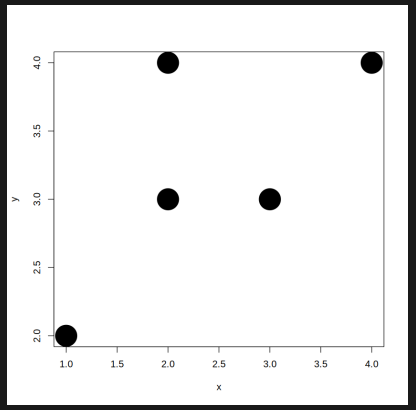
SIMPLE LINEAR REGRESSION: DESCRIPTIVE

- How strong is the linear relationship? (Correlation)
- Finding the line of best fit: theory
- Finding the line of best fit: in SAS
- How good is the fit?

HOW STRONG IS THE LINEAR RELATIONSHIP? (CORRELATION)

Problem. Draw scatterplot of the following points. Then find and interpret the strength of the linear relationship:

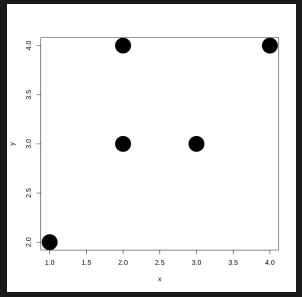
x	**1**	**2**	**2**	**3**	**4**
y	2	3	4	3	4



Answer: Draw scatterplot with PROC SGPLOT

```
data points;
    infile datalines dsd;
    input x y;
datalines;
1, 2
2, 3
2, 4
3, 3
4, 4
;

proc sgplot data=points;
    scatter x=x y=y;
run;
```



Answer: Find correlation with PROC CORR

```
proc corr data=points;
run;
```

The CORR Procedure

2 Variables:	x y

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
x	5	2.40000	1.14018	12.00000	1.00000	4.00000
y	5	3.20000	0.83666	16.00000	2.00000	4.00000

Pearson Correlation Coefficients, N = 5
Prob > |r| under H0: Rho=0

	x	y
x	1.00000	0.68139 0.2053
y	0.68139 0.2053	1.00000

Interpretation: Since correlation $r = 0.68$, there is strong (linear) association between x and y.

Different textbooks have different conventions of "strong":

Interpreting correlation coefficient values [edit]

The correlation between two variables have different associations that are measured in values such as r or R. Correlation values range from −1 to +1, where ±1 indicates the strongest possible correlation and 0 indicates no correlation between variables.[5]

r or R	r or R	Strength or weakness of association between variables[6]
+0.8 to +1.0	-1.0 to -0.8	Perfect or very strong association
+0.6 to +0.8	-0.8 to -0.6	Strong association
+0.4 to +0.6	-0.6 to -0.4	Moderate association
+0.2 to +0.4	-0.4 to -0.2	Weak association
+0.0 to 0.2	-0.2 to 0.0	Very weak or no association

https://en.wikipedia.org/wiki/Correlation_coefficient

Correlation $r = \cos(\theta_{x,y})$ measures the direction and strength of linear relationship:

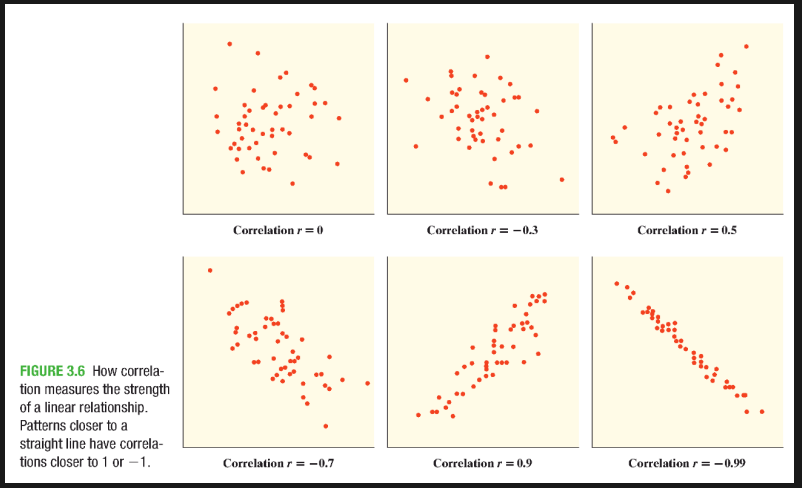
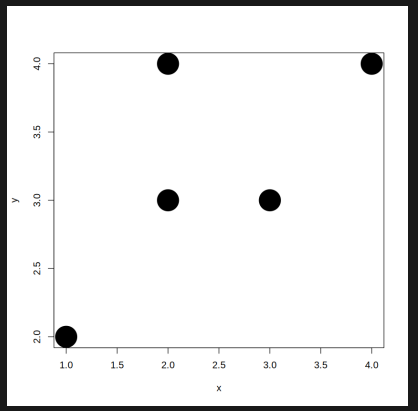


FIGURE 3.6 How correlation measures the strength of a linear relationship. Patterns closer to a straight line have correlations closer to 1 or −1.

FINDING THE LINE OF BEST FIT: THEORY

Problem. Find the best fitting line (in the least squares sense) that pass through the following points:

x	1	2	2	3	4
y	2	3	4	3	4



Problem. Find the best fitting line (in the least squares sense) that pass through the following points:

x	**1**	**2**	**2**	**3**	**4**
y	2	3	4	3	4

Model:

$$y = ax + b$$

x	**1**	**2**	**2**	**3**	**4**
y	2	3	4	3	4

Fitted model (with additive error/residual term):

$$
\begin{aligned}
2 &= a \cdot 1 + b + e_1 \\
3 &= a \cdot 2 + b + e_2 \\
4 &= a \cdot 2 + b + e_3 \\
3 &= a \cdot 3 + b + e_4 \\
4 &= a \cdot 4 + b + e_5
\end{aligned}
$$

Fitted model as matrices:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Fitted model as matrices:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix}}_{\begin{array}{c}\text{design} \\ \text{matrix } X\end{array}} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Fitted model as matrices:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Choice of $\vec{a} = (a, b)$ that minimizes Euclidean length of error vector:

$$\vec{y} = X\vec{a} \Rightarrow X^T\vec{y} = X^T X\vec{a}$$

$$\Rightarrow \boxed{(X^T X)^{-1}(X^T\vec{y}) = \vec{a}}$$

$$\vec{y} = X\vec{a} \Rightarrow X^T\vec{y} = X^TX\vec{a}$$

$$\Rightarrow \boxed{(X^TX)^{-1}(X^T\vec{y}) = \vec{a}}$$

$$\Rightarrow a = 0.5,\ b = 2$$

$$\vec{y} = X\vec{a} \Rightarrow X^T\vec{y} = X^TX\vec{a}$$

$$\Rightarrow \boxed{(X^TX)^{-1}(X^T\vec{y}) = \vec{a}}$$

$$\Rightarrow a = 0.5,\ b = 2$$

Our model is $y = 0.5x + 2$.

Our fitted model is:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Our fitted model is:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Our fitted model is:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \times 0.5 + 1 \times 2 \\ 2 \times 0.5 + 1 \times 2 \\ 2 \times 0.5 + 1 \times 2 \\ 3 \times 0.5 + 1 \times 2 \\ 4 \times 0.5 + 1 \times 2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Our fitted model is:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 3 \\ 3 \\ 3.5 \\ 4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Our fitted model is:

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 3 \\ 3 \\ 3.5 \\ 4 \end{bmatrix} + \begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -0.5 \\ 0 \end{bmatrix}$$

Our fitted model is:

$$\underbrace{\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix}}_{\text{observed } y} = \underbrace{\begin{bmatrix} 2.5 \\ 3 \\ 3 \\ 3.5 \\ 4 \end{bmatrix}}_{\text{predicted } \hat{y}} + \underbrace{\begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -0.5 \\ 0 \end{bmatrix}}_{\text{residuals}}$$

$$\underbrace{\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 2.5 \\ 3 \\ 3 \\ 3.5 \\ 4 \end{bmatrix}}_{\hat{y}} + \underbrace{\begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -0.5 \\ 0 \end{bmatrix}}_{\text{residuals}}$$

This is an orthogonal decomposition.

$$\underbrace{\begin{bmatrix} -1.2 \\ -0.2 \\ 0.8 \\ -0.2 \\ 0.8 \end{bmatrix}}_{y-\bar{y}} = \underbrace{\begin{bmatrix} -0.7 \\ -0.2 \\ -0.2 \\ 0.3 \\ 0.8 \end{bmatrix}}_{\hat{y}-\bar{y}} + \underbrace{\begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -0.5 \\ 0 \end{bmatrix}}_{\text{residuals}}$$

Centering by $\bar{y} = 3.2$ keeps it an orthogonal decomposition.

By Pythagorean theorem:

$$\left\| \begin{bmatrix} -1.2 \\ -0.2 \\ 0.8 \\ -0.2 \\ 0.8 \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} -0.7 \\ -0.2 \\ -0.2 \\ 0.3 \\ 0.8 \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -0.5 \\ 0 \end{bmatrix} \right\|^2$$

By Pythagorean theorem:

$$\begin{bmatrix} (-1.2)^2+ \\ (-0.2)^2+ \\ (0.8)^2+ \\ (-0.2)^2+ \\ (0.8)^2 \end{bmatrix} = \begin{bmatrix} (-0.7)^2+ \\ (-0.2)^2+ \\ (-0.2)^2+ \\ (0.3)^2+ \\ (0.8)^2 \end{bmatrix} + \begin{bmatrix} -0.5^2+ \\ 0^2+ \\ 1^2+ \\ -0.5^2+ \\ 0^2 \end{bmatrix}$$

By Pythagorean theorem:

$$\underset{\text{Sum of Squares Total}}{[2.8]} = \underset{\text{SS Model}}{[1.3]} + \underset{\text{SS Error}}{[1.5]}$$

By Pythagorean theorem:

$$\underset{\text{SST}}{2.8} = \underset{\text{SSM}}{1.3} + \underset{\text{SSE}}{1.5}$$

By Pythagorean theorem:

$$\underset{\text{SST}}{2.8} = \underset{\text{SSM}}{1.3} + \underset{\text{SSE}}{1.5}$$

SST is the variance of our data's y values (about the mean $\bar{y}$). It is split into two parts:

- SSM: variation in y 's explained by our model.
- SSE: variation in y 's not explained by our model.

By Pythagorean theorem:

$$\underset{\text{SST}}{2.8} = \underset{\text{SSM}}{1.3} + \underset{\text{SSE}}{1.5}$$

The proportion of total variation explained by our model is called the **coefficient of determination**:

$$R^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

By Pythagorean theorem:

$$\underset{\text{SST}}{2.8} = \underset{\text{SSM}}{1.3} + \underset{\text{SSE}}{1.5}$$

The proportion of total variation explained by our model is called the **coefficient of determination**:

$$R^2 = \frac{1.3}{2.8} = 1 - \frac{1.5}{2.8} = 0.46$$

By Pythagorean theorem:

$$\underset{\text{SST}}{2.8} = \underset{\text{SSM}}{1.3} + \underset{\text{SSE}}{1.5}$$

The proportion of total variation explained by our model is called the **coefficient of determination**:

$$R^2 = \frac{1.3}{2.8} = 1 - \frac{1.5}{2.8} = 0.46$$

Interpretation: "46% of the y's are explained by our linear model relating x's and y's."

By Pythagorean theorem:

$$\underset{\text{SST}}{2.8} = \underset{\text{SSM}}{1.3} + \underset{\text{SSE}}{1.5}$$

Root mean square error (RMSE) for simple linear regression is:

$$\text{RMSE} = \sqrt{\frac{\text{SSE}}{n-2}} = \frac{\|y - \hat{y}\|}{\sqrt{n-2}}$$

Summary:

Model	Fitted Model	Sum of Squares aka Pythagorean Decomposition
$y = ax + b$	$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$	$\underset{y-\bar{y}}{\begin{bmatrix} -1.2 \\ -0.2 \\ 0.8 \\ -0.2 \\ 0.8 \end{bmatrix}} = \underset{\hat{y}-\bar{y}}{\begin{bmatrix} -0.7 \\ -0.2 \\ -0.2 \\ 0.3 \\ 0.8 \end{bmatrix}} + \underset{\text{residual vector}}{\begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -0.5 \\ 0 \end{bmatrix}}$
		$R^2 = \dfrac{\text{SSM}}{\text{SST}} = \dfrac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2}$
		$\text{RMSE} = \sqrt{\dfrac{\text{SSE}}{n-2}} = \dfrac{\|y - \hat{y}\|}{\sqrt{n-2}}$

FINDING THE LINE OF BEST FIT: IN SAS

```
data points;
    infile datalines dsd;
    input x y;
datalines;
1, 2
2, 3
2, 4
3, 3
4, 4
;
proc reg data=points;
    model y=x / p clm alpha=0.05;
    plot y*x / NOSTAT;
run;
```

Options:

- P = predicted value;
- CLM = confidence interval

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	5
Number of Observations Used	5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.30000	1.30000	2.60	0.2053
Error	3	1.50000	0.50000		
Corrected Total	4	2.80000			

Root MSE	0.70711	R-Square	0.4643
Dependent Mean	3.20000	Adj R-Sq	0.2857
Coeff Var	22.09709		

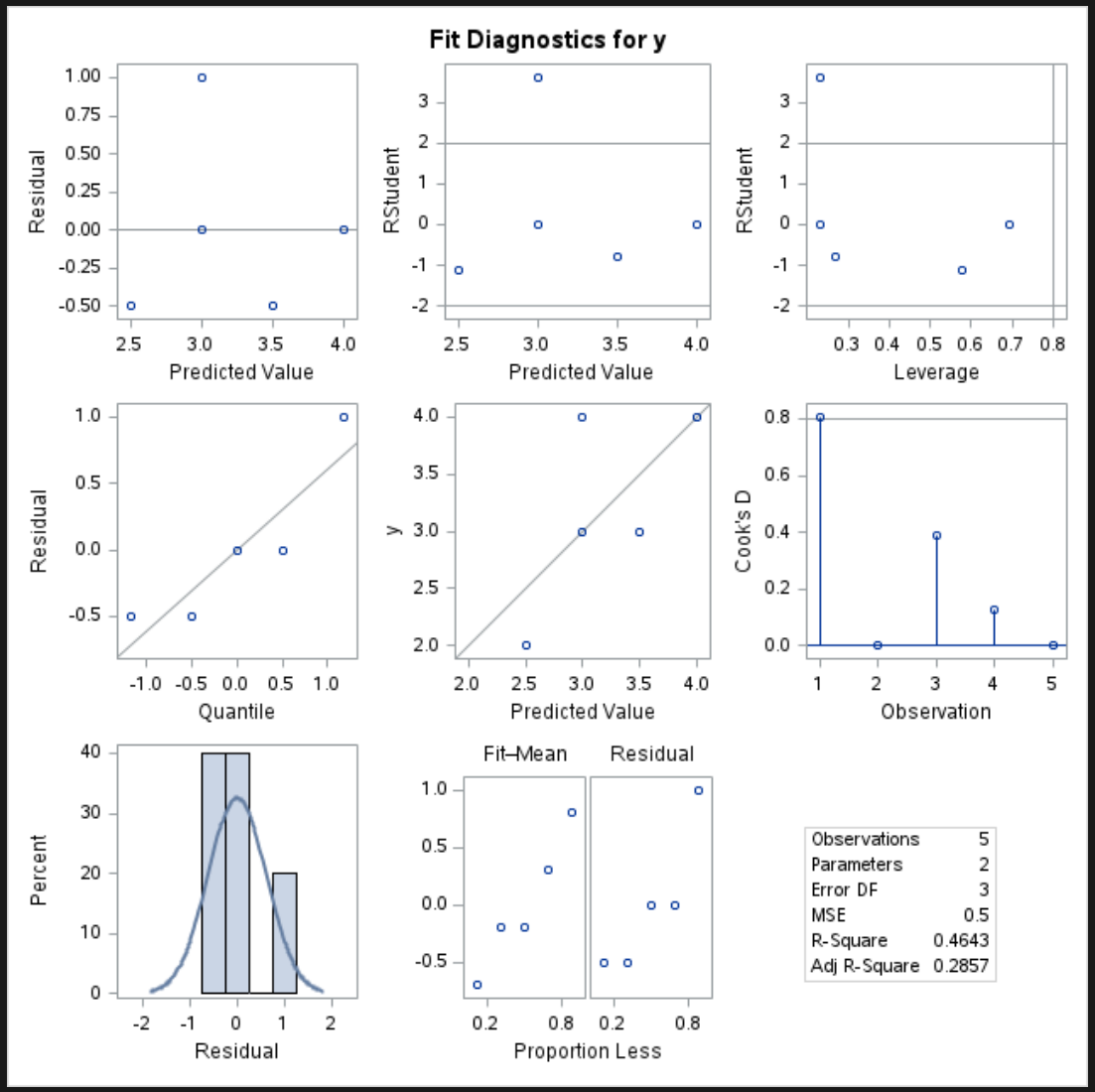
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.00000	0.80861	2.47	0.0898
x	1	0.50000	0.31009	1.61	0.2053

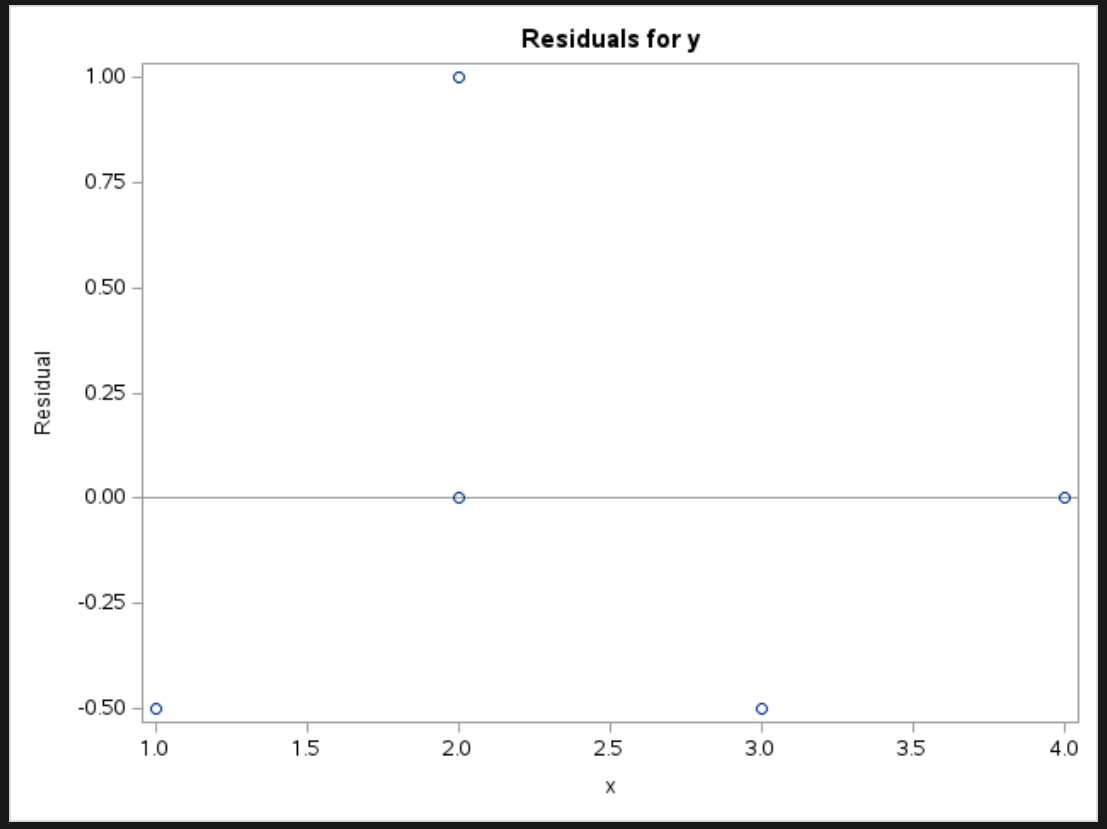
The REG Procedure
Model: MODEL1
Dependent Variable: y

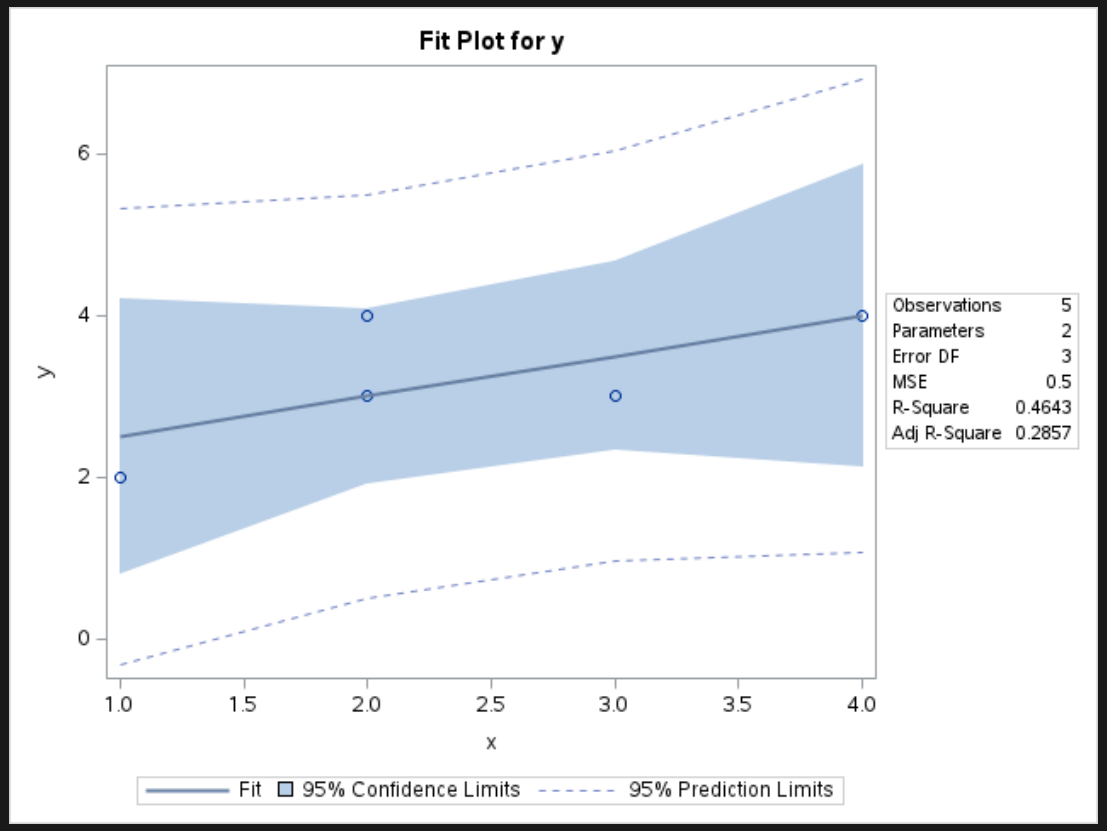
Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	2	2.5000	0.5371	0.7908	4.2092	-0.5000
2	3	3.0000	0.3397	1.9190	4.0810	0
3	4	3.0000	0.3397	1.9190	4.0810	1.0000
4	3	3.5000	0.3669	2.3324	4.6676	-0.5000
5	4	4.0000	0.5883	2.1276	5.8724	0

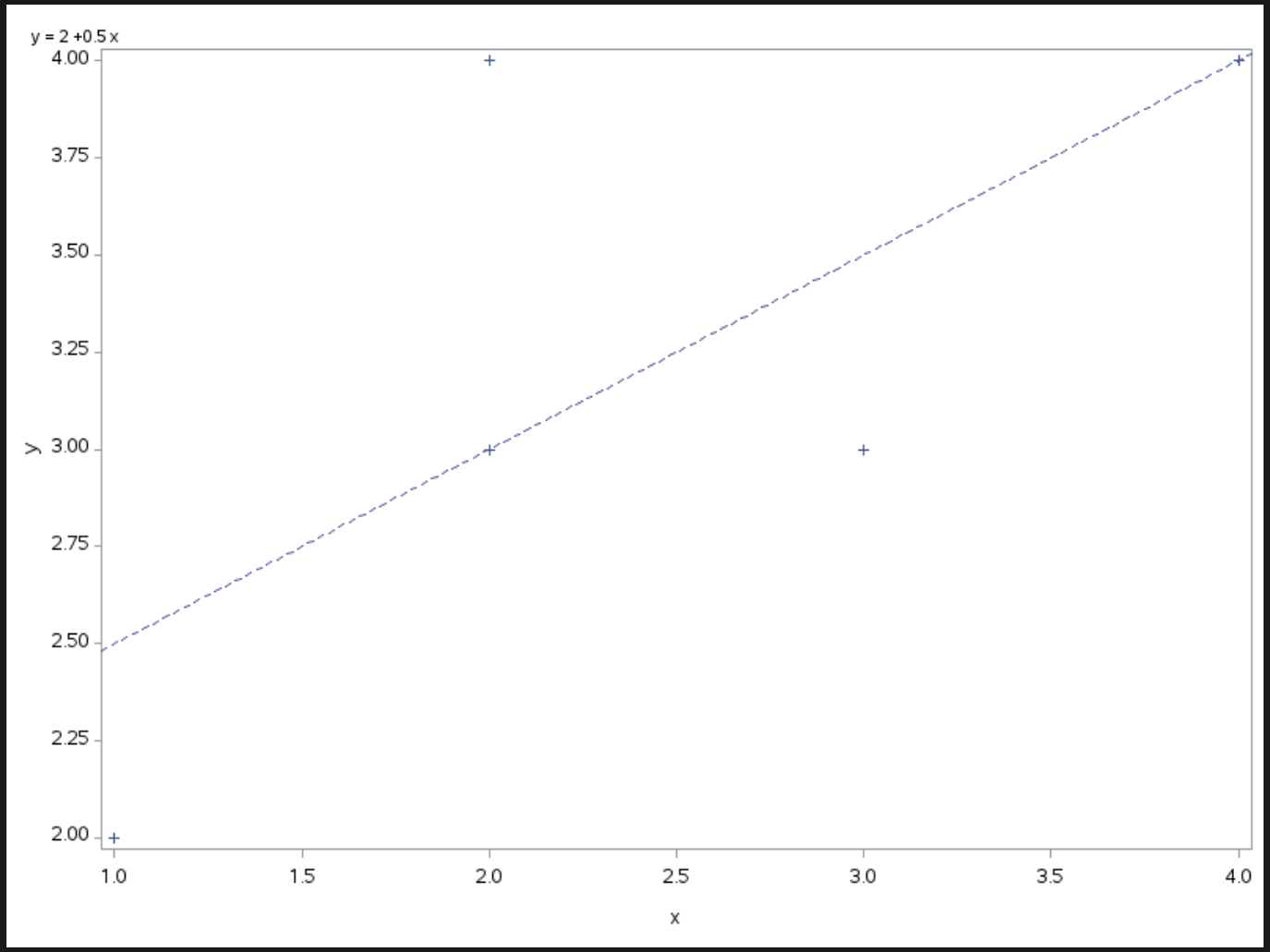
Sum of Residuals	0
Sum of Squared Residuals	1.50000
Predicted Residual SS (PRESS)	3.55484


Fit Diagnostics for y
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
y
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 5
Parameters 2
Error DF 3
MSE 0.5
R-Square 0.4643
Adj R-Square 0.2857

Residual plot of x 's vs $y - \hat{y}$'s




Fit Plot for y
y
x
6
4
2
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Observations 5
Parameters 2
Error DF 3
MSE 0.5
R-Square 0.4643
Adj R-Square 0.2857
Fit
95% Confidence Limits
95% Prediction Limits


y = 2 +0.5 x
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
y
1.0
1.5
2.0
2.5
3.0
3.5
4.0
x

How good is the fit?

- R^2 closer to 1 than to 0.
- Root mean square error (RMSE) close to 0.
- Residual plot has no systematic pattern, and no changing spread (homoscedasticity) as x varies
- No significant outliers in x- and y- values (which overly influences the line's parameters)