Mr Murphy
AP Statistics
4.4 Nonlinear Relationships and Transformations
HW Pg. 257 #5.56, 5.60, 5.61, 5.62

#Goals: 1. Perform power regression.
2. Perform transformation regression.

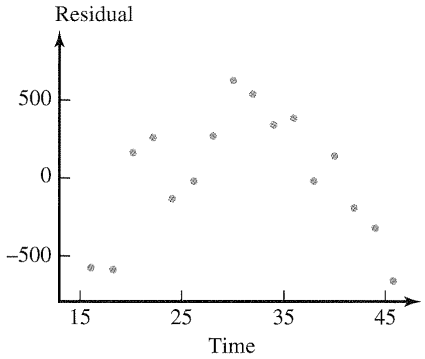
- We will look at two common approaches to fitting nonlinear relationships: power regression and transformations.

Polynomial Regression

Ex1 The focus of many agricultural experiments is to study how the yield of a crop varies with the time at which it is harvested. Accompanying data is given where the variables are x = time between flowering and harvesting (days) and y = yield of paddy, a type of grain farmed in India (in kilograms per hectare):

Plot the data. The residual plot is shown below.

X	Y
16	2508
18	2518
20	3304
22	3423
24	3057
26	3190
28	3500
30	3883
32	3823
34	3646
36	3708
38	3333
40	3517
42	3214
44	3103
46	2776



Does a linear relationship seem to be a reasonable fit? Why or why not?

What kind of fit seems more reasonable?

Look for a model of the form $\hat{y} = a + b_1 x + b_2 x^2$ by using the quadratic regression button on your calculator.

The corresponding MINITAB output is shown below.

The regression equation is
y = −1075 + 294x − 4.55x**2

Predictor	Coef	SE Coef	T	P
Constant	−1074.6	618.0	−1.74	0.106
x	293.92	42.23	6.96	0.000
x**2	−4.5464	0.6753	−6.73	0.000

S = 204.1 R-Sq = 79.4% R-Sq(adj) = 76.2%

Ex2 Researchers have examined a number of climatic variables in an attempt to understand the mechanisms that govern rainfall runoff. A study examined the relationship between x = cloud cover index and y = sunshine index.

Suppose that the cloud cover index can have values between 0 and 1. Consider the accompanying data.

Plot the data. Does a linear relationship seem to be a reasonable fit? Why or why not? What kind of fit seems more reasonable?

X	Y
0.2	10.98
0.5	10.94
0.3	10.91
0.1	10.94
0.2	10.97
0.4	10.89
0.0	10.88
0.4	10.92
0.3	10.86

Here is the following MINITAB regression.

Polynomial Regression

The regression equation is

y = 10.8768 + 1.46036x − 7.25901x**2 + 9.23423x**3

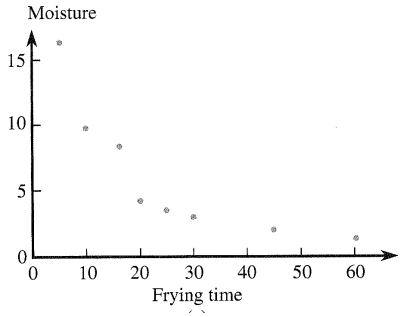
S = 0.0315265 R-Sq = 62.0% R-Sq (adj) = 39.3%

What is the predicted sunshine index for a day when the cloud cover index is 0.45?

Transformation Regression

Ex3 No tortilla chip lover likes soggy chips, so it is important to find characteristics of the production process that produce chips with an appealing texture. The following data on x = frying time (in seconds) and y = moisture content (%) are provided

Frying time, x	5	10	15	20	25	30	45	60
Moisture content, y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3



Does a linear model seem like a reasonable fit for this data?

Let's try this transformation: $y_{new} \rightarrow \log(y)$. In other words, change the list of the y values to the log of all the y values.
Is this model a better fit?

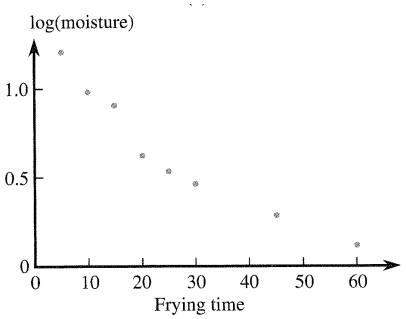
The regression equation is
log(moisture) = 1.14 − 0.0192 frying time

Predictor	Coef	StDev	T	P
Constant	1.14287	0.08016	14.26	0.000
frying t	−0.019170	0.002551	−7.52	0.000

S = 0.1246 R-Sq = 90.4% R-Sq(adj) = 88.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.87736	0.87736	56.48	0.000
Residual Error	6	0.09320	0.01553		
Total	7	0.97057			



What is our transformed regression "line"?
How can we re-write this "line" in terms of the original data?

What is the predicted moisture content when the frying time is 35 seconds?

Ex4 A response variable appears to be exponentially related to the explanatory variable. The natural logarithm of each y-value is taken and the least-squares regression line is found to be $\ln \hat{y} = 1.64 - 0.88x$. Rounded to two decimal places, what is the predicted value of y when $x = 3.1$?

(a) -1.09
(b) -0.34
(c) 0.34
(d) 0.082
(e) 1.09

Ex5 The number of a certain type of bacteria present (in thousands) after a certain number of hours is given in the following chart. Is a linear model a good fit? If yes, what would be the predicted quantity of bacteria after 3.75 hours? If not, perform a natural logarithmic transformation on the number bacteria present and use that model to predict the quantity of bacteria after 3.75 hours.

Hours	Number of Bacteria Present
1.0	1.8
1.5	2.4
2.0	3.1
2.5	4.3
3.0	5.8
3.5	8.0
4.0	10.6
4.5	14.0
5.0	18.0

Checkpoint
Multiple Choice

1. Using least-squares regression, I determine that the logarithm (base 10) of the population of a country is approximately described by the equation
Log (population) = -13.5 + 0.01(year)

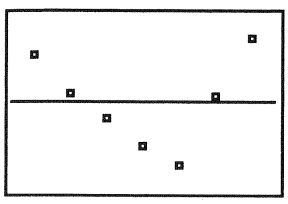
Based on this equation, the population of the country in the year 2000 should be about
(a) 6.5
(b) 665
(c) 2,000,000
(d) 3,162,277
(e) None of the above

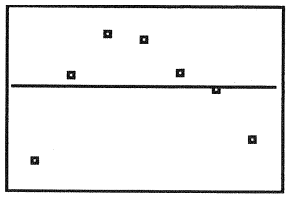
2. Suppose that the scatterplot of X and log Y shows a strong positive correlation close to 1. Which of the following is true?

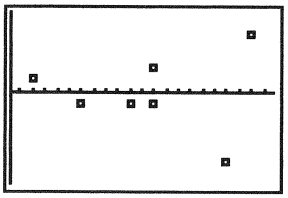
I. The variables X and Y also have a correlation close to 1.
II. A scatterplot of the variables X and Y shows a strong nonlinear pattern.
III.The residual plot of the variables X and Y shows a random pattern.

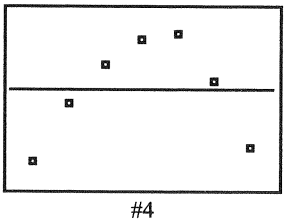
(a) I only (b) II only (c) III only (d) I and II (e) I, II, and III

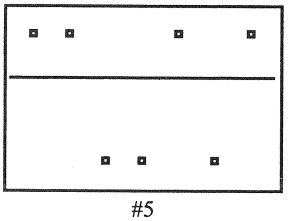
3. A researcher made a scatterplot from some previously collected data. The data was clearly nonlinear in shape. The researcher then tried a variety of transformations on the data in an attempt to linearize the results. The residual plot for each is shown below.


#1


#2


#3


#4


#5

Which of the transformations was best at linearizing the data?
(a) #1 (b) #2 (c) #3 (d) #4 (e) #5

4. A residual:

(a) is the amount of variation explained by the least-squares regression line of y on x.
(b) is how much an observed y value differs from a predicted y value.
(c) predicts how well x explains y.
(d) is the total variation of the data points.
(e) should be smaller than the mean of y.

5. Which of the following would indicate the strongest relationship between two variables?
(a) $r = 0.35$
(b) $r = -.28$
(c) $r = .21$
(d) $r^2 = .01$
(e) $r^2 = .23$

6. The coefficient of determination, r^2, between two variables is computed to be 81%. Which of the following statements must be true?

(a) Large values of the explanatory variable correspond with large values of the response variable.
(b) Large values of the explanatory variable correspond with small values of the response variable.
(c) A cause and effect relationship exists between the explanatory and response variables.
(d) There is a strong, positive, linear relationship between the explanatory and response variables.
(e) Approximately 81% of the variability in the response variable is explained by regression on the explanatory variable.

7. If the model for the relationship between the score on the AP Statistics Exam (y) and the number of hours spent preparing for the test (x) was $\log \hat{y} = 0.1 + 1.9 \log x$, determine the residual if a student studied 9 hours and earned an 85.

(a) 6.53
(b) 3.14
(c) 15.23
(d) 0
(e) -4.86