Applied Regression Modeling: A Business Approach
Chapter 4: Regression Model Building I
Section 4.1

by Iain Pardoe

4.1 Transformations

Transformations

- A *transformation* is a mathematical function applied to a variable in our dataset.
- Example: it is possible that there is a stronger relationship between Y and $\log_e(X)$ than between Y and X.
- To find out, fit two models:
  - Model 1 : $\mathrm{E}(Y) = b_0 + b_1 X$;
  - Model 2 : $\mathrm{E}(Y) = b_0 + b_1 \log_e(X)$.
- Which model fits better?
  - Smaller regression standard error, s;
  - Larger coefficient of determination, R^2;
  - Larger magnitude individual t-statistic.

4.1.1 Natural logarithm transformation for predictors

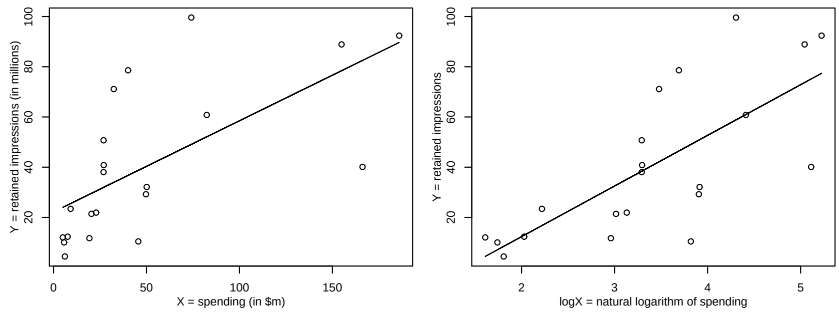
TVADS example

TV commercial data: $X =$ spending in \$m,
$Y =$ millions of retained impressions.

Firm	X (spending)	Y (retained impressions)
Miller Lite	50.1	32.1
Pepsi	74.1	99.6
Stroh's	19.3	11.7
Federal Express	22.9	21.9
Burger King	82.4	60.8
Coca-Cola	40.1	78.6
McDonald's	185.9	92.4
MCI	26.9	50.7
Diet-Cola	20.4	21.4
Ford	166.2	40.1
...	...	...
Kibbles 'n Bits	6.1	4.4

Scatterplots for models 1 and 2

Model 1 on the left: $\mathrm{E}(Y) = b_0 + b_1 X$.
Model 2 on the right: $\mathrm{E}(Y) = b_0 + b_1 \log_e(X)$.



Plots include fitted regression (least squares) lines.
Which model fits the data better?

Regression results for models 1 and 2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.651 [a]	0.424	0.394	23.5015

[a] Predictors: (Intercept), X.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	22.163	7.089	3.126	0.006
	X	0.363	0.097	3.739	0.001

[a] Response variable: Y.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.730 [a]	0.532	0.508	21.1757

[a] Predictors: (Intercept), logX.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
2	(Intercept)	−28.050	15.441	−1.817	0.085
	logX	20.180	4.339	4.650	0.000

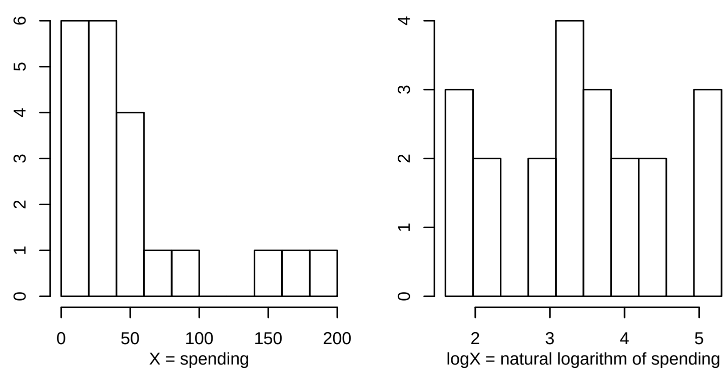
[a] Response variable: Y.

Interpretation

- Model 2 with $\log_e(X)$ more usefully describes relationship between TV commercial success and cost than model 1 with untransformed X:
  - smaller s, larger R^2, larger magnitude individual t-statistic.
- However, model interpretation less straightforward.
- For example, as X increases from 10 to 20, expect Y to increase from $-28.1 + 20.2\log_e(10) = 18.4$ to $-28.1 + 20.2\log_e(20) = 32.4$.
- But as X increases from 150 to 160, expect Y to increase from $-28.1 + 20.2\log_e(150) = 73.1$ to $-28.1 + 20.2\log_e(160) = 74.4$.
- Predictor effect plot (Section 5.4) displays this graphically.

Why it works

- Multiple linear regression models are often more effective when predictors have reasonably symmetric distributions and are not too highly skewed.
- Natural log. transformation works well for positively skewed variables with a few values much higher than the majority, since it tends to spread out lower values and pull in the higher values (i.e. makes distribution more symmetric).



Selecting transformations

- Transformations often suggested by theories about economics, consumer psychology, worker behavior, business decision-making, and so on.
- At other times, we might observe particular empirical relationships in sample datasets, and we can try out various variable transformations to see how to best model the data.
- Common transformations in business:
  - natural logarithm, e.g., $\log_e(X)$;
  - polynomial, e.g., X, X^2, X^3, $\ldots$;
  - reciprocal, e.g., $1/X$;
  - square root, e.g., $\sqrt{X}$.

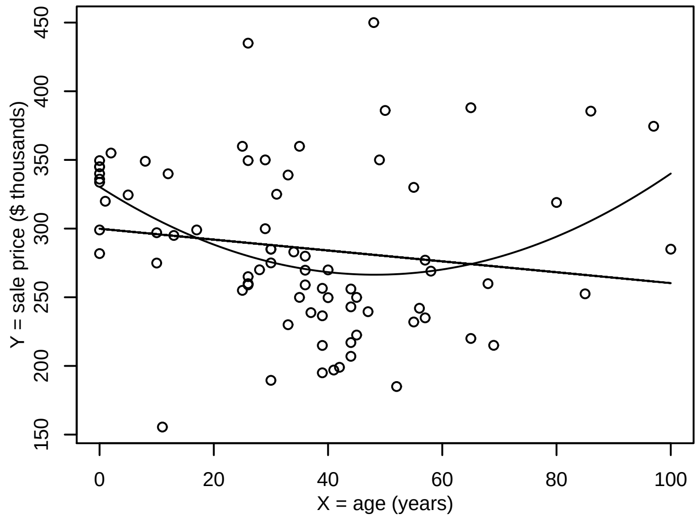
4.1.2 Polynomial transformation for predictors

HOMES4 example

- Investigate whether the age of a home factors into its sale price in a particular housing market.
- For the sake of illustration, ignore other predictors (floor size, lot size, etc.) and focus solely on $X =$ age, defined as 2005 minus year built.
- Realtor experience suggests both older and newer homes command a price premium relative to "middle-aged" homes in this market.
- Which of these models can capture such a trend?
  - Model 1 : $\mathrm{E}(Y) = b_0 + b_1 X$;
  - Model 2 : $\mathrm{E}(Y) = b_0 + b_1 X + b_2 X^2$.

Scatterplot with model 1 and 2 fitted lines

Model 1: $\mathrm{E}(Y) = b_0 + b_1 X$ (straight)
Model 2: $\mathrm{E}(Y) = b_0 + b_1 X + b_2 X^2$ (curved).



Which model fits the data better?

Regression results for models 1 and 2

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	299.883	12.554	23.887	0.000
	X	−0.396	0.295	−1.342	0.184

[a] Response variable: Y.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.383 [a]	0.147	0.123	56.486

[a] Predictors: (Intercept), X, Xsq.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
2	(Intercept)	330.407	15.103	21.876	0.000
	X	−2.652	0.749	−3.542	0.001
	Xsq	0.027	0.008	3.245	0.002

[a] Response variable: Y.

Interpretation

- Model 2 with X and X^2 more usefully describes relationship between sale price and age than model 1 with untransformed X:
  - two tail p-value for testing b_2 in model 2 is 0.002.
- However, model interpretation less straightforward.
- For example, Y decreases quite steeply as X increases between 0 and 20, levels off for X between 20 and 70, and then increases more steeply as X increases between 70 and 100.
- Quick calculations can quantify these changes more precisely.
- Predictor effect plot (Section 5.4) displays this graphically.

Polynomial transformations in practice

- General polynomial model: $\mathrm{E}(Y) = b_0 + b_1X + b_2X^2 + b_3X^3 + \cdots$.
- Rare to see powers higher than two (quadratic) or three (cubic) in linear regression models unless theoretical reasons for including higher powers.
- When using X^2, X^3, etc., in models, lower powers are often included, *regardless* of their significance—this is called preserving *hierarchy*.
  - e.g., keep X in the model if X^2 is significant (has a small p-value);
  - e.g., keep X and X^2 in the model if X^3 is significant (has a small p-value).
- When using X^2, X^3, etc., in models, common to first *rescale* values of X to have mean ≈ 0 and std. dev. ≈ 1 (example in Section 6.1).

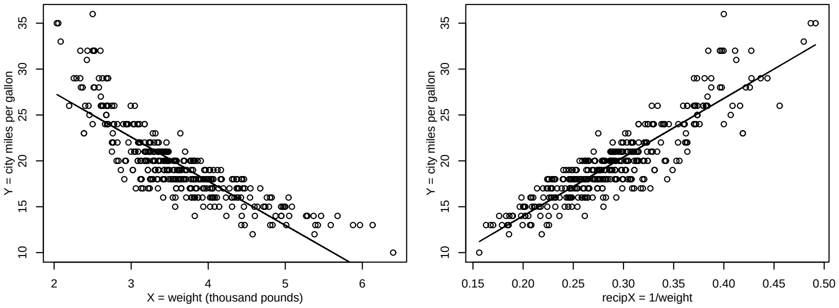
4.1.3 Reciprocal transformation for predictors

CARS3 example

- Investigate any relationship between a car's city miles per gallon (Y) and its weight (X).
- For the sake of illustration, ignore other predictors (horsepower, engine size, etc.).
- Engineering principals suggest that there is an "inverse" relationship between weight and fuel efficiency.
- Which of these models can capture such a trend?
  - Model 1 : $\mathrm{E}(Y) = b_0 + b_1X$;
  - Model 2 : $\mathrm{E}(Y) = b_0 + b_1(1/X)$.

Scatterplots for models 1 and 2

Model 1 on the left: $\mathrm{E}(Y) = b_0 + b_1 X$.
Model 2 on the right: $\mathrm{E}(Y) = b_0 + b_1(1/X)$.



Plots include fitted regression (least squares) lines.
Which model fits the data better?

Regression results for models 1 and 2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.837^a	0.700	0.700	2.291

[a] Predictors: (Intercept), X.

Parameters[a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	37.020	0.572	64.760	0.000
	X	−4.809	0.157	−30.723	0.000

[a] Response variable: Y.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.895^a	0.800	0.800	1.869

[a] Predictors: (Intercept), recipX.

Parameters[a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
2	(Intercept)	1.195	0.472	2.534	0.012
	recipX	64.019	1.590	40.251	0.000

[a] Response variable: Y.

Interpretation

- Model 2 with $1/X$ more usefully describes relationship between city miles per gallon and weight than model 1 with untransformed X:
  - smaller s, larger R^2, larger magnitude individual t-statistic.
- However, model interpretation less straightforward.
- For example, as X increases from 2 to 3, expect Y to decrease from $1.2+64.0(1/2)=33.2$ to $1.2+64.0(1/3)=22.5$.
- But as X increases from 5 to 6, expect Y to decrease from $1.2+64.0(1/5)=14.0$ to $1.2+64.0(1/6)=11.9$.
- Predictor effect plot (Section 5.4) displays this graphically.

4.1.4 Natural logarithm transformation for the response 18 / 27

WORKEXP example

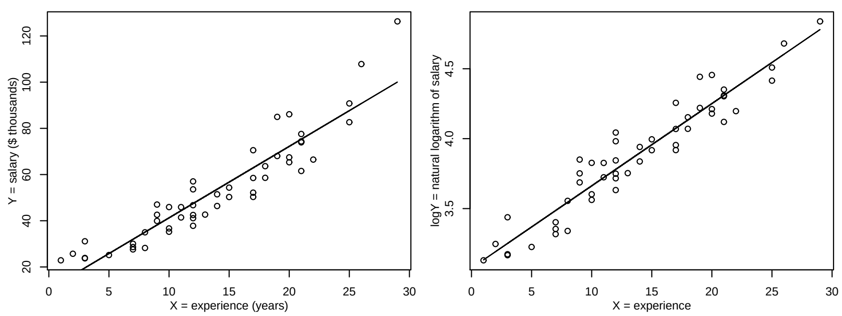
- Investigate any relationship between a worker's salary (Y) and experience (X).
- For the sake of illustration, ignore other predictors (job title, education, etc.).
- Annual salary increases are usually proportional (% increases) rather than additive ($ increases). This suggests that there should be some kind of "multiplicative" relationship between additional years of experience and salary differences.
- Which of these models can capture such a trend?
  - Model 1 : $\mathrm{E}(Y)=b_0+b_1X$;
  - Model 2 : $\mathrm{E}(\log_e(Y))=b_0+b_1X$.

Multiplicative models

- Consider Model $1 : \mathrm{E}(Y) = b_0 + b_1 X$.
- What happens to Y when X increases one unit?
  - $Y_{\text{after}} - Y_{\text{before}} = [b_0 + b_1(X+1)] - [b_0 + b_1 X] = b_1$.
- Consider Model $2 : \mathrm{E}(\log_e(Y)) = b_0 + b_1 X$.
- What happens to Y when X increases one unit?
  - $Y_{\text{after}} - Y_{\text{before}}$
    $= \exp(b_0 + b_1(X+1)) - \exp(b_0 + b_1 X)$
    $= [\exp(b_1) - 1][\exp(b_0 + b_1 X)]$.
  - $(Y_{\text{after}} - Y_{\text{before}})/Y_{\text{before}} = \exp(b_1) - 1$.
- In other words, in model 2, $\exp(b_1) - 1$ represents the proportional increase in salary (Y) when experience (X) increases by one year.
  - Example: if $\exp(b_1) - 1 = 0.06 = 6\%$, then we expect annual salary increases to be 6%, from $\$50{,}000$ to $\$53{,}000$, say $((53 - 50)/50 = 0.06)$.

Scatterplots for models 1 and 2

Model 1 on the left: $\mathrm{E}(Y) = b_0 + b_1 X$.
Model 2 on the right: $\mathrm{E}(\log_e(Y)) = b_0 + b_1 X$.



Plots include fitted regression (least squares) lines.
Which model fits the data better?

Regression results for models 1 and 2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.930 [a]	0.865	0.862	8.357

[a] Predictors: (Intercept), X.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	10.323	2.706	3.815	0.000
	X	3.094	0.177	17.515	0.000

[a] Response variable: Y.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.955 [a]	0.912	0.910	0.125

[a] Predictors: (Intercept), X.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
2	(Intercept)	3.074	0.040	76.089	0.000
	X	0.059	0.003	22.302	0.000

[a] Response variable: logY.

Interpretation

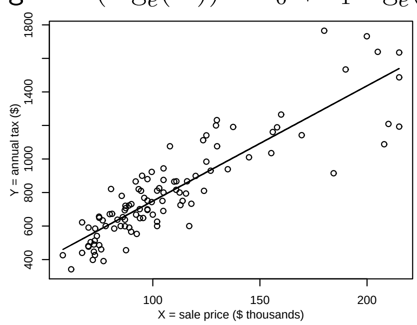
- Model 2 with $\log_e(Y)$ more usefully describes relationship between salary and experience than model 1 with untransformed Y:
  - larger magnitude individual t-statistic.
- Note that s and R^2 cannot be compared since the response variable is $ thousands for model 1 but $\log_e$($ thousands) for model 2.
- Model interpretation uses "multiplicative" idea.
- For example, since $\exp(\hat{b}_1) - 1 = \exp(0.059) - 1 = 0.0608$, we expect salary ($Y$) to increase by a multiplicative factor of 0.0608 (or 6.08%) for each additional year of experience (X).

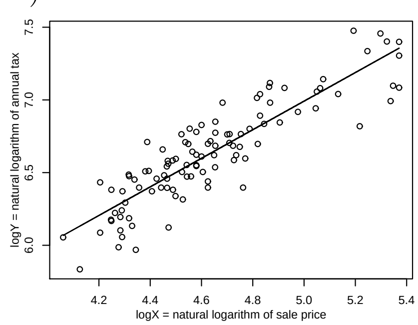
HOMETAX example

- Investigate any relationship between annual taxes (Y) and sale price (X) for homes sold in Albuquerque, New Mexico in 1993.
- These data are typical of much business and economic data in that both variables are quite skewed (a few values much higher than the majority).
- What kind of transformation works well for (positively) skewed data?
- Which of these models might work better for this dataset?
  - Model 1 : $\mathrm{E}(Y) = b_0 + b_1 X$;
  - Model 2 : $\mathrm{E}(\log_e(Y)) = b_0 + b_1 \log_e(X)$.

Scatterplots for models 1 and 2

Model 1 on the left: $\mathrm{E}(Y) = b_0 + b_1 X$.
Model 2 on the right: $\mathrm{E}(\log_e(Y)) = b_0 + b_1 \log_e(X)$.





Plots include fitted regression (least squares) lines.
Do both models satisfy the constant variance assumption?

Regression results for model 2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
2	0.886 [a]	0.785	0.782	0.162

[a] Predictors: (Intercept), logX.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(> |t|)
2	(Intercept)	2.076	0.237	8.762	0.000
	logX	0.983	0.051	19.276	0.000

[a] Response variable: logY.

So, $\widehat{\log_e(Y)} = 2.076 + 0.983 \times \log_e(X)$
and $\hat{Y} = \exp(2.076 + 0.983 \times \log_e(X))$.

Interpretation

- Model 2 with $\log_e(Y)$ and $\log_e(X)$ more usefully describes relationship between annual taxes and sale prices than model 1 with untransformed X and Y:
  - model 2 satisfies regression assumptions, but model 1 fails constant variance assumption.
- Model 2 is just as easy to use as model 1 for estimating or predicting annual taxes from home sale prices. For example, a home that sold for $\$100{,}000$ would be expected to have annual taxes of approx. $\exp(2.076 + 0.983 \times \log_e(100)) = \$737$.
- Prediction intervals should be more accurate for model 2 than for model 1 (which will tend to be too wide for low sale prices and too narrow for high sale prices)—see Problem 4.2.

Transformations in practice: example

- $\mathrm{E}(\log_e(Y)) = b_0 + b_1X_1 + b_2X_2 + b_3X_2^2 + b_4\log_e(X_3) + b_5(1/X_4)$.
  - Y: natural logarithm transformation;
  - X_1: untransformed;
  - X_2: quadratic transformation (also included X_2 to retain hierarchy);
  - X_3: natural logarithm transformation;
  - X_4: reciprocal transformation.
- Best if transformations in a model are suggested *before* looking at the data from background knowledge about the situation or from theoretical arguments about why transformations make sense.
- Used judiciously, variable transformations provide a useful tool for improving regression models.
- Dangers: overcomplicating things unnecessarily, overfitting sample data (poor generalizability).

