Statistical Data Mining
MATH 4720
Lecture 6
Data Transformation

Gaurav Gupta

GEH A401

Multiple regression

$$\beta = \left(X^T X\right)^{-1} X^T Y$$

$$X' = X^T$$

x	y
4.0	33
4.5	42
5.0	45
5.5	51
6.0	53
6.5	61
7.0	62

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.75 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} 347 \\ 1975 \end{bmatrix} \quad (X'X)^{-1} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix}$$

$$b = (X'X)^{-1}X'Y = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix} \begin{bmatrix} 347 \\ 1975 \end{bmatrix} = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

This called Least squares estimates.

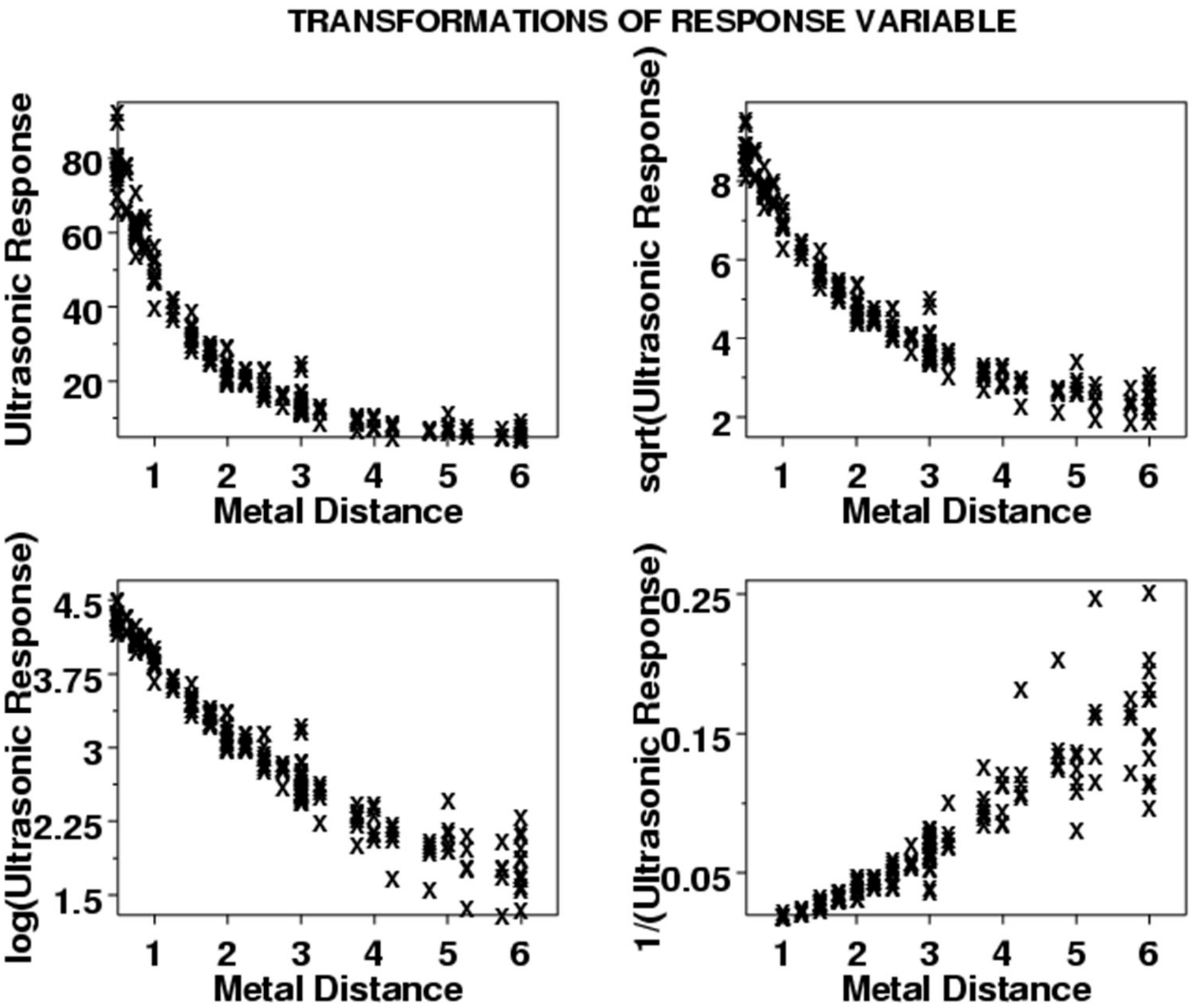
Important distinctions

- **Normalization** is the process of scaling in respect to the entire data range so that the data has a range from 0 to 1.
- **Standardization** is the process of transforming in respect to the entire data range so that the data has a mean of 0 and a standard deviation of 1. It’s distribution is now a Standard Normal Distribution.
- **Transformation** is the application of the same calculation to every point of the data separately.

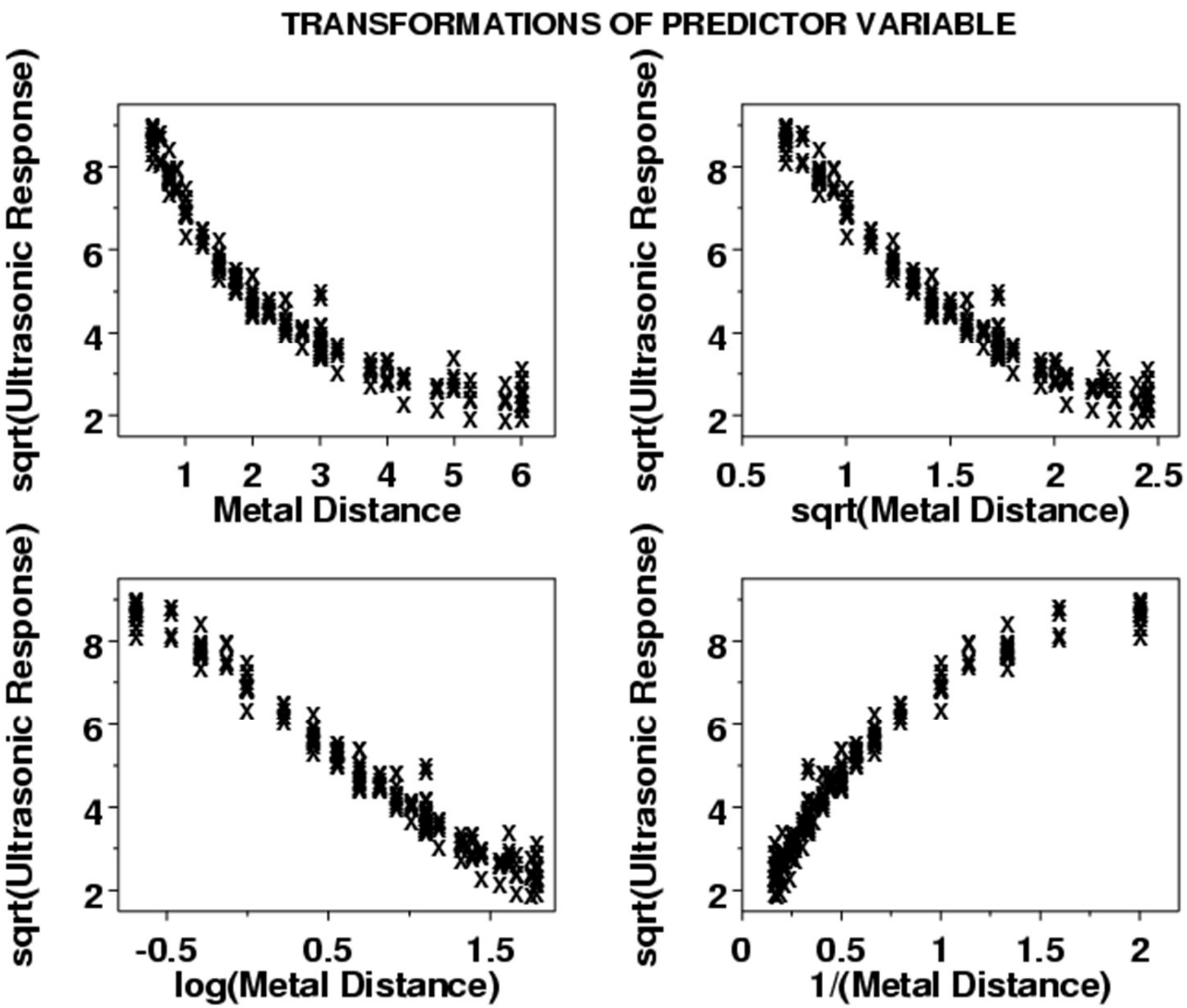
Transformations to improve fit

- If **important predictor variables are omitted**, see whether adding the omitted predictors improves the model.
- Transforming variables can be done to correct for outliers and assumption failures (normality, linearity, and homoscedasticity/homogeneity).
- If there are **unequal error variances**, try transforming the response and/or predictor variables or use "**weighted least squares regression**."
- If an **outlier** exists, try using a "**robust estimation procedure**."
- If **error terms are not independent**, try fitting a "**time series model**."

Example



TRANSFORMATIONS OF RESPONSE VARIABLE
Ultrasonic Response
80
60
40
20
1 2 3 4 5 6
Metal Distance
sqrt(Ultrasonic Response)
8
6
4
2
1 2 3 4 5 6
Metal Distance
log(Ultrasonic Response)
4.5
3.75
3
2.25
1.5
1 2 3 4 5 6
Metal Distance
1/(Ultrasonic Response)
0.25
0.15
0.05
1 2 3 4 5 6
Metal Distance


TRANSFORMATIONS OF PREDICTOR VARIABLE
sqrt(Ultrasonic Response)
8
6
4
2
1 2 3 4 5 6
Metal Distance
sqrt(Ultrasonic Response)
8
6
4
2
0.5 1 1.5 2 2.5
sqrt(Metal Distance)
sqrt(Ultrasonic Response)
8
6
4
2
-0.5 0.5 1.5
log(Metal Distance)
sqrt(Ultrasonic Response)
8
6
4
2
0 0.5 1 1.5 2
1/(Metal Distance)

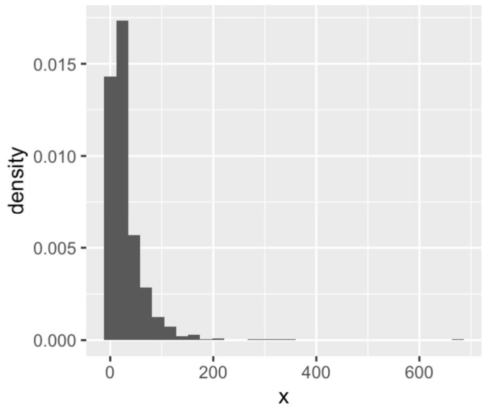
Data Transformation

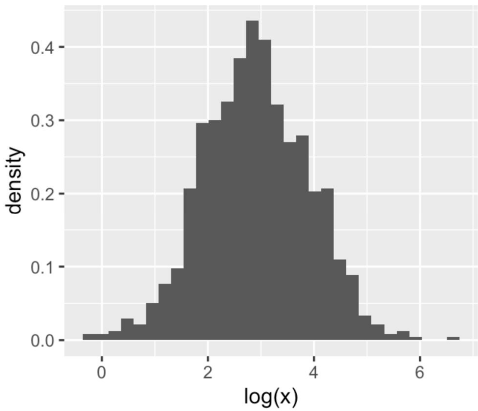
- The easiest way to learn about data transformations is by example
- Four types of log transformations:

$$\text{level} - \text{level regression: } y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$
$$\text{log} - \text{level regression: } \ln y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$
$$\text{level} - \text{log regression: } y = b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \cdots + b_k \ln x_k$$
$$\text{log} - \text{log regression: } \ln y = b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \cdots + b_k \ln x_k$$

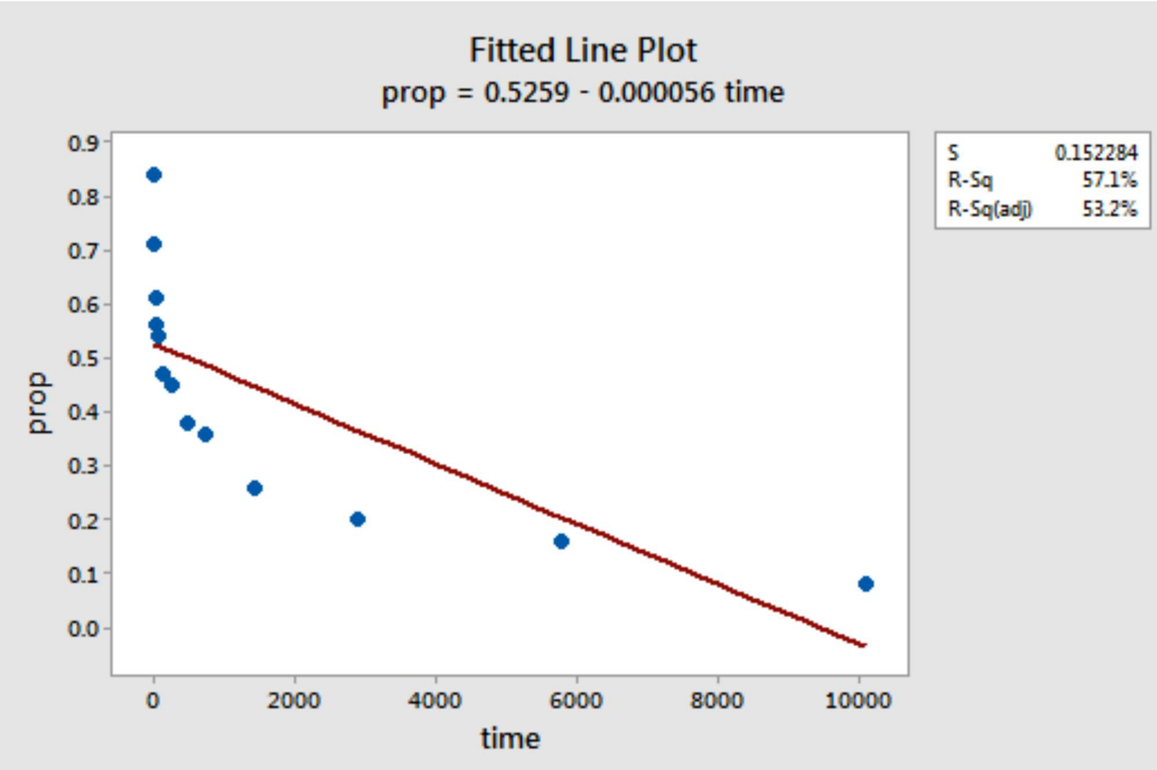
- Small values that are close together are spread further out.
- Large values that are spread out are brought closer together.

Level-level regression : Least Squares for Multiple Regression and Multiple Regression Analysis.
Log-level regression: Exponential Regression
log-log regression: Power Regression



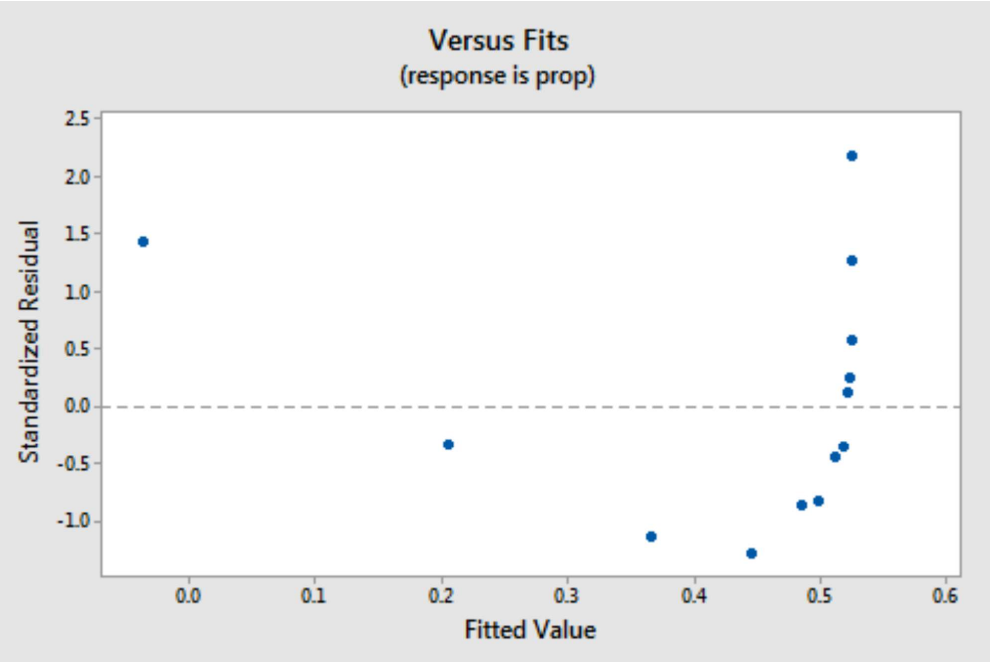


Example

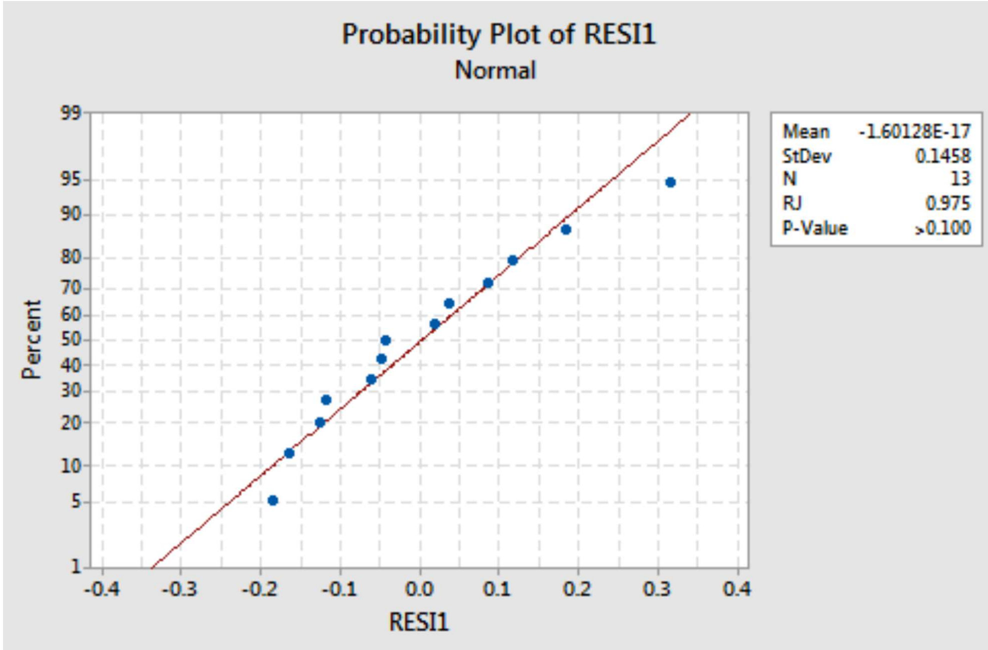


time	prop
1	0.84
5	0.71
15	0.61
30	0.56
60	0.54
120	0.47
240	0.45
480	0.38
720	0.36
1440	0.26
2880	0.20
5760	0.16
10080	0.08

Relationship is not linear

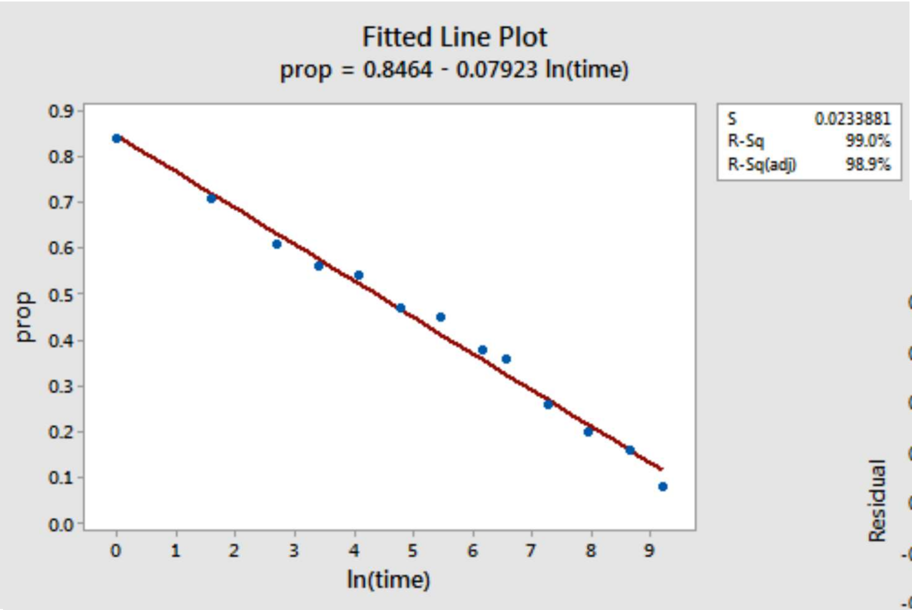


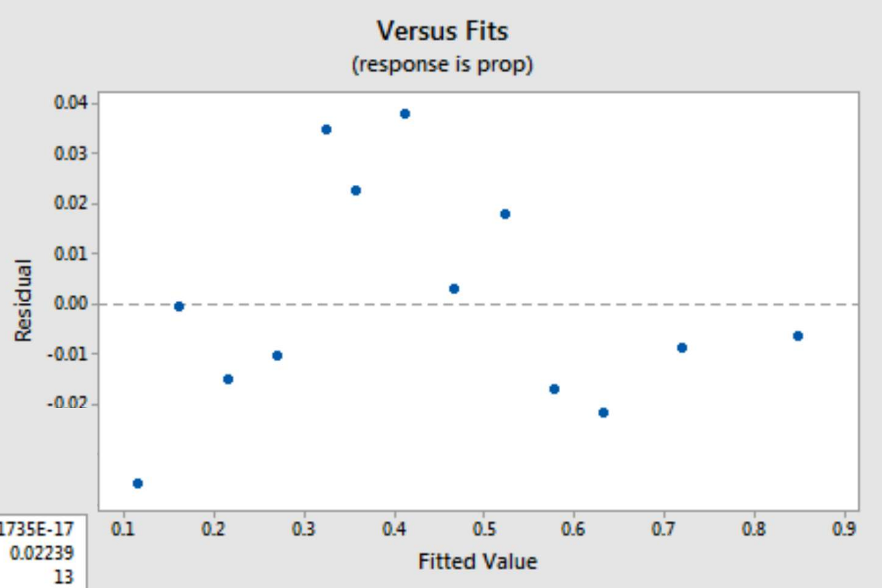
Residuals vs. fit

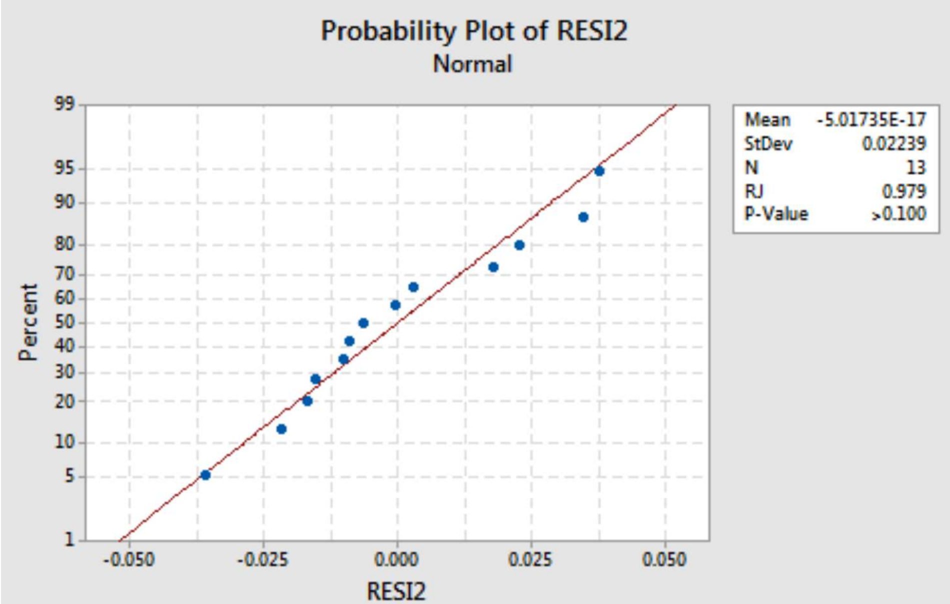


Normal probability plot of the residuals

P-value for this example is large, which suggests that we fail to reject the null hypothesis of normal error terms. There is not enough evidence to conclude that the errors terms are not normal.

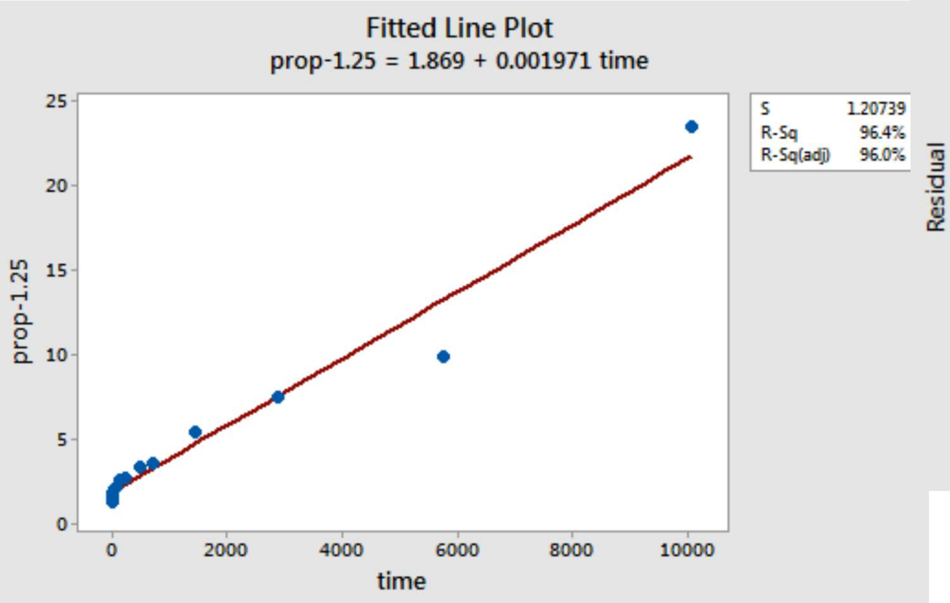


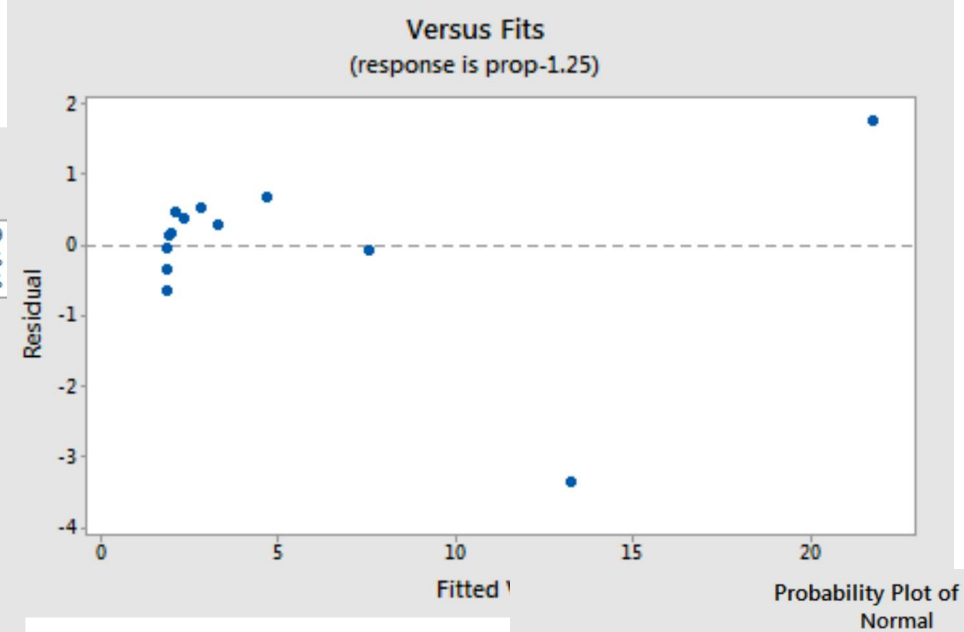


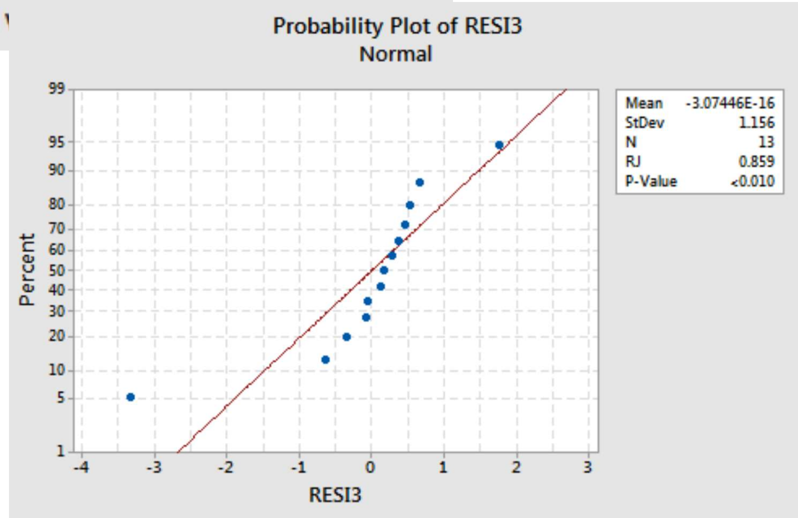


time	prop	lntime
1	0.84	0.00000
5	0.71	1.60944
15	0.61	2.70805
30	0.56	3.40120
60	0.54	4.09434
120	0.47	4.78749
240	0.45	5.48064
480	0.38	6.17379
720	0.36	6.57925
1440	0.26	7.27240
2880	0.20	7.96555
5760	0.16	8.65869
10080	0.08	9.21831

What if we had transformed the *y* values instead?







Research Question

- What is the nature of the association between time since memorized and the effectiveness of recall?

Answer: The proportion of correctly recalled words is negatively linearly related to the natural log of the time since the words were memorized. Not surprisingly, as the natural log of time increases, the proportion of recalled words decreases.

- Is there an association between time since memorized and effectiveness of recall?

The P-value is < 0.001. There is significant evidence at the 0.05 level to conclude that there is a linear association between the proportion of words recalled and the natural log of the time since memorized.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.58841	0.58841	1075.70	0.000
Residual Error	1	0.00602	0.00055		
Total	12	0.59443			

- What proportion of words can we expect a randomly selected person to recall after 1000 minutes?

Ans: We just need to calculate a prediction interval — with one slight modification,. The natural log of 1000 minutes is 6.91 log-minutes. Asking Minitab to calculate a 95% prediction interval when lntime=6.91

We can be 95% confident that, after 1000 minutes, a randomly selected person will recall between 24.5% and 35.3% of the words.

Values of Predictions for New Obervaions

New Obs	lntime
1	6.91

Prediction Values for New Obervaions

New Obs	Fit	SE Fit	95% CI	95% PI
1	0.29896	0.00766	(0.282, 0.316)	(0.245, 0.353)

Try!

	A	B	C	D	E	F	G
3	**Log-level transformation**						
4							
5	Color	Quality	Price		Color	Quality	Ln Price
6	7	5	58		7	5	4.060443
7	3	7	11		3	7	2.397895
8	5	8	24		5	8	3.178054
9	8	1	11		8	1	2.397895
10	9	3	31		9	3	3.433987
11	5	4	15		5	4	2.70805
12	4	0	5		4	0	1.609438
13	2	6	8		2	6	2.079442
14	8	7	84		8	7	4.430817
15	6	4	24		6	4	3.178054
16	9	2	21		9	2	3.044522

Which transformation to pick?

- When transforming data you will lose information about the data generation process and you will lose interpretability of the values, too.
- You can consider to back-transform the variable at a certain step in your analysis.
- Logarithm should be used if data generation effects were multiplicative and the data follows order of magnitudes. Roots should be used if the data generation involved squared effects.

Some Transformations

Right (positive) skewed data:

- **Root** $\sqrt[n]{x}$**.** Weakest transformation, stronger with higher order root. For negative numbers special care needs to be taken with the sign while transforming negative numbers:
- **Logarithm** ***log(x). T***he strength of this transformation can be somewhat altered by the root of the logarithm. It can not be used on negative numbers or 0, here you need to shift the entire data by adding at least $|min(x)|+1$.
- **Reciprocal** ***1/x.*** Strongest transformation, the transformation is stronger with higher exponents, e.g. $1/x^3$. This transformation should not be done with negative numbers and numbers close to zero, hence the data should be shifted similar as the log transform.

Left (negative) skewed data

- **Reflect Data and use the appropriate transformation for right skew.** Reflect every data point by subtracting it from the maximum value. Add 1 to every data point to avoid having one or multiple 0 in your data.
- **Square x^2.** Stronger with higher power. Can not be used with negative values.
- **Exponential e^x.** Strongest transformation and can be used with negative values. Stronger with higher base.

Light & heavy tailed data

- **Subtract the data points from the median and transform.** Deviations of the tail from normality are usually less critical than skewness and might not need transformation after all. The subtraction from the median sets your data to a median of 0. After that use an appropriate transformation for skewed data on the absolute deviations from 0 on either side. For **heavy-tailed** data use transformations for right skew to pull in on the median and for **light-tailed** data use transformations for left skew to push data away from the median.

Automatic Transformations

Power(p)	Transformation	Name
2	Y^2	Square
1	Y (No transformation)	Original Data
½	$\sqrt{Y}$	Square root
"0"	log Y or $\log_{10}$ (Y)	Logarithm
-½	$-1 / \sqrt{Y}$	Reciprocal Root
-1	-1 / Y	Reciprocal
-2	-1 / Y^2	Reciprocal Square

Tukey Ladder of Powers:

- The Tukey ladder of powers is a way to change the shape of a skewed distribution so that it becomes normalor nearly-normal. It can also help to reduce error variability (heteroscedasticity).
- Tukey (1977) created a table of powers (numbers to which data can be raised). It's possible to have an infinite number of powers, but very few are actually in common use. The following table shows the most commonly used transformations, with exponents ranging from -2 to 2.

- Going up the ladder reduces negative skew. To choose a transformation for negative skew, start with Y^2, then plot the data to see how the transformation has affected the data. An exponential function such as Y^2 will have a greater effect on larger numbers: 1,000 will become 1,000,000 while 5 will become 25. Due to the fact that Y^2 increases large numbers by such a massive amount, it's rare to see transformations above y2.
- For positive skews, start with log Y and move down the ladder, plotting as you go to see the effects. Logarithmic functions (with base 10) reduce large numbers more than small numbers. For example, 100,000 reduces to 5 and 100 reduces to 2