Class 5 Review Notes

AI & Machine Learning

Fall 2023 - Laurie Ye

5 Variable Transformations In Regression

5.1 Data Transformations

- Data transformation are functions applied to the variables in the data to make its analysis more valid and/or meaningful.
- Transformations can be applied to either the Y in a regression, one or more X's or to both the Y and the X's.
- The kinds of variables we transform are continuous variables and counts. We generally do not transform categorical variables (factors).

5.2 Purpose of Transformation (3 Primary Purposes)

1. Creating linearity
2. Creating homoscedasticity (constant variance).
3. Creating independence.

5.3 Reminder

The four assumptions of linear, least-squares regression are:

1. Linearity (i.e., the conditional expectation is linear).
2. Constant variance or the errors (homoscedasticity).
3. Independence (of the errors).
4. Normality (the errors follow a normal distribution).

In notation, the linear least-squares regression model is:

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where

$$\varepsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$$

iid mean "independent and identically distributed."

5.4 Creating Independence

- In regression, in order for the model to be valid, we need independent errors.
- Percentage changes (over time) are essentially independent whereas the original data are not. (ex. stock price)
- Since

  $$\log(x) \approx x - 1 \text{ for } x \text{ close to } 1,$$

  we have

  $$\log\left(\frac{P_{t+1}}{P_t}\right) \approx \frac{P_{t+1}}{P_t} - 1 = \frac{P_{t+1} - P_t}{P_t}.$$

  Thus, percentage changes are closely related to logs.

  Note: log() means logarithm to the base e. Basic textbooks sometimes use ln().

5.5 Primary Transformations

- We will consider four transformations: 1. Logs 2. Square Roots 3. Cube Roots 4. Logistic transformation
- If the data are multivariate normal, then the true regression of Y on X (that is, $E(Y|X)$) is linear. - If our data fits into a pattern where all variables together (multivariate) look like a classic bell curve (normal distribution), then we can draw a straight line (linear regression).
- Furthermore, the error around the true regression line will be homoscedastic. -the errors or differences between our predictions and the actual values don't spread out wider as we move along the X values.
- If we can use transformations to make the marginal distributions of the Y and X's look univariate normal, then our data will quite likely be approximately multivariate normal. - If our variables don't naturally fit a bell curve, we might be able to tweak them (transform them) so they do. Once they look more like a bell curve individually (univariate normal), it's likely that together they'll fit into a multivariate normal pattern.

Note 1. Using log, square root, or cube root, the goal is to make the variable look as normally distributed as possible or, at least, as symmetric as possible.

5.6 Rule of Thumb

- If your data are positive continuous numbers (especially if the largest is an order or magnitude or more larger than the smallest), consider logs.
- If your data are counts, consider the square root.
- If logs or square roots do not work, try cube roots.

Strength of Transforms

X (untransformed)	weakest
Square Root: $\sqrt{X}$	stronger
Cube Root: $X^{\frac{1}{3}}$	stronger still
Log: $\ln(X)$	strongest

Roughly speaking, strength has to do with how much right skewness in the data the transformation will correct.

5.7 Power Transformations and Box-Cox Transforms

- X, $\sqrt{X}$, $X^{(1/3)}$, and $\log(X)$ are all examples of the family of power transformations (also called Box-Cox transformations):
$$\frac{X^\lambda - 1}{\lambda}$$
- X corresponds to $\lambda = 1$
- $\sqrt{X}$ corresponds to $\lambda = \frac{1}{2}$
- $X^{(1/3)}$ corresponds to $\lambda = \frac{1}{3}$
- $\log(X)$ corresponds to $\lambda = 0$

Note:
$$\lim_{\lambda \to 0} \frac{X^\lambda - 1}{\lambda} = \log(X)$$

”Undoing” Transformations

- You need to know how to ”undo” the transformations for the purposes of prediction.
- To ”undo”:
$$\text{If } x_{\text{trans}} = \sqrt{x} \text{ use } x = (x_{\text{trans}})^2$$
$$\text{If } x_{\text{trans}} = (x)^{(1/3)} \text{ use } x = (x_{\text{trans}})^3$$
- In Excel:
$$\text{If } x_{\text{trans}} = \ln(x) \text{ use } x = \exp(x_{\text{trans}})$$
$$\text{If } x_{\text{trans}} = \log(x) \text{ use } x = 10^{(x_{\text{trans}})}$$
- Elsewhere:
$$\text{If } x_{\text{trans}} = \log(x) \text{ use } x = \exp(x_{\text{trans}})$$

“Undoing” the Logistic Transformation

To revert a logistic transformation, proceed as follows:

If

$$x_{\text{trans}} = \log\left(\frac{p}{1-p}\right)$$

then

$$p = \frac{e^{x_{\text{trans}}}}{1+e^{x_{\text{trans}}}}$$

5.8 Handling Zeros in Logarithmic and Logistic Transformations

- You can’t take a log of 0. If your data has a small percentage of 0’s, you can try to fix this problem by adding a small number (say c) to all of the X’s and then take $\log(X+c)$. A reasonable value for c might be $\frac{1}{3}$ or something close to $\frac{1}{3}$ of the smallest non-zero value.
- If you have fractions that are 0’s or 1’s, the logistic transformation will not work. If you have a small percentage of 0’s and 1’s in your data, you might try to ”lengthen” the interval a bit. Since p is in $[0,1]$, you might try the interval $[-c, 1+c]$.
- The logistic becomes: $\log\left(\frac{p+c}{1-p+c}\right)$. Here, c is small (like 0.01 or close to $\frac{1}{3}$ of the distance between the nearest non-zero or non-one value to 0, or 1, respectively).

Note 2. Even when you have a small percentage of 0’s (and 1’s in the logistic case) and need to use logs, the objective is to end up with marginal distributions that are approximately normal. This is how you judge your success at transformation.

5.9 Sample Size Matters

- Least squares regression experiences a version of the Central Limit Theorem.
- In large sample sizes, you can tolerate data that is less normal than in small sample sizes.
- The X variables must, in a sense, remain bounded as the sample size increases in order for the Central Limit Theorem to work.
- How big a sample is big enough is even more ambiguous than in the univariate statistics case.

5.10 Large Fractions of 0’s and 1’s

- If you have a large fraction of 0’s (and 1’s in the case of fractions), you should not use least-squares regression.
- Do not use regular regression for binary dependent variables Y.

Note 3. For binary dependent variable Y, you might consider logistic regression. For counts, you might consider Poisson regression.

5.11 Reference

Goizueta Business School-Emory University: Professor George S. Easton