Total mark: 80

STAT 2509
Assignment #2
SOLUTION

1. A study was conducted to examine the quality of fish after seven days in ice storage. Ten raw fish of the same kind and approximately the same size were caught and prepared for ice storage. Two of the fish were placed in storage immediately after being caught, two were placed in storage 3 hours after being caught, and two each were placed in storage at 6, 9 and 12 hours after being caught. Let y denote a measurement of fish quality (on a 10-point scale) after the seven days of storage and let x denote the time after being caught that the fish were placed in ice packing. The sample data are shown below:

y	8.5	8.4	7.9	8.1	7.8	7.6	7.3	7.0	6.8	6.7
x	0	0	3	3	6	6	9	9	12	12

$$\sum y_i = 76.1 \qquad \sum x_i = 60$$
$$\sum y_i^2 = 582.85 \qquad \sum x_i^2 = 540$$
$$\sum x_i y_i = 431.1$$

[3] Plot a scatter diagram (using SAS, see part (i)) to get an idea of the form of the relationship between the number of hours when fish were placed in ice storage after being caught and the fish quality. Does the scatter diagram indicate an approximately straight line?

See scatter plot in SAS output on Page 12. (Give a full mark if this scatter plot is found anywhere in SAS output)

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

Model: $y = \beta_0 + \beta_1 x + \varepsilon$, $n = 10$

Assumptions: (i) ***x*'s are observed without error**

(ii) ***y*'s (or ε's) are independently distributed with mean** $E(y) = \beta_0 + \beta_1 x$ **(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*)**

(Or, give 5 marks if it was indicated that ϵ_i's are independent and identically distributed with $N(0, \sigma^2)$)

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

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{431.1 - \frac{(60)(76.1)}{10}}{540 - \frac{(60)^2}{10}} = \frac{-25.5}{180} = \underline{-0.14167}$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{76.1}{10} - (-0.14167)\left(\frac{60}{10}\right) = 7.61 - (-0.85) = \underline{8.46}$$

∴ **the least squares fitted regression line is given by:** $\hat{y} = \underline{8.46 - 0.14167}\, x$

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

[6] (c) Find s^2, an estimate of σ^2.

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

$$= \frac{\left[582.85 - \frac{(76.1)^2}{10}\right] - \frac{(-25.5)^2}{180}}{8} = \frac{3.729 - 3.6125}{8} = \frac{0.1165}{8} = \underline{\underline{0.014563}}$$

$$\therefore s = \sqrt{s^2} = 0.120675$$

[7] (d) Use the t-test to test whether there is a significant linear relationship between the number of hours when fish were placed in ice storage after being caught and the fish quality. Use $\alpha = 0.05$.

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

test-statistics: $t = \dfrac{b_1}{s/\sqrt{S_{xx}}} = \dfrac{-0.14167}{0.120675/\sqrt{180}} =$ **–15.7502**

R.R: **we reject** H_0 **if** $t < -t_{\alpha/2;n-2} = -t_{0.025;8} =$ **–2.306**
or $t > t_{\alpha/2;n-2} = t_{0.025;8} =$ **2.306**

Since *t* = –15.7502 < –2.306, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the number of hours when fish were placed in ice storage after being caught and the fish quality exists.

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

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

$$\beta_1 \in \left(b_1 \pm t_{\alpha/2;n-2} \, s/\sqrt{S_{xx}} \right) = \left(-0.14167 \pm t_{0.025;8} \, 0.120675/\sqrt{180} \right) =$$

$$= \left(-0.14167 \pm 2.306(0.008995) \right) =$$

$$= \left(-0.14167 \pm 0.020742 \right) = \left(-0.16241, -0.12093 \right) \cong \left(-0.162, -0.121 \right)$$

i.e. We are 95% confident that in repeated sampling the true value of the population slope would lie in the interval (–0.162, –0.121).

[19] (f) Set up the ANOVA table and hence test whether there is a significant linear relationship between the number of hours when fish were placed in ice storage after being caught and the fish quality. Use α = 0.05.

$$TSS = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i \right)^2}{n} = \underline{3.729}$$ **(as calculated in part (d))**

$SSR = \frac{S_{xy}^2}{S_{xx}} =$ **3.6125** (as calculated in part (d))

$SSE = TSS - SSR =$ **0.1165** (calculated in part (d))

$MSR = \frac{SSR}{1} =$ **3.6125**

$MSE = \frac{SSE}{n-2} = \frac{0.1165}{8} =$ **0.014563** $(= s^2)$

$F = \frac{MSR}{MSE} =$ **248.0687**

one mark for each column if values are entered correctly

Source	d.f.	SS	MS	F
Regression	1	3.6125	3.6125	248.0687
Error	8	0.1165	0.014563	
Total	9	3.729		

$H_0 : \beta_1 = 0$ $\alpha = 0.05$

$H_a : \beta_1 \neq 0$

test-statistics: $F = \dfrac{MSR}{MSE} =$ **248.0687**

R.R: **we reject** H_0 **if** $F > F_{\alpha(1,n-2)} = F_{0.05(1,8)} =$ **5.32**

Since *F* = 248.0687 > 5.32, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the number of hours when fish were placed in ice storage after being caught and the fish quality exists.

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

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-25.5}{\sqrt{(180)(3.729)}} = -0.98426$$

i.e. the number of hours when fish were placed in ice storage after being caught and the fish quality are very strongly negatively correlated (related) with the strength of their relationship approx. 98.43%.

$$r^2 = \frac{SSR}{TSS} = 0.968758 \cong 96.87\%$$

i.e. approximately 96.87% of the total variation in the data is explained by the regression line (and only 3.13% is due to error).

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

See SAS output on Page 13

2. Refers to question 1.

[10] (a) Find a 95% Confidence Interval for the mean value of the response variable (i.e. of the quality of fish) and a 95% Prediction Interval for an individual value of this response variable when the number of hours when fish were placed in ice storage after being caught is 4 (i.e. $x_p = 4$).

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

95% C.I. for *E(y)* when $x_p = 4$:

$\hat{y} = 8.46 - 0.14167(4) = 7.89332$ **and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(7.89332 \pm t_{0.025;8} (0.120675) \sqrt{\frac{1}{10} + \frac{(4-6)^2}{180}} \right) =$$

$$= (7.89332 \pm 2.306(0.042188)) = (7.89332 \pm 0.097286) = (7.796034, 7.990606) \cong$$

$$\cong (7.796, 7.991)$$

i.e. We are 95% confident that (in repeated sampling) the average fish quality when fish were placed in ice storage 4 hours after being caught would be between 7.796 and 7.991

and

95% P.I. for y when $x_p = 4$:

$\hat{y} = 8.46 - 0.14167(4) = 7.89332$ and $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore y \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(7.89332 \pm t_{0.025;8} (0.120675) \sqrt{1 + \frac{1}{10} + \frac{(4-6)^2}{180}} \right) =$$

$$= (7.89332 \pm 2.306(0.127837)) = (7.89332 \pm 0.294793) = (7.598527, 8.188113) \cong$$

$$\cong (7.599, 8.188)$$

i.e, We are 95% confident that (in repeated sampling) the fish quality when fish were placed in ice storage 4 hours after being caught would be between 7.599 and 8.188

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*. (1)

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

See SAS output on page 14

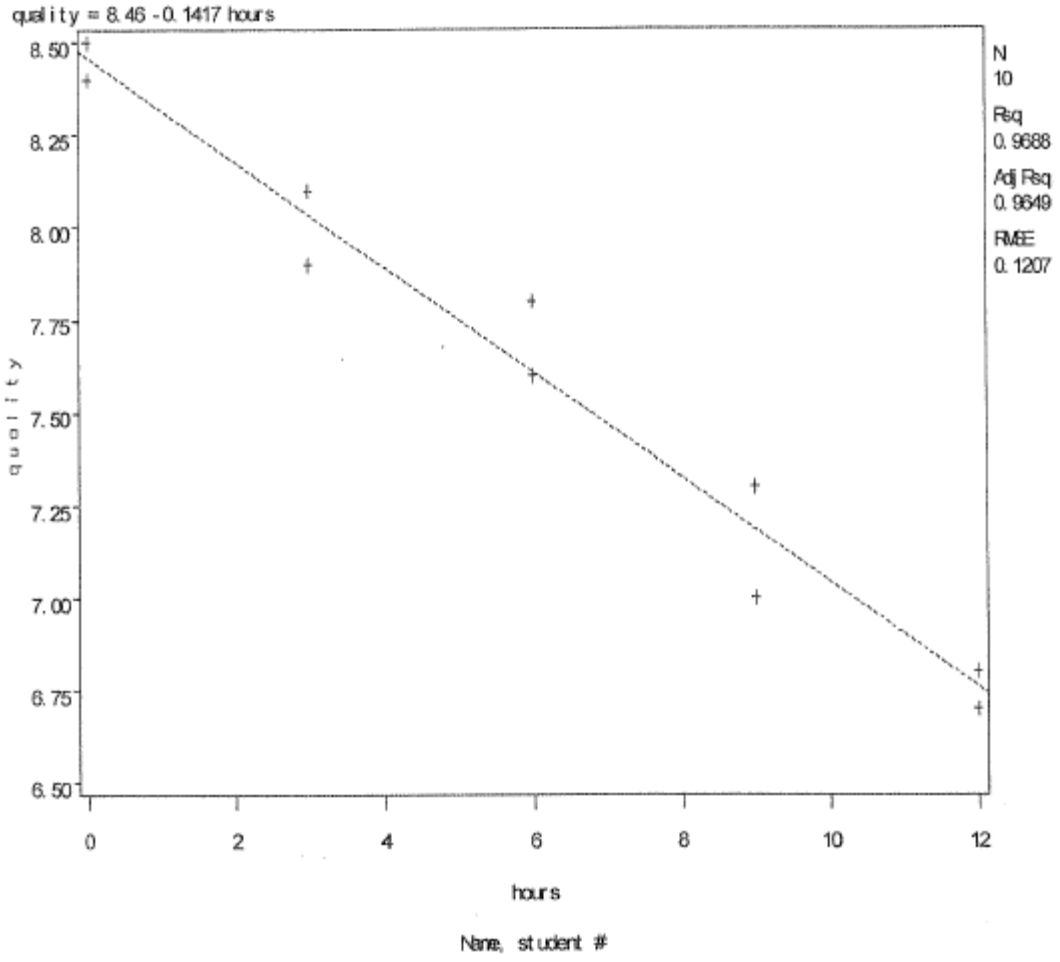
[4] **3.** Refers to question 1.

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

- The residual plots suggest no violations of the model assumptions and given that $r^2 = 96.87\%$ is very high, we may conclude that the model is appropriate. (1)

See SAS output on pages 15, 16 and 17 (3)

[4] SAS program (See page 18)



- **Scatter plot indicates a straight line with negative slope**

The SAS System 19:26 Tuesday, January 27, 2015 1

The REG Procedure
Model: MODEL1
Dependent Variable: quality

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.61250 (SSR)	3.61250 (MSR)	248.07 (F)	<.0001
Error	8	0.11650 (SSE)	0.01456 (MSE)		
Corrected Total	9	3.72900 (TSS)			

Root MSE	0.12068 (s)	R-Square	0.9688 (r^2)
Dependent Mean	7.61000	Adj R-Sq	0.9649
Coeff Var	1.58574		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.46000 ($\hat{\beta}_0$)	0.06610	127.99	<.0001
hours	1	-0.14167 ($\hat{\beta}_1$)	0.00899	-15.75 (t)	<.0001

3

The REG Procedure
Model: MODEL1
Dependent Variable: quality

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	8.5000	8.4600	0.0661	8.3076	8.6124	8.1427	8.7773	0.0400
2	8.4000	8.4600	0.0661	8.3076	8.6124	8.1427	8.7773	-0.0600
3	7.9000	8.0350	0.0467	7.9272	8.1428	7.7366	8.3334	-0.1350
4	8.1000	8.0350	0.0467	7.9272	8.1428	7.7366	8.3334	0.0650
5	7.8000	7.6100	0.0382	7.5220	7.6980	7.3181	7.9019	0.1900
6	7.6000	7.6100	0.0382	7.5220	7.6980	7.3181	7.9019	-0.0100
7	7.3000	7.1850	0.0467	7.0772	7.2928	6.8866	7.4834	0.1150
8	7.0000	7.1850	0.0467	7.0772	7.2928	6.8866	7.4834	-0.1850
9	6.8000	6.7600	0.0661	6.6076	6.9124	6.4427	7.0773	0.0400
10	6.7000	6.7600	0.0661	6.6076	6.9124	6.4427	7.0773	-0.0600
11	.	7.8933	0.0422	7.7960	7.9906	7.5985	8.1881	.

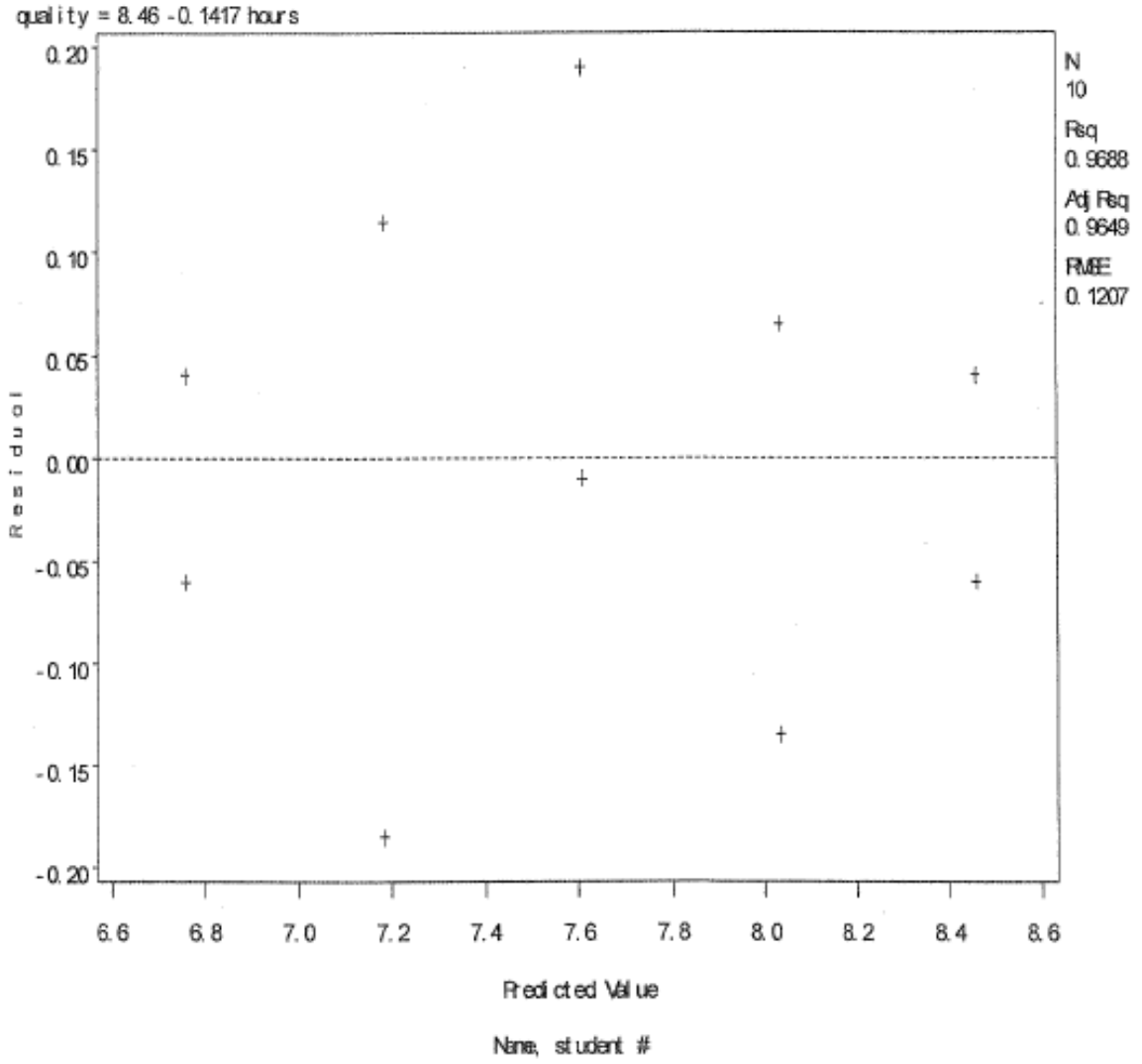
3

$\hat{y}$ when $x_p = 4$

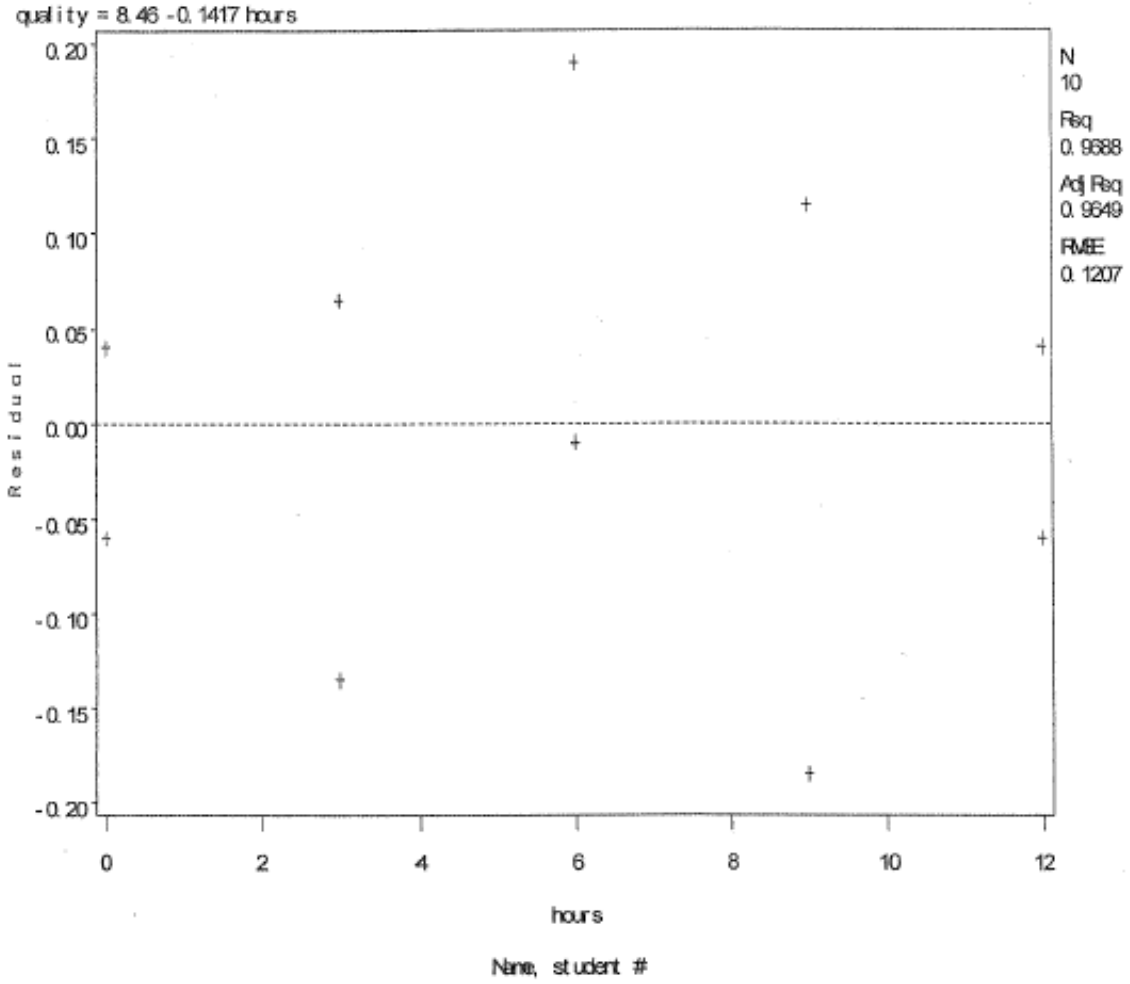
95% C.I. for $E(y)$ when $x_p = 4$

95% P.I. for y when $x_p = 4$

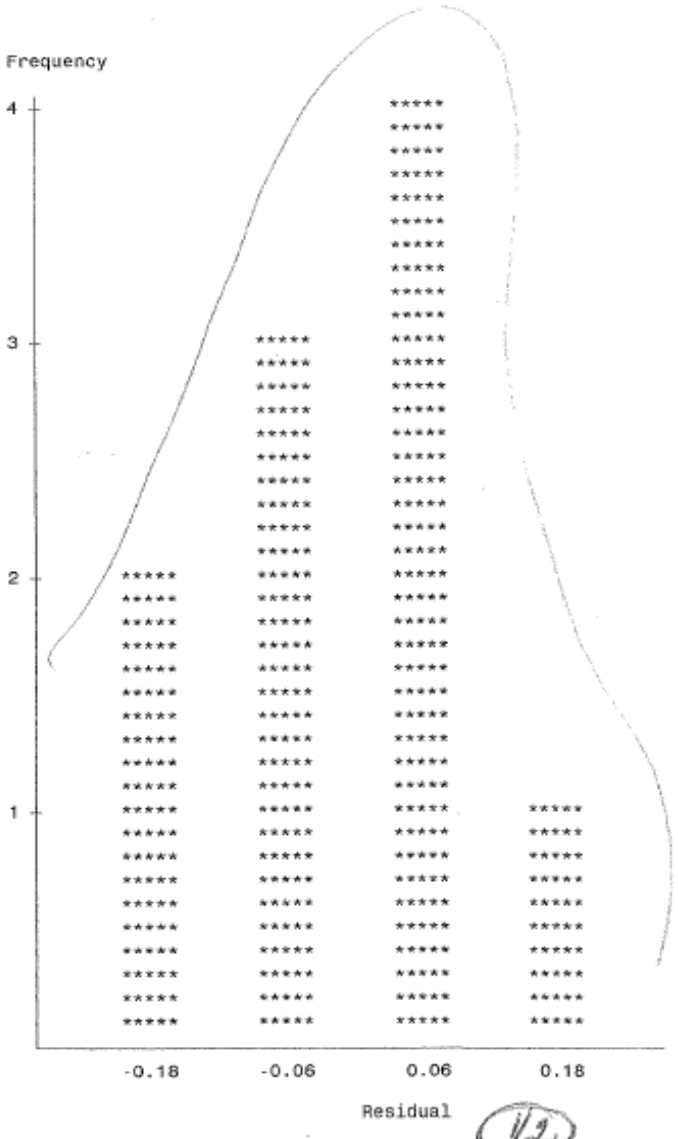
Sum of Residuals	0
Sum of Squared Residuals	0.11650
Predicted Residual SS (PRESS)	0.16266



- **Residuals do not seem to follow any pattern, i.e. no obvious violations of the assumption of independence/linearity**



- **Residuals seem to be randomly scattered around zero ⇒ variance is constant (i.e. no violations of the assumption of constant variance)**



1/2 1/2

- **histogram of residuals is approximately bell-shaped and approximately symmetric (i.e no violations of the assumption of normality of the errors)**

Please keep in mind that sample size is very small ($n = 10$)

Name, student #

```
Footnote 'Name, student #';
Data Fish_Quality;
Input hours quality @@;
Cards;
      0 8.5 0 8.4 3 7.9 3 8.1 6 7.8
      6 7.6 9 7.3 9 7.0 12 6.8 12 6.7
Run;
Proc Reg;
      Model quality=hours;
      Plot quality*hours;
Run;

Data Predict;
Input hours quality;
Cards;
      4 .
Run;
Data Join;
      Set Fish_Quality Predict;
Run;
Proc Reg;
      Model quality=hours/CLM CLI;
Run;

Proc Reg;
      Model quality=hours;
      Plot R.*P.;
      Plot R.*hours;
      Output out=res R=resids;
Run;
Proc Chart;
      vbar resids;
Run;
```

Q.1 (i)

Q.2 (a), (b)

Q.3

4

(Total)