Helen of Troy Example: MRC04

Bumble and Bimble argued over the possible relationship between physical attractiveness and intelligence. Bumble thought the relationship must be negative, because he knew examples of the stereotypical ‘dumb hunky jock’ and ‘air-headed blonde bombshell,’ but Bimble said she knew a lot of smart, attractive people. B&B decided to collect data to address the question.

Measurement was their first challenge. Intelligence was easy - they found standard measures of IQ which they could use to measure intelligence. But standardized measures of physical attractiveness are not as readily available. Bimble decided to use a panel of her friends to make ratings of attractiveness on a 1000 point scale. Bumble did some Google research and he found an ancient scale of beauty in units of milliHelens. This scale is based on the legendary Helen of Troy who had “the face that launched a thousand ships.” Thus, a milliHelen is the amount of beauty required to launch just one ship!

Bimble went to Muscle Beach and collected data on ten contestants for Mr. Muscle, while Bumble went to the same beach to a beauty contest for the title of Miss Mussel. Amazingly, the data they each collected on ten contestants were exactly the same (as shown in the table).

Contestant	X (Looks)	Y (IQ)
A	380	91
B	500	107
C	440	86
D	460	102
E	600	117
F	530	110
G	410	111
H	470	99
I	480	98
J	570	119
Mean	484.0	104.0
SD	68.5	10.8

Bumble and Bimble remembered the important first step for any data analysis – LOOK at your data. They used SPSS to generate a diagnostic plot to see if linear regression might be a reasonable model. Then they used SPSS to carry out a regression analysis whereby they predicted appearance (Y) using IQ (X) as a predictor.

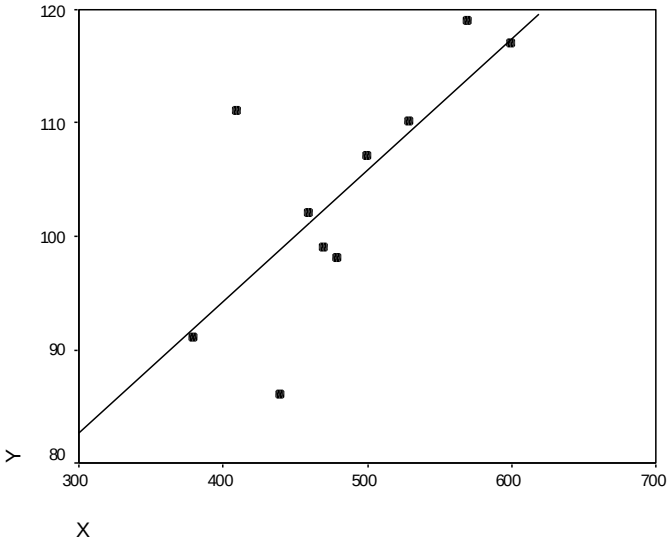
The SPSS output is attached. What questions would you pose of the data? For each of the questions I posed, see if you can find and interpret relevant information from the output before reading my answers.

```
GRAPH
  /SCATTERPLOT(BIVAR)=x WITH y
  /MISSING=LISTWISE .
```

Double-click on the scatterplot to open the Chart Editor. Click Element, Fit Line at Total, select Linear, Apply, to generate the regression line.

Graph

Q1:
Does the bivariate distribution of X and Y values shown in the figure satisfy assumptions for regression analysis? In particular, are there outliers that will distort findings? Are the data points distributed reasonably normally around the regression line? Is the variance of points around the line about the same for all values of X? Is the relationship reasonably linear?



A1:
The data set is very small so findings are unstable, but there is nothing in the figure to suggest that the assumptions needed for regression analysis have been violated substantially.

B&B decided to proceed with a regression analysis to address their research questions.

Here is their regression syntax, with output attached. A summary of key statistics is in the table.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Y
  /METHOD=ENTER X
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(3).
```

$r_{xy} = .736$ (beta in attached SPSS)
$s_y = 10.78$
$s_x = 68.508$
$\bar{y} = 104.0$
$\bar{x} = 484.0$
$\hat{y}_i = a + bx_i$
$b = r\frac{s_y}{s_x} = (.736)(\frac{10.78}{68.508}) = .1158$
$a = \bar{y} - b\bar{x} = 104 - (.1158)(484) = 47.94$
$\hat{y}_i = 47.94 + .1158x_i$

Q2: Is the correlation (r=.736) statistically significant?

A2: $t_{df=8} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.736\sqrt{10-2}}{\sqrt{1-.736^2}} = 3.075; \quad From\ StatWISE: p = .0152\ \ (two-tailed)$

$From\ table: t_{8,.05/2} = 2.306;\ so\ p < .05\ two-tailed$

Yes, the correlation is statistically significant. If the assumptions are valid, we can be reasonably confident that the population correlation is greater than zero.

Q3: What is the confidence interval for the population correlation?

A3: $95\%\ CI\ for\ \rho: \quad r \to r';\ .736 \to .9417;\ \sigma_{r'} = \sqrt{1/n-3} = \sqrt{1/7} = .3780$

$r' \pm (Z_{.05/2})(\sigma_{r'}) \Rightarrow .9417 \pm (1.96)(.3780) \quad \Rightarrow \quad prob[.2008 < \rho' < 1.6826] = 95\%$

$Convert\ to\ correlation: prob[.198 < \rho < .933] = 95\%$

If the assumptions are valid, the probability is 95% that this confidence interval includes the actual population correlation.

Q4: Why is this confidence interval so wide?

A4: The sample size of only 10 is much too small for most practical applications of regression. Not only is the confidence interval wide, but with such a small sample we cannot use our data to make confident judgments about the shape of the population distribution and whether we have satisfied assumptions. Small departures can have substantial impact on calculations.

Bumble had a flash of inspiration. He realized that the beauty of Helen of Troy is known (she set the standard of 1000 milliHelens). Thus, Bumble could use his data to estimate the IQ of Helen of Troy! He was delighted because he was sure he would be famous for establishing the new field of Archeo-psychometrics.

Q5: What is the estimated IQ for Helen of Troy?

A5: $\hat{y}_i = 47.94 + .1158x_i$ = 47.94 + (.1158)(1000) = 163.74

Helen not only was a great beauty, she was a genius!

Q6: What is the Standard Error of Estimate for this set of data?

$$s_{y-\hat{y}} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} = s_y\sqrt{\left(\frac{(n-1)}{(n-2)}\right)(1-r^2_{xy})} = 10.78\sqrt{\left(\frac{(9)}{(8)}\right)(1-.736^2)} = 10.78(.718) = 7.74$$

See the attached SPSS output: MSres = 59.987 = $(7.74)^2$

Q7: What is the standard error of prediction for the individual score, Xi = 1000?

A7: $s_{y_i-\hat{y}_i} = s_{y-\hat{y}}\sqrt{1+\frac{1}{n}+\frac{z_x^2}{n-1}}$ *where* $z_x = \frac{x_i - \bar{x}}{s_x} = \frac{1000-484}{68.51} = 7.53$

$$= 7.74\sqrt{1+\frac{1}{10}+\frac{(7.53)^2}{10-1}} = 7.74\sqrt{7.403} = 21.06$$

Q8: What is the 95% confidence interval for Y_i when X_i = 1000?

A8: $CI(y_i) = \hat{y}_i \pm (t_{df=n-2;\,\alpha/2})(s_{y_i-\hat{y}_i})$

$$163.7 \pm (2.306)(21.06)$$

$$163.7 \pm (48.6)$$

$$prob[115 \le IQ \text{ for Helen of Troy} \le 213] = 95\%$$

If all assumptions are valid, we can be 95% confident that the interval from 115 to 213 includes the IQ for Helen of Troy. Why is this interval so very wide? Because our sample is very small and also because Helen is very far from the mean on the predictor variable.

Q9. What assumptions are needed for the estimate of Helen's IQ to be valid? Are they satisfied?

A9: Consider validity of measures across time and cultures, sampling, implementation of measures, great risk in extrapolating a linear trend far beyond the range of observed data, ….

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Y	104.00	10.781	10
X	484.00	68.508	10

← $Sy = sd_y = 10.78$
← $Sx = sd_x = 68.51$

Correlations

		Y	X
Pearson Correlation	Y	1.000	.736
	X	.736	1.000
Sig. (1-tailed)	Y	.	.008
	X	.008	.
N	Y	10	10
	X	10	10

← $r_{xy} = .736$

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	566.101	1	566.101	9.437	.015[a]
	Residual	479.899	8	59.987		
	Total	1046.000	9			

a. Predictors: (Constant), X

b. Dependent Variable: Y

←MSres = 59.987

The square root of MSres is the Standard Error of Estimate = 7.74, the SD of points around the line

$SS_{TOT} = SS_{reg} + SS_{resid}$; $SE_{Y-\hat{Y}} = \sqrt{MS_{resid}} = \sqrt{59.987} = 7.745$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	47.969	18.403		2.607	.031	5.531	90.407
	X	.116	.038	.736	3.072	.015	.029	.203

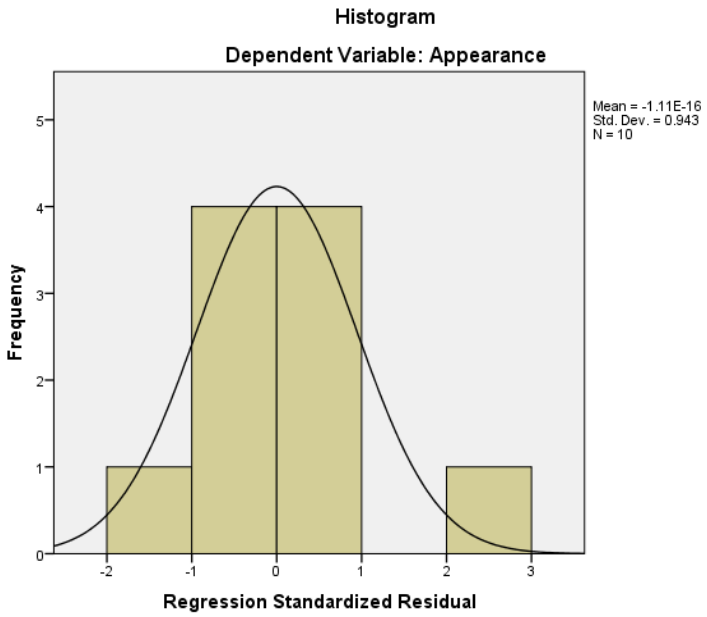
a. Dependent Variable: Y

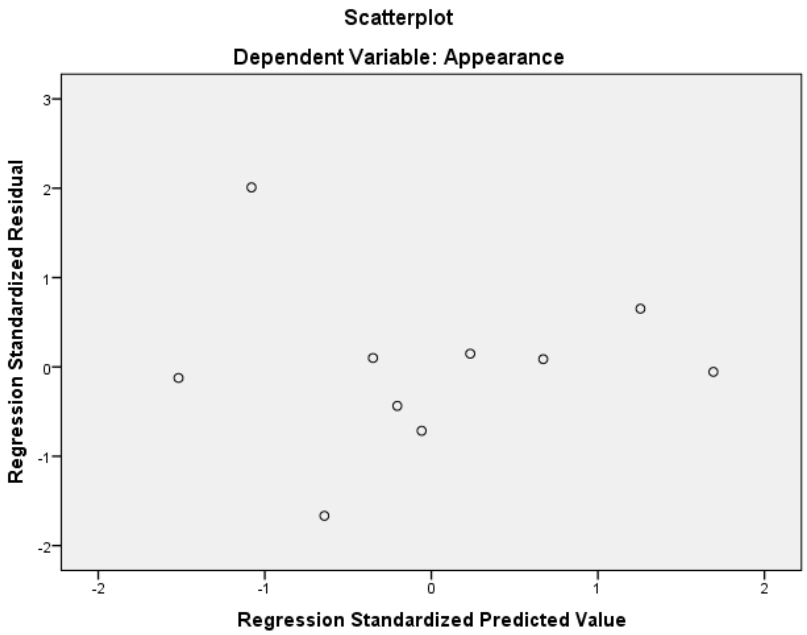
From the B coefficients: $\hat{Y}_i = 47.969 + .116\, X_i$

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	91.96	117.43	104.00	7.931	10
Residual	-12.91	15.57	.00	7.302	10
Std. Predicted Value	-1.518	1.693	.000	1.000	10
Std. Residual	-1.666	2.010	.000	.943	10

a. Dependent Variable: Y





With only ten data points, one should be caution in generalizing about the population distribution of residuals, but there are no serious problems with the observed data.

The standardized residuals do not reveal any extreme outliers,and the scatterplot of residuals shows reasonable homoscedasticity.