AP Statistics
Unit 02 REDO – Day 02 Notes

Name Key 2018
Period 1

ransforming Data!

Sometimes our data doesn't appear in a straight line. Since we want to use a least squares regression LINE, we may need to adjust our data. Additionally, r measures the strength of the LINEAR relationship between two variables, which is yet another reason to transform the data. We can use transformations like square roots, cube roots, squaring, cubing, etc., but the two common transformation we will focus on include logs.

*for the sake of my sanity (and not having to use ten billion special symbols) the hats have been omitted.

LINEAR MODEL	EXPONENTIAL MODEL	POWER MODEL
$\hat{y} = a + bx$	$\hat{y} = 10^a 10^{bx}$	$\hat{y} = 10^a x^b$
	$\widehat{\log(y)} = a + bx$	$\widehat{\log(y)} = a + b\log(x)$
L1 vs. L2 (x, y)	L1 vs. L4 (x, log y)	L3 vs. L4 (log(x), log(y))

L1 = x
L2 = y
L3 = log(x)
L4 = log(y)

Things to check when comparing models:

1. R^2. We want this to be close to 1. The closer, the better. If you look at r instead, make sure it's close to -1 or +1.

2. The scatterplots. We want to choose the model that looks most linear.

3. The residual plots. If our comparison of R^2 is too close to call, and our plots look linear making it difficult to choose one, check out the residual plot! No pattern is good. Usually one plot has less of a pattern than the others.

EXAMPLE: Below is a table of data representing bacteria growth (in hundreds) after a certain amount of time (in minutes).

Time	Count
1	15
2	19
3	21
4	32
5	36
6	38
7	56
8	60
9	104
10	106
11	142
12	166
13	197
14	211
15	355

original: curved $R^2 = 0.9163$

exponential: linear $R^2 = 0.9870$ $\hat{y} = a \cdot b^x$

power: curved $R^2 = 0.8867$

I will transform using an exponential model because the scatterplot of x vs. log y looks linear and R^2 is closest to 1.

Is a linear model a good fit? If not, suggest a better model.

No. the scatterplot of time vs. bacteria count has a strong non-linear pattern (curved), so it is not a good fit.

Write the equation for the LSRL of your chosen model.

$\widehat{\log(\text{count})} = 1.0955 + 0.09127(\text{time})$

$\widehat{\text{count}} = 10^{1.0955 + 0.09127(\text{time})}$

$\widehat{\text{count}} = 12.4595 \cdot 1.2339^{(\text{time})}$

Calculate and interpret the residual value at 7 minutes.

(7, 56) (7 = time in minutes, 56 = actual # of bacteria (in hundreds))

$\hat{y} = 12.4595 \cdot 1.2339^{(7)} = 54.2576$ hundreds of bacteria (predicted)

residual $= y - \hat{y}$

$= 56 - 54.2576 = 1.742$ hundreds of bacteria

At 7 minutes, there are 174.2 more bacteria than we predicted there to be.