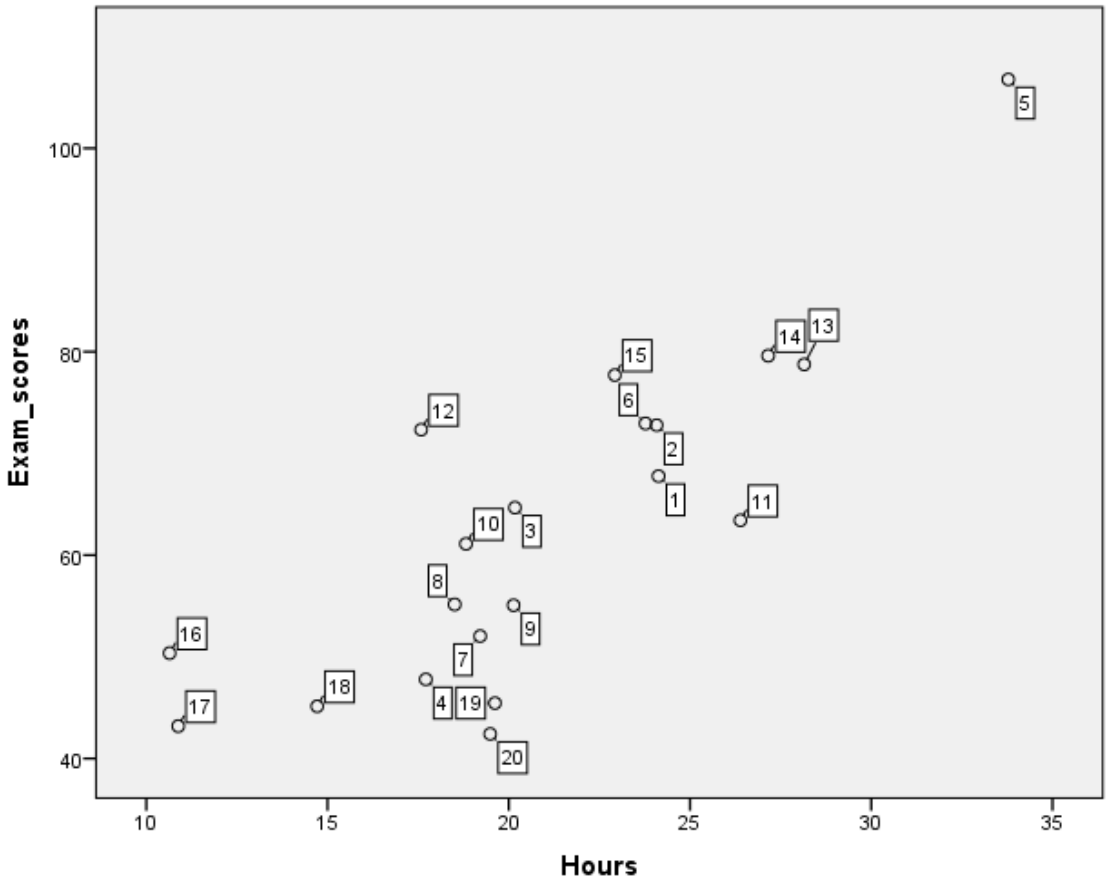
1.

- ***Graphs -> Legacy Dialogs -> Scatter/Dot...***
- Choose ***Simple Scatter -> Define***.
- Transfer *Exam_scores* to ***YAxis*** and *Hours* to ***X Axis***.

  OK

- Double click on the scatter plot

- ***Elements -> Show Data labels -> Close***



It *appears* that the more time students study, the higher the exam scores and the relation looks linear.

- ***Analyze -> Regression -> Linear...***

- From the list on the left, select the variable *Exam_scores* as ***Dependent*** and the variable *Hours* as ***Independent(s)***.

  OK.

Output and Interpretation:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3450,712	1	3450,712	38,959	,000[b]
	Residual	1594,308	18	88,573		
	Total	5045,020	19			

a. Dependent Variable: Exam_scores

b. Predictors: (Constant), Hours

We have $p-value = 0 < 0.05 = \alpha$. Therefore, the null hypothesis that there is no relationship between the hours of study and the exam scores will be rejected.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,827[a]	,684	,666	9,411

a. Predictors: (Constant), Hours

The correlation coefficient 0.827 tells us that the study hour is positively correlated with exam scores and the relation is pretty strong.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	12,762	8,276		1,542	,140
	Hours	2,391	,383	,827	6,242	,000

a. Dependent Variable: Exam_scores

Finally, we obtain the simple linear regression

$$y^* = 12.762 + 2.391x$$

2.

1.

See the data file *carprice.sav*.

2.

- Select

  Graphs -> Legacy Dialogs -> Scatter/Dot...

- Select ***Simple Scatter***, and then click the ***Define*** button.
- Move the variable *price* to ***Y Axis*** and the variable *age* to ***X Axis***.

  Click ***Titles***. Choose the title "Car Prices by Age".

  Continue.

- ***OK.***

3.

- Select

  Analyze -> Correlate -> Bivariate

- Select *age* and *price* as ***Variables***.

  Select ***Pearson*** as the ***Correlation Coefficients***.

  OK.

Output and Interpretation:

Correlations

		age	price
age	Pearson Correlation	1	-,9679**
	Sig. (2-tailed)		,000
	N	10	10
price	Pearson Correlation	-,968**	1
	Sig. (2-tailed)	,000	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

The coefficient correlation is -0.9679. This value suggests a strong negative linear correlation. The data points should be clustered closely about a negatively sloping regression line. This is consistent with the graph obtained above.

4. – 7.

Since we eventually want to predict the price of a 4-year old car (see parts 11 – 14), enter the number "4" in the *age* variable column of the ***Data*** window after last row. Enter a ".". For the corresponding *price* variable value (this lets *SPSS* know that we want a prediction for this value and not to include the value in any other computations)

- Select

  Analyze -> Regression -> Linear…

- Select *price* as the ***Dependent*** variable and *age* as the ***Independent*** variable.

  Click ***Statistics***…

- Select ***Estimates*** and ***Confidence Intervals*** for the ***Regression Coefficients*** ,

  Select ***Model fit*** to obtain r^2 ,

  Continue

- Click ***Plots…***

- Select ***Normal probability plot*** of the residuals,

  Continue.

- Click ***Save…***

- Select ***Unstandardised*** in ***Predicted Values*** .

  Select ***Unstandardised*** and ***Studentized*** in ***Residuals***,

  Select ***Mean*** (to obtain a confidence interval…output in the ***Data*** window) and ***Individual*** in ***Prediction Intervals***.

  Continue

- ***OK.***

Output and Interpretation:

4.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	291,602	11,433		25,506	,000	265,238	317,966
	age	-27,903	2,563	-,968	-10,887	,000	-33,813	-21,993

a. Dependent Variable: price

$$y^* = 291.602 - 27.903x$$

From within the ***Output*** window, double-click on the scatterplot to enter ***Chart Editor*** mode. From the ***Elements*** menu, select ***Fit Line at Total***. Close the ***Close*** box.

6.

There do not appear to be any points that lie far from the cluster of data points or far from the regression line; thus there are no possible outliers or influential observations.

7.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,968[a]	,9368	,929	14,24653

a. Predictors: (Constant), age

b. Dependent Variable: price

The coefficient of determination is 0.9368. Therefore, about 93.68% of the variation in the price data is explained by age. The regression equation appears to be very useful for making predictions.

8.

The residuals and standardized values (as well as the predicted values, the confidence interval points, and the prediction interval endpoints) can be found in the ***Data*** window.

We now create residual plots:

- Select

  Graphs -> Legacy Dialogs -> Scatter/Dot…

- Select

  Simple Scatter -> Define

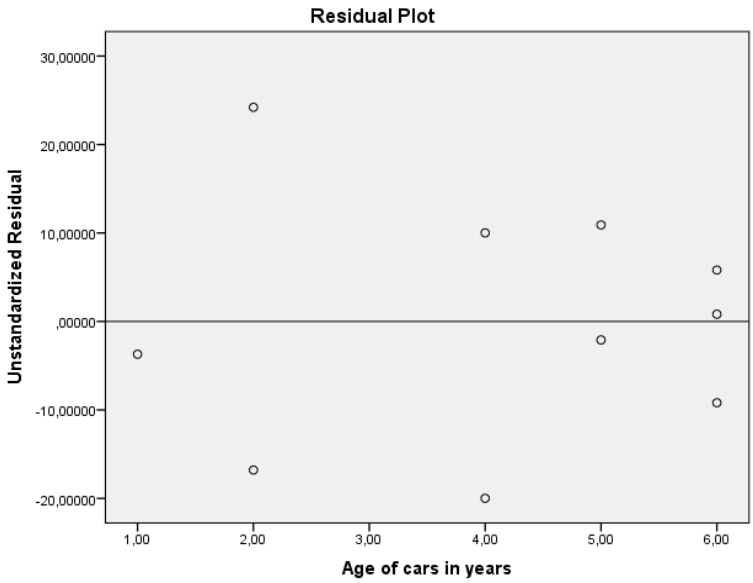
  Move the residual ***RES_1*** to ***Y Axis*** and *age* to ***X Axis***.

  Click ***Titles*** to enter "Residual Plot**"** as the title for your graph.

  Continue

- ***OK.***

- Double-click the resulting graph in the ***Output*** window.
- Select ***Options -> Y Axis Reference Line***.
- Select the ***Reference Line*** tab in the ***Properties*** window, add ***Position*** of line 0, and click ***Apply***.
- Click the ***Close*** box to exit the ***Chart Editor***.



- Select

  Graphs -> Legacy Dialogs -> Scatter/Dot…

- Select

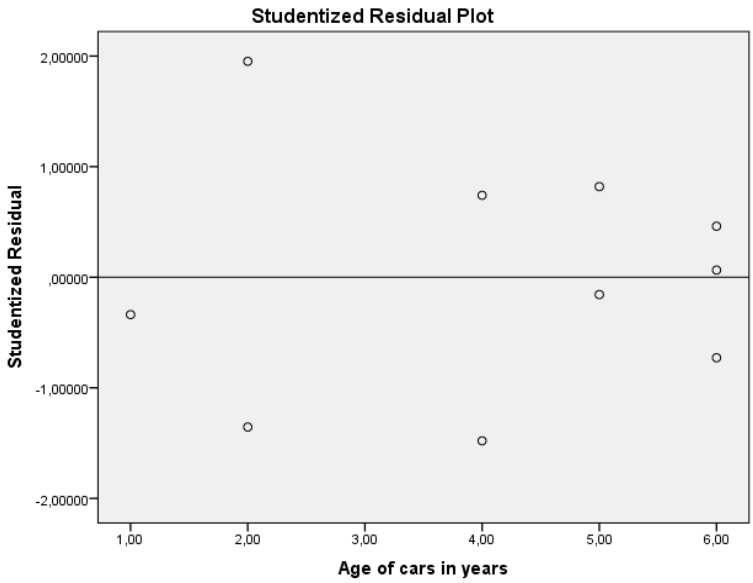
  Simple Scatter -> Define

  Move the residual ***SRE_1*** to ***Y Axis*** and *age* to ***X Axis***.

  Click ***Titles*** to enter "Studentized Residual Plot**"** as the title for your graph.

  Continue

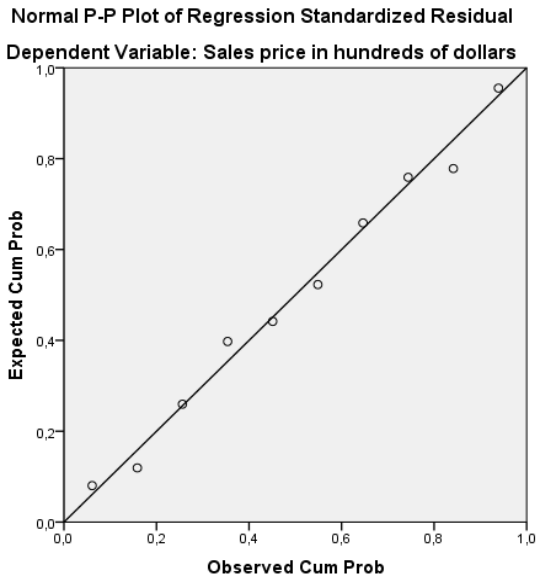
- ***OK.***
- Double-click the resulting graph in the ***Output*** window.
- Select ***Options -> Y Axis Reference Line***.
- Select the ***Reference Line*** tab in the ***Properties*** window, add ***Position*** of line 0, and click ***Apply***.
- Click the ***Close*** box to exit the ***Chart Editor***.



If 2 and/or -2 are in the range covered by the y-axis, repeat the last steps to add a reference line at 2 and -2 (see the last plot; any points that are not between these lines are considered potential outliers.

If 3 and/or -3 are in the range covered by the y-axis, repeat the last steps to add a reference line at 3 and -3; any points that are not between these lines are considered potential outliers.

To assess the normality of the residuals, consult the ***P-P Plot***:



The residual plot shows a random scatter of the points (independence) with a constant spread (constant variance).

The studentized residual plot shows a random scatter of the points (independence) with a constant spread (constant variance) with no values beyond the ± 2 standard deviation reference (no outliers).

The normal probability plot of the residuals shows the points close to a diagonal line. Therefore, the residuals appear to be approximately normally distributed.

Thus, the assumptions for regression analysis appear to be met.

9 -10.

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)	291,602	11,433		25,506	,000	265,238	317,966
	Age of cars in years	-27,903	2,563	-,968	-10,887	,000	-33,813	-21,993

a. Dependent Variable: Sales price in hundreds of dollars

$$p-value = 0.000 < 0.05 = \alpha$$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24057,891	1	24057,891	118,533	,000[b]
	Residual	1623,709	8	202,964		
	Total	25681,600	9			

a. Dependent Variable: Sales price in hundreds of dollars

b. Predictors: (Constant), Age of cars in years

$$F = 118.533 > 0.00000448 = p-value$$

Therefore, at 0.05 level of significance, there exists enough evidence to conclude that the slope of the population regression line is not zero and, hence, that age is useful as a predictor of price for cars.

We are 95% confident that the slope of population regression line is somewhere between -33813 and -21993.

In other word, we are 96% confident that for every year older the cars get, their average price decreases between $3.3812946 and $2.1992880.

11.
The point estimate (*PRE_1)* is $17999.0291.

12.
We are 95% confident that the mean sales price of all 4-year old cars is somewhere between $16958.46 (*LMCI_1*) and $19039.60 (*UMCI_1*).

13.
The predicted sales price is $17999.0291.

14.
We are 95% certain that the individual sales price of this particular person will be somewhere between $14552.9173 (*LICI_1*) and $21445.1410 (*UICI_1*)

(Last revised: 09.12.2019)