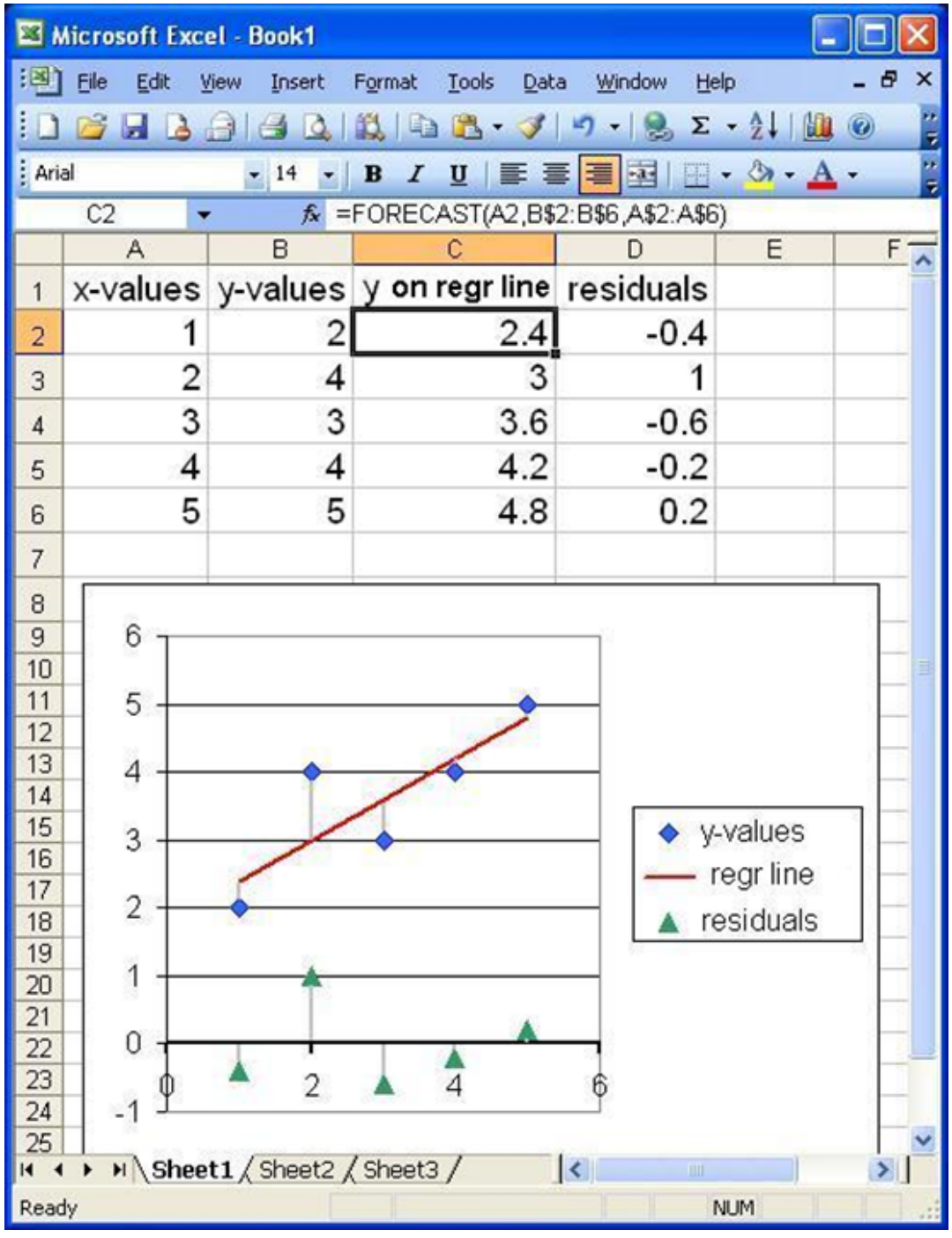
Unit 5: Regression


Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
Arial 14
C2
=FORECAST(A2,B$2:B$6,A$2:A$6)
x-values
y-values
y on regr line
residuals
1 2 2.4 -0.4
2 4 3 1
3 3 3.6 -0.6
4 4 4.2 -0.2
5 5 4.8 0.2
6
5
4
3
2
1
0
-1
0
2
4
6
y-values
regr line
residuals
Sheet1 Sheet2 Sheet3
Ready
NUM

Psych Dept's version

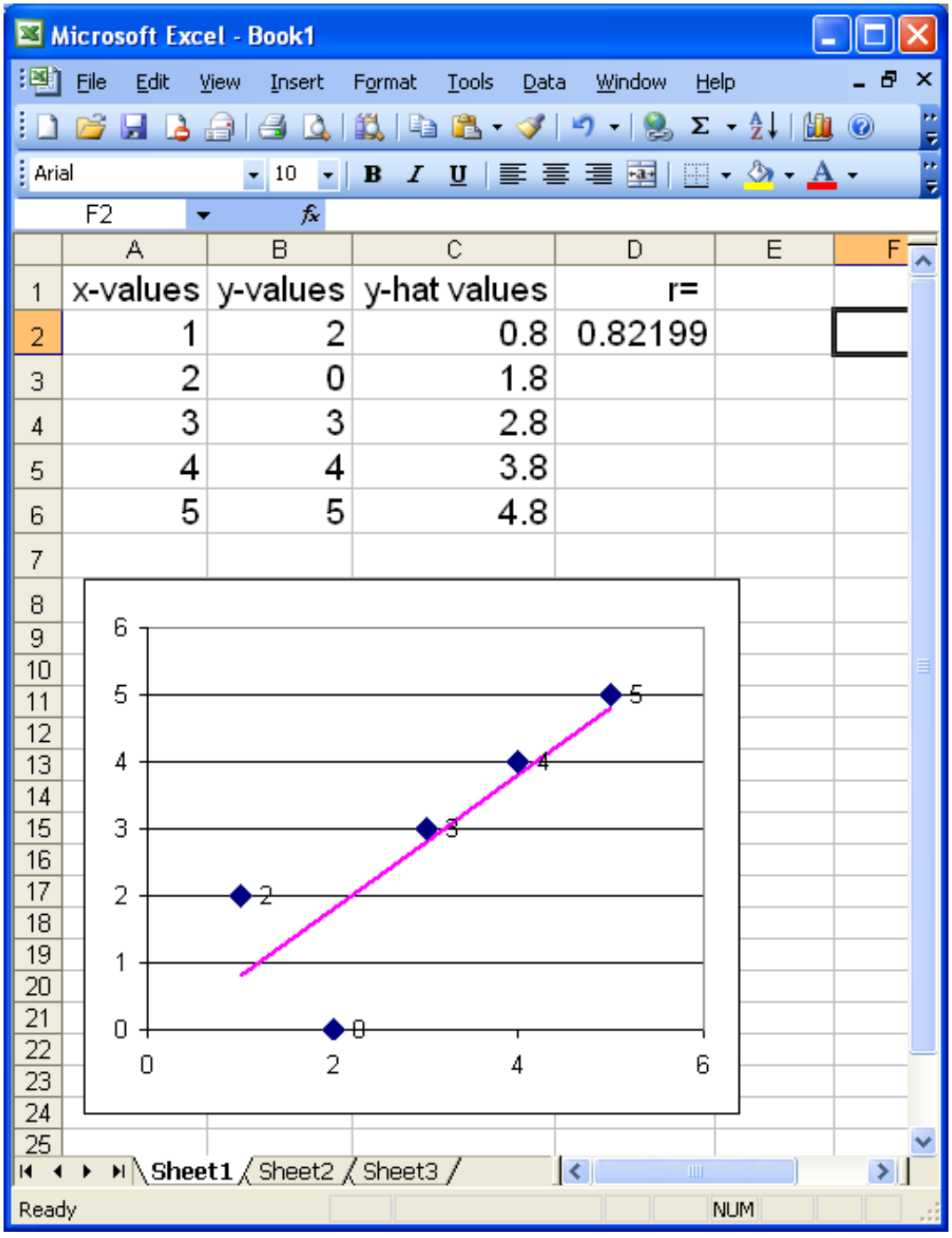
Regression line is $z_y = rz_x$

i.e., change in x by 1 std unit changes y by the fraction r std unit.

Algebraically equivalent:

$$\frac{y - \mu_y}{\sigma_y} = r\frac{x - \mu_x}{\sigma_x}$$

$$y - \mu_y = r\left(\frac{\sigma_y}{\sigma_x}\right)(x - \mu_x)$$


Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
Arial 10
F2
x-values
y-values
y-hat values
r=
0.82199
Sheet1 Sheet2 Sheet3
Ready
NUM

Projecting %iles with regression

(Assumes both variables are normal, ...
but avgs and std devs aren't needed)

Suppose # of hairs on a man's head and his IQ have $r = -.7$:

- If he is at 80th %ile in # hairs, about what is his IQ %ile?
- If he is at 10th %ile in IQ, about what is his hair %ile?

Normal approx within a vertical slice (???)

Ex (cntd): SAT scores vs. 1st sem GPA: $\mu_s = 1200$, $\sigma_s = 200$, $\mu_g = 2.5$, $\sigma_g = 1$, $r = .3$

(and assume homoscedastic):

Suppose a student has a 1300 SAT and a 3.7 GPA. What %ile does that make her GPA among the students who got 1300 SATs?

- In that group (as in all the other vertical slices), best guess for σ is RMS error for regr, $\approx .95$,
- and by regr, best guess for avg within slice is 2.65,
- so in slice, her GPA in std units is $(3.7 - 2.65)/.95 \approx 1.10$
- and by normal table, that's 86th %ile.

Warning on regression:

Don't extrapolate.

Ex: Grade inflation at CU: GPA avg in F1993 was 2.91, increasing 0.007/sem ($r = 0.95$). By S2074, avg GPA will be 4.00.(?)

HERE BE DRAGONS

From this point on in this presentation, the material is not in the text.

In fact, our authors specifically warn against some methods (like transforming data), but the methods are commonly used.

This material will not be on an exam, ...

but multivariate regression is used in Midterm Project I.

What does r^2 measure?

It says how much better for predicting y is using regr line (i.e., using the y-value $\hat{y}$ on the regression line at that x-value) than just always using $\overline{y}$.

Difference of SSE (sum of squares of errors) using avg [i.e., $\sum(y - \overline{y})^2$] vs. SSE using regr [i.e., sum of squares of residuals$\sum(y - \hat{y})^2$], divided by SSE using avg ...

... which $= r^2$.

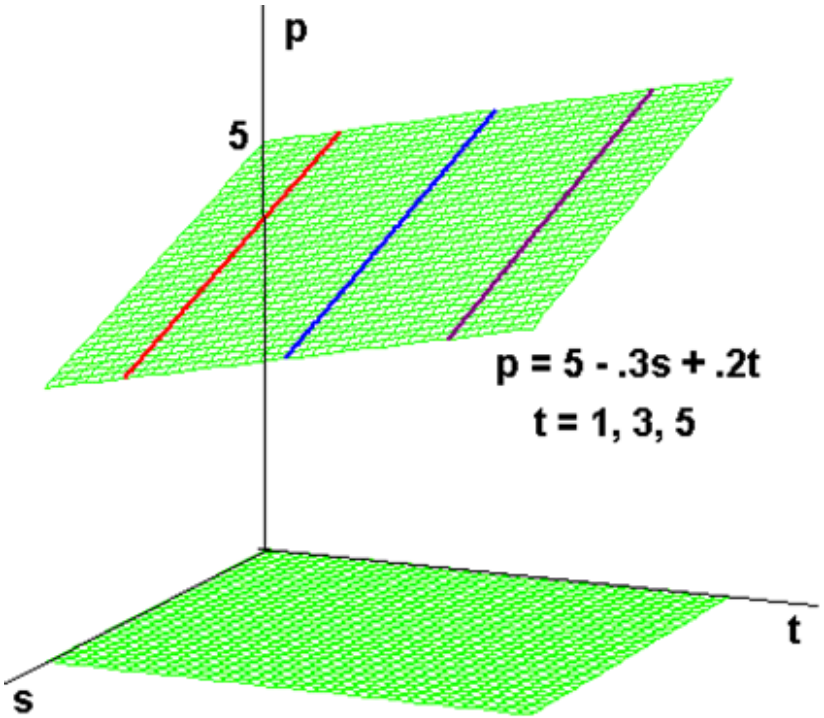
so, if $r^2 = 0.4$, say, "regression results in a 40% improvement in projection".

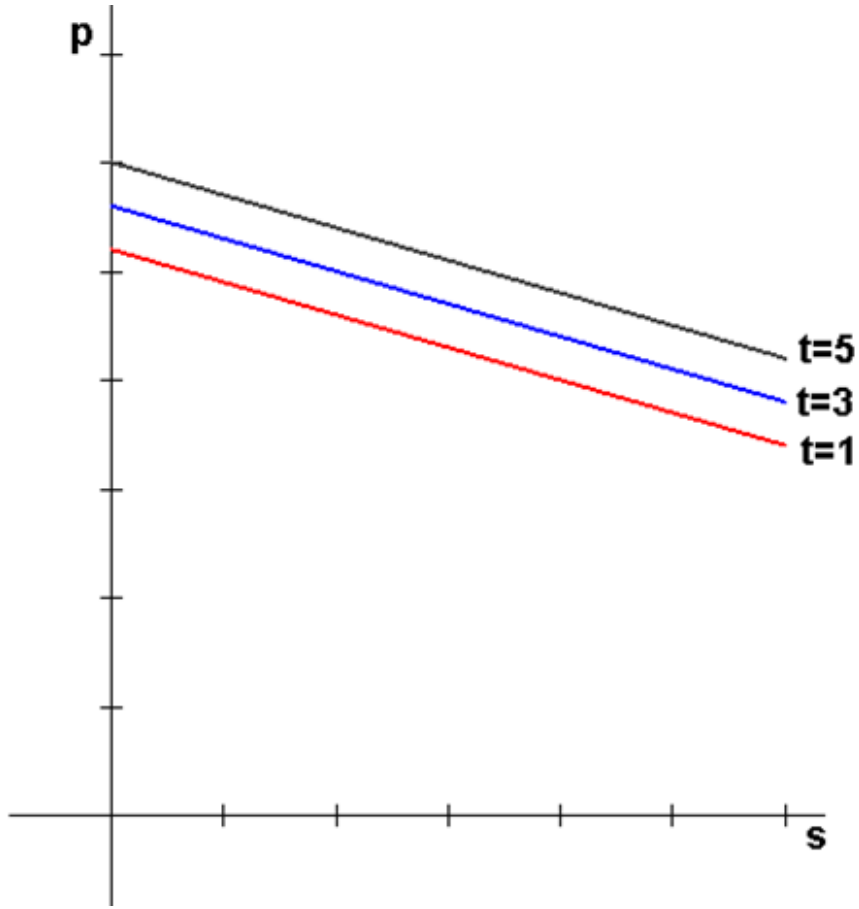
Multiple (linear) regression

If there is more than one explanatory variable (x_1, x_2, x_3 say) and one response variable (y), it may be useful to model it as
$y = a + b_1x_1 + b_2x_2 + b_3x_3$

Ex: Aspirin is so acidic that it often upsets the stomach, so it is often administered with an antacid — which limits its effect. Suppose the pain, measured by the rating of headache sufferers, is given by $p = 5 - .3s + .2t$ where s is the aspirin dose and t is the antacid dose.

Graphs of aspirin example





Multiple regression

As with simple regression, there is a (multiple) correlation R (indep of units) that measures how closely the data points (in 3-space or higher dims) follow a (hyper)plane.

(What does the sign mean, because y can go up when x_1 goes up or when x_2 goes down?)

In this case R^2 is easier to understand (and means the same as before), so it appears in the computer outputs as well.

The next page is Excel output from a (fictional) economic multiple regression. (Bold italics = added)

Example of Multiple Regression (Words in ***bold italics*** are added)

Y-data Inflation	X1-data Unemployment	X2-data Per Capita GNP	<-***Labels***
1	3	1	
2	2	2	
3	2	6	
4	5	4	
5	6	6	
6	4	3	

Tools->Data Analysis -> Regression: Y-range a4:a10, X-range b4:c10, Labels On, Output Range A14

^
Command

SUMMARY OUTPUT

Regression Statistics		
Multiple R	0.717300207	<- ***Multivariable Correlation Coefficient***
R Square	0.514519587	<- ***squared (so that it is between 0 and 1)***
Adjusted R Square	0.190865978	<- ***adjusted for the number of variables***
Standard Error	1.68284553	<- ***Standard Error of Y-estimate***
Observations	6	

Sum of Squares ***Mean of the sum of squares (RMS)***

ANOVA		v	v			
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	9.004092769	4.502046385	1.589723003	0.338265394	
Residual	3	8.495907231	2.831969077		^	
Total	5	17.5			***P-value that all slopes are 0***	

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.198499318	1.974901446	0.100510999	0.926279215	-6.086524387	6.483523023
Unemployment	0.613915416	0.497258543	1.234600038	0.30486542	-0.968584681	2.196415513
Per Capita GNP	0.286493861	0.393117395	0.728774316	0.518878217	-0.964582315	1.537570037
^	^	^	^	^	^	^
Label	***Y = A0 + A1*X1 + A2*X2***	***SE for coefficient***	***t for NH that coefficient is 0***	***P-value for NH that coefficient is 0***	***95% conf interval for coefficient***	

Polynomial regression

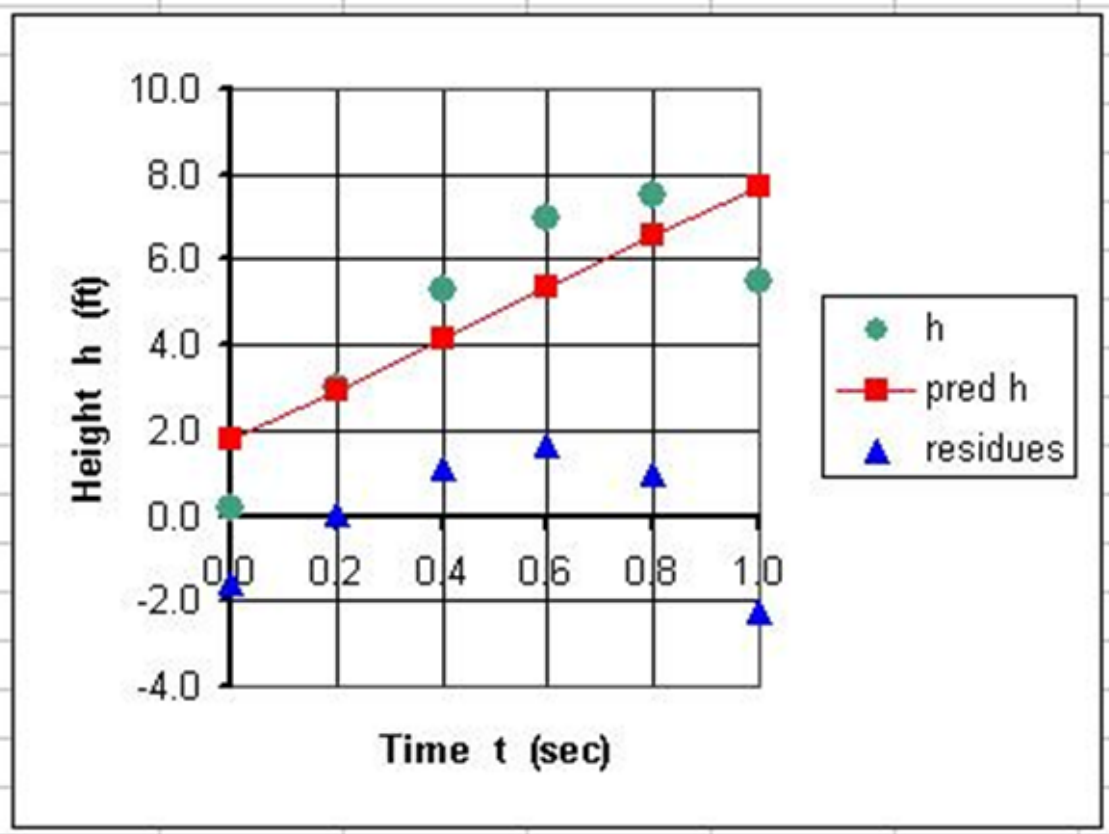
If theory or scatterplot (or plot of residues) suggests a higher-degree polynomial would fit data better than linear regression of y on x , add cols of x^2 (and x^3 and $\ldots$) and do multiple regression.

Ex of theory: path of projectile under gravity, weight vs. height

Ex of fitting: Boston poverty level vs. property values (Midterm Project I)

D3 = =B3-C3

	A	B	C	D	E	F
1	A ball is thrown upwards. The height h in ft after t sec:					
2	**t**	**h**	**pred h**	**residues**		
3	0.0	0.2	1.771	-1.571		
4	0.2	3.0	2.963	0.037		
5	0.4	5.3	4.154	1.146		
6	0.6	7.0	5.346	1.654		
7	0.8	7.5	6.537	0.963		
8	1.0	5.5	7.729	-2.229		
9						



Clipboard | Font | Alignment | Number | Formatting as Table

D2 | =B$19+B$20*A2+B$21*B2

	A	B	C	D
1	t	t^2	h	h-hat
2	0.0	0.00	0.2	-0.08571
3	0.2	0.04	3.0	3.334286
4	0.4	0.16	5.3	5.64
5	0.6	0.36	7.0	6.831429
6	0.8	0.64	7.5	6.908571
7	1.0	1.00	5.5	5.871429

SUMMARY OUTPUT

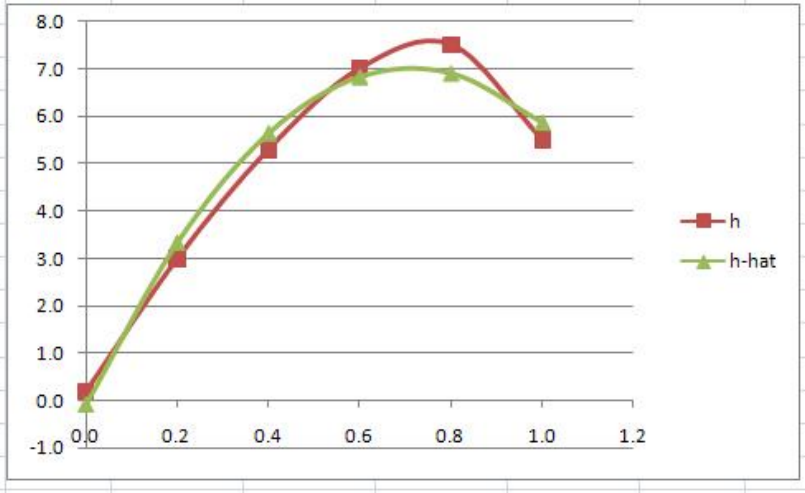
Regression Statistics	
Multiple R	0.988863735
R Square	0.977851487
Adjusted R Square	0.963085811
Standard Error	0.524449825
Observations	6

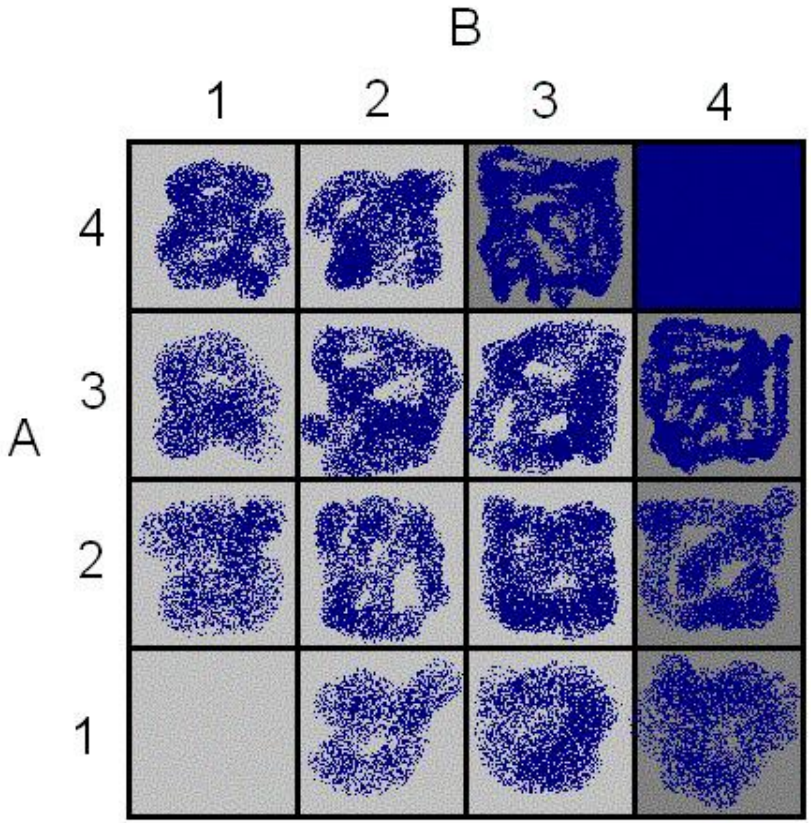
Deleted rows

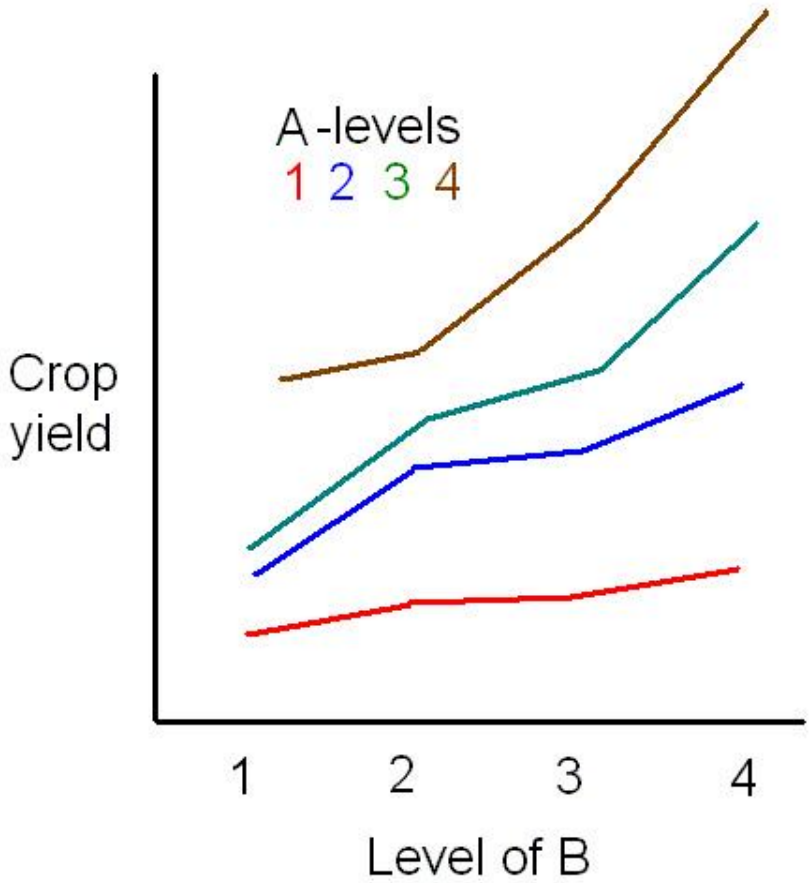
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.08571429	0.475323019	-0.18033	0.868388	-1.59840427	1.4269757	-1.59840427	1.4269757
t	19.88571429	2.235512694	8.895371	0.002996	12.7713152	27.0001134	12.7713152	27.0001134
t^2	-13.9285714	2.145831269	-6.49099	0.007424	-20.7575642	-7.09957863	-20.7575642	-7.09957863

RESIDUAL OUTPUT

Observation	*Predicted h*	*Residuals*
1	-0.08571429	0.285714286
2	3.334285714	-0.334285714
3	5.64	-0.34
4	6.831428571	0.168571429
5	6.908571429	0.591428571
6	5.871428571	-0.371428571







Model with $Y = pA + qB + rAB$?

Correlations in multiple regression

If we add more x-variables in an attempt to approximate a y-variable, the absolute R-value (or R^2-value) cannot go down.

It will probably go up, unless there is no relation at all between the new x-variables and y.

But the correlations between the old x-variables and y may change — may even change sign! — as new x-values are added.

Ex: In the thrown ball example, both t^2 and h go up, at least initially, so without t their correlation is positive. But if we add t , it's a better determiner of h , and t_2 becomes a negative influence on h, namely in the gravity term.

Curvilinear associations

(Linear) regression, as we have studied it, is built to find the best line approximating data. But sometimes the theory, or just a curviness of the data cloud, hints that the equation that best relates the x- and y-values is not a straight line.

In this case, experimenters often "transform" the data, replacing the x-, or y-, or both values by their logs, or their squares, or ... and use regression on the new data to find slopes and intercepts, which translate to exponents or other constants in equations for the original data.

Our authors object!

The next few slides give some common examples of curves used to approximate data.

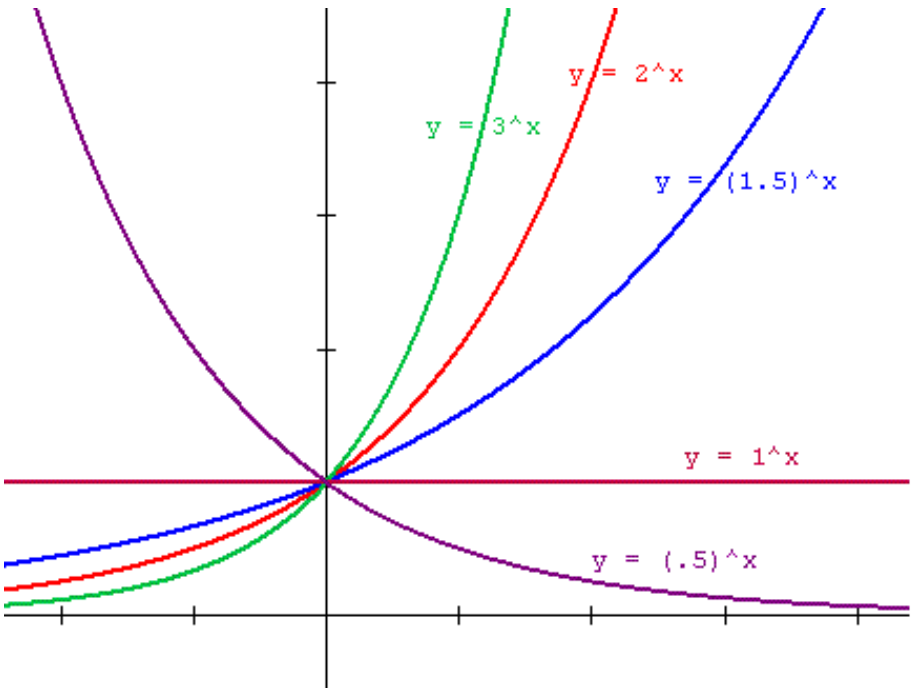
Our authors don't trust them, because the errors in transformed data are no longer normally distributed, and the theory (of regression, for example) expects normally distributed errors.

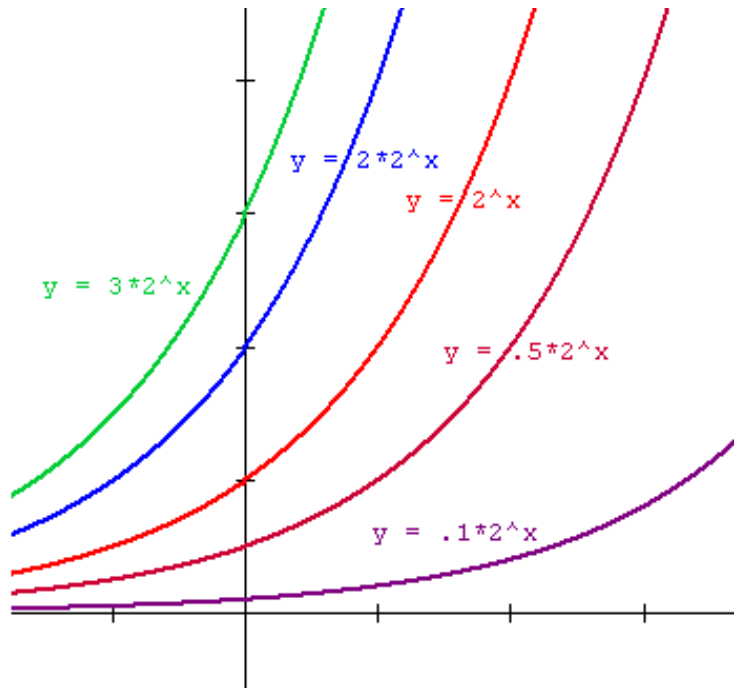
Exponential regression

In many situations, an exponential function fits data better than a linear one.

- population
- radioactive decay

Form: $y = ab^x$ for some constants a, b





Logarithms

$y = b^x$ and $x = \log_b(y)$ say the same thing

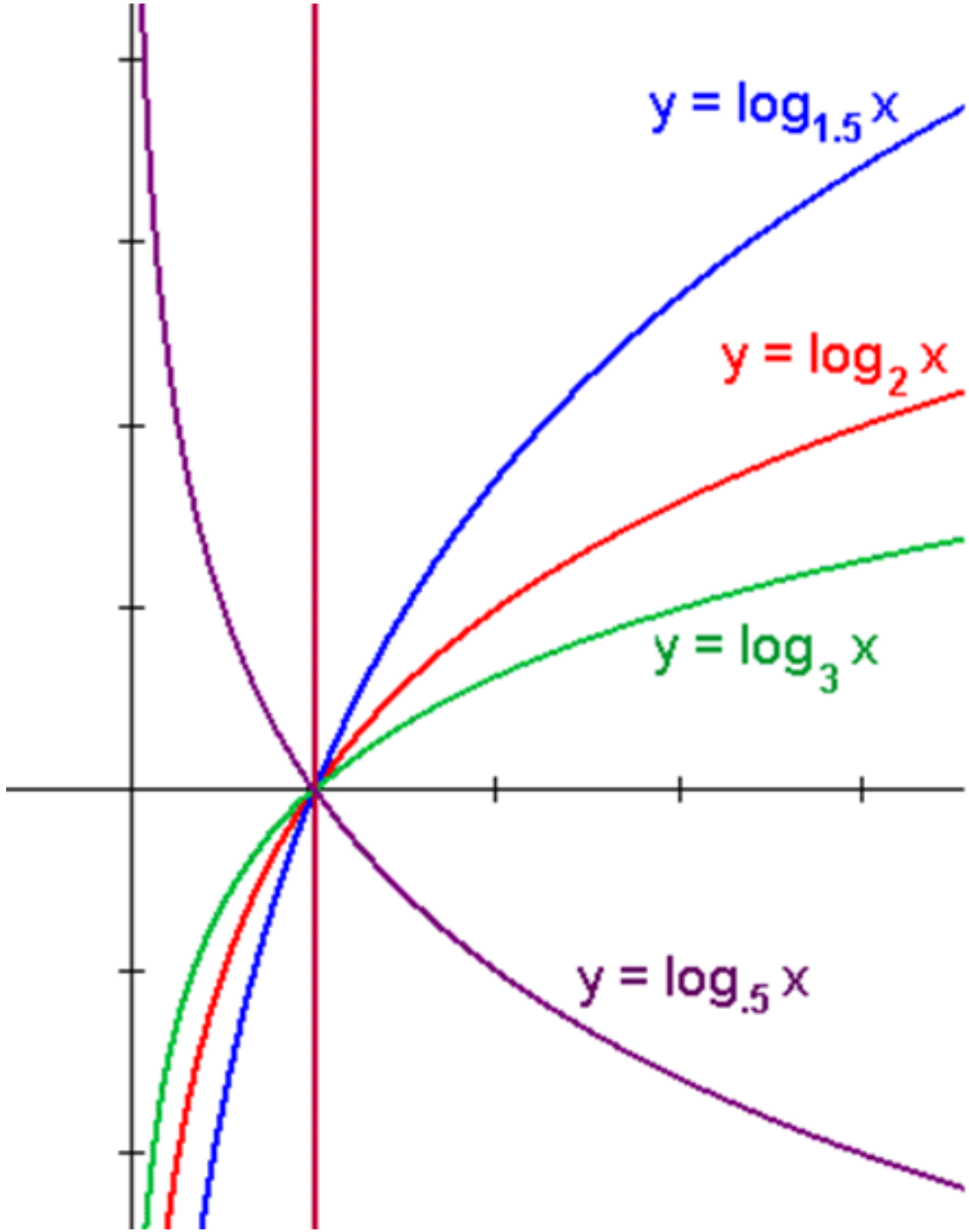
From $c^x c^y = c^{x+y}$: $\log_c(uv) = \log_c(u) + \log_c(v)$

From $(c^y)^x = c^{yx}$: $\log_c(a^x) = x \log_c(a)$

So $y = ab^x$ can be written as $\log_c(y) = \log_c(a) + x \log_c(b)$.

Thus, x and $\log_c(y)$ are linearly related.

So maybe replace ("transform") y by $\log_c(y)$.


y = log1.5 x
y = log2 x
y = log3 x
y = log.5 x

Exp and log notation

$e = 2.71828\ldots$ (more convenient logarithm base for calculus reasons)

$\exp(x) = e^x$

- [Note: $a^x = (b^{\log_b(a)})^x = b^{x \log_b(a)}$
- so switching bases is just a linear change of variable (sorta)]

$\ln = \log_e$; $\log = \log_{10}, \log_2, \ldots$

Logistic models

Several applications fit "logistic" models better than linear, exp or log:

- $y = K \frac{ea+bx}{1+ea+bx}$
- For large x, y is close to K.
  - In population models, K is "carrying capacity", i.e., max sustainable pop.
  - But y may be proportion p of pop, so $K = 1$.
- For large neg x , y is close to 0.

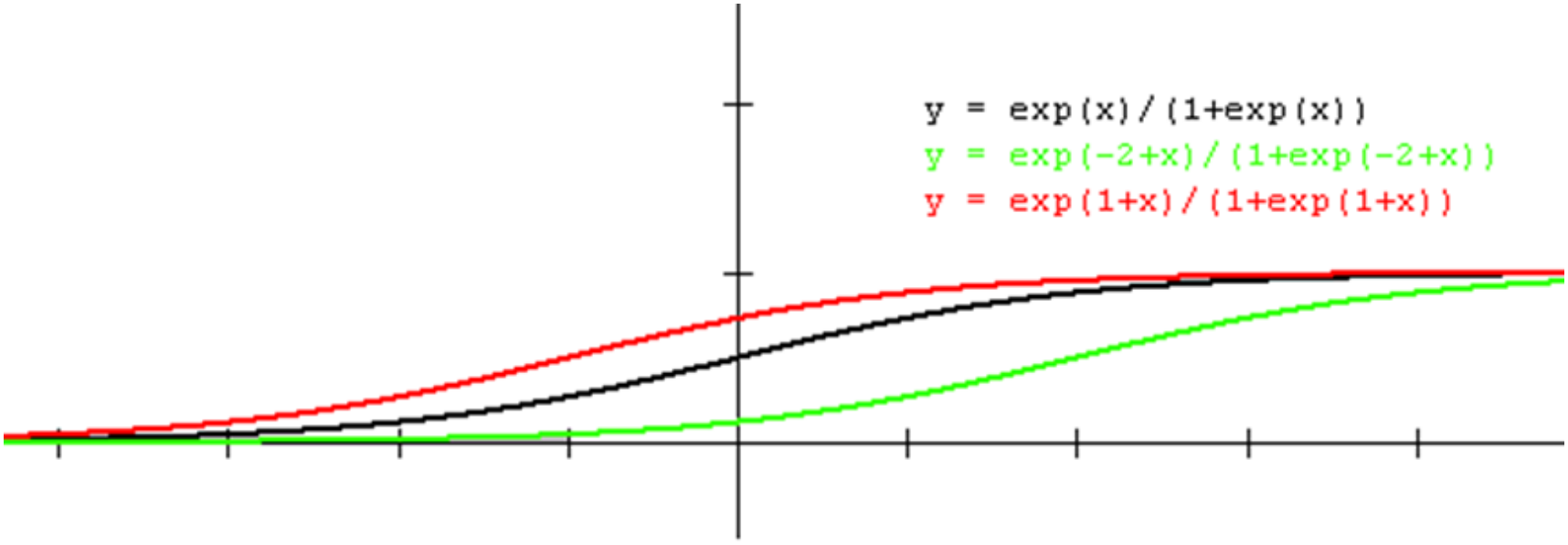
Ex: Smokers: $x = \#$ packs/day, $p = \%$ who smoke that much and have a cough

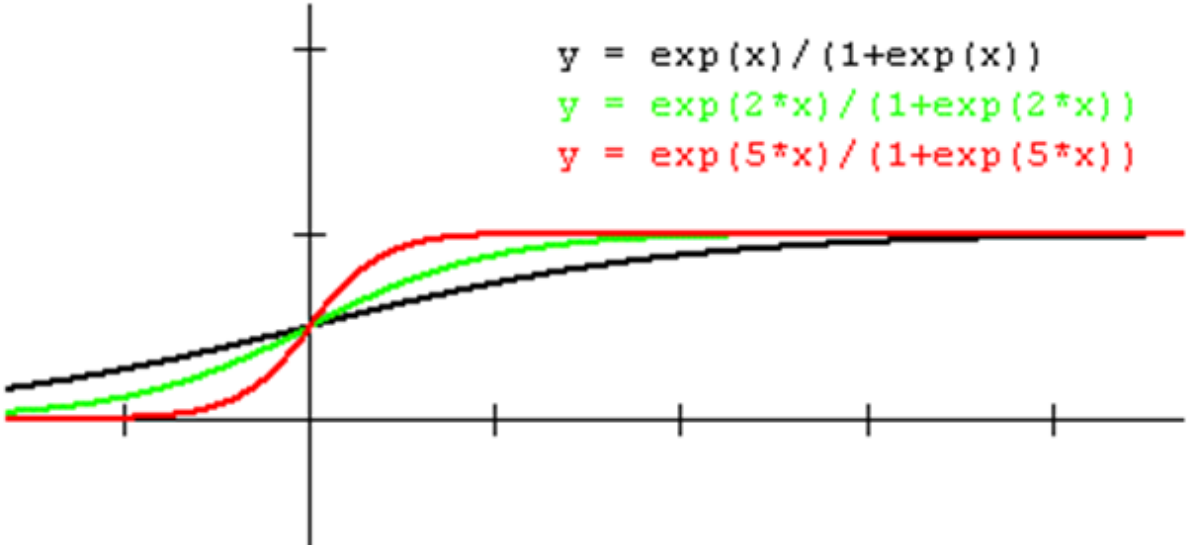
For logistic with $K = 1$,

x and $\ln(\frac{y}{1-y})$ are related linearly:

$$
\begin{aligned}
y &= \frac{e^{a+bx}}{1+e^{a+bx}} \\
y + y(e^{a+bx}) &= e^{a+bx} \\
y &= e^{a+bx} - y(e^{a+bx}) = (1-y)e^{a+bx} \\
\frac{y}{1-y} &= e^{a+bx} \\
\ln(\frac{y}{1-y}) &= a + bx
\end{aligned}
$$

so maybe "transform" y to $\ln(\frac{y}{1-y})$


y = exp(x)/(1+exp(x))
y = exp(-2+x)/(1+exp(-2+x))
y = exp(1+x)/(1+exp(1+x))


y = exp(x)/(1+exp(x))
y = exp(2*x)/(1+exp(2*x))
y = exp(5*x)/(1+exp(5*x))