Testing and Linear Regression

Siddartha Devic, Dipnil Chakraborty
Elements of Statistical Learning (Friedman et. al)

December 20, 2019

Contents

1 Preliminaries

The population set contains all possible values generated by a set of a parameters. Conversely, the "sampled" set may only be a subset of the population. In statistics, we are mainly dealing with sample sets, and base all estimation on their properties.

Given some probability distribution, we can calculate it's mean and variance as:

$$\mu = \sum_{i=1}^{n} p_i x_i$$
$$Var(x) = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

, where $p_i = f(x_i)$. In fact, these are related to the first and second moment, where the k^{th} moment is given by $\mathbb{E}(x^k)$:

$$\mu = \mathbb{E}(x)$$
$$Var(x) = \mathbb{E}(x^2) - (\mathbb{E}(x))^2$$

. Given a sample S, we can calculate some sample statistics from it:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

, Where X_i is the ith sampled observation, n is the sample size, $\bar{X}$ is the sample mean, and $Var(X)$ is the sample variance. We will come to see that this method of calculating variance is known as a "biased estimator".

1.1 Estimation

We are given a sample from a population. There are three general methods for determining parameters which characterize the underlying distribution: moment, unbiased, and maximum likelihood methods. The moment method approach will not be covered.

1.2 Unbiased Estimation

If $\mathbb{E}[u(X_1, X_2, ..., X_n)] = \theta$ for the underlying parameter *theta*, then u is said to be an unbiased estimator of θ. Otherwise, u is biased. Bias can be measured by seeing how well our estimated parameter approximates the true one (if known): $Bias(\theta) = \theta - \mathbb{E}_\theta(T)$.

Theorem 1.1. *$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an unbiased estimator of $\theta = \mu$, the mean.*

Proof. To show that $\bar{X}$ is an unbiased estimator, we must prove that in expectation, our sample mean will converge to our true mean, that is: $\mathbb{E}(\bar{X}) = \mathbb{E}(X)$.

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu \\ &= \frac{n\mu}{n} = \mu = \mathbb{E}(X).\end{aligned}$$

□

Theorem 1.2. $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *is an unbiased estimator of variance.*

Proof.

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\ &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i) - (\bar{X} - \mathbb{E}(X_i)))^2\right] \\ &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}((X_i - \mathbb{E}(X_i))^2 - 2(X_i - \mathbb{E}(X_i))(\bar{X} - \mathbb{E}(X_i)) + (\bar{X} - \mathbb{E}(X_i))^2)\right] \\ &= \mathbb{E}\left[\frac{1}{n-1}(\sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\bar{X} - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2)\right]\end{aligned}$$

Since the last term $\sum_{i=1}^{n}(\bar{X} - \mu)^2 = (\bar{X} - \mu)^2\sum_{i=1}^{n} 1 = n(\bar{X} - \mu)^2$,

$$= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2\right] \tag{1}$$

Lemma 1.1. $\bar{X} - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$

Proof. Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then

$$\begin{aligned}\bar{X} - \mu &= \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{n\mu}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\end{aligned}$$

□

Substituting $\bar{X} - \mu$ from **Lemma 1** into 1:

$$= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)\sum_{i=1}^{n}(X_i-\mu) + n(\bar{X}-\mu)^2\right]$$

$$= \frac{1}{n-1}\left[n\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\Big] - \frac{2n}{n}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\Big] + n\mathbb{E}\Big[(\bar{X}-\mu)^2\Big]\right]$$

Once again by **Lemma 1**, the final term: $n\mathbb{E}\Big[(\bar{X}-\mu)^2\Big] = \mathbb{E}\Big[\frac{n}{n^2}\sum_{i=1}^{n}(X_i-\mu)^2\Big] = \sigma^2$

$$= \frac{1}{n-1}\left[n\sigma^2 - 2\sigma^2 + \sigma^2\right]$$

$$= \frac{1}{n-1}(n\sigma^2 - \sigma^2)$$

$$= \frac{1}{n-1}[\sigma^2(n-1)]$$

$$= \sigma^2$$

Completing our proof that $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of variance.

□

1.3 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimators start from the sample $\bar{X}$ and derive the parameters θ which maximize the probability of $\bar{X}$ being sampled.

Theorem 1.3. *The MLE of the mean of a normally distributed RV $X = X_1, X_2, ..., X_n$ with known variance σ^2 is $\frac{1}{n}\sum_{i=1}^{n} X_i$.*

Proof. Since X is normally distributed, $f_X(X_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$. The *likelihood* of the set of samples X is then:

$$L(X_1, ..., X_n) = \prod_{i=1}^{n} f_X(X_i)$$

Note that we usually consider the *log-likelihood* to manipulate expressions easier. We can do this since log is monotonically increasing and one-to-one (for bases ≥ 1) and preserves the global extrema of our likelihood function.

$$\begin{aligned}\ln(L(X_1, ..., X_n)) &= \ln\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\right)\\ &= \sum_{i=1}^{n}\left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\right)\right)\\ &= \sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(X_i-\mu)^2}{2\sigma^2}\right)\end{aligned}$$

Since we wish to estimate μ, we take the partial derivative of the log-likelihood function w.r.t μ, and set it equal to zero to solve for the extrema value.

$$\frac{\partial \ln(L)}{\partial \mu} = \sum_{i=1}^{n} \frac{(X_i - \mu)}{\sigma^2} = 0$$
$$\sum_{i=1}^{n}(X_i) - n\mu = 0$$
$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

□

Theorem 1.4. *The MLE of the variance of a normally distributed RV $X = X_1, X_2, ..., X_n$ with known mean $\mu = \hat{\mu}_{MLE}$ is $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, or the biased estimator of variance.*

Proof. The likelihood is given by:

$$L(X_1, ..., X_n) = \prod_{i=1}^{n} f_X(X_i) = (\frac{1}{\sigma\sqrt{2\pi}})^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

Let $\mathcal{L} = \ln L(X_1, ..., X_n)$, the log-likelihood of L. Then

$$\mathcal{L} = n \ln \frac{1}{\sigma\sqrt{2}\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = -n\ln\sigma - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Since we are looking for extrema of σ^2, we take the partial $\frac{\partial \mathcal{L}}{\partial \sigma^2}$ and set it equal to zero.

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \frac{-n}{2\sigma^2} - 0 + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$
$$\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \text{ then using our known mean } \hat{\mu}_{MLE}:$$
$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_{MLE})^2$$

□

2 Linear Regression

2.1 Introduction

Linear regression is the process of fitting a model that is linear in its *parameters*. That is, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$, and we solve for parameters β to minimize some loss

function (e.g. mean squared error). As a remark, this means that we can fit functions not completely linear if we square certain inputs before applying our regression formula. The resultant model will still be linear in the parameters.

If we have more than one data point, we can use matrix form to describe the system:

$$Y = X\beta$$

$$\text{, where } X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}, \; Y = [Y_1, Y_2, ..., Y_m]^T \text{ , and } \beta = [\beta_1, \beta_2, ..., \beta_n]^T$$

That is, X is an $m \times n$ matrix representing m data points, with each data point having n predictors. Then we have m predictions (stored in Y) and n β-parameters to fit.

The standard regression problem is fitting $f(x)$ in $Y = f(x) + \epsilon$, where ϵ is noise in the data usually assumed to be $\epsilon \sim \mathcal{N}(0,\, 1)$.

Let us derive our parameters β. We begin with the assumption that we want to minimize $\sum_{i=1}^{m} \epsilon_i^2 = (y_i - \hat{y}_i)^2$ the squared of the sum of the residuals (RSS). Assuming we fit a model perfectly, then the only error left will be ϵ^2 from our original data $Y = X\beta + \epsilon$. This is irreducible error. Rewriting,

$$\begin{aligned} \min_{\beta} \epsilon^T\epsilon &= (Y - X\beta)^T(Y - X\beta) \\ \min_{\beta} \epsilon^T\epsilon &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta \\ \frac{\partial \epsilon^T\epsilon}{\partial \beta} &= -2X^TY + 2X^TX\beta = 0, \text{ then} \\ X^TX\beta &= X^TY \\ \text{and } \beta &= (X^TX)^{-1}X^TY \end{aligned}$$

So we can estimate β using the projection matrix $(X^TX)^{-1}X^T$.

2.2 Simple Linear Regression

Consider the simple case $y_i = a + bx_i + \epsilon_i$ for $i \in [1, n]$ (a is essentially b_0). Let

$$S = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - q - bx_i)^2$$

From here, we can derive explicit expressions for estimations of $\hat{a}, \hat{b}$.

$$\frac{\partial S}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\frac{1}{n}\sum_{i=1} n = a + b\bar{x} \text{ , then } \hat{a} = \bar{y} - \hat{b}\bar{x}$$

For $\hat{b}$, we have

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i) = b\sum_{i=1}^{n} x_i^2$$

Note:

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

and that as $n \to \inf$, $cov(x,y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Theorem 2.1. $corro(x,y) = \frac{cov(x,y)}{(var(x)var(y))^2}$, $\hat{b} = \frac{cov(x,y)}{var(x)}$.

2.3 SLR: Step by Step

We list out some steps to perform simple linear regression.

1. Fit the regression coefficients $\hat{a}, \hat{b}$ using definitions derived in [2.2]. Let the i'th residual be $\epsilon_i = y_i - \hat{y}_i$, which measures our fit at the i'th data point.

2. Create a hypothesis test for $\hat{b}$ to test for regression. Let b_c be some constant and true b value. Then we set up the test as:

   $$H_0 \ : \ \hat{b} = b_c$$

   $$H_1 \ : \ \hat{b} \neq b_c$$

   We create the test statistic $T_0 = \frac{\hat{b} - b_c}{se(\hat{b})}$, where squared error:

   $$se(\hat{b}) = \sqrt{\frac{\frac{\sum_{i=1}^{n}\epsilon_i^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

   The test statistic T_0 follows a t-distribution with $n-2$ degrees of freedom where n is the total number of data points we used when fitting $\hat{b}$. We test significance, accepting the null hypothesis if

   $$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$$

   If we set $b_c = 0$, we are essentially testing for significance of regression and failing to reject the H_0 implies the data cannot be fit with a linear model.

3. We also test for the intercept $\hat{a}$ using $T_0 = \frac{\hat{a} - a_c}{se(\hat{a})}$ where

   $$se(\hat{a}) = \sqrt{\frac{\sum_{i=1}^{n}\epsilon_i^2}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

   Testing for the same t-distribution range as $\hat{b}$.

2.4 Testing Case Study: Risk Factors for Cardiovascular Disease

We obtain data from the Framingham Heart Study. It includes $n = 4434$ participants who completed one of the regularly scheduled examinations from 1956 – 1968. The following table shows variable names, as they appear in the csv, along with brief descriptions and coding details for each variable

Variable Name	Description	Coding
AGE	Age at exam, in years	32-70
TOTAL CHOL	Total cholesterol, mg/dL	107-696
SBP	Systolic blood pressure, mmHg	83.5-295
DBP	Diastolic blood pressure, mmHg	48-142.5
BMI	Body mass index, kg/meters2	15.54-56.8
CIGS PER DAY	Number of cigarettes smoked per day	0-70
GLUCOSE	Serum glucose mg/dL	40-394
HEART RATE	Heart rate, beats/minute	44-143
CVD	Cardiovascular disease over 24 year follow-up	0=no, 1=yes
HYPERTENSION	Hypertension over 24 year follow-up	0=no, 1=yes

Exercise 1. *Is hypertension associated with CVD? Carry out an appropriate test. Include the appropriate hypotheses, test statistic value, p-value, and conclusion.*

Proof. We must perform a χ^2 test for significance between two categorical variables. We lay out our hypotheses:

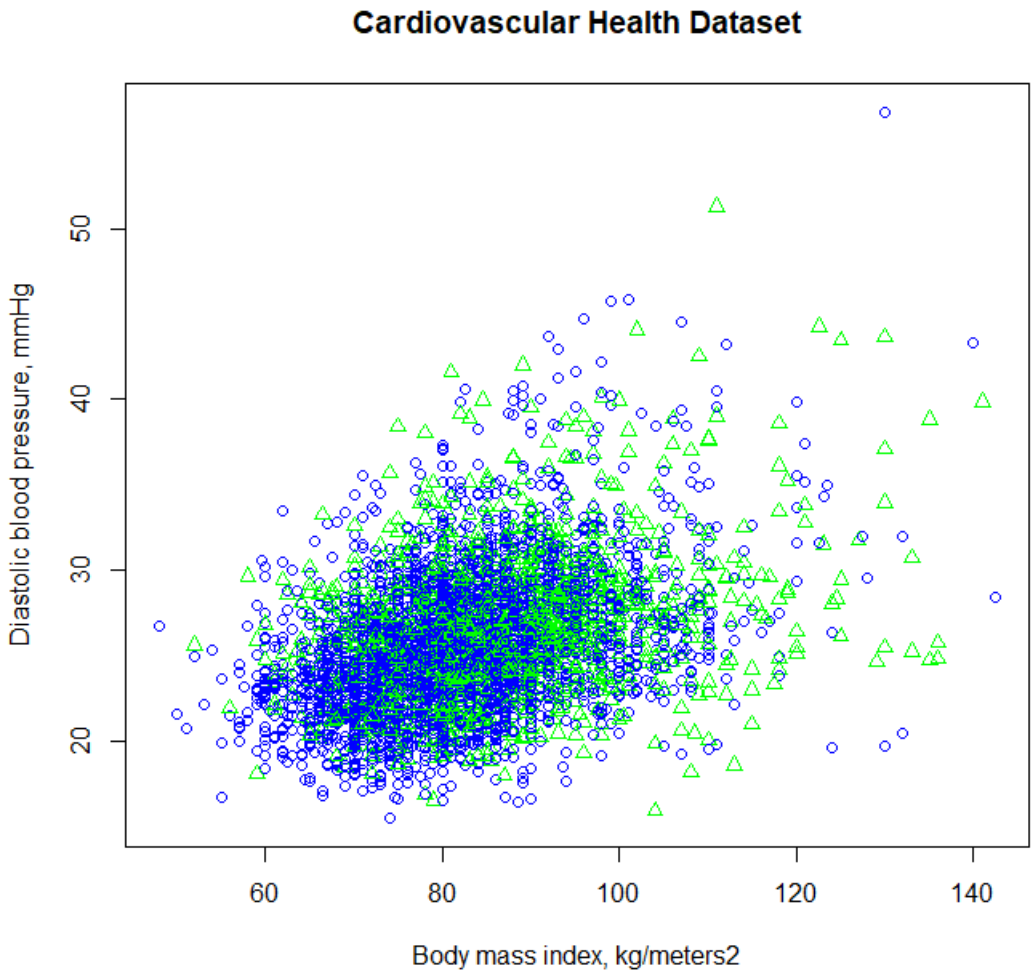
$$H_0: \text{ CVD and HYPERTENSION are independent variables}$$
$$H_1: \text{ CVD and HYPERTENSION are dependent variables}$$

Running the associated R code snippit:

```
## heart_study.R
data = data<-read.csv(file.choose(),header=TRUE)
# We have two categorical variables
data$CVD <- as.factor(data$CVD)
data$HYPERTENSION <- as.factor(data$HYPERTENSION)

chi2 = chisq.test(data$CVD, data$HYPERTENSION)
c(chi2$statistic, chi2$p.value)
...
   X-squared
1.230504e+02 1.359530e-28
```

Since our $p = 1.359530e - 28 < 0.05$, we reject our null hypothesis and conclude that the variables CVD and HYPERTENSION are dependent on each other. □

Exercise 2. *Plot BMI vs DBP with different plotting symbols for the two CVD groups. Describe the relationship between BMI and DBP. Does CVD appears to have an effect on these variables?*



Proof. R Code used to plot the data:

```
data = data<-read.csv(file.choose(),header=TRUE)
...
plot(data$BMI ~ data$DBP,
     xlab = "Body mass index, kg/meters2",
     ylab = "Diastolic blood pressure, mmHg",
     pch = as.numeric(data$CVD)
     col = c("blue", "green")[as.numeric(data$CVD)])
```

There seems to generally be a linear relation between BMI and DBP. However, due to the distribution of the CVD's through all areas of the plot, it does not have discernable influence on the other two variables. □

2.5 Linear Regression Case Study: Optical Gravitational Lensing

We obtain data (linked) from the Optical Gravitational Lensing Experiment (Ogle-II). Stars in this data set have been identified as Cepheid variables located in the Small Magellenic

Cloud (SMC). Two types of Cepheids are included here, fundamental mode (FU) and first overtone (FO). Measurements include the logarithm of the periods of variability and stellar magnitudes at various wavelengths of light. The first column contains ID's of these stars. For this problem we are interested in the following quantities:

MW: an extinction-free measure of stellar magnitude (since these stars are all essentially the same distance from Earth, this is directly related to their luminosity).

Type: type of Cepheid, FU or FO. These classifications were obtained by a fourier analysis of the light curve and so may be wrong.

VI: a color index (difference in brightness between visual and infra-red).

logPeriod: log base 10 of the period of variability.

We first fit a full linear regression model, taking into account all given variables:

```
# stardata.R
con = url("http://www.utdallas.edu/~ammann/stat6341scripts/OgleLMCSMCWIII.RData")
load(con)
data = OgleLMCSMCW.df

# full model
Full =
    lm(data$MW~data$Galaxy*data$logPeriod+data$Type*data$logPeriod+data$Type*data$VI+
    data$Galaxy*data$VI)
```

Next, we do outlier removal according to: $|rstudent|$ and *dffits*:

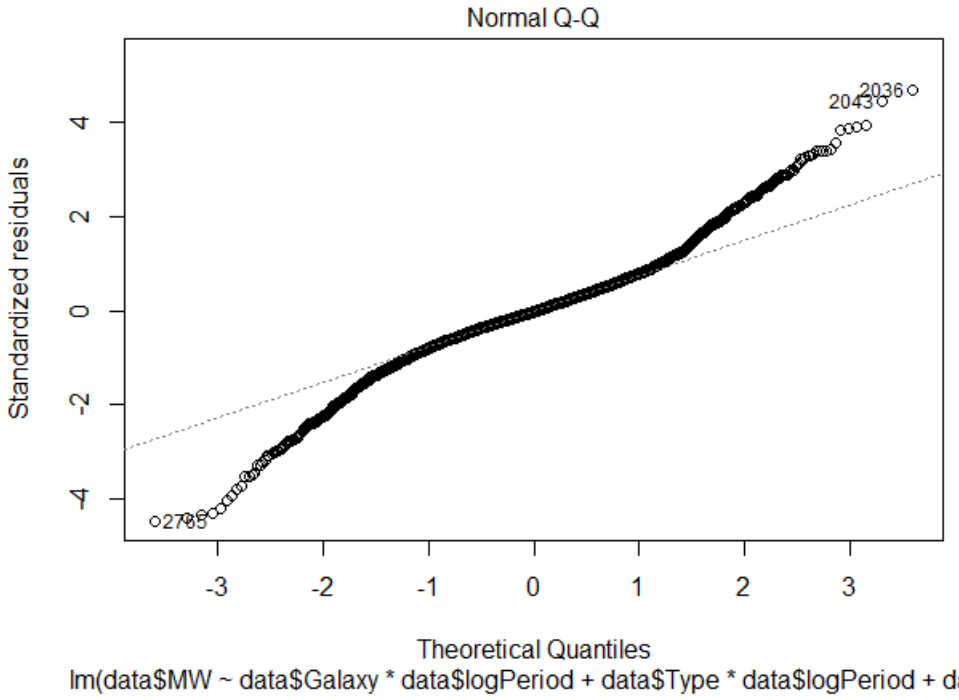
```
# Outlier removal
# get number of predictors, should be 9 including intercept
p = length(Full$coefficients)
# remove points with rstudent value > 3
rstudent = (abs(rstudent(Full)))
count <- 1
for (val in rstudent){
  print(val)
  print(count)
  if (val > 3){
    data <- data[-c(count), ]
  } else {
    count = count + 1
  }
}

# remove points with > 2 * sqrt((p+1)/(n-p-1)) dffits value
dffits = (dffits(Full))
count <- 1
for (val in dffits){
  print(val)
  print(count)
```

```
  if (val > 2 * sqrt((p+1)/ (3991-p-1)))
  {
    data <- data[-c(count), ]
  } else {
    count = count + 1
  }
}
```

We have reduced the number of data points from 3991 $\rightarrow$ 3084. We refit our model and produce summary results:

```
# re-fit model on data without outliers
Full = lm(data$MW~data$Galaxy*data$logPeriod+ data$Type*data$logPeriod+
    data$Type*data$VI+ data$Galaxy*data$VI)
summary(Full)
...
Residuals:
     Min       1Q   Median       3Q      Max
-0.53439 -0.05710 0.00308 0.05826 0.50655

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -3.09165   0.02331 -132.607 < 2e-16 ***
data$GalaxySMC               -0.18544   0.02607   -7.113 1.41e-12 ***
data$logPeriod               -3.42603   0.02230 -153.646 < 2e-16 ***
data$TypeFU                   0.52854   0.02476   21.351 < 2e-16 ***
data$VI                      -0.03500   0.03263   -1.072   0.284
data$GalaxySMC:data$logPeriod -0.27213 0.01911 -14.236 < 2e-16 ***
data$logPeriod:data$TypeFU  0.14371   0.02110    6.812 1.15e-11 ***
data$TypeFU:data$VI         -0.06362   0.03934   -1.617   0.106
data$GalaxySMC:data$VI       0.64221   0.03991   16.093 < 2e-16 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 0.1145 on 3075 degrees of freedom
Multiple R-squared: 0.9859, Adjusted R-squared: 0.9859
F-statistic: 2.696e+04 on 8 and 3075 DF, p-value: < 2.2e-16
```

Since we obtain a very high R^2 value, we check our model assumptions by showing the QQ-plot (quantile-quantile).



It seems our data follows a heavy-tail distribution after outlier removal.

Perfoming a shapiro wilks test with

```
shapiro.test(data$MW)
...
   Shapiro-Wilk normality test

data:  data$MW
W = 0.96891, p-value < 2.2e-16
```

verifies that with low p-value, MW, which we are attempting to predict, is normally distributed.

Continuing, we perform cross validation, sampling train/test data with a 70/30 split.

```
SPSE = 0
for (i in 1:2000){
  ## 70% of the sample size
  smp_size <- floor(0.7 * nrow(data))

  train_ind <- sample(seq_len(nrow(data)), size = smp_size)
  * need to sample without replacement

  train <- data[train_ind, ]
  test <- data[-train_ind, ]

  Full = lm(MW~Galaxy*logPeriod+ Type*logPeriod+ Type*VI+ Galaxy*VI, data=train)
```

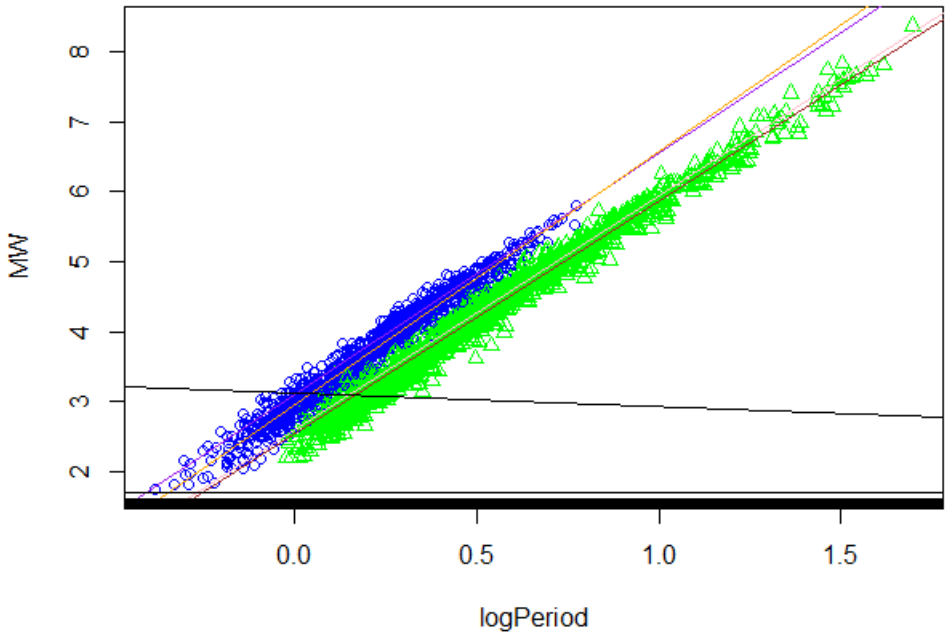
```
  SPSE = SPSE + mean((data.frame(test$MW)- predict.lm(Full,
      newdata=list(X=test))) ^ 2)
}

# Find MPSE
MPSE = SPSE / 2000
print(MPSE)
...
[1] 0.01285563
```

Obtaining a mean square predicted error (MPSE) of 1.843719. We repeat this for each of the models, obtaining the MPSE for each:

Model	MPSE	Adj R^2
Full (PLC)	0.01285563	0.9859
GalVI (PLC)	0.01284548	0.9859
VI (PLC)	0.01873347	0.9797
Gal (PL)	0.01442136	0.9842
Gal0 (PL)	0.01465222	0.984
Base (PL)	0.0187344	0.9797
PL (PL)	0.08138108	0.9123

We can clearly see the correlation between model complexity and expressibility with the R^2 column of the table. Among the PLC models, GalVI takes the crown with high expressibility and the best MPSE fit. Among the PL models, Gal seems to be a reasonable balance, having the lowest MPSE as well as the highest Adjusted R^2. There does not seem to be any advantage to using a PLC model over a PL one, given that the MPSE of the PL model is clearly up to par with that of the PLC models.

Refitting our full, outliers removed, data with the best PL model (Gal), we obtain an Adjusted R^2 of 0.9842. Here is a comparison with the methods published in the paper:



the black regression line corresponds to the fitted Gal line, which does not line up since it is not only dependent on the logPeriod. The four other lines are published results.