Transformations

* In many cases we do not analyze the original data.

* If Y and X are the observed variables we may fit the model

$$g_y(Y) \;=\; \beta_0 + \beta_1 g_x(X) + \varepsilon$$

* The functions g_y and g_x are transformations applied to the original data.

* We will look at what transformations are useful and why we may want to make transformations.

* For simplicity we will mainly focus on the simple linear regression case but everything also applies to multiple regression.

* Also note that transformations make no sense for categorical variables!

Reasons for Transformations

* We will usually consider transformations when there is a problem with one or more of the assumptions underlying the linear model when applied to the original data.

* The main reasons for applying transformations are to correct:
  1. Non-linearity in the original model.
  2. Dependence of the error variance on the covariate (Heteroscedasticity).

* Transformations can also improve the (approximate) normality of the error terms.

Non-linearity

* In simple linear regression we will usually plot Y against X to see if a linear model is reasonable.

* If it is not then we should not use the methods that we have considered so far.

* Many relationships, however, are linearizable through transformations.

Linearizable Functions

* Suppose, for example, that we expect an exponential growth model

$$\mathsf{E}\Big(Y \mid X\Big) = \alpha\beta^X$$

We can linearize this by taking logs of each side so we get

$$\log \mathsf{E}\Big(Y \mid X\Big) \;=\; \log(\alpha) + \log(\beta)X$$

* For a first order approximation we can then say that

$$\mathsf{E}\Big(\log(Y) \mid X\Big) \;\approx\; \log \mathsf{E}\Big(Y \mid X\Big) \;=\; \log(\alpha) + \log(\beta)X$$

* Thus we could fit a linear model of the form

$$\log(Y) \;=\; \beta_0 + \beta_1 X + \varepsilon$$

in which we have $\beta_0 = \log(\alpha)$ and $\beta_1 = \log(\beta)$.

Linearizable Functions

$\mathrm{E}\left(Y \mid X\right)$	Transformation	Linear Model
$\alpha\beta^X$	$Y' = \log(Y)$	$Y' = \log(\alpha) + \log(\beta)X + \varepsilon$
αX^β	$Y' = \log(Y)$ $X' = \log(X)$	$Y' = \log(\alpha) + \beta X' + \varepsilon$
$\alpha + \beta\log(X)$	$X' = \log(X)$	$Y = \alpha + \beta X' + \varepsilon$
$\dfrac{X}{\alpha X + \beta}$	$Y' = \dfrac{1}{Y}$ $X' = \dfrac{1}{X}$	$Y' = \alpha + \beta X' + \varepsilon$
$\dfrac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$	$Y' = \log\left(\dfrac{Y}{1-Y}\right)$	$Y' = \alpha + \beta X + \varepsilon$

Non-linearizable Functions

* It is important to note that not all functions as linearizable.
* For example the function

$$\mathsf{E}\Big(Y \mid X\Big) \;=\; \alpha + \beta\delta^{X}$$

  cannot be linearized.
* Methods for such functions are beyond the scope of this course.

Heteroscedastic Errors

* A major assumption of the linear model is that the variance of the errors does not depend on the value of the covariate.

* This can often be violated, particularly for count data.

* Count data can often be modeled using a Poisson distribution for which

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$$

* Hence if the mean varies with a covariate X so will the variance.

* Transformations can also help with this problem.

Delta Method

Theorem 7

Let Y be a random variable with finite mean and variance and let g be a function. If we define the random variable $Y' = g(Y)$ then

$$\begin{aligned} \mathsf{E}\left(Y'\right) &\approx g\Big(\mathsf{E}(Y)\Big) \\ \mathsf{Var}(Y') &\approx \Big(g'\Big(\mathsf{E}(Y)\Big)\Big)^2 \mathsf{Var}(Y) \end{aligned}$$

where g' is the derivative of g.

Heteroscedastic Errors

- Returning to the Poisson example with $\mathsf{E}(Y) = \mathsf{Var}(Y) = \mu$.

- If we take $Y' = g(Y) = \sqrt{Y}$ then we have

$$\mathsf{E}\left(Y'\right) \approx \sqrt{\mu}$$
$$\mathsf{Var}(Y') \approx \left(\frac{1}{2\sqrt{\mu}}\right)^2 \mu = \frac{1}{4}$$

- Hence the variance of Y' is now (approximately) independent of its mean and so independent of any covariates on which the mean depends.

- As an added bonus, the distribution of Y' is often also closer to normal.

Detecting Heteroscedasticity

* Usually we do not have a known probabilistic model so we rely on our diagnostic plots of the studentized residuals.

* In simple regression, the best plot is the plot of the studentized residuals against the covariate.

* When the spread of the studentized residuals clearly increases (or decreases) as the covariate increases or decreases, we need to take some remedial action.

Removing Heteroscedasticity

* In many cases we have that the standard deviation seems to be be proportional to the covariate. That is Var$(\varepsilon_i) \approx x_i^2 \sigma^2$.

* Hence we have that Var$(\varepsilon_i / x_i) \approx \sigma^2$.

* This suggests the linear model

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i}$$

* We fit this by considering the transformed variables

$$Y' = \frac{Y}{X} \qquad X' = \frac{1}{X}$$

Removing Heteroscedasticity

* Note that the intercept and slope of this fitted model correspond to the slope and intercept of the original model.

* Generally we will need to express our results in terms of the original model so care needs to be taken with the interpretation.

* This example is a special case of a method known as Weighted Least Squares which we shall examine in the next chapter.

Power Transformations

* Transformations are most useful with strictly positive quantities.

* Since most variables are measurements, this is generally not a problem in practice.

* We have seen that the logarithm and square root are useful transformations.

* Other useful transformations are other powers of the variables.

Power Transformations

* When the power that should be used is not known from theoretical considerations, we need to use the data to find an appropriate transformation.

* Define the (scaled) power family of transformations to be

$$\psi_\lambda(X) \;=\; \frac{X^\lambda - 1}{\lambda}$$

* When $\lambda = 0$ we can define

$$\psi_0(X) \;=\; \lim_{\lambda \to 0} \psi_\lambda(X) \;=\; \log(X).$$

* We then need to find λ_y and λ_x such that the regression of $Y' = \psi_{\lambda_y}(Y)$ on $X' = \psi_{\lambda_x}(X)$ is approximately linear with normal homoscedastic errors.

Selecting a Power Transformation

* Usually we do not need to find λ very accurately and so we will generally only consider the set of transformations

$$\frac{1}{X^2},\ \frac{1}{X},\ \frac{1}{\sqrt{X}},\ \log(X),\ \sqrt{X},\ X,\ ,X^{1.5},\ X^2$$

* We will usually either transform both the response and co-variate using the same transformation or we just transform one of them.

* Transformations of the response are more likely to help with normality concerns.

* Transformations of the covariate or both will often help with linearity or heteroscedasticity.

Selecting a Power Transformation

* The choice of power is usually done by examining graphical diagnostics.

* Comparing scatterplots of Y' against X', the residuals against the fitted values for the model and the normal qq plots of the residuals are helpful.

* The transformation which makes all of these most acceptable is usually chosen.

* When transforming only the covariate, some books suggest using the transformation which minimizes the error sum of squares.

Other Comments about Transformations

* Similar techniques can be used in multiple regression but the presence of many covariates complicates the search.

* Power transformations are not suitable for variables that can take on non-positive values.

* Some researchers have examined this problem but there has not been any very suitable solution proposed to date.