REGRESSION ANALYSIS ON PHOTOMETRIC REDSHIFT

Vinicio Haro

A thesis submitted in partial fulfillment of the requirements for, the degree of Master of Arts in Applied Mathematics and Statistics, Concentration in Statistics. Summer 2016

Hunter College CUNY

ABSTRACT

Regression Analysis on Photometric Redshift

By Vinicio Haro

Faculty Supervisor: Professor Jingtao Gou

The purpose of this case study is to highlight an efficient regression model to estimate the photometric redshift of a galaxy based on its photometry. For the purpose of this study, the knowledge of the underlying physical processes behind redshift do not need to be known as the emphasis is on data analysis techniques and their overall performance. The analysis will be done using R open source programing language on two separate datasets. One data set is simulated redshifts from a gamma distribution; the other dataset is the observed de-reddened redshifts from the Sloan Digital Satellite Survey.

TABLE OF CONTENTS

GLOSSARY

Redshift. When light or other forms of electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.

Photometry. A technique of astronomy concerned with measuring the flux or intensity of an astronomical object's electromagnetic radiation

GLM. General Linear Model

Galaxy. A system of stars independent from all other systems

Kurtosis. The sharpness of the peak of a frequency distribution curve

Noise. Unexplained variation in a sample

Chapter 1

Introduction

Background

Astrostatistics is a fairly new discipline which bridges the gap between astrophysics and statistical analysis, commonly believed to be disjoint. Observing the universe via satellites or radio waves produces a large quantity of data. Methods of statistical analysis and machine learning seek to interpret large datasets and make predictions. It only seems natural to take these statistical techniques and apply them to astronomical data. One of the very first statistical techniques to be used in astronomy was the classical least squares regression to model phenomenon such as luminosity. Regression is going to be the focus of this study due to the ability to be able to train regression models using any standard modern day desktop. This element opens the door to amateur astronomers who are looking to try their hand at using statistical techniques on the vast quantity of astronomical data available to the public.

John Nelder and Robert Wedderburn introduced the GLM in 1972. Julian Faraway refers to GLM's as being central to the practice of statistics. The GLM can be extended into several different variations. The most basic is the least squares linear model, which to some, is considered to be the standard when the residuals have a normal distribution. The extension of the ordinary least squares model is the generalized linear model also known as GLM. GLM's cover a larger class of distributions and data frames containing categorical, discrete, and continuous variables. Some examples of common GLM's include the binomial model, Poisson model, gamma model, and inverse Gaussian model. Another variation would be the mixed effect model. Mixed effect models come into play when data is grouped or contains a hierarchal structure. The error term of a mixed effect model contain a correlation component. The third variation is Non-Parametric Regression models. These models are used when linearity is not enough to capture a data structure and require some smoothing parameters. Examples of non-parametric regression are additive models. (Faraway, Julian. "Extending the Linear Model in R." Boca Raton, Florida: Chapman & Hall, 2006)

Problem Statement

Using the programing language R, regression models will be implemented in order to make an estimate of photometric redshift of a galaxy. The analysis guidelines in *"The Overlooked Potential of Generalized Linear Models in Astronomy II Gamma Regression and Photometric Redshifts"* for the COIN collaboration by Elliot, Souza, Krone-Martins, Cameron, Ishida, Hilbe, will be loosely followed. The next section takes a look at other GLM's popular in astrostatistics from *"Modern Statistical Methods in Astronomy"* by Eric D. Feigelson and G. Jogesh Babu. This is followed by a look into the use of additive models from *"Extending the Linear Model with R"* by Julian Faraway. The data analysis portion consists of an examination of data processing to deal with potential irregularities in the data frame. The next section of the analysis is implementing the gamma regression model for reference, and selecting additional regression models from those discussed in this case study. The goal is to find the most efficient regression model out of the models discussed.

There are two datasets that will be used in this study. The first dataset is called PHAT0, which simulates galaxies with redshifts from a gamma distribution. PHAT0 is available through the R package CosmoPhotoz in the CRAN repository. The other dataset is the actual observed data from SDSS otherwise known as the Sloan Digital Sky Survey. The SDSS was obtained from the sky server using a simple SQL query on a web-based interface.

For further reading, I highly recommend *"Astrostatisical Challenges for the New Astronomy"*, edited by Joseph M. Hilbe. The literature highlights studies by some of the worlds leading Astrostatiscians and the text is a great way to see how much the field has grown.

Chapter 2

Generalized Linear Models

The Gamma Model

We will closely examine the gamma model, which is the model used in "The overlooked Potential of GLM's." The Gamma distribution takes the following form with a slight modification for the purpose of the GLM. For y greater than zero, and mean u:

$$f(y) = \frac{1}{\Gamma(v)}\left(\frac{v}{u}\right)^{v} y^{v-1}e^{-\left(\frac{yv}{u}\right)}$$

The response variable of the Gamma model is part of the exponential family, consisting of a canonical parameter theta and dispersion parameter phi, which respectively represents the location and scale. The exponential family takes the general form:

$$f(y|\theta,\phi) = \exp\,[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)]$$

Taking the log of both sides of the gamma distribution will show that it can be re written in the general exponential family form. This also produces the necessary link function, which is crucial for the regression analysis. In the case of the gamma, it can be shown that the link function is the inverse of the canonical parameter (Faraway, Julian. "Extending the Linear Model With R." Boca Raton, Florida: Chapman & Hall, 2006).

In the context of gamma regression, the following summarizes the mean, variance, and the link function:

$$var(y_i) = \sigma^2\big(E(y_i)\big)^2 \qquad u_i = E(y_i) = e^{B_0}x_{i1}^{B_1}\dots x_{ip}^{B_p}$$

If the model has normally distributed errors in the log scale, then consider a linear model on the log scale. If the response variable has a gamma distribution, then use the gamma regression model with the following parameters:

$$u_i = \frac{v}{\lambda_i} \qquad\qquad \sigma^2 = \frac{1}{v}$$

In the analysis of redshift data, we will consider only a positive and continuous redshift. For this reason, it makes sense that the Gamma Model was chosen despite the fact that Gamma is not the most popular of the GLM's in the exponential family.

How do we gauge the performance of the Gamma? In R, we can produce diagnostics plots that are specific for GLM's including the gamma. We can also check the predicted values and see how close they are to the observed values in a QQ line plot. In combination with prediction, we can gauge how well the model is performing against the observed values in a test subset by taking its correlation. The higher the correlation, the better our model predicts the observed values.

Other GLM's

In this section, GLM's from *"Modern Statistical Methods for Astronomy With R Applications"* by Eric D. Feigelon and G. Jogesh Babu are going to be discussed. The following GLM's are already implemented in the analysis of astrological data.

Regression is a powerful tool that can model phenomena in astronomy with the use of any desktop computer by today's standards. At the most simplest case, a linear model has been used to fit the SDSS quasar data using i and z band photometry. The effectiveness of the linear model can be gauged using the adjusted R squared value or confidence intervals. Linear models operate with the assumption that the errors are normally distributed. Using R, diagnostics can be produced quickly. Common diagnostics for linear models are Cook's statistic, Q-Q plot, and square root transformation of residuals. These plots provide a visual tool to check for non-constant variance and potential transformation of the response variables. Linear models have its weaknesses especially when it comes to outliers. There are alternate models that can help reduce the effect of such outliers.

Robust regression comes into play when there exist outliers that are not a result of large measurement errors. Robust regression seeks to minimize the effect of such outliers and reduce their overall influence. R allows the user to implement this model using the MASS function. Within the package MASS, the fit is computed using iteratively reweighted least squares algorithm. The model is based on Huber's function, which yields a default M-estimator with the value of k being 1.345. This model does a better job at investigating the relationships in the mentioned quasar data.

A modification of the linear model is called Quantile Regression, also known as Q regression for short. When the epsilon errors are not normally distributed, we can find crucial information hidden along the quantiles. An example mentioned inside the literature is the case when there exists a distribution that has a rather narrow peak and has a wide symmetric tail. In R, Q regression is implemented using the quantreg package. Once implemented, non-linearity is much easier to identify.

Another model used in the field is the MARs regression model also known as multivariate adaptive regression splines developed by J. Friedman and J. Turkey in the 1970's. This is a very flexible regression model that is able to get around the problem of co-linearity. According to the literature, MARS fits piecewise polynomials over specific regions called "Knot" locations. MARS deletes the terms that yield the smallest contribution to the residual square errors and implements general cross validation. MARS regression is part of the non-parametric regression models family known as additive models. In general, additive models are useful when the distribution of the response variable doesn't closely follow a well-defined distribution such as normal or gamma. There are not many assumptions attached to additive models. Given the nature of the additive models, an entire section is dedicated to taking a closer look at them.

As shown, there already exists some variety in the type of regression models currently used on astrological data. There are more GLM's of course but these seem to be the most popular ones. The literature goes into much deeper detail along with sample code for the reader to try at home. This is highly recommended read for someone who is starting out in Astrostatistics such as myself. (Feighelson, E. and Babu, G. "Modern Statistical Methods for Astronomy with R Applications" University of Cambridge: University Printing House, 2012).

Additive Models

This section provides a behind the scenes look at the nature of the additive model and what makes them a potentially better model to use for the nature of the data at hand. From *Extending the Linear Model With R*, by Julian Faraway, the additive model is defined as follows: Try a non-parametric approach:

$$y = f(x_1, \dots, x_p) + \epsilon$$

With this form, the necessity for parametric assumptions is bypassed, however when p becomes large, it becomes impractical to fit such models, hence the additive model provides a good remedy that bypasses parametric assumptions and is practical in nature to fit. The additive model takes the form:

$$y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

The literature points out that function f, for all j greater than or equal to 1, is a collection of smooth arbitrary functions.

Stone first introduced the additive model in 1985. They have much more flexibility than their parametric counterparts but still very straight forward to interpret. Plots of additive models are functional when it comes to visualizing any marginal relationships between the response variable and the predictor variables. The additive model does a good job when exploring interaction terms between the predictors such as the factorial form, which we will see in more detail later on.

Using R, the additive model can be fit using three different approaches. The package “gam” fits an additive model using a back fitting algorithm. The other approach for additive models is using the “mgcv” package in R. This package implements penalized smoothing splines where parameters for smoothing can be adjusted. The third approach is found in the package “gss” which is based on splines.

Additive models can de gauged with the usual diagnostic plots for GLM’s such as the QQ line plot. They are very functional for prediction and also have a metric called, adjusted R-squared, which indicates if the model is a good fit or not. In this case study, additive models will be explored with the SDSS dataset. They show signs of being more efficient and faster to compute than parametric regression models. Their ease of interpretation makes them more intuitive and can be adapted by amateur astronomers curious on finding a quick way to quantify astronomical data, be it quasar data or redshift data.

In the previous section, the MARS regression model, formally known, as multivariate adaptive regression splines model is another popular variation of the additive model. It should be noted that Faraway defines the MARS model slightly different than in *Modern Statistical Methods for Astronomy with* R *applications.* Faraway defines MARS as follows:

$$\hat{f}(x) = \sum_{j=1}^{k} c_j B_j(x) \text{ Where } B_j(x) \text{ are the basis functions}$$

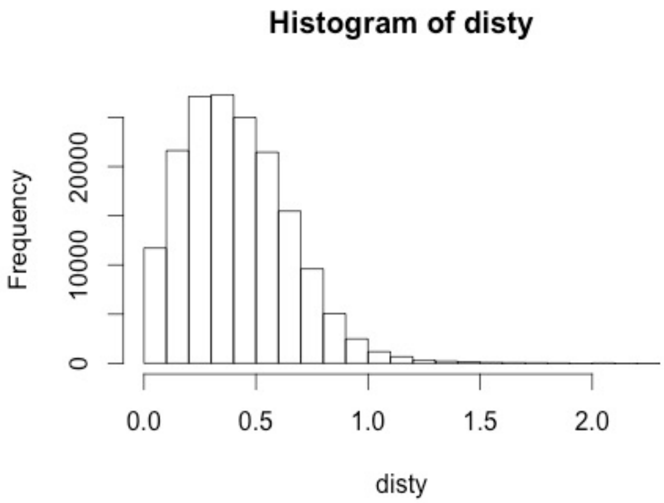
For the complete breakdown on additive models along with examples, look for *Extending the Linear Model With* R*,* by Faraway. For the sake of this case study, it was only necessary to briefly summarize Faraway's literature however; it is recommended to read the entirety of chapter 12 for a much deeper understanding.

Chapter 3

Data and analysis

The Data

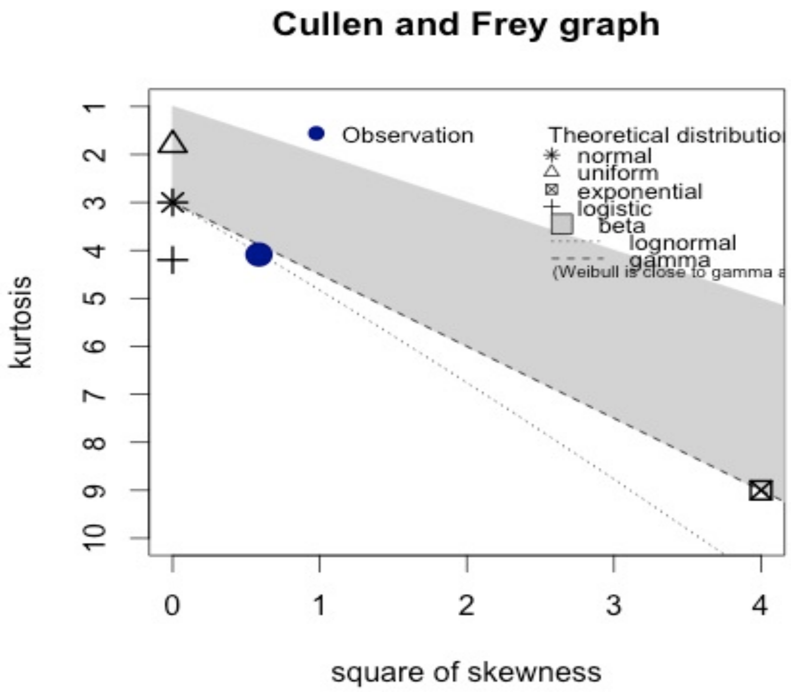
The first dataset comes from the R package CosmoPhotoz found in the CRAN repository. The package contains 169,520 observations simulated using gamma distribution with a log link function. The 12 variables are redshift, u, g, r, i, z, Y, J, H, K, IRAC_1, and IRAC_2. Redshift is the response variable and the other variables are the photometric magnitudes of galaxies for that respective band. It includes a pre-split training and test subset of the data. The test set PHAT0test contains 161042 observations and the training set PHAT0train contains 8478 observations. The simulated dataset however is not the focus of this case study but rather used to see the behavior of the gamma regression model before applying it to the SDSS data. The distribution of the response variable "redshift" can be seen with a simple histogram. Note: "disty" stands for distribution of Y (response variable)



The histogram reveals that the distribution of the response variable is pretty close to a gamma distribution, which shouldn't come as a surprise since these observations are simulated from a gamma distribution. We should take into consideration that the values for redshift in this data frame are continuous, positive, and bounded between 0.02 and 2.24. Using R, we produced a simple table of descriptive statistics.

Descriptive Statistics for PHAT0

	Redshift	u	g	r	i	z	Y	J	H	K	IRAC_1	IRAC_2
Min	0.0200	15.37	14.88	14.70	14.4	14.29	14.12	14.07	13.84	14.03	14.80	15.22
Median	0.4000	24.16	23.74	23.13	22.7	22.53	22.39	22.18	22.02	21.95	22.78	23.27
Mean	0.4244	24.05	23.48	22.82	22.4	22.29	22.14	21.93	21.76	21.69	23.35	23.69
Max	2.2400	28.94	26.60	24.14	24.1	23.97	23.92	23.74	23.70	23.89	26.37	26.37

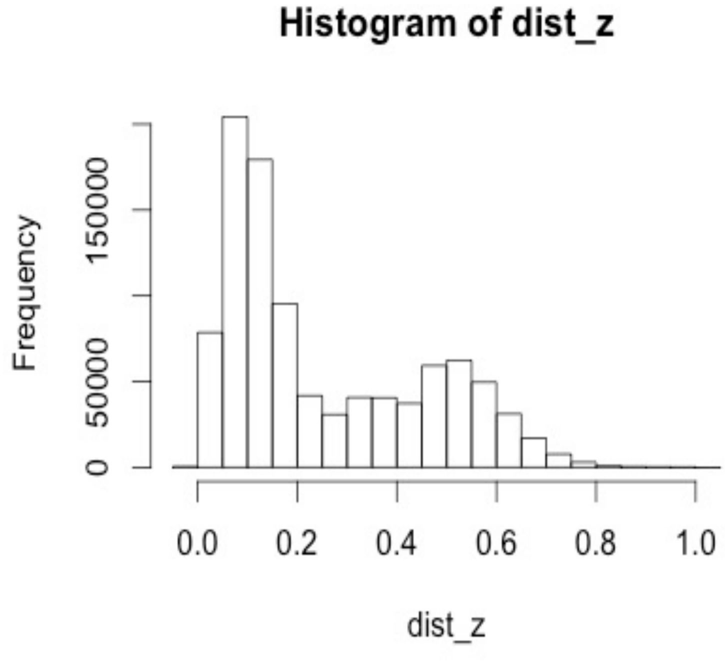


The above Cullen and Frey graph solidifies the notion that the simulated data is consistent with what we would expect to find in a gamma distribution for the response variable redshift. Rafael S. de Souza, Alberto Krone-Martins, Jonathan Elliott, and Joseph Hilbe are the authors of "CosmoPhotoz", released on The 20th of August 2014.

The second dataset used in this case study is the actual observed data from the Sloan Digital Satellite Survey (SDSS) and can be retrieved from the Skyserver website.

The data is accessed from the skyserver using a modified version of the sample SQL query from casjobs(public.mikhailway.ashok2015). The data contains 981,270 observations and 25 variables. Only the variables z, dered_u, dered_g, dered_r, dered_i, and dered_z will be considered from this data frame. In this dataset, z, is the response variable. By restricting the data frame to just these six variables, we can mirror the bands used in the simulated dataset as closely as possible. Rows with missing observations will be omitted from the data frame. Please note that the predictor variables consist of de-reddened magnitudes, meaning they have a reduced redshift.

A histogram of the response variable is shown below to gauge the distribution. For the SDSS data, dist_z is the distribution of the response variable.



Upon first glance, it looks like the redshift has an overall gamma distribution with some discrepancy between 0.5 and 0.6 on the x-axis. The data frame contains some extreme magnitudes relative to the other magnitudes for each variable. This is confirmed with a quick outlier test:

Extreme Outliers per Variable on SDSS data

Dered_u	Dered_g	Dered_z	Dered_i	Dered_r	z
9999.000000	9999.000000	9999.000000	9999.000000	9999.000000	1.002166

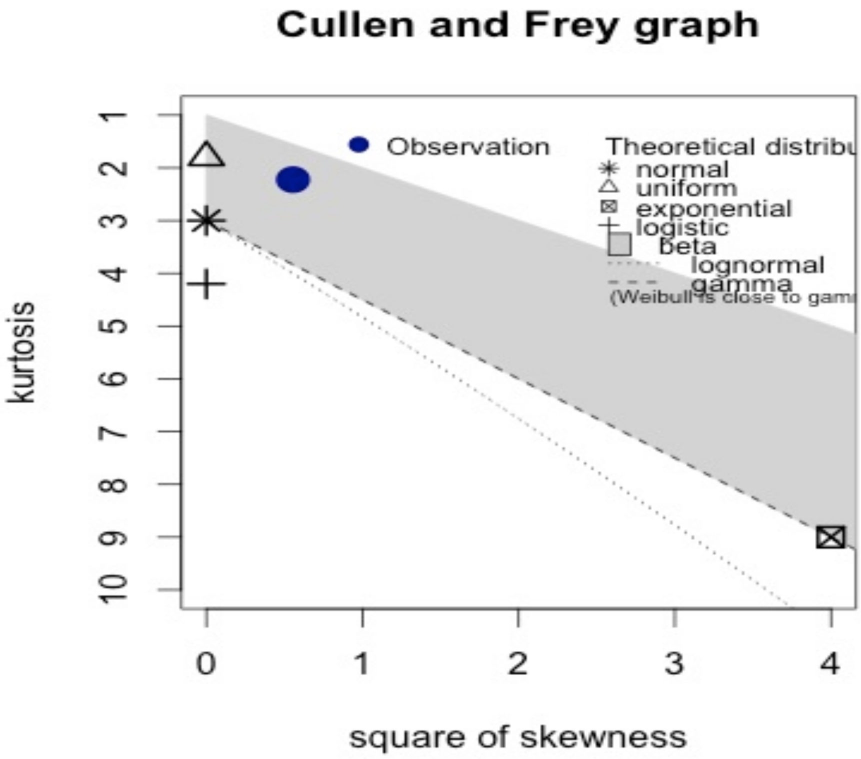
These values were removed from the data frame. Some possibilities onto why they are there could be some sort of error in the data entry or perhaps something immeasurable. In the scope of this study, they can certainly have a negative effect on our results seeing as how the other relative magnitudes are roughly below 50.

A simple table of descriptive statistics is provided. As shown on the table below, the redshift z is bounded between 0.00 and 1.0021. It should be noted that the absolute value of the magnitudes is considered in this data frame, Removing extreme values and considering the absolute value of magnitudes does not change the overall distribution.

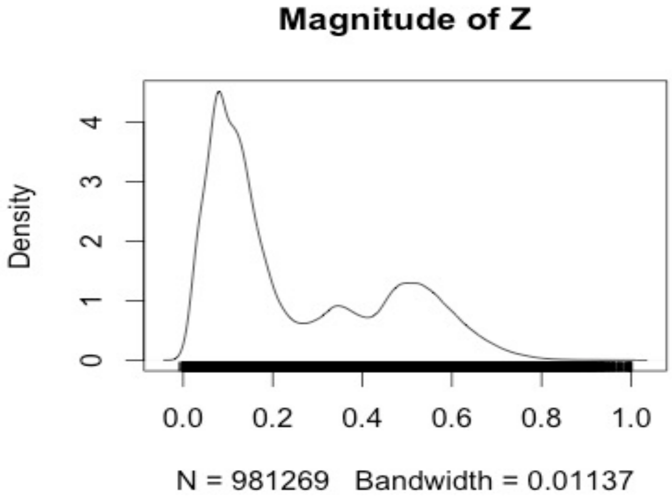
Descriptive Statistics for SDSS observed data

	dered_u	dered_g	dered_r	dered_i	dered_z	z
Min	0	7.446	8.51	0	10.14	0.0000
Median	20.35	18.558	17.62	17.22	16.95	0.1610
Mean	21.05	19.274	18.14	17.58	17.26	0.2532
Max	30.64	29.947	31.13	29.52	29.89	1.0021

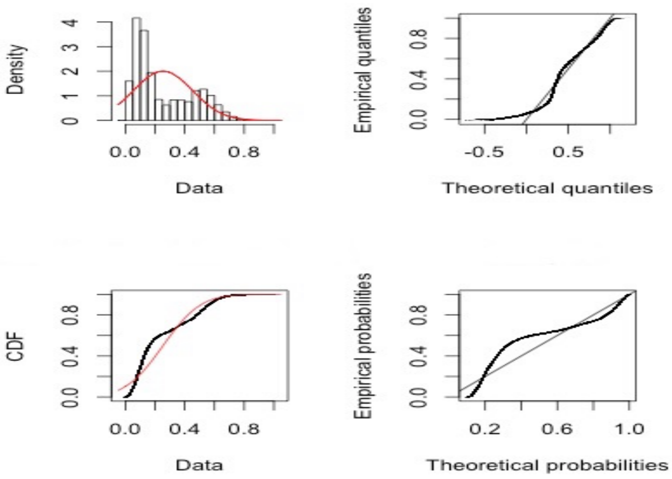
The histogram above only gives us an idea of what the distribution may be. It is not as clear as the histogram from the simulated dataset. A closer look is recommended. The Cullen and Frey graph is a useful visual utility that uses the kurtosis and square of skew to identify the type of distribution the data may be more consistent with. Since the redshift is continuous, the parameter is adjusted to cross check with other continuous distributions and not discrete distributions.



The estimated standard deviation is 0.1994624. The estimated skew is 0.745037 and the estimated kurtosis is 2.223756. The location of the blue dot seems to indicate that the distribution of the response variable may be more consistent with a beta distribution. It should be mentioned that real world data almost never actually follows the form of a known distribution perfectly. 0 and 1 bound beta distribution but in the case of redshift, our values are outside that range. The beta distribution is a continuous distribution of probabilities so trying to compare this to redshift, would not be such a good move. Instead, perhaps it's best to look at the other distributions that can be somewhat associated with redshift such as gamma or normal based on the information uncovered. Another useful plot is the density plot for our dependent variable. Such a plot can help identify if a transform on the response variable is necessary to "center" the distribution.



In the case of the redshift in the SDSS data, it's observed that there exists some discrepancy around the middle consistent with what was seen in the histogram. This could also be known as "noise." The left portion of the graph is roughly consistent with what a gamma distribution density could look like. There are some distributions that can be ruled out right off the bat such as a normal distribution. If the distribution of redshift is superimposed on a normal density, the results are as follows:



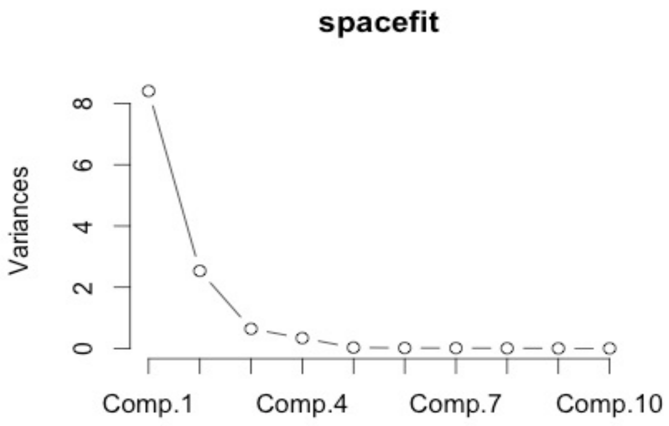
From the four mini plots, it's clear that the distribution of redshift does not share much relation with a normal distribution. The KS test confirms this with a very low p value, much lower than 0.5.

Data Processing

Data processing is a vital element of analysis. Real world data sometimes contains errors or irregularities that can have a big effect on a regression model. Sometimes there are outliers that can throw off the results of some test or other times, a response variable may need some sort of transformation to account for skew. The most common issue is the problem of multi-linearity. Data processing also includes Principal Component analysis and cross validation.

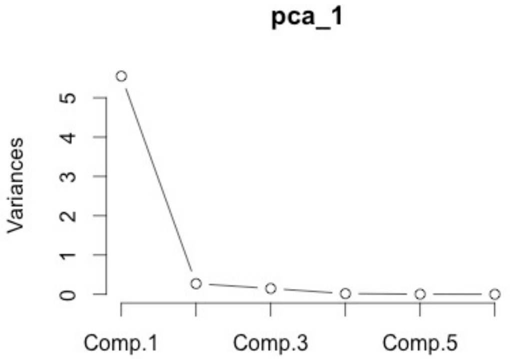
Before going into the data processing, it is important to know about the motivation behind PCA. The purpose of the PCA is to apply some sort of normalizing transformation to variables that may have correlation. By doing such transformation, these variables are converted into linearly uncorrelated values called principal components. Karl Pearson introduced the PCA technique in 1901. PCA's play a pivotal role when making predictive models. The long story short is that PCA's reduce the overall dimension of a data matrix and the results are gauged using the component scores. Producing component scores is simple to do using R.

For the simulated data PHAT0, the data processing guidelines highlighted in *"The Overlooked Potential of Generalized Linear Models in Astronomy II Gamma Regression and Photometric Redshifts, 2015"* will be followed. The data will be split into a test and training sample such that the training sample will contain 10 percent of the observations and the test set 90 percent. Please note that these subsets are of a different size than the pre-split subsets already included in the package. After splitting the data into its respective subsets, Principal Component analysis is implemented on the complete dataset, test set, and training set. In the literature, PCA decomposition revealed that at least 6 components needed to be used to retain roughly 99% variance. This is confirmed with the following plot:



For the SDSS dataset, we will implement an additional technique for processing the data. Lets use k-fold cross validation in addition to PCA decomposition. The purpose of this technique is to retain a single test subset of the original sample and the remaining k-1 subsets are going to be used as my training subsets. The data is partitioned into k equally sized subsets. In this study, let k=10.

The general motivation behind cross validation is to avoid problems such as over fitting and to arrive at a more accurate prediction. After creating a partition of the data with 10 fold cross validation, implement principal component decomposition. Six principal components were needed to achieve a variance threshold of 98%. The following plot is the PCA decomposition before 10 fold cross validation.



The two plots generated above only give a rough idea of how many principal components to retain, however, the best indicator of the number of PC's would be the loadings table showing the scores for each PC. The predictors are transformed into PC's where the scores become the new coefficients. The regression model will consist of the response variable written as a function of principal components.

Fitting the GLM Models on Simulated Data and SDSS data

Before fitting a GLM model, it is essential to justify the use of PCA. This can be done using a metric called the variance inflation factor commonly also known as VIF numbers. The VIf measures the degree of multi co-linearity within the variables of the data frame. If we check the VIF numbers for the simulated dataset, the following table is obtained:

VIF Numbers (Simulated dataset)

	Variables	VIF
1	Redshift	7.710026
2	up	26.138473
3	gp	87.223416
4	rp	93.218401
5	ip	201.004773
6	zp	203.345134
7	Y	195.864835
8	J	475.977235
9	H	538.471690
10	K	511.549180
11	IRAC_1	23.042958
12	IRAC_2	25.158006

As shown on the table, there are variables that have a substantially large VIF compared to the other variables for example, the band K has a VIF of over 500 compared to the up band which has a VIF under 30. This is evidence of multi co-linearity. Certain steps need to be done in order to remedy this problem and one of the most useful remedies is principal component decomposition otherwise known as PCA. Using a PCA method from Faraway's book, the following cumulative variances are calculated:

Cumulative Variance Table for PCA's (Simulated CosmoPhotoz data)

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Cumulative Proportion of Variance	0.7005268	0.9117499	0.96521330	0.99353193	0.996075024	0.997416270

The generated table states that either 5 or 6 components are necessary to account for at least 99.5 percent of the data variance, which agrees with the number of components found in the Gamma Regression study. In accordance to following the guidelines in that study, we will select six components.

The VIF numbers and the correlation matrix below are good indicators if interaction terms should be included in the regression model. If we observe the correlation matrix, it is shown that there exists high correlation amongst the predictor variables, for example the zp band and the ip band have a correlation of 0.99. We have quantifiable evidence of co-linearity amongst predictors based on the correlation matrix and VIF numbers. The literature suggests that interactions are necessary for complex data of this nature; hence they use the factorial form for interactions. The factorial PCA model is shown below. In our case, n=6 components and just to reiterate, this formulation is included in the analysis guidelines of the Gamma Regression study. It should also be mentioned that the poly function is applied to PC 1 and PC 2 in order to return orthogonal polynomials of degree 2.

$$Z_{phot} \sim PC_1^2 + PC_1 * \dots * PC_n$$

If we think about what redshift actually is, it's the Doppler effect analog for light waves. They represent a galaxy moving away due to the expenditure of space causing light waves to go towards the red section of the light spectrum, hence the name redshift. It makes sense to use interaction terms. As redshift changes, so do the magnitudes of the other photometry bands. There is a proportionality that is not so easily modeled with a linear dependency due to the complexity of the varying photometric bands. This is essentially what is being said in the gamma regression study.

Correlation Matrix (Simulated CosmoPhotoz dataset)

	Redshift	up	gp	rp	ip	zp	Y	J	H	K	IRAC_1	IRAC_2
Redshift	1.00	0.33	0.44	0.34	0.18	0.12	0.10	0.07	-0.02	-0.08	-0.38	-0.30
up	0.33	1.00	0.91	0.74	0.60	0.53	0.49	0.44	0.37	0.32	-0.09	-0.08
gp	0.44	0.91	1.00	0.93	0.82	0.78	0.75	0.70	0.63	0.58	0.10	0.14
rp	0.34	0.74	0.93	1.00	0.96	0.94	0.92	0.90	0.85	0.81	0.35	0.39
ip	0.18	0.60	0.82	0.96	1.00	0.99	0.99	0.97	0.95	0.93	0.51	0.55
zp	0.12	0.53	0.78	0.94	0.99	1.00	0.99	0.99	0.97	0.96	0.58	0.61
Y	0.10	0.49	0.75	0.92	0.99	0.99	1.00	0.99	0.98	0.97	0.60	0.64
J	0.07	0.44	0.70	0.90	0.97	0.99	0.99	1.00	0.99	0.98	0.65	0.69
H	-0.02	0.37	0.63	0.85	0.95	0.97	0.98	0.99	1.00	1.00	0.70	0.74
K	-0.08	0.32	0.58	0.81	0.93	0.96	0.97	0.98	1.00	1.00	0.73	0.76
IRAC_1	-0.38	-0.09	0.10	0.35	0.51	0.58	0.60	0.65	0.70	0.73	1.00	0.97
IRAC_2	-0.30	-0.08	0.14	0.39	0.55	0.61	0.64	0.69	0.74	0.76	0.97	1.00

In order to successfully apply a gamma regression model post PCA, the PCA scores must be transformed to play the role of "new" coefficients. The gamma regression model is redshift as a function of PCA scores with the factorial form. The complete tables of coefficients are left out due to the sheer size, however reproducible r code is included in the appendix. To reiterate the steps highlighted in the gamma regression study, split the dataset into two subsets with a 1:9 ratio, apply principal component analysis to the whole data and to each subset, and transform PC's into predictors using the factorial form of the minimum number of components necessary to achieve at least 99.5 percent variance using the scores.

The performance of the gamma model is quickly be gauged using a correlation between the trained gamma model and the test subset. The correlation value came out to 0.91377729, which indicates that the trained model is performing well in predicting the observations from the test subset. Full diagnostics will be shown in the following chapter.

We now turn to the SDSS dataset. The goal is to follow similar guidelines as the simulated dataset, however, include additional data processing techniques that could optimize the results of the fitted model. As in the simulated dataset, we begin the investigation with a table of VIF numbers and the correlation matrix.

VIF Numbers (SDSS Dataset)

	Variables	VIF
1	dered_u	7.348662
2	dered_g	68.403840
3	dered_r	289.802325
4	dered_i	307.685878
5	dered_z	122.065305
6	z	11.257299

Correlation Matrix (SDSS dataset)

	dered_u	dered_g	Dered_z	Dered_i	dered_r	z
dered_u	1.00	0.92	0.82	0.84	0.87	0.85
dered_g	0.92	1.00	0.94	0.96	0.98	0.93
dered_r	0.82	0.94	1.00	0.99	0.98	0.82
dered_i	0.84	0.96	0.99	1.00	0.99	0.85
dered_z	0.87	0.98	0.98	0.99	1.00	0.89
z	0.85	0.93	0.82	0.85	0.89	1.00

From looking at the correlation matrix, there is some correlation with the highest being at .99. The VIF numbers paint a similar picture where the highest VIF numbers pertain to dered_i and dered_z. There are several things that can be done with this information. Given that only two of the five predictor variables exhibit signs of multi co-linearity, they can be dropped all together from the model, interaction can be explored with those two variables, or some additional data processing such as cross validation can be used. Lets adopt the use of k fold cross validation.

PCA decomposition is applied in addition to k fold cross validation where k=10. The following cumulative variance proportions are generated:

Cumulative Variance for PCA's (SDSS data)

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Cumulative Proportion of Variance	0.9250037	0.97051505	0.99543559	0.998719473	0.9996193625	1.00000000

The table above tells us that 3 PC's are necessary to retain at least a 99.5 variance threshhold. To reiterate, the data was split into ten subsets of equal size where a single subset was picked to be test data and the other nine as training data. Apply a gamma regression model with the following coefficients, also using the factorial formulation from the simulated dataset including returning orthogonal polynomials of degree 2 for the first two PC's:

Coefficients of Gamma Model on SDSS data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.269e-01	2.525e-04	-3275.41	<2e-16 ***
poly(Comp.1, 2)1	-1.117e+02	7.432e-02	-1503.27	<2e-16 ***
poly(Comp.1, 2)2	-2.809e+01	7.472e-02	-375.90	<2e-16 ***
poly(Comp.2, 2)1	3.786e+01	9.428e-02	401.55	<2e-16 ***

poly(Comp.2, 2)2	-2.896e+00	1.364e-01	-21.23	<2e-16 ***
Comp.3	-2.945e-01	5.188e-04	-567.63	<2e-16 ***
poly(Comp.2, 2)1:Comp.3	1.404e+01	1.141e-01	123.07	<2e-16 ***
poly(Comp.2, 2)2:Comp.3	-1.216e+01	1.304e-01	-93.29	<2e-16***

With 3 PC's used including orthogonal polynomials of degree two for the first two PC's, the Dispersion parameter is taken at 0.005209606. This model produces Null Deviance of 14251.64 on 98125 degrees of freedom in addition to 921.16 on 98118 degrees of freedom for Residual deviance. AIC=-339832

How well did the trained model perform? After doing PCA and cross validation on subsets, we check the performance of the trained model against test data using a correlation metric. A value of 0.9525174 is produced, which indicates that the trained model is doing very well in predicting the z magnitudes in the test subset. Diagnostics will show any irregularities that may exist within the data and the model that can't be seen on a standard summary. We now explore the argument that perhaps this data would be best modeled with a more flexible type of regression model, however this does not imply that the gamma model is not a good choice either.

Additive Regression Models

This section will now focus on fitting two additive models on the SDSS dataset. In the previous section, we explored the usage of the gamma model on both the simulated and actual data. The additive models will be constructed using the same factorial form since there was evidence of high multi co-linearity amongst predictors. PCA decomposition and k fold cross validation is preserved in order to remedy co-linearity. The choice of additive models allows the user a bit more flexibility when modeling data especially when the distribution of the response variable is more on the ambiguous side.

The first of the additive model family to be used is MARS regression known formally as multivariate adaptive regression splines. The "earth" package in R facilitates the use of MARS regression. It should be noted that there will not be any tweaking done on the parameters in the earth function since the goal is just to provide a broad idea of the potential of this additive models. The following coefficients are generated:

Coefficients of MARS Regression Model on SDSS Data

	Coefficients
(Intercept)	0.212694
h(0.000463318-poly(Comp.1,2)1)	165.865147
h(poly(Comp.1,2)1-0.000463318)	-164.902071
h(-0.000985432-poly(Comp.2,2)1)	-44.668694
h(poly(Comp.2,2)1- -0.000985432)	50.608246
h(-0.790403-Comp.3)	0.170851
h(Comp.3- -0.790403)	-0.140477

The built in algorithm in the earth function selected 7 out of 7 terms, and 3 out of 7 predictors. At 7 terms, the R-Squared changes by less than 0.001. This model gives an RSS of 48.36754 and an R-squared of 0.9984716. The MARS model has indications of being a good fit based on its high R squared value, however the model performance should not be strictly dependent on a high or low R square value. This is a common pitfall when doing this type of analysis. It's easy to get into the habit of observing a very high R squared value and concluding that the model is "correct." The correlation metric yields a value of .9671756, meaning that the model is predicting about 96% of the test observations. This model is much simpler in structure since it contains less predictor terms and the coefficients are more intuitive to interpret.

The next additive model we will look at is the additive model built using the back fitting algorithm from the "gam" package in R. The gamma family with log link is still preserved in building this type of additive model. Specifying the family and link function within the parameters of the gam function does this for us. The following Coefficients are generated:

Coefficients of Additive Model using Gamma with log link on SDSS Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.807e+00	2.330e-04	-7758.49	<2e-16 ***
poly(Comp.1, 2)1	-7.455e+02	2.285e-01	-3262.06	<2e-16 ***
poly(Comp.1, 2)2	-2.189e+02	2.775e-01	-788.70	<2e-16 ***
poly(Comp.2, 2)1	2.972e+02	2.547e-01	1166.93	<2e-16 ***
poly(Comp.2, 2)2	3.217e+01	2.680e-01	120.04	<2e-16 ***
Comp.3	-5.847e-01	7.529e-04	-776.58	<2e-16 ***
poly(Comp.2, 2)1:Comp.3	-2.210e+01	4.321e-01	-51.15	<2e-16 ***
poly(Comp.2, 2)2:Comp.3	-2.737e+01	2.583e-01	-105.95	<2e-16 ***

This type of model has 91.3% of deviance explained. The R squared value comes out to .701 with just three PC's. Adding additional PC's into the model improves the adjusted R squared at the cost of increasing the complexity of the model. The correlation metric comes out to 0.9319969. If we look at the coefficients, they are considerably small meaning the additive model is more sensitive to marginal changes not so easily captured by a gamma GLM. Taking a look at the coefficients would indicate that any small change in component 1 would drastically change the magnitude of redshift.

Chapter 4

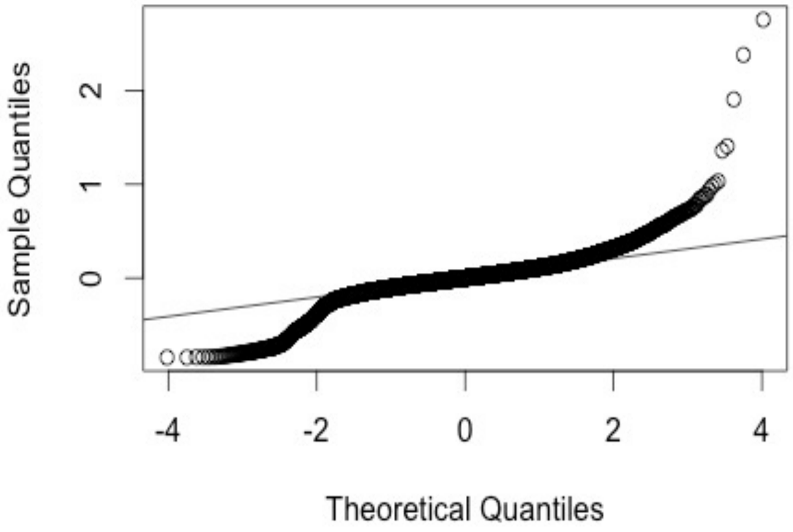
Results

Diagnostics

This section is the final stage of a typical regression analysis. The first involves interpretation of statistics given in the output table but model validation is confirmed with the use of diagnostic plots. Diagnostic plots often tell of irregularities between a trained model and observed values not so easily seen in regular output tables. This sections looks at diagnostics for the gamma model on the simulated dataset, the gamma model on the SDSS data, the MARS model on the SDSS data, and the back fitting additive model using a gamma with log link on SDSS data.

The gamma model built on the simulated dataset was constructed using 6 PC's based on the cumulative proportion to retail a certain variance threshold. The correlation value came out to 0.9137729, which indicates that the trained model is performing well in predicting the observations from the test subset, but how is this reflected in diagnostic plot?

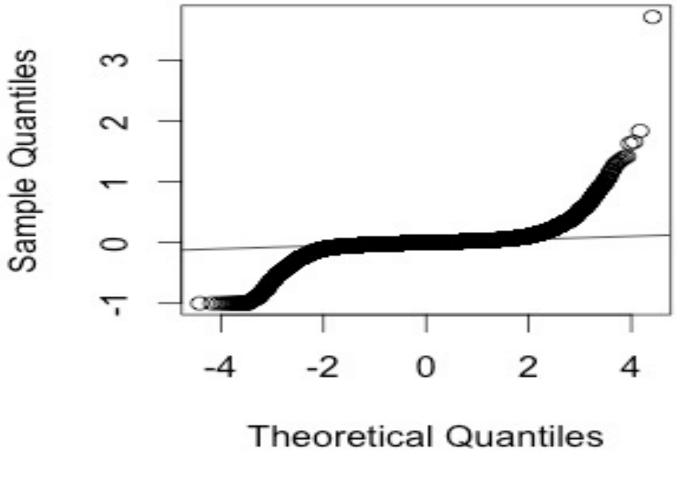
QQ Plot (Gamma Predicted Values vs. Observed Values for Simulated Data)



The QQ line plot demonstrates that the model does a good job at modeling a good majority of the data with the exception of the left and right tail end. The left tail has a higher concentration of observations while the right tail may indicate the existence of outliers or extreme values in the data frame. This model was built using PCA without any form of cross validation. The model took 0.700207 seconds seconds to compute which is pretty efficient given the amount of observations and PC's used. It should be noted that this diagnostic plot is very similar to the plot produced in the gamma regression case study. We won't include much more diagnostics for the simulated dataset in order to focus our attention on the SDSS dataset.

The next model built in this case study was the gamma regression model with a log link on the SDSS dataset with extreme values removed. The correlation between the trained model and test observations is 0.9525174 and took 1.402535 seconds to compute. The following diagnostics are produced:

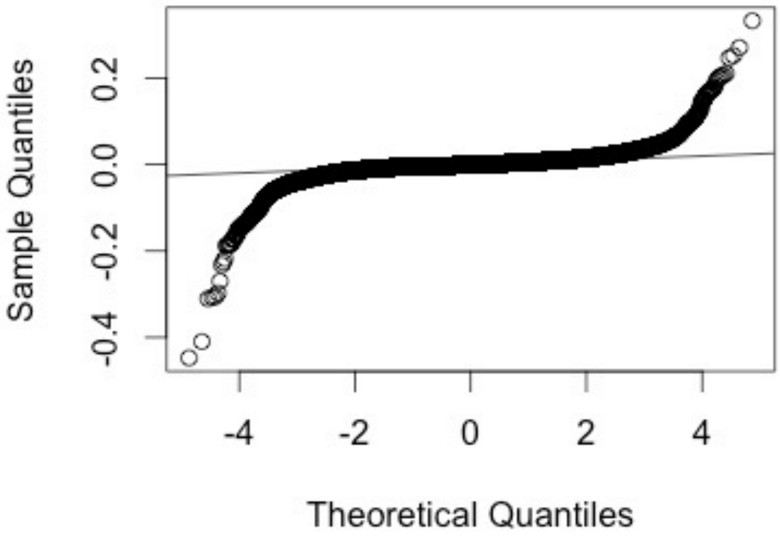
QQ Plot (Gamma Predicted Values vs. Observed Values for SDSS Data)



The behavior found in the plot here is similar to that of the simulated data QQ plot. The same type of irregularities seems to be present despite removing extreme values from the actual observed data frame. The difference is that the tail ends on the right and left have a much deeper concentration than in the simulated data. One potential reason may be that there are marginal changes within magnitudes that are not caught by the gamma glm, which would not be too surprising given that photometric magnitudes are already generally small. The standard deviation between residuals and the fit is 0.07217227, which indicates that there is not much deviation from a gamma distribution with a log link.

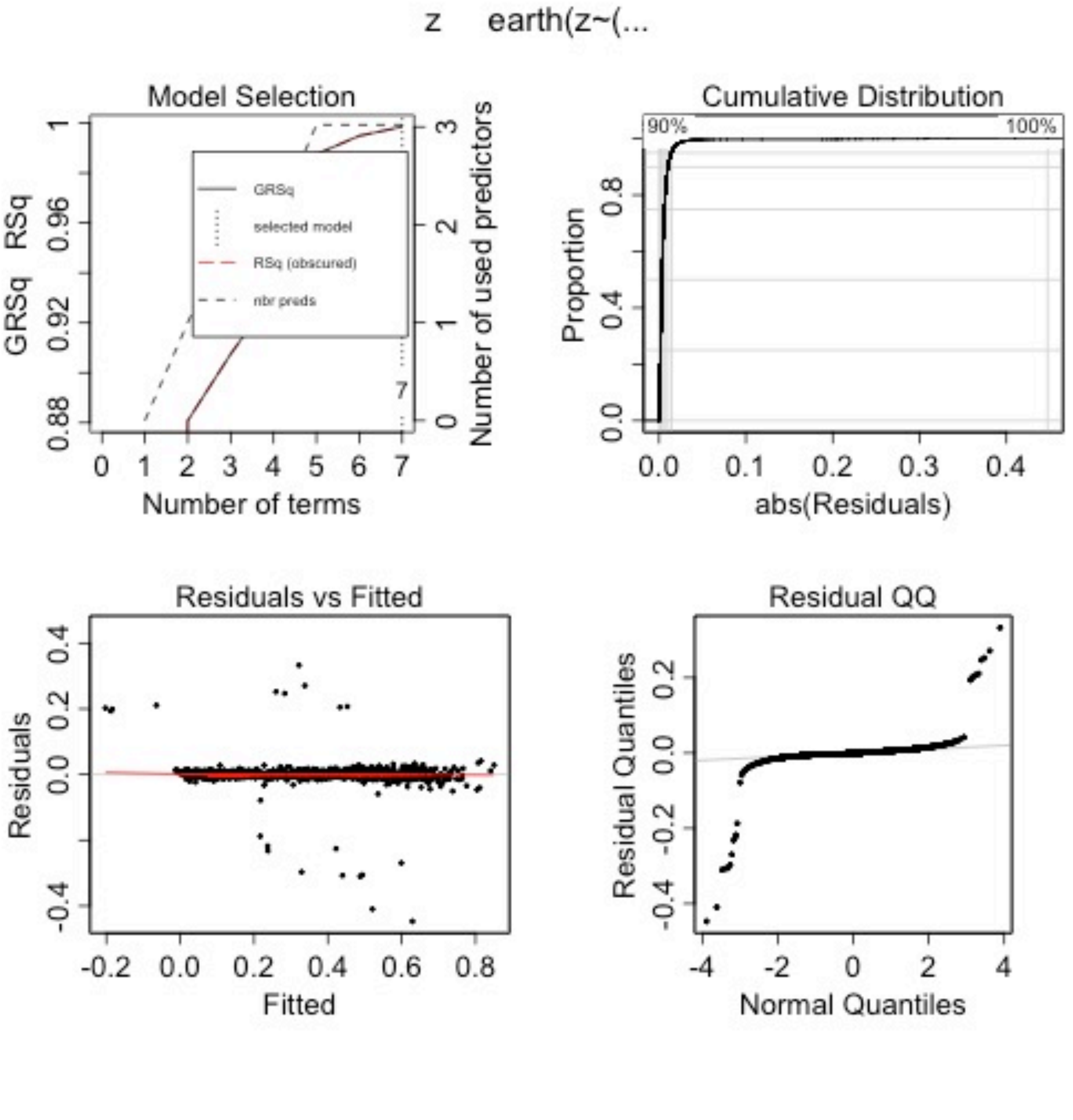
The next model built was the MARS regression model on the SDSS dataset. PCA and cross validation was preserved, however only the default parameters in the earth function were considered in order to test the flexibility of the MARS model. The following diagnostics are produced:

QQ Plot (MARS Predicted Values vs Observed Values on SDSS)



The MARS model yielded an R Square of 0.9984716. The correlation metric yields a value of .9671756 between the trained model and observed values. The MARS model was built in 8.868595 seconds. The time to compute was longer than the gamma model, however the MARS regression model was able to reflect more marginal changes amongst the photometric band missed by the gamma glm. The standard deviation of the residuals is 0.007366022, which is quite small, in other words there is little deviation from the default normal parameter in the earth function. The coefficients of the MARS model are in the form of basis polynomials and tell us the subtle change that can affect the magnitude of redshift for interaction within the different photometric bands. The MARS model looks promising hence it is a good reason to look at additional plots, displayed below:

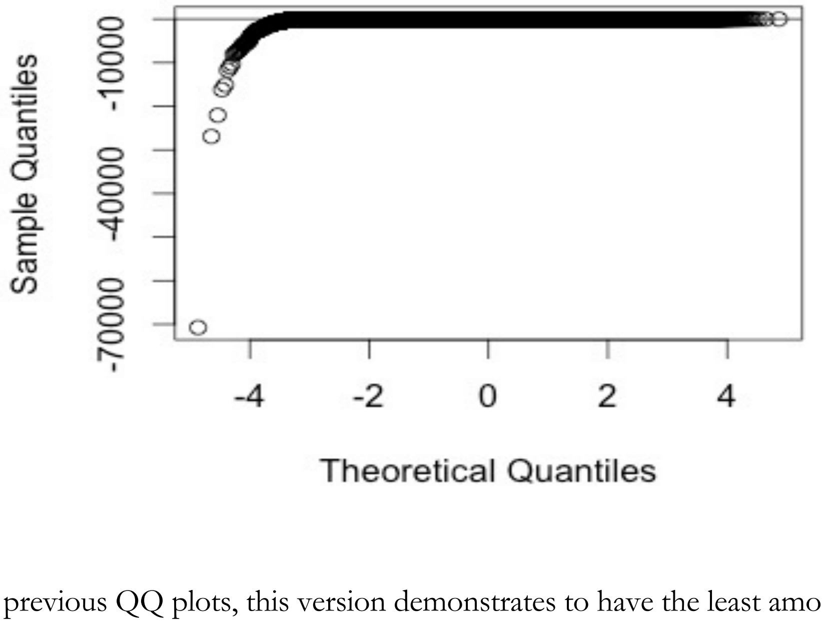
Additional Diagnostic Plots for MARS in the earth function



The top left corner is the model selection graph. Shown in the graph are R squared lines, and GR squared lines. Although partially obscured by the legend, there is convergence at 7 terms with 3 predictors. At 7 terms, the R squared value goes up with a penalty on the GCV however the selected model converges. The bottom left graph is the residuals vs. fitted graph. The red line represents a lowess fit and in our case, there is very little divergence and constant variance is shown due to the symmetry around the lowess fit line but this notion is not as important as with a linear model. The cumulative distribution graph on the top right shows the cumulative distribution of the residuals using their absolute value. In our case, the graph starts off at 0 and shoots off to one rather quickly, which is exactly what we want to see in this type of graph. The graph on the lower right corner is a smaller version of the QQ plot that was already discussed.

The final model analyzed in this case study is the gamma additive model built by back fitting using a log link function. The R squared value comes out to .701 with just three PC's. Adding additional PC's into the model improves the adjusted R squared at the cost of increasing the complexity of the model. The correlation metric comes out to 0.9319969. The diagnostic plots are as follows:

QQ Plot (Gamma Additive Predicted Values vs. Observed Values on SDSS)



Unlike the previous QQ plots, this version demonstrates to have the least amount of discrepancy when it comes to the predicted vs. observed values. There are some irregularities on the left tail end but they are not as high in concentration. The standard deviation between fitted model and distribution is much higher than the other models coming in at 95.83605. This model was built in 11.32 seconds. It should be noted that this model does not use the “lo” or the “s” smoothers on the predictors. Using such smoothers led to a plethora of rank incompatibility errors with the code in R.

Summary

The purpose of this case study was to build and compare different types of regression models and make a sound judgment on which model would be the most efficient and accurate to use for estimating photometric redshift. We utilized two separate datasets, one from the "CosmoPhotoz" r package in the CRAN repository and the other from the Sloan Digital Satellite Survey. In a way, the simulated dataset was a warm-up for training the gamma glm before using it on the actual SDSS data. An efficient outlier detection test informed us that the SDSS dataset contained "extreme" values much larger than the other observations. These values could have thrown off the effectiveness of the models; hence they were removed from the data frame. We proceeded forward by considering the absolute value of magnitudes, which did not change the shape of the distribution for the redshift in the SDSS data. By doing this, the positive and continuous structure of the data frame was preserved and optimized for analysis.

The first model built was a gamma regression model on a simulated dataset from a gamma distribution. The Cullen and Frey plot indeed confirmed that the redshift was most closely associated with a gamma distribution. Before building the gamma regression model, PCA decomposition was performed and yielded a minimum of 6 PC's to retain the desirable variance threshold. This was done in order to remedy multi co-linearity. There was quantifiable evidence of such phenomenon shown in the VIF and correlation matrix. The PC scores took the role of "new" coefficients. The regression model was built using the factorial form of PC's with the first two PC's returning an orthogonal polynomial of degree 2. Interactions amongst predictors had to be considered. The gamma model was built in 0.7024269 seconds and the correlation between the predicted and observed values was 0.9137729. Diagnostics revealed some irregularities on the right and tail ends.

The second model built was the gamma model with a log link function on the treated SDSS dataset. PCA decomposition revealed that only three PC's were necessary to preserve the desired variance threshold. We also employed an additional k fold cross validation technique on PC's in order to improve the chances of optimizing the results. The model was built using the factorial form for interactions in addition to returning degree 2 orthogonal polynomials for the first 2 PC's. This model was built in 1.149134 seconds. The standard deviation between the fit and the gamma distribution was 0.07217227, which is quite low. The correlation between the predicted vs. observed values was .9525174. Diagnostics revealed a QQ plot similar to the one found from the simulated dataset hinting that the same irregularities might exist in this data frame as well.

The next group of models comes from the additive family of regression models. The first of these models to be used was the MARS regression model with the default parameters in the earth function. The MARS model was built in 8.868595 seconds. The standard deviation between the fit and the default Gaussian distribution included in the earth function was 0.007366022, also quite low. It was previously stated that we could disregard any relation with the normal distribution but due to the flexibility and freedom from assumptions, the default distribution was a fair choice. The MARS model had an R square at .9984717 and the correlation metric between the predicted vs. observed values came in at .9671756. This model had all the indications of being a good model to estimate photometric redshift and the coefficients reflected marginal changes in photometric not found in the gamma glm. Extended diagnostics revealed favorable plots that capture the relationship in the data frame well.

The final model in this case study is the additive gamma regression model with a log link built using the back fitting algorithm included in the "mgcv" package. This model took 11.32 seconds to compute. The standard deviation between the fit and the gamma distribution came in at a whopping 95.83605. This is a much larger deviation than with the previous models. The correlation metric between predicted and observed values was .9405625 and was accompanied by an adjusted R square of .701. Diagnostic plot revealed the least amount of discrepancy between the predicted and observed values. It should be noted that this model was built without using the "s" smoother or the "lo" smoother on predictors. It was still built using 3 PC's and cross-validated subset.

Conclusion

Based on our findings, the most efficient model of the four used, was the MARS regression model. The gamma models with log link function were not necessarily bad; however there may be some ways to optimize their usage for this type of data. Transformations on the response variables could prove to be a good way of improving the model. The right kind of transformation can "center" the distribution of the response variable thus making the gamma glm much more effective.

Additive models in general proved to be more efficient in reflecting marginal changes in photometry. Given that the MARS model was more effective than the gamma additive model, there are some improvements that can be done on the additive gamma model. As mentioned before, a transformation of the response variable could be a very viable and effective way to better capture the relationships found in the data frame. It is also recommended to expand the additive model to include the spline smoother or the loess smoother. This however can be a difficult task due to the sheer amount of rank incompatibility errors that arise when I tried it. This would be better left to someone more adept in debugging problems within the R framework, which is out of the scope of this case study. It is also recommended to try different base distributions for redshift, for example the Weibull distribution, which is not far from a gamma.

Bibliography

Faraway, Julian James. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton: Chapman & Hall/CRC, 2006. Print.

Feigelson, Eric D., and Gutti Jogesh Babu. *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge: Cambridge UP, 2012. Print.

Hilbe, Joseph M. *Astrostatistical Challenges for the New Astronomy*. New York: Springer, 2012. Print.

Hogg, Robert V., and Allen T. Craig. *Introduction to Mathematical Statistics*. New York: Macmillan, 1978. Print.

"Sky Server SDSS Redshift Data." N.p., n.d. Web. <public.mikhailway.ashok2015i>.

"R Cran Repository." N.p., n.d. Web. <https://cran.r-project.org/web/packages/CosmoPhotoz/CosmoPhotoz.pdf>.

Elliot, J., R. S. De Souza, E. Cameron, E.E.O Ishida, and J. Hilbe. "The Overlooked Potential of Generalized Linear Models in Astronomy-II: Gamma Regression and Photometric Redshifts." (2015): n. pag. Web.

Appendix: R Code

```
mydata<-subset(Sky_server, select=c(dered_u,dered_g,dered_z,dered_i,dered_r,z))
summary(mydata)

#make positive and continous
mydata1<-abs(mydata)
summary(mydata1)

#remove missing values
mydataclean <- na.omit(mydata1)
summary(mydataclean)
View(mydataclean)

#outliers
library(outliers)
chisq.out.test(mydataclean$dered_u, variance=var(mydataclean$dered_u), opposite =
FALSE)
chisq.out.test(mydataclean$dered_g, variance=var(mydataclean$dered_g), opposite =
FALSE)
chisq.out.test(mydataclean$dered_z, variance=var(mydataclean$dered_z), opposite =
FALSE)
chisq.out.test(mydataclean$dered_i, variance=var(mydataclean$dered_i), opposite = FALSE)
chisq.out.test(mydataclean$z, variance=var(mydataclean$z), opposite = FALSE)

#outlier test to the whole dataframe
outlier(mydataclean, opposite = FALSE, logical = FALSE)

#remove extreme value
# install.packages('data.table') may need to be run if you don't have the
# package
library(data.table)
outlierReplace = function(dataframe, cols, rows, newValue = NA) {
  if (any(rows)) {
    set(dataframe, rows, cols, newValue)
  }
}
outlierReplace(mydataclean, "dered_u",  which(mydataclean$dered_u >  1000),NA)
outlierReplace(mydataclean, "dered_i",  which(mydataclean$dered_i >  1000),NA)
outlierReplace(mydataclean, "dered_g",  which(mydataclean$dered_g >  1000),NA)
outlierReplace(mydataclean, "dered_r",  which(mydataclean$dered_r >  1000),NA)
outlierReplace(mydataclean, "dered_z",  which(mydataclean$dered_z >  1000),NA)
mydataclean<-na.omit(mydataclean)
View(mydataclean)
summary(mydataclean)

#VIF
library(usdm)
```

```
dd = mydataclean
vif(dd)

#distribution check
dist_z<-mydataclean$z
hist(dist_z)

#correlation matrix
#correlation
round(cor(mydataclean, use="pairwise.complete.obs"),2)
#shows high correlation

#faraway PCA
pca_1<-princomp(mydataclean,select=c(-z), cor=T)
pca_data<-data.frame(z=mydataclean$z, pca_1$scores)
summary(pca_1)
loadings(pca_1)
plot(pca_1,type="lines")

#10 fold cross validation
#k-fold cross validation
#Create 10 equally size folds
folds <- cut(seq(1,nrow(pca_data)),breaks=10,labels=FALSE)

#Perform 10 fold cross validation
for(i in 1:10){
  #Segement your data by fold using the which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)
  testDataPC <- pca_data[testIndexes, ]
  trainDataPC <- pca_data[-testIndexes, ]
  #Use the test and train data partitions however you desire...
}
summary(testDataPC)

#apply Gamma
#apply a GLM model Time to compute (1.149134 seconds)
start.time <- Sys.time()
glm_pc<-glm(z~poly(Comp.1,2)+poly(Comp.2,2)*Comp.3, family=Gamma(link="log"),
data=testDataPC,maxit = 200)
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken

summary(glm_pc)

#model performance .9525174
cor(predict(glm_pc,testDataPC),testDataPC$z)
```

```
#diagnostics
#residuals on gamma model
#check residuals on the PCA gamma model
plot(predict(glm_pc),residuals(glm_pc),xlab="Fitted", ylab="Residuals")
qqnorm(residuals(glm_pc))

#square root error (residuals plot)
#square root test for GLM on PCA
plot(fitted(glm_pc), sqrt(abs(residuals(glm_pc))), xlab="fitted",
ylab=expression(sqrt(hat(epsilon))))
summary(lm(sqrt(abs(residuals(glm_pc)))~fitted(glm_pc)))

#QQ line
qqline(glm_pc$residuals)

#QQ plot
sd(glm_pc$residuals)
qqnorm(glm_pc$residuals)
qqline(glm_pc$residuals)
qqplot(glm_pc$residuals)

#MARS regression
#Lets consider MARS non linear regression
install.packages('earth')
library(earth)

# 8.868595 seconds
start.time <- Sys.time()
MARSfit<-earth(z~(poly(Comp.1, 2))^2 + poly(Comp.2, 2)*Comp.3, data=trainDataPC)
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken
#R square at .9984717
summary(MARSfit)
print(MARSfit)

#corrections
#diagnostics on additive model
sd(MARSfit$residuals)
qqnorm(MARSfit$residuals)
qqline(MARSfit$residuals)

#model performance (.9671756)
cor(predict(MARSfit,testDataPC),testDataPC$z)


predict(MARSfit,cleantest$z)
```

```
names(MARSfit)

plot(MARSfit, se=TRUE)

#back fitting algorithm 11.32 seconds
library(mgcv)
start.time <- Sys.time()
#Adj R square of .701
back_hm<-gam(z~(poly(Comp.1, 2))^2 + poly(Comp.2,
2)*Comp.3*Comp.4,family=Gamma(link="log"), data=trainDataPC)
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken
summary(back_hm)

#model performance .9405625
cor(predict(back_hm,testDataPC),testDataPC$z)

#diagnostics
sd(back_hm$residuals)
qqnorm(back_hm$residuals)
qqline(back_hm$residuals)

plot(back_hm,residuals=TRUE , se=TRUE, pch=".")


#now with spline smoother 4.36829 minutes
start.time <- Sys.time()
back_hm_2<-gam(z~s(poly(Comp.1,2)^2
+poly(Comp.2,2)*Comp.3),family=Gamma(link="log"), data=trainDataPC)
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken
#r squared .877
summary(back_hm_2)

#plot
plot(back_hm_2, se=TRUE)

#model performance .9405625
cor(predict(back_hm_2,testDataPC),testDataPC$z)

#new subset without PCA
#10 fold cross validation
#k-fold cross validation
#Create 10 equally size folds
folds <- cut(seq(1,nrow(pca_data)),breaks=10,labels=FALSE)
```

```
#Perform 10 fold cross validation
for(i in 1:10){
  #Segement your data by fold using the which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)
  testData <- mydataclean[testIndexes, ]
  trainData <- mydataclean[-testIndexes, ]
  #Use the test and train data partitions however you desire...
}
testData<-na.omit(testData)
trainData<-na.omit(trainData)
summary(trainData)

#additive model without PCA
View(trainData)
start.time <- Sys.time()
new_hm<-
gam(log(z+1)~s(dered_u*dered_i*dered_g*dered_r*dered_z),family=Gamma(link="log"),
data=trainData)
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken
#r squared Adj. .805
summary(new_hm)

#model performance .5828884
cor(predict(new_hm,testData),testData$z)

#plot
plot(new_hm, se=TRUE)

#diagnostics
sd(new_hm$residuals)
qqnorm(new_hm$residuals)
qqline(new_hm$residuals)
```