Logarithmic Transformations

In the following “Regression Modeling” listing, the last two (optional) points, involving logarithmic transformations, are “the next things I’d cover if we had a bit more time.”

Regression Modeling

The list below summarizes steps which should be taken after you've preliminarily explored a regression model. The steps can be taken in any order, and can be tried repeatedly as you continue to improve your model.

1. Use sets of dummy variables to represent qualitative variables in your mode. An analysis of variance (ANOVA) will tell you if the data supports inclusion of these variables.

2. Plot the residuals against each explanatory variable. If you see a "U" (bending upwards or downwards), try adding the square of that explanatory variable to your model. Then look at "c" and "-b/(2c)" to see the nature of the nonlinearity you've captured.

3. For each explanatory variable in turn, ask yourself whether its impact on the dependent variable might vary as some other explanatory variable varies. If so, try adding the product of those two explanatory variables to your model (in order to capture a possible interaction). Interpret the regression results in terms of the "conceptual" model in which the coefficient of the first variable explicitly incorporates the second.

4. Find the sample observations with the largest positive residuals, and those with the largest (in magnitude) negative residuals. If some as-yet-not-in-your-model factor seems to differentiate the two groups, collect data on that factor and try including it as a new explanatory variable in your model.

5. Do a "model analysis," and examine any outliers that turn up. Check that the data was entered correctly. If it was, see if you can identify something "special" about the outliers (ideally, new explanatory variables which will yield a model where the observations are no longer outliers).

6. [Plot the residuals against the predicted values of the dependent variable. If the "scatter" of the residuals grows as the predicted values grow, consider using the logarithm of the dependent variable as the dependent variable in a new model.]

7. [If you suspect that the effects of the explanatory variables are "scale" effects (for example, if you think that changes in an explanatory variable are associated with percentage changes in the dependent variable, rather than additive changes), consider using the logarithms of the explanatory variables in a new model, instead of the original explanatory variables themselves.]

Here are a couple of examples which illustrate points (6) and (7).

The first is pulled from the Session-4 section of the course materials (where you can find all of the data):

I collected player performance data on National Basketball Association guards for the 1997-8 season, and matched that data to their 1998-9 salaries. From the performance data, I determined two performance indices: How many points they scored per minute of playing time, and how many other contributions to their team they made per minute (the "other" contributions give them credit for assists, free throws made, rebounds, and the like, while reducing credit for fouls committed, turnovers, and other "bad" things). I also included their age in the study, as well as whether they were primarily a "shooting" guard (sg=1) or a "point" guard (sg=0).

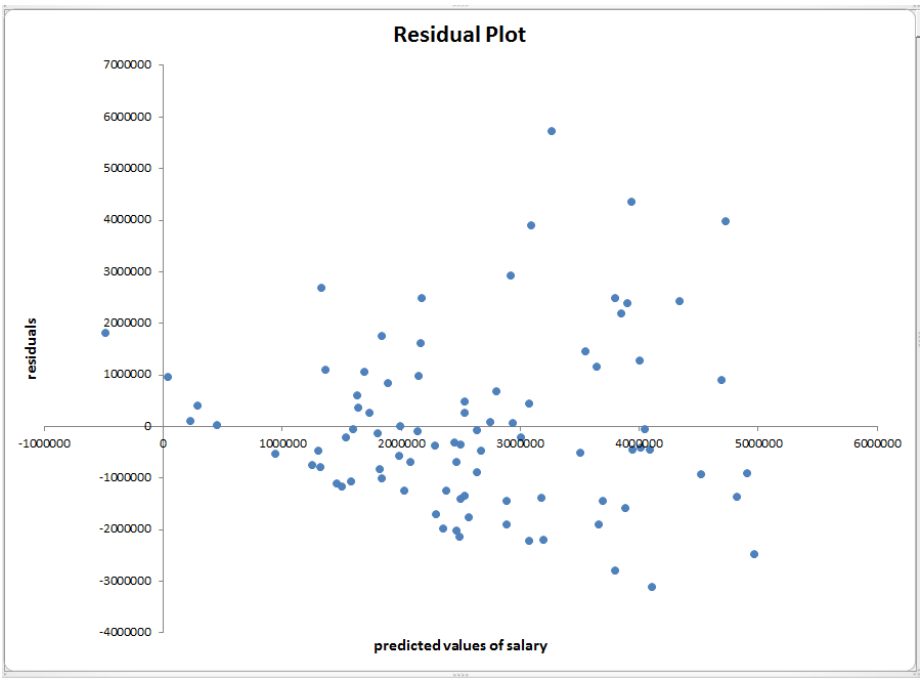
Here's the regression of salary onto the four explanatory variables:

Regression: salary

	constant	ppm	cpm	age	sg
coefficient	-4422980.4	5109892.49	10968564.6	108678.696	852094.908
std error of coef	1609690.64	1549029.42	2880311.02	53200.2007	441455.612
t-ratio	-2.7477	3.2988	3.8081	2.0428	1.9302
significance	0.7377%	0.1438%	0.0269%	4.4282%	5.7040%
beta-weight		0.3131	0.3903	0.1879	0.2017

standard error of regression	1703713.89
coefficient of determination	33.18%
adjusted coef of determination	29.92%

I then plotted the residuals against the predicted salaries:



Clearly, the errors (residuals) in my predictions grew, on average, as predicted salaries increased. This is an instance of what's known as ***heteroskedasticity*** (a fun word to pronounce, and sometimes spelled ***heteroscedasticity*** although the "k" leads the "c" in Google, 691,000 to 607,000).

Generally, heteroskedasticity refers to any situation where the residuals vary systematically with the size of the dependent variable, and a common type is when the dependent variable varies over a wide range, and there's more "room" for error for larger values of the dependent variable.

Heteroskedasticity doesn't distort coefficient estimates, but it does throw off the estimates of the standard errors of the coefficients and the standard error of the regression, as well as the standard errors of predictions.

KStat offers a test for heteroskedasticity (on the "Model Analysis" page) known as the Breusch-Pagan test. The null hypothesis is that the residuals have equal variance for all values of the dependent variable, and significance levels near 0% indicate that the data strongly contradicts that hypothesis, i.e., values of the significance level near 0% indicate strong evidence of the presence of heteroskedasticity:

Predicted values and residuals

11.8247	0.058%	**Breusch-Pagan heteroskedasticity test**
18.6203	0.009%	**Jarque-Bera non-normality test**

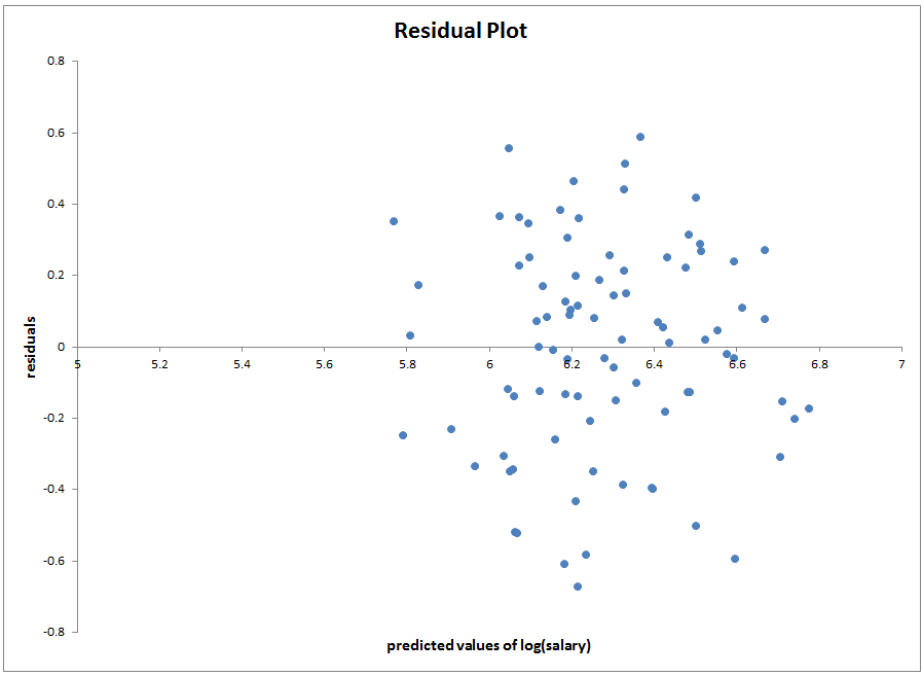
When you see the type of "fanning-outwards" residual plot we saw above, one common modeling approach is to recale the dependent variable, and a common rescaling is to use its logarithm as a new dependent variable. Regressing log(salary) onto the same explanatory variables yields:

Regression: logsal

	constant	**ppm**	**cpm**	**age**	**sg**
coefficient	4.79270892	0.77434739	2.13552137	0.02842371	0.1675415
std error of coef	0.28599623	0.27521846	0.51174932	0.00945216	0.0784341
t-ratio	16.7579	2.8136	4.1730	3.0071	2.1361
significance	0.0000%	0.6131%	0.0074%	0.3501%	3.5655%
beta-weight		0.2601	0.4166	0.2694	0.2174

standard error of regression	0.30270149
coefficient of determination	36.61%
adjusted coef of determination	33.51%

The plot of the residuals against the predicted values (of log(salary)) looks much more "in control."



And the Breusch-Pagan statistic has a significance level far above 0, indicating no remaining evidence of heteroskedasticity!

Predicted values and residuals

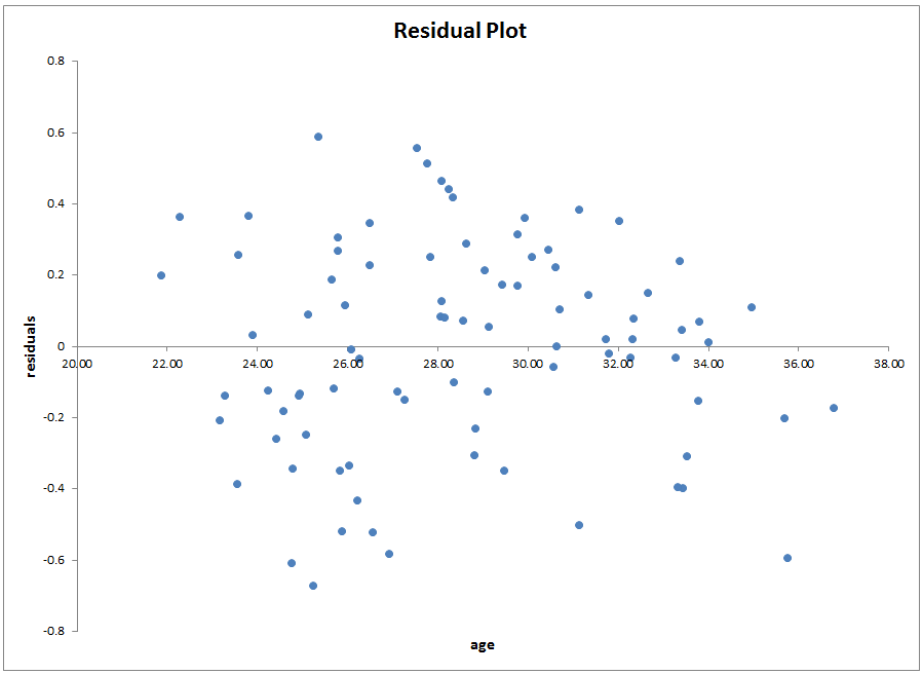
0.5347	46.465%	**Breusch-Pagan heteroskedasticity test**
2.1191	34.661%	**Jarque-Bera non-normality test**

(The Jarque-Bera test also shows no evidence that the distribution of the residuals is non-normal. Together, these tell us that our various standard errors can be properly used to determine confidence intervals for the estimated coefficients and for predictions.)

To now predict a player's salary, you'd first predict his log(salary), and then raise 10 to that power (to "unlog" the prediction). For a 95%-confidence interval for your prediction, you'd take the endpoints of the confidence interval for log(salary) and unlog them as well. This yields a somewhat-asymmetric interval around the prediction, but it's the best you can do.

Final notes on basketball salaries: The adjusted coefficient of determination indicates that we've potentially explained about 1/3 of the overall variation in salary levels using these four explanatory variables. Other variables that likely play a role in the relationship are the player's health (is he prone to injury?) and his position in a multi-year contract (was his salary determined after the previous playing season, or several years earlier?) As well, the model itself can be further improved.

For example, “age” actually enters the relationship in a non-linear fashion, as is seen in this residual plot:



If you have a few beers and then stare at the plot of the residuals against age, you’ll eventually see a downward-bending “U”.

The regression below shows strong evidence that this is a real nonlinearity. The age effect tops out at 31.6 years, and then begins to taper off. Try to come up with your own explanation!

Regression: logsal

	constant	ppm	cpm	age	age^2	sg
Coefficient	0.37206996	0.83187619	2.37428449	0.33426693	-0.0052846	0.18015783
std error of coef	2.13146893	0.2711184	0.51434661	0.14647899	0.00252592	0.07710339
t-ratio	0.1746	3.0683	4.6161	2.2820	-2.0922	2.3366
significance	86.1861%	0.2927%	0.0014%	2.5112%	3.9556%	2.1933%
beta-weight		0.2795	0.4631	3.1678	-2.9131	0.2338

standard error of regression	0.29665438
coefficient of determination	39.86%
adjusted coef of determination	36.14%

Moving onwards: If you believe that an explanatory variable has a scaling effect (instead of an additive effect) on the dependent variable, you might consider regressing its logarithm onto the dependent variable.

For example, in a study I did for McDonalds of the relationship between the approximate retail value of a prize offered in the McDonald's "Monopoly" game, and the likelihood that the prize would actually be claimed, I found that

$$\text{Redemption rate} = 0.367524 + 0.062011 \cdot \log(\text{prize ARV}) .$$

The regression yields these predictions:

36.75%	$1
42.95%	$10
49.15%	$100
55.36%	$1,000
73.96%	$1,000,000

In a direct linear model, the increment from the $1,000 case to the $1,000,000 case would have to be 100 times as large as the increment (6.2%) from the $10 case to the $100 case. This is clearly ridiculous!

A final example: If several of your explanatory variables have scaling effects (instead of additive effects) on the dependent variable, you might even consider regressing their logarithms onto the logarithm of the dependent variable.

A standard model in marketing is: $\text{Sales} = a \cdot \text{Price}^{b_1} \cdot \text{Adv}^{b_2} \cdot \text{Promo}^{b_3}$

In order to estimate the coefficients of this model, recast it as

$$\log(\text{Sales}) = \log(a) + b_1\log(\text{Price}) + b_2\log(\text{Adv}) + b_3\log(\text{Promo})$$

(Typically, you'll find that b_1 is negative, and b_2 and b_3 lie between 0 and 1.)