Project Report

This project aims to analyze the efficiency of Inverse Normal Transformation and Yeo-Johnson Power Transformation in transforming extreme data to normal distribution.

Least Squares Estimation

Construct a matrix $A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 1 \end{bmatrix}_{n\times 2}$. We have that $A^T A = \begin{bmatrix} \frac{n}{2} & 0 \\ 0 & \frac{n}{2} \end{bmatrix}$ and$(A^T A)^{-1} = \begin{bmatrix} \frac{2}{n} & 0 \\ 0 & \frac{2}{n} \end{bmatrix}$.

To perform a least squares estimation, we use the simple linear regression model $Y = \beta X + \varepsilon$. Let $X = A$ and Y be our simulated data, we can fit the matrix A to our simulated data Y and calculate the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ as:

$$\hat{\beta} = \begin{bmatrix} \frac{2}{n} & \frac{2}{n} & \dots & 0 & 0 \\ 0 & 0 & \dots & \frac{2}{n} & \frac{2}{n} \end{bmatrix} Y = \begin{bmatrix} \frac{2}{n}y_1 + \frac{2}{n}y_2 + \dots + \frac{2}{n}y_{\frac{n}{2}} \\ \frac{2}{n}y_{\frac{n}{2}+1} + \dots + \frac{2}{n}y_{n-1} + \frac{2}{n}y_n \end{bmatrix} = \begin{bmatrix} \frac{2}{n}(y_1 + y_2 + \dots + y_{\frac{n}{2}}) \\ \frac{2}{n}(y_{\frac{n}{2}+1} + \dots + y_{n-1} + y_n) \end{bmatrix}$$

Now, we can get our test statistic $t = \frac{\sqrt{n}\hat{\beta}}{\hat{\sigma}}$. In this case, we are testing whether there is a difference in the mean of the first half elements compare to the mean of the second half elements of our data. We can use the t-distribution to find the associated p-value. We can then observe the distribution of the p-values, expected to be uniform if the original data is distributed in normal shape, and calculate the type I error rate.

Inverse Normal Transformation

In order to perform the Inverse Normal Transformation, we first replace the data values by their fractional ranks. We then use the probit function Φ^{-1} to map these probabilities to Z-scores. If W is any continuous random variable with CDF F_W, then the transformed random variable $U = F_W(W)$ is uniformly distributed in large samples. Consequently, we know that $INT(W) = \Phi^{-1}\{F_n(W)\} \sim N(0,1)$, regardless of the initial distribution F_W. Therefore, for an observed W_i for each of n independent subjects, the formula for performing the Inverse Normal Transformation is $INT(W_i) = \Phi^{-1}\{\frac{rank(W_i) - c}{n + 1 - 2c}\}, c \in [0,1/2]$. Based on the paper, we choose $c = \frac{3}{8}$ in our transformations. (McCaw *et al.*, 2019).

For our simulated data Y, we perform the Inverse Normal Transformation to it so that $\tilde{Y} = INT(Y)$. Use this $\tilde{Y}$ to replace the original Y in the least squares estimation model above and let $X = A$ remain unchanged. We can get our new $\hat{\beta}$ and use it to obtain the p-values and the type I error rate for the transformed data and compare the results to those of the original data to examine how the transformation performs.

Yeo-Johnson Power Transformation

We use the existed $yeojohnson()$ function in R to perform the Yeo-Johnson power transformation. The code behind this function can be found at github.com/petersonR/bestNormalize/blob/master/R/yeojohnson.R.

We then use the same method above to fit $X = A$ to the data transformed by the Yeo-Johnson Power Transformation $\tilde{Y} = yeojohnson(Y)$. We then get another $\hat{\beta}$ and use it to obtain the p-values and the type I error rate for the power transformed data and compare the results to those of the original data to examine how the transformation performs.

The type I error rates of the original data, the INT transformed data, and the power transformed data are listed in Table 1.

Two Ways of Calculating P-Values

In simple regressions like what we performed above, we obtain the p-values by the inverse CDF of the t distribution $p = \Phi^{-1}(T)$ where $\Phi(t) = \mathbb{P}(T \leq t)$. Now, we want to calculate the p-values by just using a Gaussian approximation $\Phi(t) = \mathbb{P}(N(0,1) \leq t)$ which uses the Gaussian distribution rather than the t distribution. We obtain the type I error rates for our simulated data, the INT transformed data, and the power transformed data using the least squares estimation method above with these two ways of p-value calculation. The results are listed in Table 2.

Table 1. Type I Error Rate

	Distribution	Sample Size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Original	$N(0,1)$	6	0.0497	0.0102	0.0008
INT	$N(0,1)$	6	0.1029*	0.0000^	0.0000^
Power	$N(0,1)$	6	0.0659*	0.0174*	0.0009
Original	$N(0,1)$	20	0.0504	0.0093	0.0011
INT	$N(0,1)$	20	0.0503	0.0092	0.0010
Power	$N(0,1)$	20	0.0507	0.0090	0.0011
Original	$N(0,1)$	50	0.0485	0.0100	0.0008
INT	$N(0,1)$	50	0.0467	0.0100	0.0006
Power	$N(0,1)$	50	0.0487	0.0099	0.0010
Original	$Exp(1)$	6	0.0378^	0.0095	0.0015
INT	$Exp(1)$	6	0.0964*	0.0000^	0.0000^
Power	$Exp(1)$	6	0.0687*	0.0192*	0.0027*
Original	$Exp(1)$	20	0.0443^	0.0071	0.0004^
INT	$Exp(1)$	20	0.0495	0.0108	0.0006
Power	$Exp(1)$	20	0.0519	0.0114	0.0012
Original	$Exp(1)$	50	0.0478	0.0080	0.0005
INT	$Exp(1)$	50	0.0509	0.0107	0.0014
Power	$Exp(1)$	50	0.0501	0.0112	0.0012
Original	$\chi^2(1)$	6	0.0353^	0.0096	0.0011
INT	$\chi^2(1)$	6	0.1018*	0.0000^	0.0000^
Power	$\chi^2(1)$	6	0.0745*	0.0262*	0.0035*
Original	$\chi^2(1)$	20	0.0396^	0.0044^	0.0004^
INT	$\chi^2(1)$	20	0.0523	0.0094	0.0013
Power	$\chi^2(1)$	20	0.0532	0.0112	0.0018*

	Distribution	Sample Size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Original	$\chi^2(1)$	50	0.0436^	0.0054	0.0004^
INT	$\chi^2(1)$	50	0.0507	0.0109	0.0007
Power	$\chi^2(1)$	50	0.0501	0.0110	0.0009
Original	$Laplace(0,1)$	6	0.0359^	0.0062	0.0006
INT	$Laplace(0,1)$	6	0.0977*	0.0000^	0.0000^
Power	$Laplace(0,1)$	6	0.0549	0.0112	0.0008
Original	$Laplace(0,1)$	20	0.0466	0.0075	0.0009
INT	$Laplace(0,1)$	20	0.0519	0.0097	0.0013
Power	$Laplace(0,1)$	20	0.0488	0.0089	0.0009
Original	$Laplace(0,1)$	50	0.0459	0.0096	0.0008
INT	$Laplace(0,1)$	50	0.0466	0.0114	0.0012
Power	$Laplace(0,1)$	50	0.0468	0.0102	0.0011
Original	$Rayleigh(1)$	6	0.0544	0.0101	0.0013
INT	$Rayleigh(1)$	6	0.1003*	0.0000^	0.0000^
Power	$Rayleigh(1)$	6	0.0680*	0.0166*	0.0018*
Original	$Rayleigh(1)$	20	0.0495	0.0090	0.0011
INT	$Rayleigh(1)$	20	0.0489	0.0096	0.0014
Power	$Rayleigh(1)$	20	0.0506	0.0093	0.0014
Original	$Rayleigh(1)$	50	0.0483	0.0095	0.0010
INT	$Rayleigh(1)$	50	0.0465	0.0098	0.0011
Power	$Rayleigh(1)$	50	0.0474	0.0098	0.0010
Original	$Weibull(1,0.5)$	6	0.0386^	0.0096	0.0009
INT	$Weibull(1,0.5)$	6	0.0957*	0.0000^	0.0000^
Power	$Weibull(1,0.5)$	6	0.0694*	0.0197*	0.0014

	Distribution	Sample Size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Original	$Weibull(1,0.5)$	20	0.0464	0.0057	0.0001^
INT	$Weibull(1,0.5)$	20	0.0486	0.0107	0.0010
Power	$Weibull(1,0.5)$	20	0.0502	0.0119	0.0008
Original	$Weibull(1,0.5)$	50	0.0495	0.0081	0.0003^
INT	$Weibull(1,0.5)$	50	0.0529	0.0100	0.0012
Power	$Weibull(1,0.5)$	50	0.0543	0.0104	0.0012
Original	$Cauchy(0,1)$	6	0.0170^	0.0037^	0.0004^
INT	$Cauchy(0,1)$	6	0.0964*	0.0000^	0.0000^
Power	$Cauchy(0,1)$	6	0.0369^	0.0078	0.0005
Original	$Cauchy(0,1)$	20	0.0195^	0.0014^	0.0000^
INT	$Cauchy(0,1)$	20	0.0497	0.0117	0.0012
Power	$Cauchy(0,1)$	20	0.0365^	0.0035^	0.0001^
Original	$Cauchy(0,1)$	50	0.0188^	0.0015^	0.0000^
INT	$Cauchy(0,1)$	50	0.0523	0.0120	0.0009
Power	$Cauchy(0,1)$	50	0.0381^	0.0041^	0.0000^
Original	$Lognormal(0,3)$	6	0.0098^	0.0025^	0.0015
INT	$Lognormal(0,3)$	6	0.0990*	0.0000^	0.0000^
Power	$Lognormal(0,3)$	6	0.0665*	0.0234*	0.0034*
Original	$Lognormal(0,3)$	20	0.0077^	0.0005^	0.0000^
INT	$Lognormal(0,3)$	20	0.0514	0.0099	0.0013
Power	$Lognormal(0,3)$	20	0.0508	0.0112	0.0022*
Original	$Lognormal(0,3)$	50	0.0114^	0.0002^	0.0000^
INT	$Lognormal(0,3)$	50	0.0489	0.0098	0.0012
Power	$Lognormal(0,3)$	50	0.0512	0.0108	0.0016*

*10,000 simulations are used. * marks the values that are greater than expectation. ^ marks the values that are smaller than expectation.*

Table 2: Type I Error Rate with Two Ways of Calculating P-Values

	Distribution	Sample Size	T Distribution	Gaussian Approximation
Original	$N(0,1)$	6	0.0504	0.1327
INT	$N(0,1)$	6	0.1073	0.1073
Power	$N(0,1)$	6	0.0692	0.1485
Original	$N(0,1)$	20	0.0477	0.0627
INT	$N(0,1)$	20	0.0465	0.0627
Power	$N(0,1)$	20	0.0474	0.0639
Original	$N(0,1)$	50	0.0492	0.0560
INT	$N(0,1)$	50	0.0510	0.0556
Power	$N(0,1)$	50	0.0500	0.0547
Original	$N(0,1)$	150	0.0485	0.0499
INT	$N(0,1)$	150	0.0483	0.0510
Power	$N(0,1)$	150	0.0472	0.0494
Original	$Exp(1)$	6	0.0412	0.1008
INT	$Exp(1)$	6	0.1010	0.1010
Power	$Exp(1)$	6	0.0715	0.1364
Original	$Exp(1)$	20	0.0438	0.0588
INT	$Exp(1)$	20	0.0512	0.0675
Power	$Exp(1)$	20	0.0526	0.0688
Original	$Exp(1)$	50	0.0451	0.0517
INT	$Exp(1)$	50	0.0480	0.0532
Power	$Exp(1)$	50	0.0480	0.0524
Original	$Exp(1)$	150	0.0467	0.0487
INT	$Exp(1)$	150	0.0499	0.0516
Power	$Exp(1)$	150	0.0515	0.0530

	Distribution	Sample Size	T Distribution	Gaussian Approximation
Original	$\chi^2(1)$	6	0.0333	0.0858
INT	$\chi^2(1)$	6	0.1032	0.1032
Power	$\chi^2(1)$	6	0.0748	0.1398
Original	$\chi^2(1)$	20	0.0367	0.0534
INT	$\chi^2(1)$	20	0.0514	0.0643
Power	$\chi^2(1)$	20	0.0506	0.0651
Original	$\chi^2(1)$	50	0.0465	0.0525
INT	$\chi^2(1)$	50	0.0505	0.0569
Power	$\chi^2(1)$	50	0.0513	0.0574
Original	$\chi^2(1)$	150	0.0516	0.0541
INT	$\chi^2(1)$	150	0.0498	0.0524
Power	$\chi^2(1)$	150	0.0518	0.0539
Original	$Laplace(0,1)$	6	0.0376	0.1119
INT	$Laplace(0,1)$	6	0.1027	0.1027
Power	$Laplace(0,1)$	6	0.0596	0.1377
Original	$Laplace(0,1)$	20	0.0456	0.0623
INT	$Laplace(0,1)$	20	0.0499	0.0664
Power	$Laplace(0,1)$	20	0.0473	0.0644
Original	$Laplace(0,1)$	50	0.0493	0.0546
INT	$Laplace(0,1)$	50	0.0506	0.0552
Power	$Laplace(0,1)$	50	0.0480	0.0539
Original	$Laplace(0,1)$	150	0.0524	0.0544
INT	$Laplace(0,1)$	150	0.0515	0.0533
Power	$Laplace(0,1)$	150	0.0518	0.0543
Original	$Rayleigh(1)$	6	0.0535	0.1216

	Distribution	Sample Size	T Distribution	Gaussian Approximation
INT	$Rayleigh(1)$	6	0.1027	0.1027
Power	$Rayleigh(1)$	6	0.0703	0.1375
Original	$Rayleigh(1)$	20	0.0506	0.0661
INT	$Rayleigh(1)$	20	0.0509	0.0673
Power	$Rayleigh(1)$	20	0.0517	0.0676
Original	$Rayleigh(1)$	50	0.0511	0.0572
INT	$Rayleigh(1)$	50	0.0514	0.0582
Power	$Rayleigh(1)$	50	0.0515	0.0590
Original	$Rayleigh(1)$	150	0.0505	0.0520
INT	$Rayleigh(1)$	150	0.0516	0.0537
Power	$Rayleigh(1)$	150	0.0512	0.0531
Original	$Weibull(1,0.5)$	6	0.0399	0.0984
INT	$Weibull(1,0.5)$	6	0.0976	0.0976
Power	$Weibull(1,0.5)$	6	0.0706	0.1321
Original	$Weibull(1,0.5)$	20	0.0442	0.0591
INT	$Weibull(1,0.5)$	20	0.0512	0.0654
Power	$Weibull(1,0.5)$	20	0.0518	0.0665
Original	$Weibull(1,0.5)$	50	0.0432	0.0480
INT	$Weibull(1,0.5)$	50	0.0462	0.0525
Power	$Weibull(1,0.5)$	50	0.0470	0.0523
Original	$Weibull(1,0.5)$	150	0.0485	0.0506
INT	$Weibull(1,0.5)$	150	0.0498	0.0511
Power	$Weibull(1,0.5)$	150	0.0468	0.0480
Original	$Cauchy(0,1)$	6	0.0198	0.0691
INT	$Cauchy(0,1)$	6	0.0948	0.0948

	Distribution	Sample Size	T Distribution	Gaussian Approximation
Power	$Cauchy(0,1)$	6	0.0420	0.1174
Original	$Cauchy(0,1)$	20	0.0195	0.0306
INT	$Cauchy(0,1)$	20	0.0522	0.0679
Power	$Cauchy(0,1)$	20	0.0370	0.0562
Original	$Cauchy(0,1)$	50	0.0212	0.0247
INT	$Cauchy(0,1)$	50	0.0488	0.0553
Power	$Cauchy(0,1)$	50	0.0370	0.0430
Original	$Cauchy(0,1)$	150	0.0185	0.0193
INT	$Cauchy(0,1)$	150	0.0496	0.0511
Power	$Cauchy(0,1)$	150	0.0372	0.0390
Original	$Lognormal(0,3)$	6	0.0112	0.0358
INT	$Lognormal(0,3)$	6	0.0989	0.0989
Power	$Lognormal(0,3)$	6	0.0696	0.1308
Original	$Lognormal(0,3)$	20	0.0089	0.0149
INT	$Lognormal(0,3)$	20	0.0520	0.0666
Power	$Lognormal(0,3)$	20	0.0536	0.0690
Original	$Lognormal(0,3)$	50	0.0100	0.0131
INT	$Lognormal(0,3)$	50	0.0478	0.0529
Power	$Lognormal(0,3)$	50	0.0501	0.0557
Original	$Lognormal(0,3)$	150	0.0135	0.0151
INT	$Lognormal(0,3)$	150	0.0495	0.0513
Power	$Lognormal(0,3)$	150	0.0490	0.0517

10,000 simulations and a significance level $\alpha = 0.05$ *are used.*