STAT 2509 B
Assignment#2
SOLUTION

1. Recently, research efforts have focused on the problem of predicting a manufacturer's market share by using information on the quality of its product. Suppose that the following data are available on the market share, in percentage (Y), and product quality, on scale of 0 to 100, determined by an objective evaluation procedure (X).

Product	Evaluation procedure (X)	% of market shares (Y)
1	27	2
2	39	3
3	73	10
4	66	9
5	33	4
6	43	6
7	47	5
8	55	8
9	60	7
10	68	9
11	70	10
12	75	13
13	82	12

$$\sum y_i = 98 \qquad \sum x_i = 738$$
$$\sum y_i^2 = 878 \qquad \sum x_i^2 = 45\,580$$
$$\sum x_i y_i = 6\,251$$

[3] (a) Plot a scatter diagram (using SAS, see part (i)) to get an idea of the form of the relationship between the evaluation procedure and the percentage of market shares. Does the scatter diagram indicate an approximately straight line?

See SAS output on page 7.

[6] (b) State a SLR model making sure you give all assumptions necessary for statistical inference.

Model: $y = \beta_0 + \beta_1 x + \varepsilon$, $n = 13$ (1)

Assumptions: (i) ***x*'s are observed without error** (1)
(ii) ***y*'s (or ε's) are independently distributed with mean** $E(y) = \beta_0 + \beta_1 x$ (1)
(or $E(\varepsilon) = 0$**)**

(iii) variance of *y*'s (or ε's) is constant, σ^2 for all *x*'s

(iv) $y \sim N(E(y), \sigma^2)$ for any value of *x* (or $\varepsilon \sim N(0, \sigma^2)$ for any value of *x*)

[6] (c) Find the least squares estimates of β_0 and β_1. Find the least squares fitted regression line.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{6251 - \frac{(738)(98)}{13}}{45580 - \frac{(738)^2}{13}} = \frac{687.6154}{3684.308} = 0.186634 \cong \underline{0.18663}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{98}{13} - (0.186634)\left(\frac{738}{13}\right) = 7.538462 - 0.186634(56.76923) =$$

$$= -3.05658 \cong \underline{-3.0566}$$

∴ **the least squares fitted regression line is given by:** $\hat{y} = \underline{-3.0566 + 0.18663}\, x$

For the rest of the question assume that the assumptions hold.

[6] (d) Find s^2, an estimate of σ^2.

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2} = \frac{\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] - \frac{\left[\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right]^2}{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}}{n-2} =$$

$$= \frac{\left[878 - \frac{(98)^2}{13}\right] - \frac{(687.6154)^2}{3684.308}}{11} = \frac{139.2308 - 128.3321}{11} = \frac{10.89868}{11} = \underline{\mathbf{0.990789}} \cong \underline{\mathbf{0.99}}$$

$\therefore s = \sqrt{s^2}$ **=0.995384 ≅ 1**

(e) Use the t-test to test whether there is a significant linear relationship between the evaluation procedure and the percentage of market shares. Use $\alpha = 0.05$. [7]

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

test-statistics: $t = \dfrac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \dfrac{0.186634}{0.995384/\sqrt{3684.308}} = \underline{\mathbf{11.38091}} \cong \underline{\mathbf{11.38}}$

R.R: **we reject** H_0 **if** $t < -t_{\alpha/2;n-2} = -t_{0.025;11} =$ **- 2.201**
or $t > t_{\alpha/2;n-2} = t_{0.025;11} =$ **2.201**

Since *t* = 11.38 > 2.201, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that the evaluation procedure and the percentage of market shares are linearly related.

(f) Find a 95% confidence interval for β_1. [4]

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;n-2}\ s/\sqrt{S_{xx}}\right) = \left(0.186634 \pm t_{0.025;11}\ 0.995384/\sqrt{3684.308}\right) = (0.186634 \pm 2.201(0.016399)) =$

$= (0.186634 \pm 0.036094) = (0.15054, 0.222727) \cong (0.1505, 0.2227)$

i.e. We are 95% confident that in repeated sampling the true value of the population slope would lie in the interval (0.1505, 0.2227).

(g) Set up the ANOVA table and hence test whether there is a significant linear relationship between the evaluation procedure and percentage of market shares. Use $\alpha = 0.05$. [19]

$$TSS = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \underline{139.2308} \cong \underline{139.23}$$ **(as calculated in part (d))**

$$SSR = \frac{S_{xy}^2}{S_{xx}} = \underline{128.3321}$$ **(as calculated in part (d))**

$$SSE = TSS - SSR = \underline{10.89868}$$ **(calculated in part (d))**

$$MSR = \frac{SSR}{1} = \underline{128.3321}$$

$$MSE = \frac{SSE}{n-2} = \frac{10.89868}{11} = \underline{0.990789}\ (= s^2)$$ **(as calculated in part (d))**

$$F = \frac{MSR}{MSE} = \underline{129.5252}$$

Source	d.f.	SS	MS	F
Regression	1	128.3321	128.3321	129.5252
Error	11	10.89868	0.990789	
Total	12	139.2308		

$$H_0 : \beta_1 = 0 \qquad \alpha = 0.05$$
$$H_a : \beta_1 \neq 0$$

Test-statistics: $F = \frac{MSR}{MSE} = \underline{129.5252}$

R.R: **we reject** H_0 **if** $F > F_{\alpha(1,n-2)} = F_{0.05(1,11)} =$ **4.84**

Since *F* = 129.52 > 4.84, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the evaluation procedure and percentage of market shares exists.

(h) Find the values of the coefficient of correlation, *r*, and coefficient of determination, r^2, and interpret their meanings in this problem

[5]

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{687.6154}{\sqrt{(3684.308)(139.2308)}} = 0.960064 \cong \underline{0.96}$$

i.e. the evaluation procedure and the percentage of market shares are highly positively correlated (related) with the strength of their relationship approx. 96%.

$$r^2 = \frac{SSR}{TSS} = \mathbf{0.921722 \cong \underline{0.9217}}$$

i.e. approximately 92.17% of the total variation in the data is explained by the regr. line (and approx. 7.83% is due to error).

(i) Verify your above results using SAS (Proc REG in SAS Manual or see the handout, SAS program (a)).

See SAS output on page 8.

2. Refers to question 1.

[10]

(a) Find a 95% Confidence Interval for the mean value of the response variable (i.e. of the percentage of market shares) and a 95% Prediction Interval for an individual value of this response variable when the value of evaluation procedure is 79 (i.e. $x_p = 79$).

What is your conclusion about the widths of these two intervals?

95% C.I. for *E(y)* when $x_p = 79$:

$\hat{y} = -3.05658 + 0.18663(79) = 11.68747$ **and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(11.68747 \pm t_{0.025;11} (0.995384) \sqrt{\frac{1}{13} + \frac{(79 - 56.76923)^2}{3684.308}} \right) =$$

$$= (11.68747 \pm 2.201(0.4572931)) = (11.68747 \pm 1.006503) = (10.68097, 12.69397) \cong$$

$$\cong (10.681, 12.694)$$

i.e. We are 95% confident that in repeated sampling the mean value of the percentage of market shares for evaluation procedures of 79 will fall in the interval (10.681,12.694).

and

95% P.I. for *y* when $x_p = 79$:

$\hat{y} = -3.05658 + 0.18663(79) = 11.68747$ **and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore y \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(11.68747 \pm t_{0.025;11} (0.995384) \sqrt{1 + \frac{1}{13} + \frac{(79 - 56.76923)^2}{3684.308}} \right) =$$

$$= (11.68747 \pm 2.201(1.0954023)) = (11.68747 \pm 2.41098) = (9.276488, 14.09845) \cong$$

$\cong (9.276, 14.098)$

i.e. We are 95% confident that in repeated sampling the value of the percentage of market shares when the evaluation procedure is 79 will lie in the interval (9.276,14.098).

Conclusion:

- **The P.I. is wider than C.I. (as expected), since the variability in the error for predicting a single value of y is always greater than the variability of the error for the estimation of the mean/average value of y.**

(b) Use SAS to compare your results with part (a) (see handout, SAS program (c)).

See SAS output on page 9.

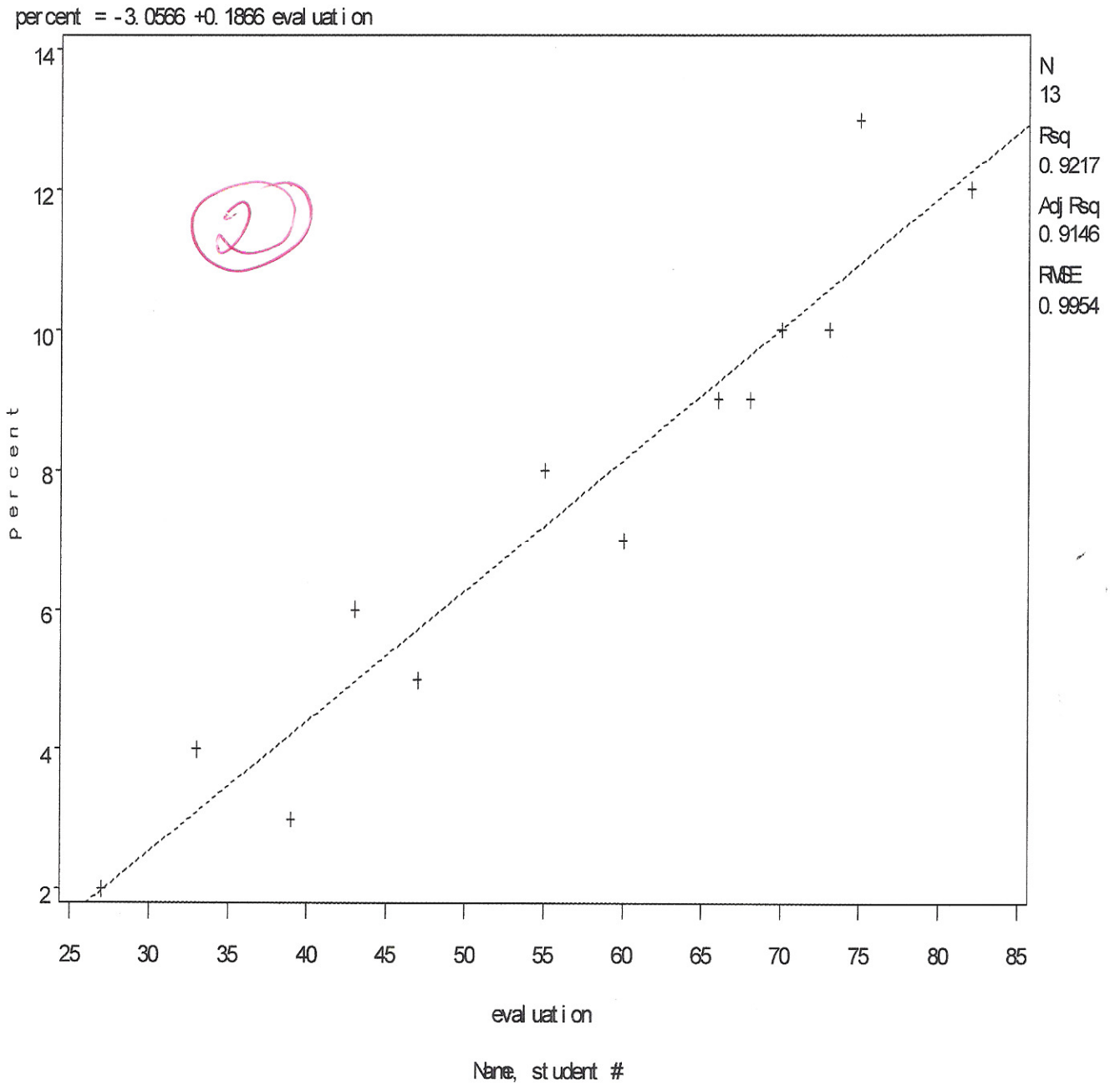
3. Refers to question 1.

Perform a residual analysis using SAS (see handout, SAS program (b)). What can you say about the model?

Given that the residual plots indicate no violations of the model assumptions and given that r^2 = 92.17% is very high, we may conclude that the model is an appropriate one.

See SAS output on pages 10, 11 & 12 (and/or 13).

Optional



2

Scatter plot indicates approximately straight line with positive slope 1

The REG Procedure
Model: MODEL1
Dependent Variable: percent

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	128.33209	128.33209	129.53	<.0001
Error	11	10.89868	0.99079		
Corrected Total	12	139.23077			

Root MSE	0.99538	R-Square	0.9217
Dependent Mean	7.53846	Adj R-Sq	0.9146
Coeff Var	13.20407		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.05658	0.97102	-3.15	0.0093
evaluation	1	0.18663	0.01640	11.38	<.0001

3

Name, student #

The REG Procedure
Model: MODEL1
Dependent Variable: percent

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	2.0000	1.9825	0.5608	0.7481	3.2169	-0.5321	4.4972	0.0175
2	3.0000	4.2221	0.4014	3.3386	5.1056	1.8599	6.5844	-1.2221
3	10.0000	10.5677	0.3835	9.7236	11.4117	8.2199	12.9155	-0.5677
4	9.0000	9.2612	0.3148	8.5683	9.9542	6.9634	11.5590	-0.2612
5	4.0000	3.1023	0.4776	2.0510	4.1536	0.6723	5.5323	0.8977
6	6.0000	4.9687	0.3567	4.1837	5.7536	2.6415	7.2959	1.0313
7	5.0000	5.7152	0.3192	5.0127	6.4177	3.4145	8.0159	-0.7152
8	8.0000	7.2083	0.2776	6.5973	7.8192	4.9338	9.4827	0.7917
9	7.0000	8.1414	0.2811	7.5227	8.7601	5.8649	10.4179	-1.1414
10	9.0000	9.6345	0.3319	8.9041	10.3649	7.3251	11.9439	-0.6345
11	10.0000	10.0078	0.3511	9.2349	10.7806	7.6846	12.3309	-0.007767
12	13.0000	10.9409	0.4069	10.0453	11.8366	8.5741	13.3078	2.0591
13	12.0000	12.2474	0.4974	11.1526	13.3421	9.7982	14.6965	-0.2474
14	.	11.6875	0.4573	10.6810	12.6940	9.2765	14.0984	.

$\hat{y}$ when $x_p = 79$

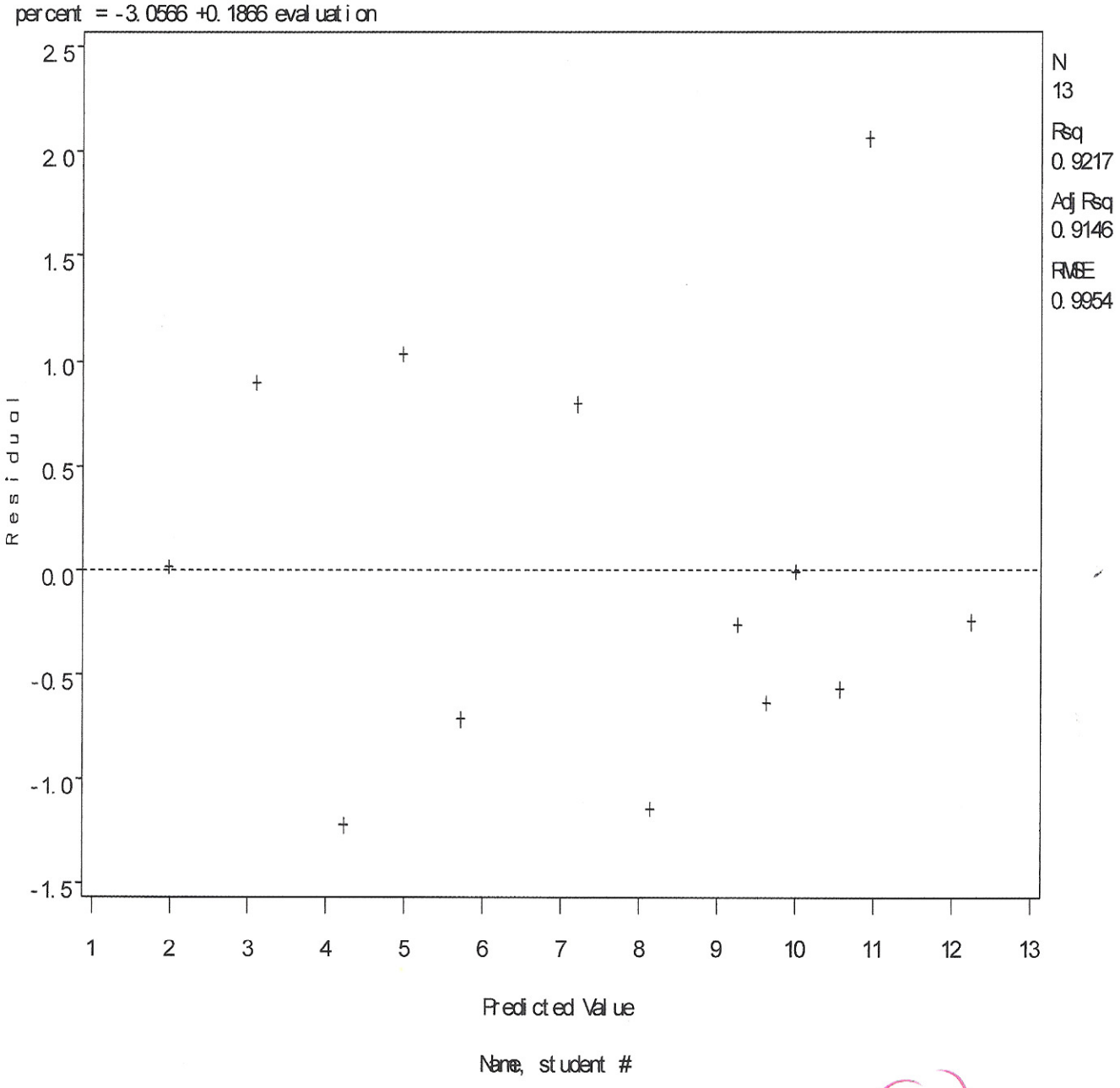
95% C.I. for E(y) when $x_p = 79$

95% P.I. for y when $x_p = 79$

Sum of Residuals	0
Sum of Squared Residuals	10.89868
Predicted Residual SS (PRESS)	15.06091

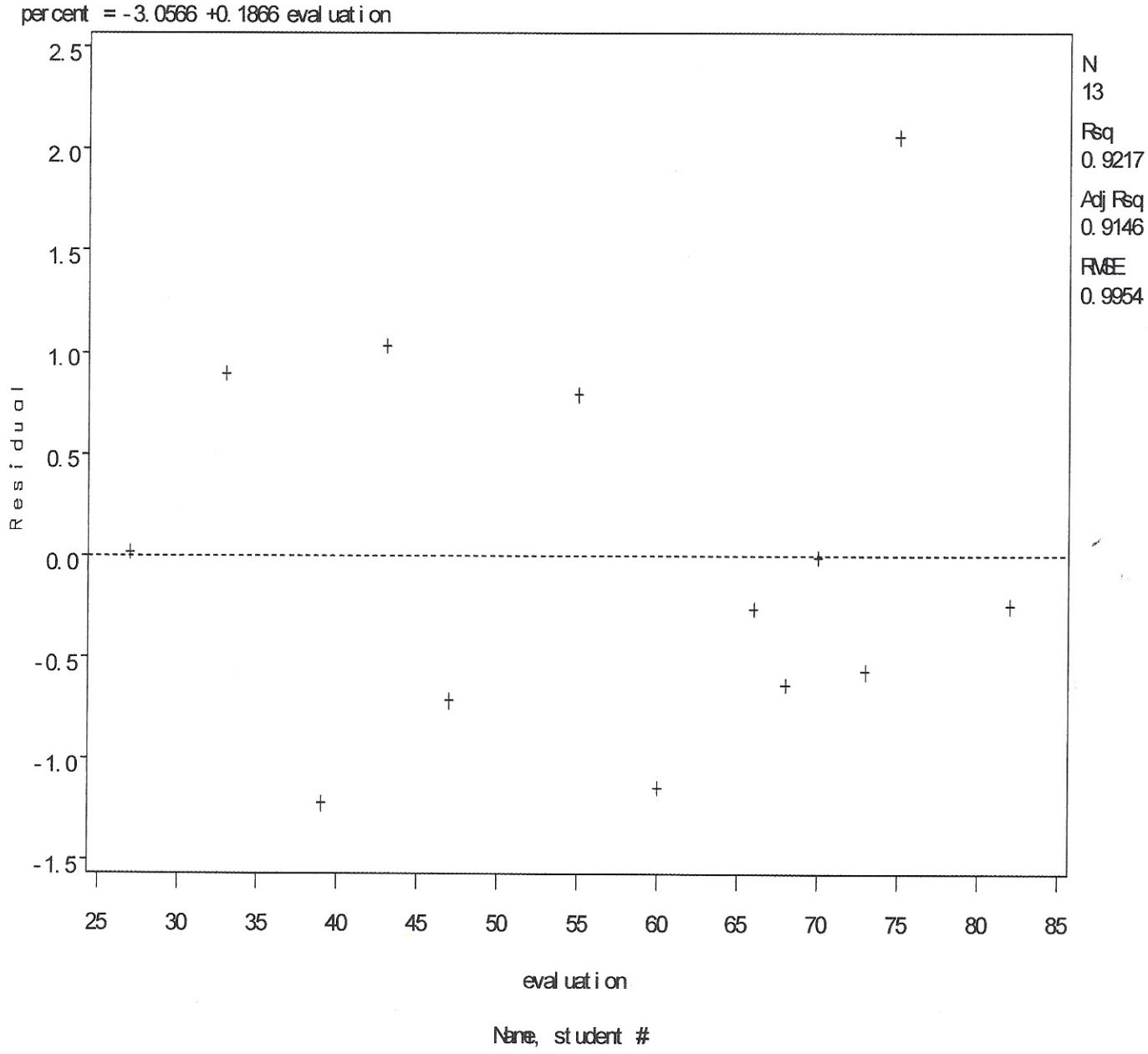
3

Name, student #



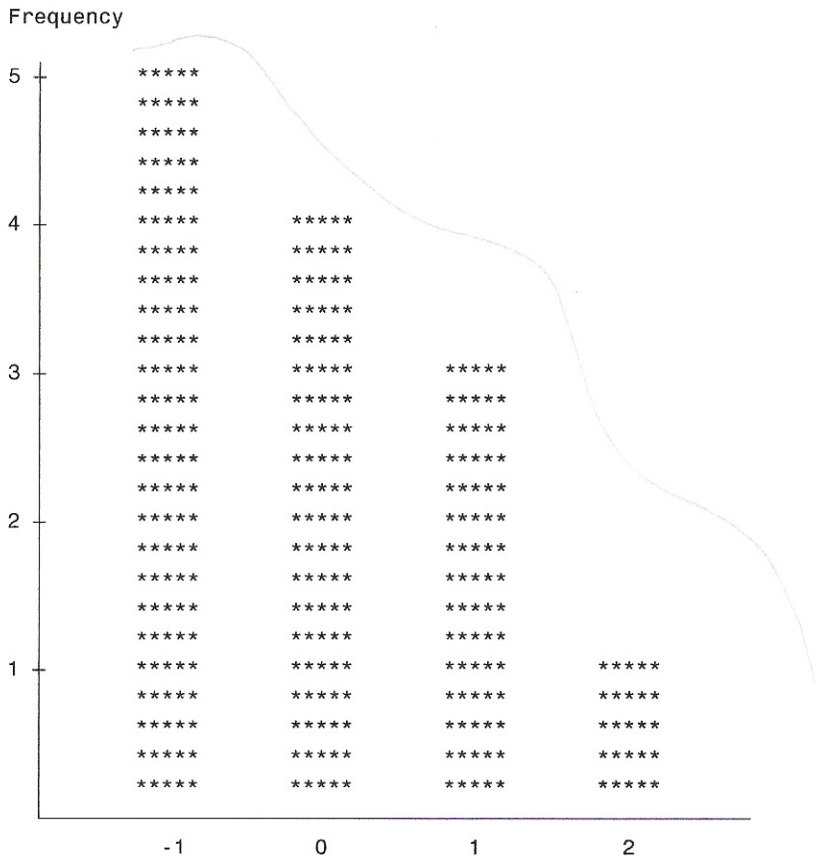
(1)

Residuals seem to be randomly scattered around zero (i.e. no pattern) ⇒ no violations of independence (and linearity)



Residuals seem to be randomly scattered around zero (i.e. no pattern) ⇒ no violations of constant variance

(1)



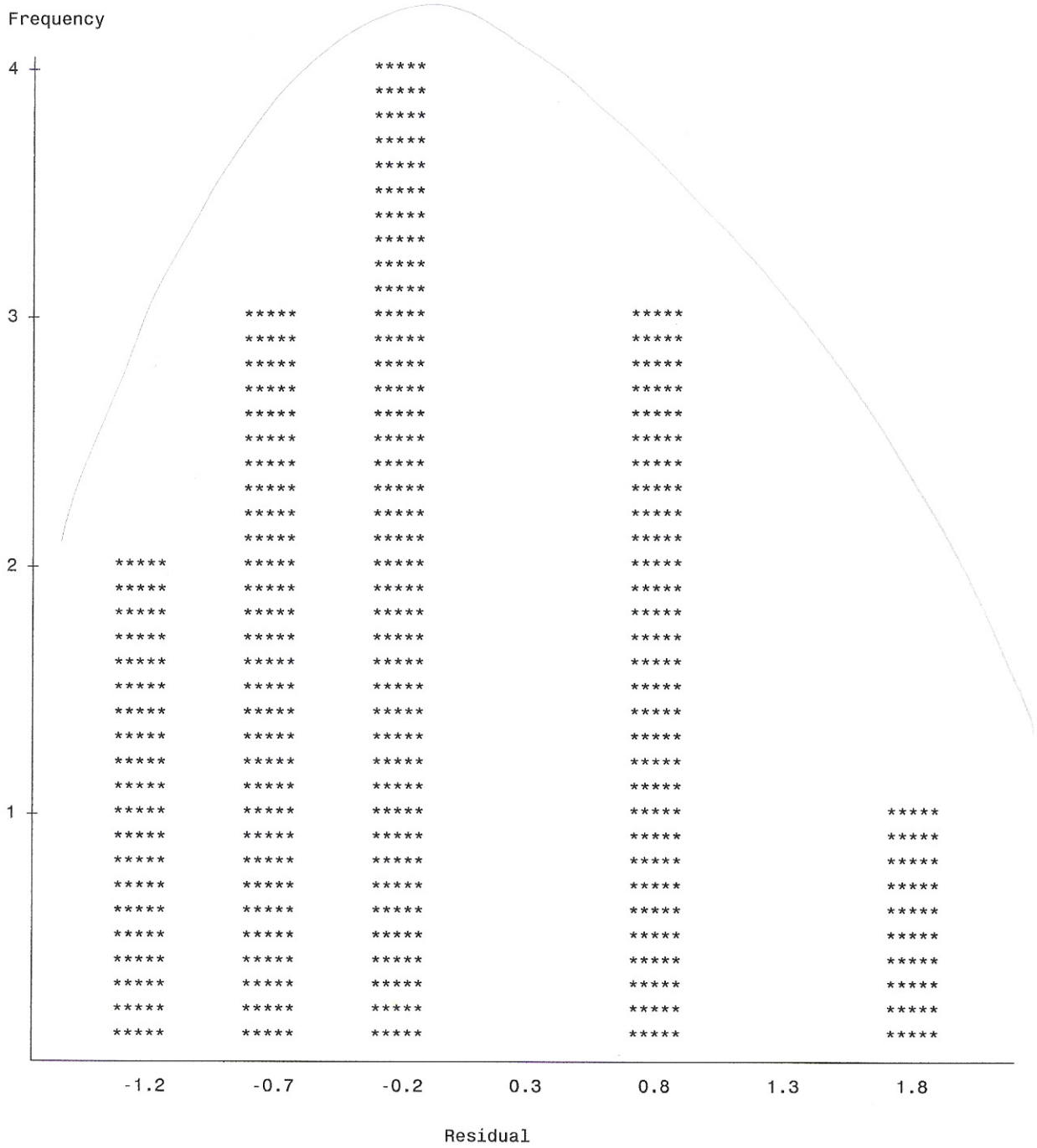
Histogram of errors appears to be skewed to the right (i.e. is not symmetric and not bell-shaped).

Given that all other plots suggest no problems with the model, also r^2 = 92.17% (very high) and taking into consideration that n = 13 (<30), we may not have problem with the model assumptions after all.

We can try to change frequency boundaries (see p. 13)

Optional

Name, student #



After changing frequency boundaries, the errors seem to be bell-shaped and approx.. symmetric

∴ the model is appropriate

Name, student #

```
Footnote 'Name, student #';
Data Market;
Input evaluation percent @@;
Cards;
            27 2 39 3 73 10 66 9 33 4 43 6 47 5
            55 8 60 7 68 9 70 10 75 13 82 12
Run;
Proc Reg;
            Model percent=evaluation;
            Plot percent*evaluation;
Run;

Data Predict;
Input evaluation percent;
Cards;
            79 .
Run;
Data Join;
            Set Market Predict;
Run;
Proc Reg;
            Model percent=evaluation/CLM CLI;
Run;

Proc Reg;
            Model percent=evaluation;
            Plot R.*P.;
            Plot R.*evaluation;
            Output out=res R=resids;
Run;
Proc Chart;
            vbar resids;
            vbar resids/midpoint=-1.2 to 2.2 by 0.5;
Run;
```

Q.1 parts a)-i)

Q.2 parts a), b)

Q.3

4