Exoplanet Hunting: A Statistical Approach

Sagnik Roy - MB1901
Soham Bonnerjee -MB1907
Chandrachur Banerjee - MB1925
Soumen Ghosh - MB1940

March 28, 2020

M.Stat 1st Year
Indian Statistical Institute, Kolkata

Content

Introduction
Exoplanets

Simply put, an Exoplanet or Extrasolar planet is a planet outside the Solar System.The first exoplanet was detected in 1917, using ground-based telescopes.

Introduction
Exoplanets

Simply put, an Exoplanet or Extrasolar planet is a planet outside the Solar System.The first exoplanet was detected in 1917, using ground-based telescopes.

Exoplanet hunting in deep space is a very important undertaking in our quest to find new earth. To this end, NASA launched the satellite ***The Kepler Space Observatory*** in 2009. The satellite is dedicated to searching for exoplanets in star systems besides our own, with the ultimate goal of possibly finding other habitable planets besides our own.

Data Description

This dataset is a cumulative record of all observed Kepler "objects of interest" — basically, all of the approximately 10,000 exoplanet candidates Kepler has taken observations on. Among them, Kepler, using extensive and complex mathematical formulae, predicts 4496 to be *Candidate* for Exoplanet, and designate rest 5068 to be *False Positive*. We also have some other astronomical variables that will be descried shortly.

Data Description

This dataset is a cumulative record of all observed Kepler "objects of interest" — basically, all of the approximately 10,000 exoplanet candidates Kepler has taken observations on. Among them, Kepler, using extensive and complex mathematical formulae, predicts 4496 to be *Candidate* for Exoplanet, and designate rest 5068 to be *False Positive*. We also have some other astronomical variables that will be descried shortly.

Research Goal

To find a suitable and simple statistical prediction method for Kepler data.

Data Description
Variables

There are 17 variables in total, from which we describe the important few.

Data Description
Variables

There are 17 variables in total, from which we describe the important few.

- `koi_pdisposition` : The designation that Kepler gives to the exoplanet. It's a binary variable with levels `Candidate` and `False Positive`. This is our response variable.

Data Description
Variables

There are 17 variables in total, from which we describe the important few.

- `koi_pdisposition` : The designation that Kepler gives to the exoplanet. It's a binary variable with levels `Candidate` and `False Positive`. This is our response variable.
- `koi_score` : A continuous random variable taking values in $[0, 1]$. A designation of `Candidate` with high confidence , should indicate a high value of `koi_score`, and vice-versa. This is another response variable.

Data Description
Variables

There are 17 variables in total, from which we describe the important few.

- `koi_pdisposition` : The designation that Kepler gives to the exoplanet. It's a binary variable with levels `Candidate` and `False Positive`. This is our response variable.
- `koi_score` : A continuous random variable taking values in $[0, 1]$. A designation of `Candidate` with high confidence , should indicate a high value of `koi_score`, and vice-versa. This is another response variable.
- `koi_duration` : The duration of the observed transits, i.e in simple words, the time the planet takes to rotate around itself.

Data Description
Variables

There are 17 variables in total, from which we describe the important few.

- `koi_pdisposition` : The designation that Kepler gives to the exoplanet. It's a binary variable with levels `Candidate` and `False Positive`. This is our response variable.
- `koi_score` : A continuous random variable taking values in $[0, 1]$. A designation of `Candidate` with high confidence , should indicate a high value of `koi_score`, and vice-versa. This is another response variable.
- `koi_duration` : The duration of the observed transits, i.e in simple words, the time the planet takes to rotate around itself.
- `koi_impact` : The sky-projected distance between the center of the stellar disc and the center of the planet disc. This has been normalized by Stellar Radius.

Data Description
Variables

- `koi_prad` : The radius of the Exoplanet, in terms of Earth radii.
- `koi_teq` : Approximation for the temperature of the planet, in Kelvin(K).
- `koi_steff` : The photospheric temprature of the sun of the exoplanet, in Kelvin.
- `koi_slogg` : Log-transformed accelaration due to gravity on the planet surface.

Data Description
Variables

- `koi_prad` : The radius of the Exoplanet, in terms of Earth radii.
- `koi_teq` : Approximation for the temperature of the planet, in Kelvin(K).
- `koi_steff` : The photospheric temprature of the sun of the exoplanet, in Kelvin.
- `koi_slogg` : Log-transformed accelaration due to gravity on the planet surface.

The other variables are, `koi_time0bk` (Transit Epoch), `koi_insol` (Insulation Flux), `koi_model_snr` (Normalized Transit Depth), `koi_period`, `koi_depth`, `koi_srad` (Solar radius), `koi_ra`, `koi_dec`, `koi_kepmeg` (Kepler Band parameters). Note that our response variable is `koi_pdisposition`, and `koi_score`, and the rest are Predictor variables.

Variable Selection
Multicollinearity

We have some missing values in some columns. In order to impute them we first perform the variable selection method.

Variable Selection
Multicollinearity

We have some missing values in some columns. In order to impute them we first perform the variable selection method.

First we try to find if any Multicollinearity is present among the predictors. The *Variance Inflation Factor(VIF)* for the predictors is given by:

koi_period	koi_time0bk	koi_impact	koi_duration	koi_depth
1.010538	1.236553	1.854218	1.185540	1.566536
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
1.86551	2.225391	1.574242	1.567405	1.345416
koi_slogg	koi_srad	ra	dec	koi_kepmag
2.685013	2.369972	1.033121	1.006093	1.505617

Table: VIF of Predictors

- We can see as for no predictor *VIF*>10, we can safely assume that there is no multicollinearity between the predictor variables.

Variable Selection
Model Fitting

Next we take the complete datset available for all variables, and fit a logistic model on `koi_pdisposition`.

Next we take the complete datset available for all variables, and fit a logistic model on `koi_pdisposition`.

Our model is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{15} x_{ij}\beta_j$$

where, π_i is the probability that the disposition is `False Positive`, and x_{ij} is the value of j-th predictor for i-th exoplanet.

Next we take the complete datset available for all variables, and fit a logistic model on `koi_pdisposition`.

Our model is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{15} x_{ij}\beta_j$$

where, π_i is the probability that the disposition is `False Positive`, and x_{ij} is the value of j-th predictor for i-th exoplanet.

We perform Backward Stepwise Regression to find the model giving minimum AIC.

Variable Selection
Stepwise Regression

The initial model has coefficients:

koi_period	koi_time0bk	koi_impact	koi_duration	koi_depth
-1.577910e-06	5.273012e-03	4.489425e-01	1.179662e-01	5.447226e-05
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
-2.160622e-04	1.396988e-03	-1.817013e-06	-2.473957e-04	8.784349e-05
koi_slogg	koi_srad	ra	dec	koi_kepmag
3.452714e-01	1.091986e-02	4.437354e-02	-4.269540e-02	5.898983e-02

Table: **Coefficient of initial model**

and the backward selection procedure chooses the model with coefficients:

koi_period	koi_time0bk	koi_impact	koi_duration	koi_depth
–	5.276350e-03	4.485437e-01	1.177286e-01	5.450274e-05
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
-2.153918e-04	1.396963e-03	-1.699252e-06	-2.485537e-04	7.884044e-05
koi_slogg	koi_srad	ra	dec	koi_kepmag
2.721519e-01	–	4.439881e-02	-4.274748e-02	6.205313e-02

Table: **Coefficient of Selected Model in stepwise regression**

Variable Selection
Stepwise Regression

The initial model has coefficients:

koi_period	koi_time0bk	koi_impact	koi_duration	koi_depth
-1.577910e-06	5.273012e-03	4.489425e-01	1.179662e-01	5.447226e-05
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
-2.160622e-04	1.396988e-03	-1.817013e-06	-2.473957e-04	8.784349e-05
koi_slogg	koi_srad	ra	dec	koi_kepmag
3.452714e-01	1.091986e-02	4.437354e-02	-4.269540e-02	5.898983e-02

Table: **Coefficient of initial model**

and the backward selection procedure chooses the model with coefficients:

koi_period	koi_time0bk	koi_impact	koi_duration	koi_depth
–	5.276350e-03	4.485437e-01	1.177286e-01	5.450274e-05
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
-2.153918e-04	1.396963e-03	-1.699252e-06	-2.485537e-04	7.884044e-05
koi_slogg	koi_srad	ra	dec	koi_kepmag
2.721519e-01	–	4.439881e-02	-4.274748e-02	6.205313e-02

Table: **Coefficient of Selected Model in stepwise regression**

As we can see from the above two tables, the stepwise regression procedure has removed the variables `koi_period` and `koi_srad` from the model. Hence we proceed with imputation of predictor values in rest of the predictor variables.

Imputation of Missing Values

For each of the Predictor Variables with missing values, we estimate a density based on the complete data available on that variable using a **Standard Normal** Kernel. Some of the estimated densities are as follows:

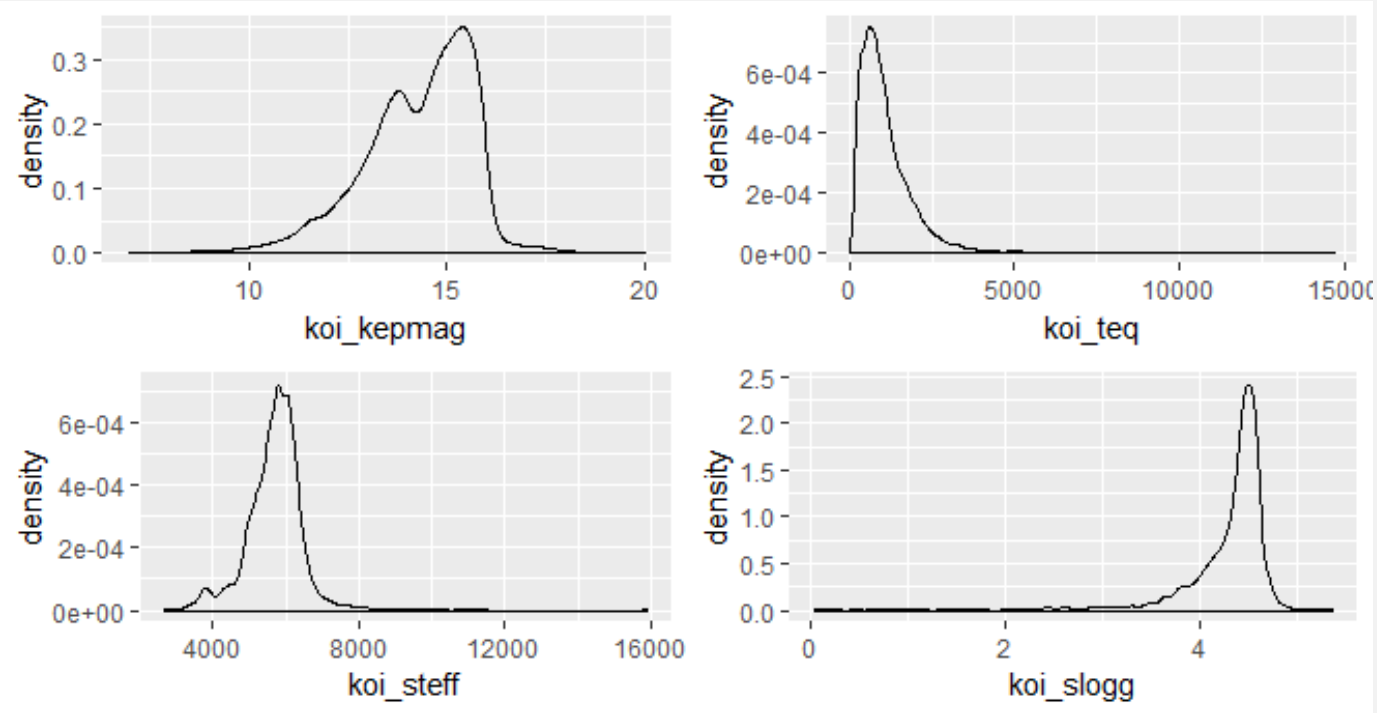


Figure: Estimated Densities

Imputation of Missing Values

We generate random observations from these densities multiple times, take their average, and impute them in place of the missing values, thus completing the imputation.

Fitting Logistic Model

Now we fit the logistic model to the imputed data, to predict `koi_pdisposition`. We divide the data randomly in 3 parts, fit the logistic model in the first two parts, and test it in the remaining part. This was done 5 times.

Fitting Logistic Model

Now we fit the logistic model to the imputed data, to predict `koi_pdisposition`. We divide the data randomly in 3 parts, fit the logistic model in the first two parts, and test it in the remaining part. This was done 5 times.

- The model has predicted the designation of `Candidate` or `False Positive` approximately 76% times.
- The model has predicted a designated `Candidate` exoplanet to be `Candidate` approximately 84% times.
- The model has predicted a designated `False Positive` exoplanet to be a `False Positive` one approximately 69% times.

Quantile Regresssion
Why should we do this?

- We have a curious observation that high values of `koi_impact` or `koi_prad` result in `koi_disposition` being "1", i.e `False Positive`. This is evident from the corresponding scatter plots.

Quantile Regresssion

Why should we do this?

- We have a curious observation that high values of `koi_impact` or `koi_prad` result in `koi_disposition` being "1", i.e `False Positive`. This is evident from the corresponding scatter plots.

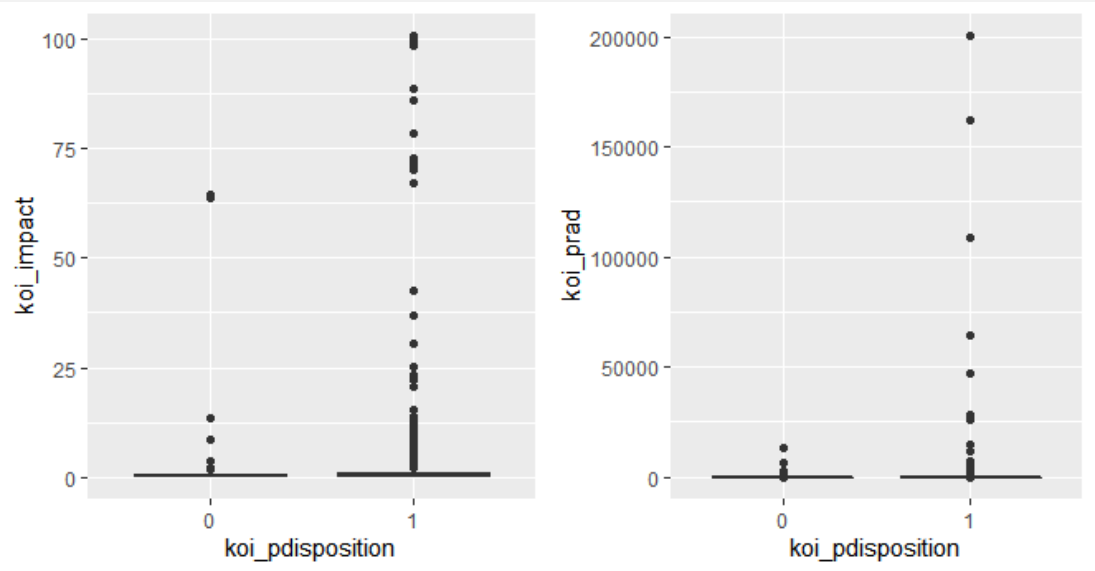


Figure: Box Plots

Quantile Regresssion
Why should we do this?

- Calculating *Cook's Distance* indicates presence of an outlier:

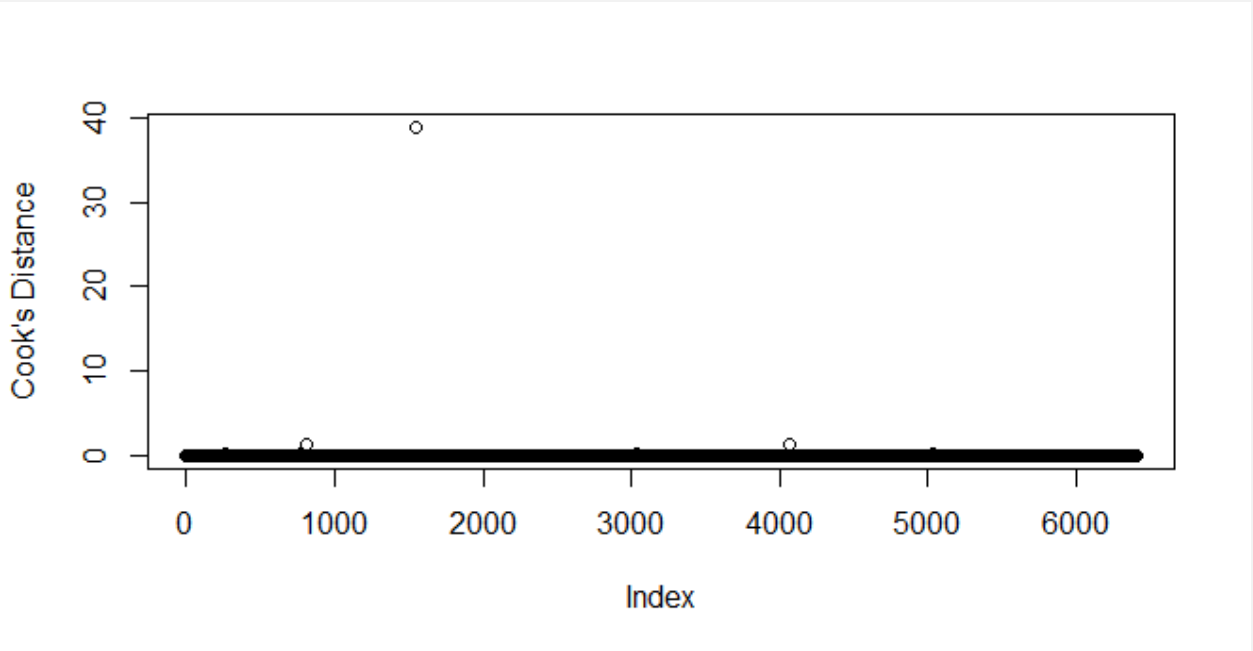


Figure: Cook's Distance Plot

Quantile Regresssion

Why should we do this?

- Thus to get a clearer picture, we should do a Quantile Regression.

Quantile Regresssion
Why should we do this?

- Thus to get a clearer picture, we should do a Quantile Regression.
- But since `koi_pdisposition` is a Binary Random Variable, we instead do Quantile Regression on `koi_score`.

Quantile Regresssion
Why should we do this?

- Thus to get a clearer picture, we should do a Quantile Regression.
- But since `koi_pdisposition` is a Binary Random Variable, we instead do Quantile Regression on `koi_score`.

Note that we can use the same variables for fitting this model, as the above two responses are inherently related.

Quantile Regression
Fitting

For $\tau = 0.5$, the model coefficients are as follows:

koi_time0bk	koi_impact	koi_duration	koi_depth	koi_kepmag
8.108637e-06	3.047438e-04	-9.276725e-04	-1.015740e-06	-3.115233e-05
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
-1.104849e-06	-3.245209e-07	-8.862506e-09	9.983946e-06	-1.335775e-07
koi_slogg	ra	dec		
1.241563e-03	-7.505649e-06	2.326195e-05		

Table: **Coefficient of Quantile Regression Model**

Quantile Regression
Fitting

For $\tau = 0.5$, the model coefficients are as follows:

koi_time0bk	koi_impact	koi_duration	koi_depth	koi_kepmag
8.108637e-06	3.047438e-04	-9.276725e-04	-1.015740e-06	-3.115233e-05
koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff
-1.104849e-06	-3.245209e-07	-8.862506e-09	9.983946e-06	-1.335775e-07
koi_slogg	ra	dec		
1.241563e-03	-7.505649e-06	2.326195e-05		

Table: **Coefficient of Quantile Regression Model**

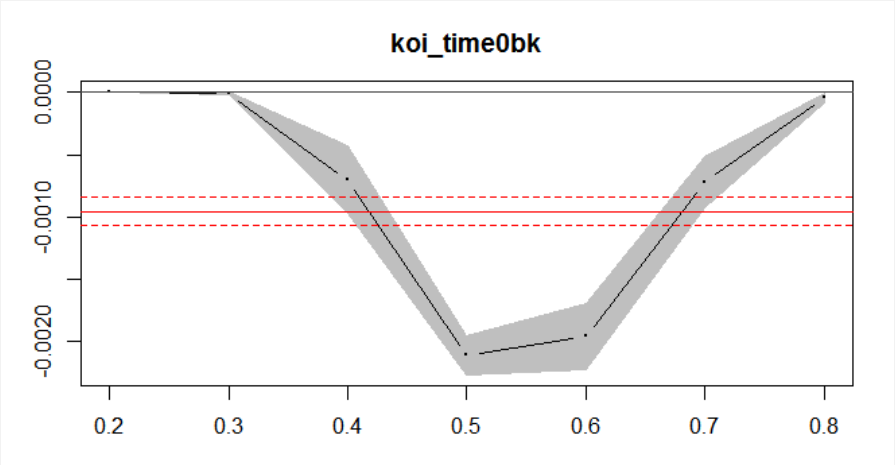
Similaly we do Quantile Regression for $\tau = 0.3$, and $\tau = 0.7$.

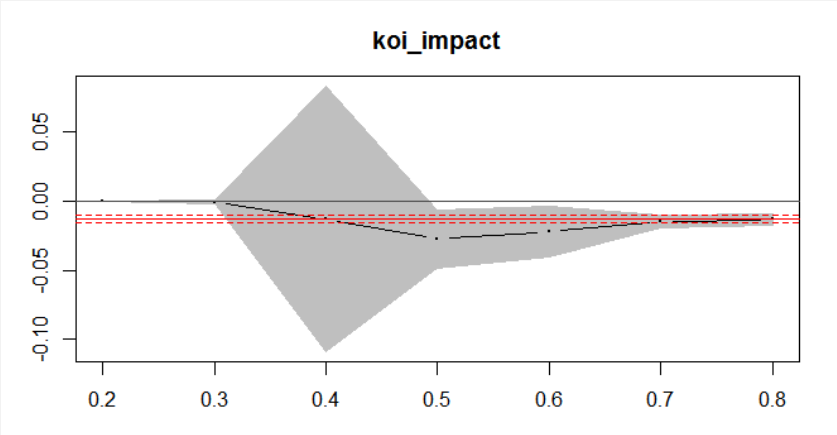
Quantile Regression

We do a test to see, if the regression coefficients are equal at $\tau = 0.3$, and $\tau = 0.7$, using the `anova` function, we observed that the null hypothesis that the coefficients are not different, is *rejected* at 5% level of significance. This justifies our Quantile Regression.
We can also visualize it in the following plots of regression coefficients vs tau:

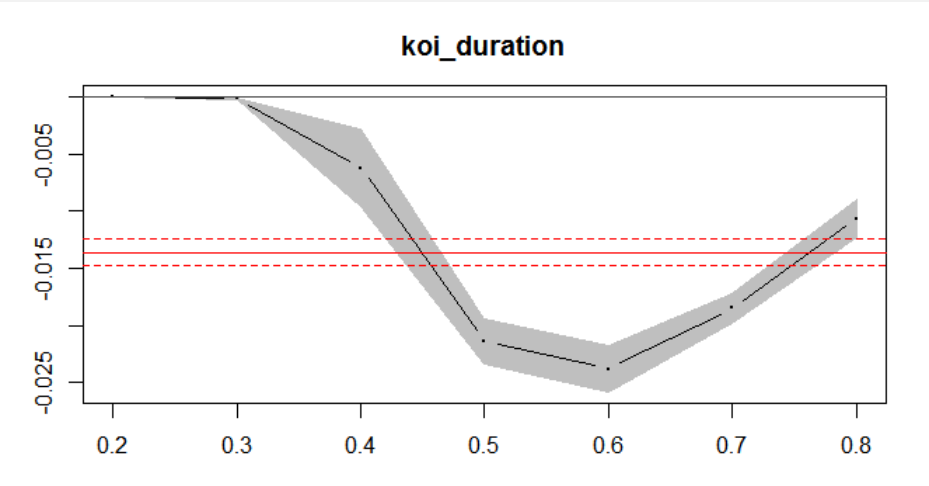
Quantile Regression

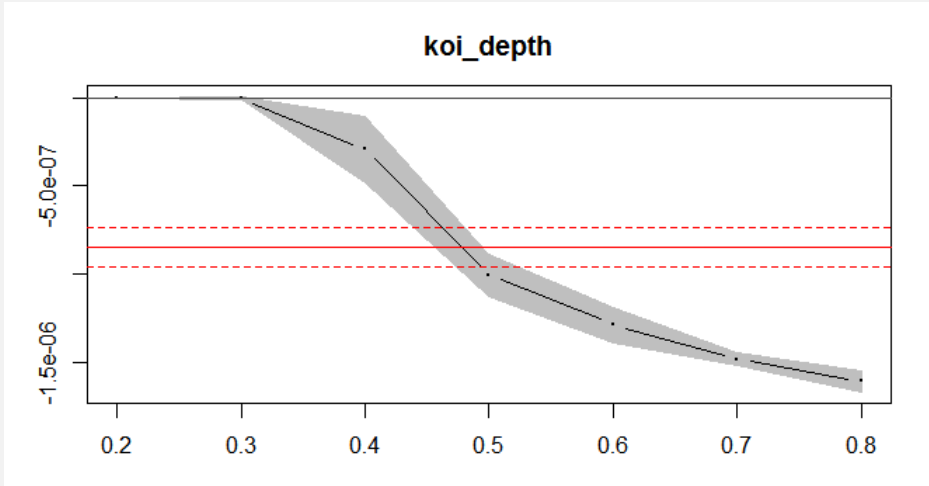
We do a test to see, if the regression coefficients are equal at $\tau = 0.3$, and $\tau = 0.7$, using the `anova` function, we observed that the null hypothesis that the coefficients are not different, is *rejected* at 5% level of significance. This justifies our Quantile Regression.
We can also visualize it in the following plots of regression coefficients vs tau:



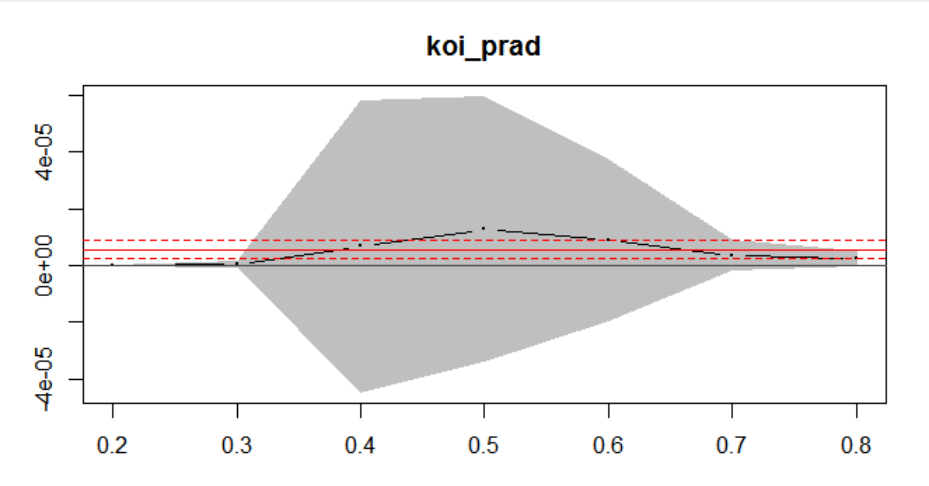


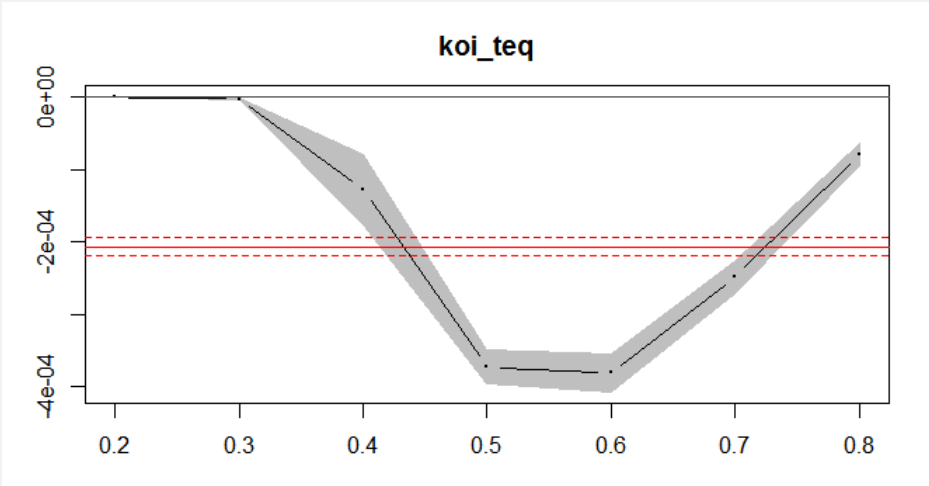
Quantile Regression



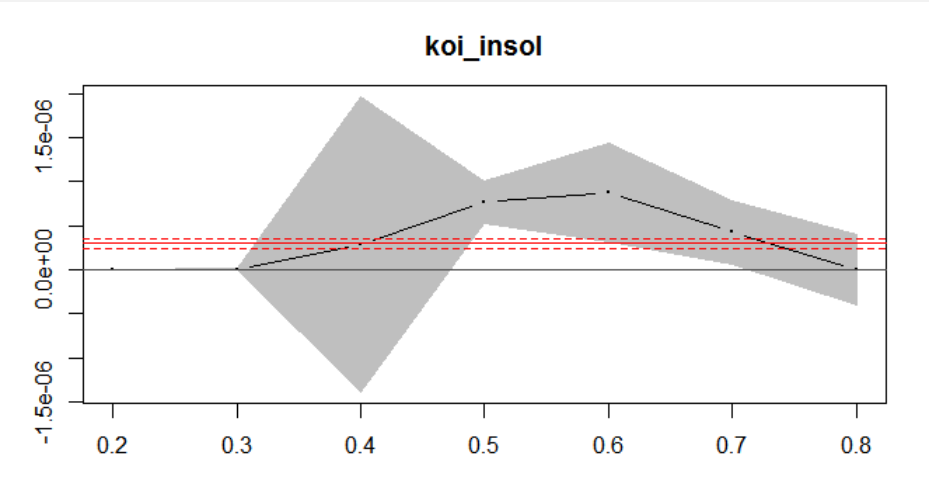


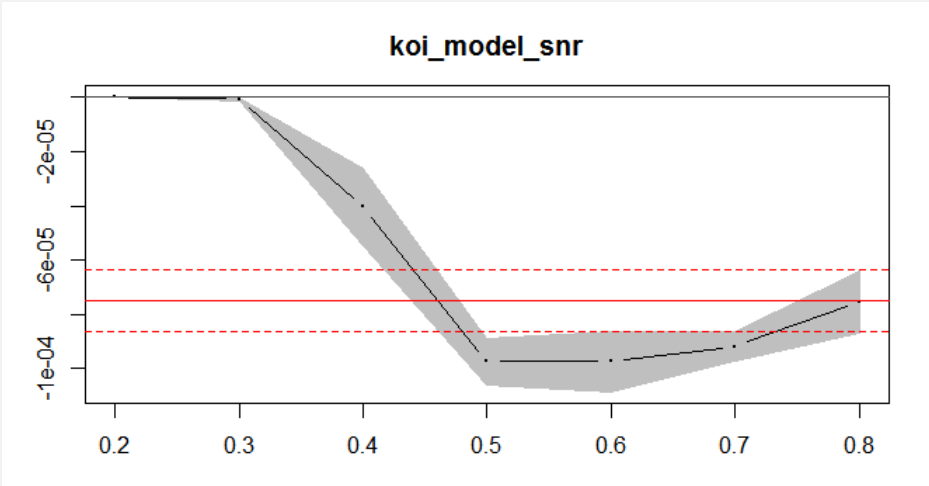
Quantile Regression



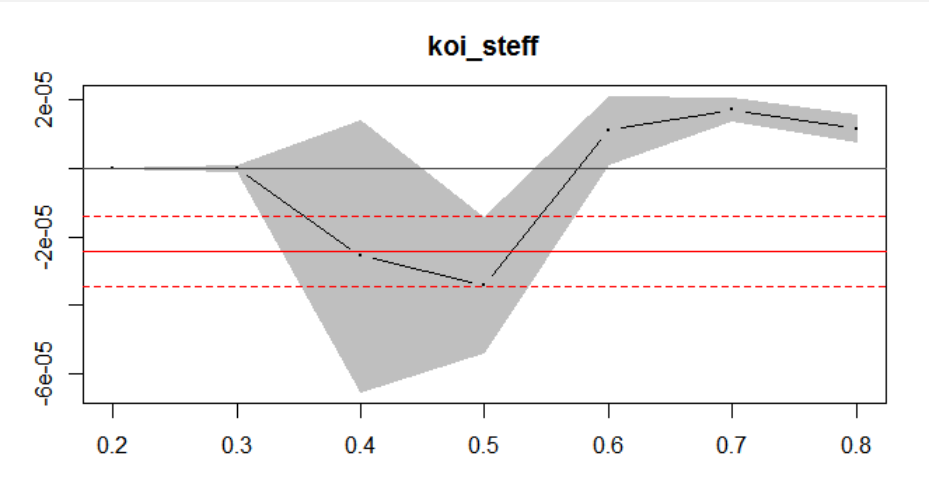


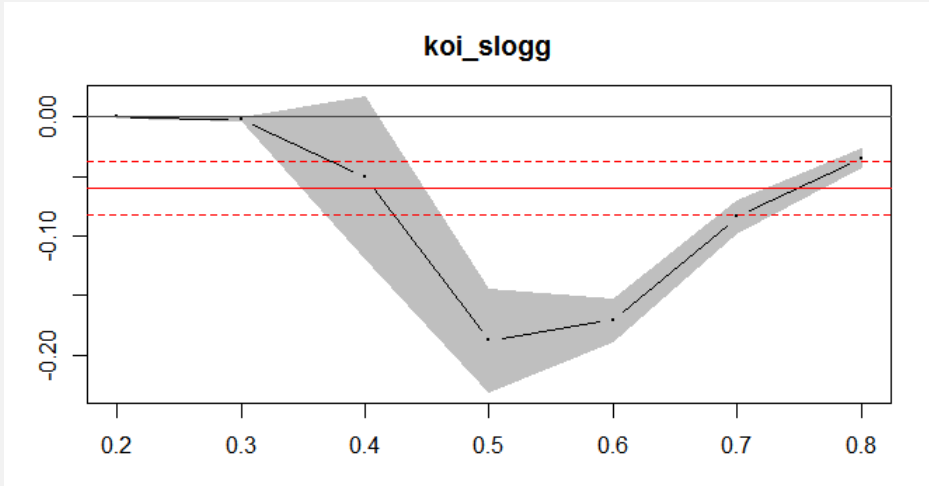
Quantile Regression



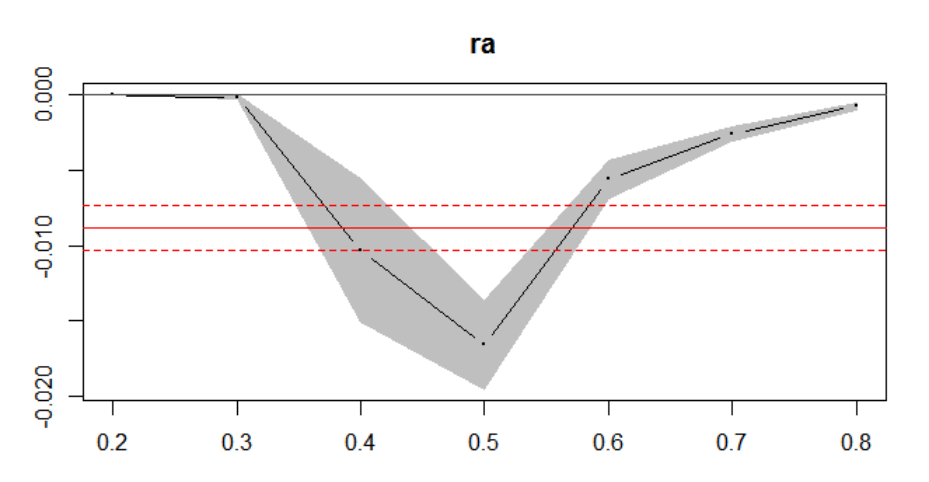


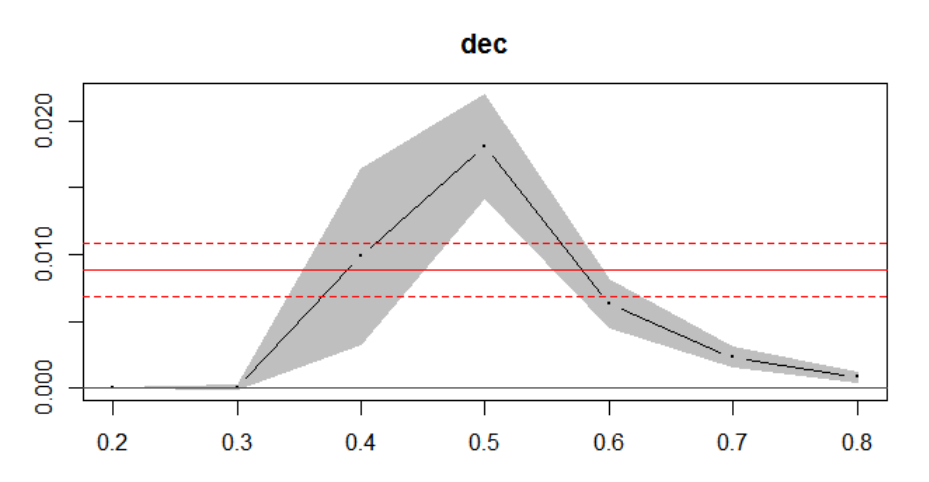
Quantile Regression

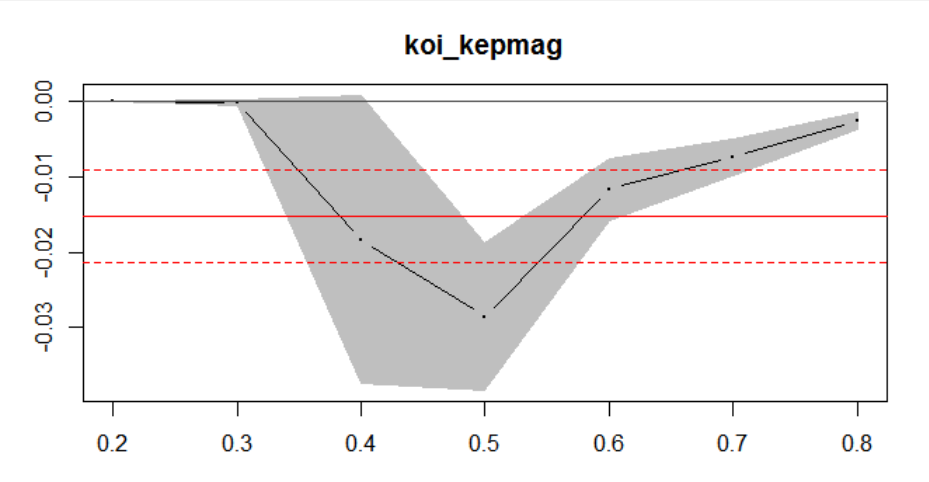




Quantile Regression







Quantile Regression

We can also compare with the linear regression line for predicting purposes:

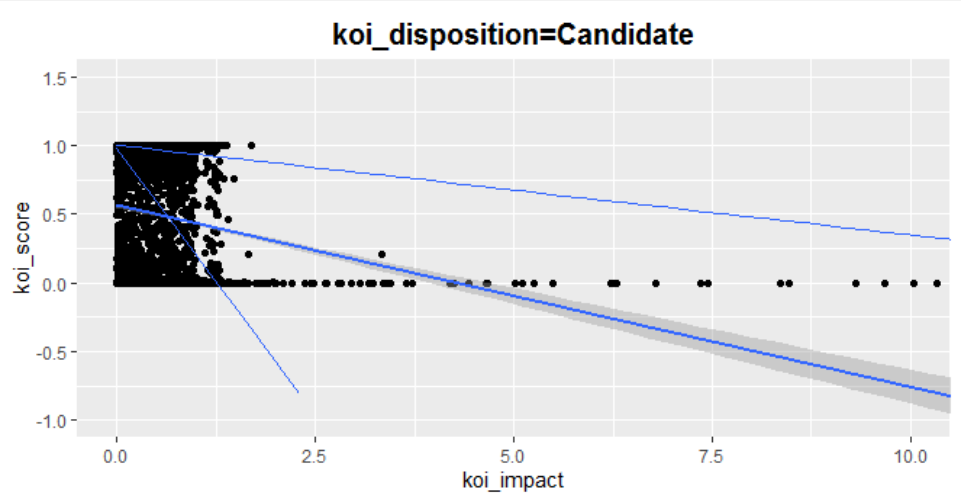


Figure: **Quantile Regression for score vs impact**

Quantile Regression

We can also compare with the linear regression line for predicting purposes:

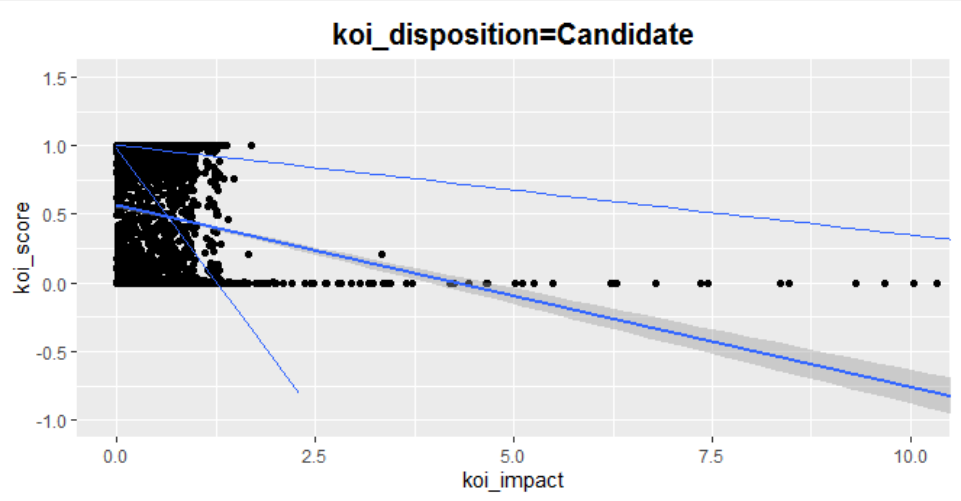


Figure: **Quantile Regression for score vs impact**

Note we should get high value of `koi_score` when `koi_pdisposition=0`, i.e `Candidate`, and a low value when `koi_pdisposition=1`, i.e `False Positive`.

Quantile Regression

We can also compare with the linear regression line for predicting purposes:

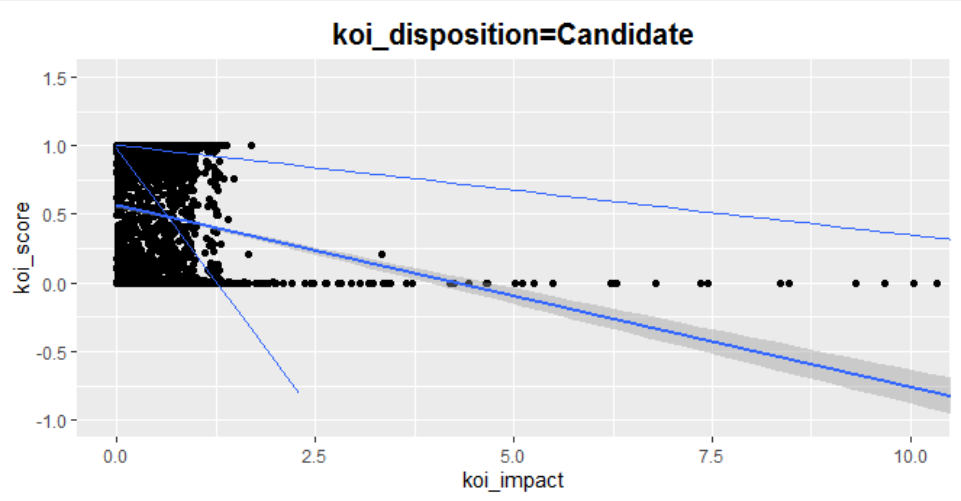


Figure: **Quantile Regression for score vs impact**

Note we should get high value of `koi_score` when `koi_pdisposition=0`, i.e `Candidate`, and a low value when `koi_pdisposition=1`, i.e `False Positive`. Thus we can see the linear regression line is a bad predictor of `koi_score`.

Quantile Regression

We can also compare with the linear regression line for predicting purposes:

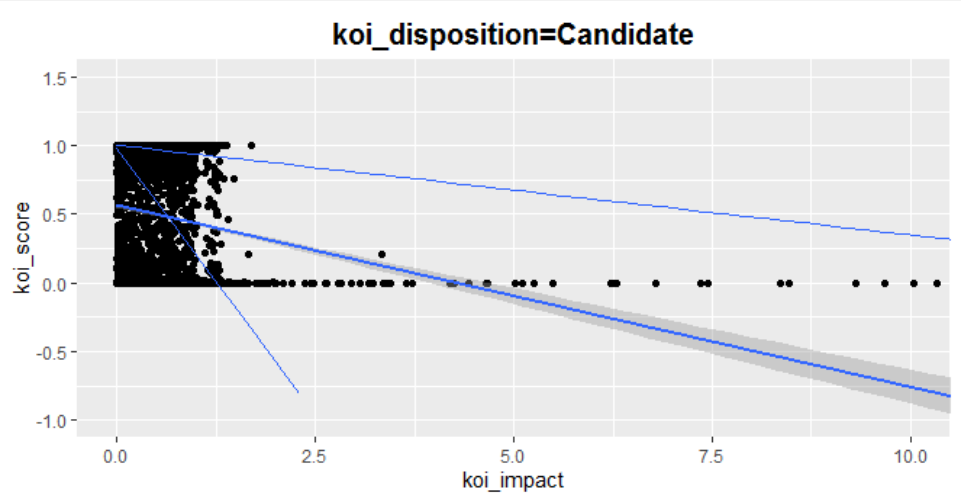


Figure: **Quantile Regression for score vs impact**

Note we should get high value of `koi_score` when `koi_pdisposition`=0, i.e `Candidate`, and a low value when `koi_pdisposition`=1, i.e `False Positive`. Thus we can see the linear regression line is a bad predictor of `koi_score`.

Linear regression model predicts the mean. Thus instead of getting either a high or a low value of `koi_score`, we get a prediction near 0.5, when `koi_impact` is near zero.

Quantile Regression

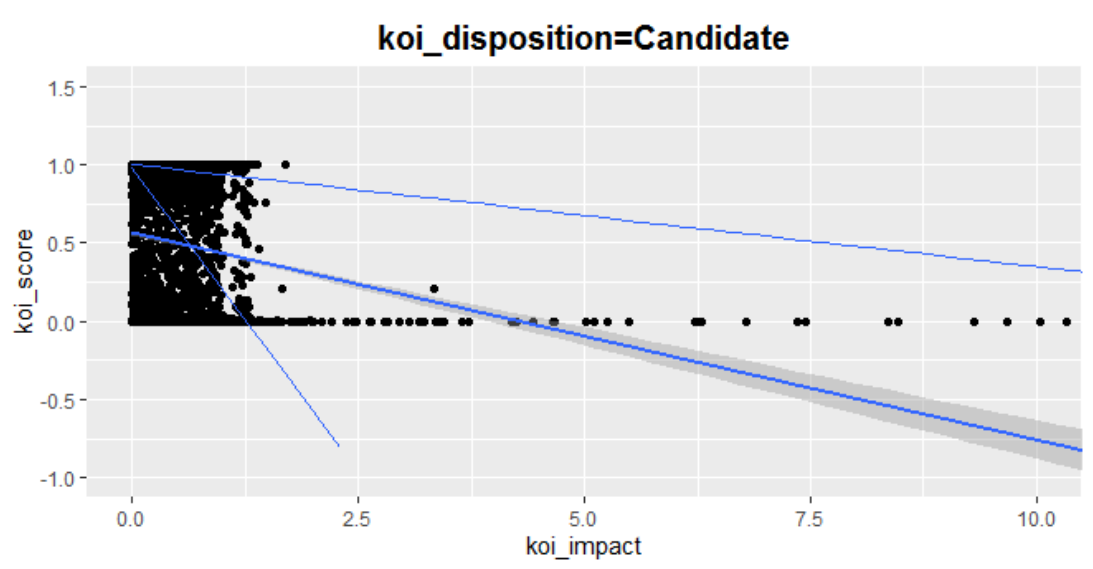


Figure: Quantile Regression for score vs impact

But the quantile regression lines for $\tau = 0.5$ and $\tau = 0.7$ predict `koi_score` to be near 1 when `koi_impact` is very low.

Quantile Regression

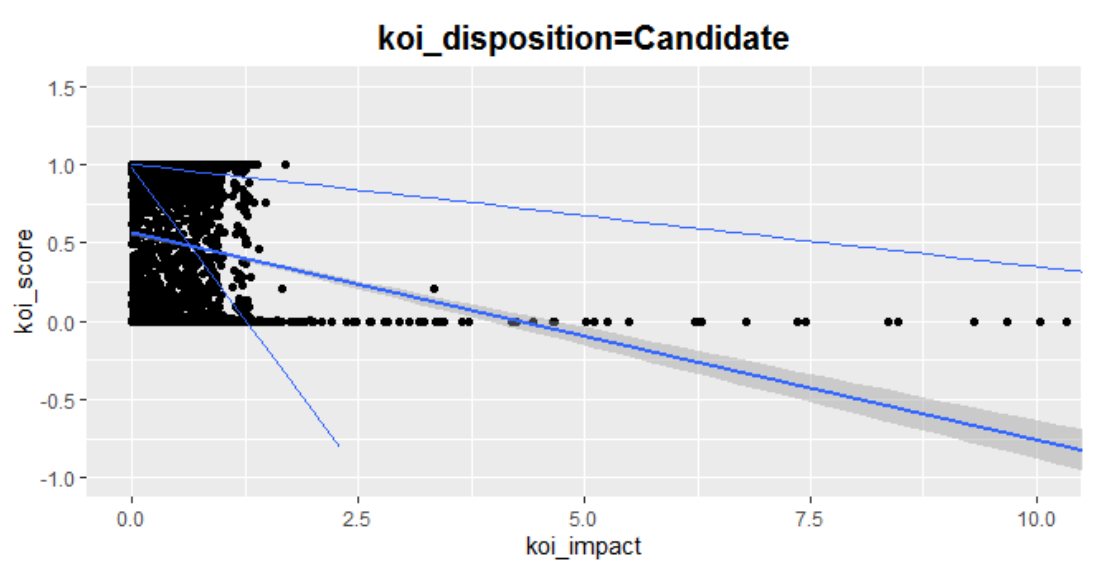


Figure: Quantile Regression for score vs impact

But the quantile regression lines for $\tau = 0.5$ and $\tau = 0.7$ predict `koi_score` to be near 1 when `koi_impact` is very low.
All the regression lines predict `koi_score` to be low when `koi_impact` is high.

Conclusion

- The variables related to an exoplanet or its corresponding sun, hence there was a chance that multicollinearity would be present. Thus firstly we checked multicollinearity, and then we proceeded to do variable selection using backward regression. We use backward regression because given the full model, it automatically tells me which variables are not important, thereby helping in imputation.
- Missing data/data with huge uncertainty is a frequent phenomena in astro-statistics data, thus we have done imputation using *Kernel Density Estimation*.
- Without using complex mathematical formulae, using the logistic model we fitted, we can predict the designation by which Kepler would denote an exoplanet with more than 76% precision on the average. This also validates our imputation procedure.
- Finally we used quantile regression to estimate `koi_score`, and further explained why linear model won't have worked here.

Further Investigations

- Quantile regression could not be done for categorical response `koi_disposition`. It would have been good to see how more precision we would achieve if we could do that.
- In general, probit model are somewhat better than the logit models. But in this case, the probit model gave only 59% precision on the average. It'd be good to further investigate how to model `koi_pdisposition` so as to further increase the precision.

THANK YOU!