INDIAN STATISTICAL INSTITUTE
Saptarshi Sinha(bs2031)

## Acknowledgement

I want to thank Prof. Arnab Chakraborty for providing detailed explanations with R for this topic. R codes are inside the custom environment **“boxed2”**, like this one.

# 1 Multicollinearity: R Lab- Creating data

We will illustrate a cooked-up example where we will demonstrate the dangers of multi-collinearity in linear models. First, let's assume that a response variable $Y$ depends functionally on a regressor variable $X$ as,

$$Y \;=\; 1 \;+\; 2X \;+\; 0.05X^2 \tag{1}$$

Now, assume that the above relation is known **"only to Nature"**. We just know(*from Domain knowledge, Prior information,*$\cdots$) that $Y$ is quadratically related to $X$ and will model our data according to this knowledge, in order to find those coefficients. From the basic assumptions of regression, since the $X$ values are fixed, we define the variables;

```
x1= 1:100                        #regressor variable values(assumed to be fixed)
x2= 3*x1 + rnorm(100)/5          #x1 measured in another unit with measurement error
x3= x1*x1                         #we suspected quadratic relation
```

The **"x1"** is the $X$ in (1), role of $X^2$ is played by the **"x3"**. The **"x2"** term is generally "x1" measured in different unit*(like from inches to cm it will be 2.54)*. We added a little perturbation with the help of the **rnorm(n,mean=0,sd=1)** function, to indicate the error in measurement.[1] Actually, here we are trying to mimic a real-life scenario, where we have a sample size of $n =$ 100, and our $X$-values(x1) $1, 2, \cdots, 100$ will remain fixed(assumption) across any samples.

As statisticians, we suspect $Y$ is somehow related to x1,x2,x3, since we are unaware of (1). We tend to find the coefficients of x1,x2,x3 based on our sample.[2] For this purpose we use the Gauss-Markov estimators $\hat{x1}, \hat{x2}, \hat{x3}$. But in this scenario, we will prove how our estimators actually betray us (and they does that because of guess what, multi-collinearity).

## 1.1 Sample distribution

To know the behaviour of the estimators, we will see the sampling distribution of them. To know that, we will **"create"**(generally we won't be having these God-like powers) many samples and find the corresponding estimates. Then we will study and plot those values per estimate across the tailor-made samples and then calculate their variance. We will be making 200 samples in R as;

```
temp= c()                                 #initialization of matrix
j= rep(1,100)                              #vector with n(sample size) 1s
for(i in 1:200){                           #i iterates for each sample
y= j + 2*x1 + 0.05*x3 + 4*rnorm(100)       #recorded Y values(as vector) per sample
temp= rbind(temp,lm(y~x1+x2+x3)$coef)      #matrix with 200 sample estimates
}
```

Now, we plot the value of each estimator across samples and calculate the co-variance matrix of these values. The following R code is useful,

[1] For eg. if you measure a 1-inch pencil in cm scale, you will get something like 2.538 cm or 2.541 cm (depending on the instrument precision), but not always you will get the exact 2.54 cm

[2] Though it's over-smartness to include "x2" in real life since we know x1,x2 are linearly related, but for the sake of demonstration assume we did this mistake unknowingly

```
par(mfrow=c(4,1))         #dividing the graphic screen into 4 rows
hist(temp[,1])            #histogram of values of intercept estimator
hist(temp[,2])            #histogram of values of x1 estimator
hist(temp[,3])            #histogram of values of estimator of x2
hist(temp[,4])            #histogram of values of estimator of x3
var(temp)                 #variance-covariance matrix of the values
```

We used the **par(mfrow=c(4,1))** to divide R's graphic screen into 4 rows. Then we inserted the histograms one-by-one. We did one instance in our machine and got these plots,

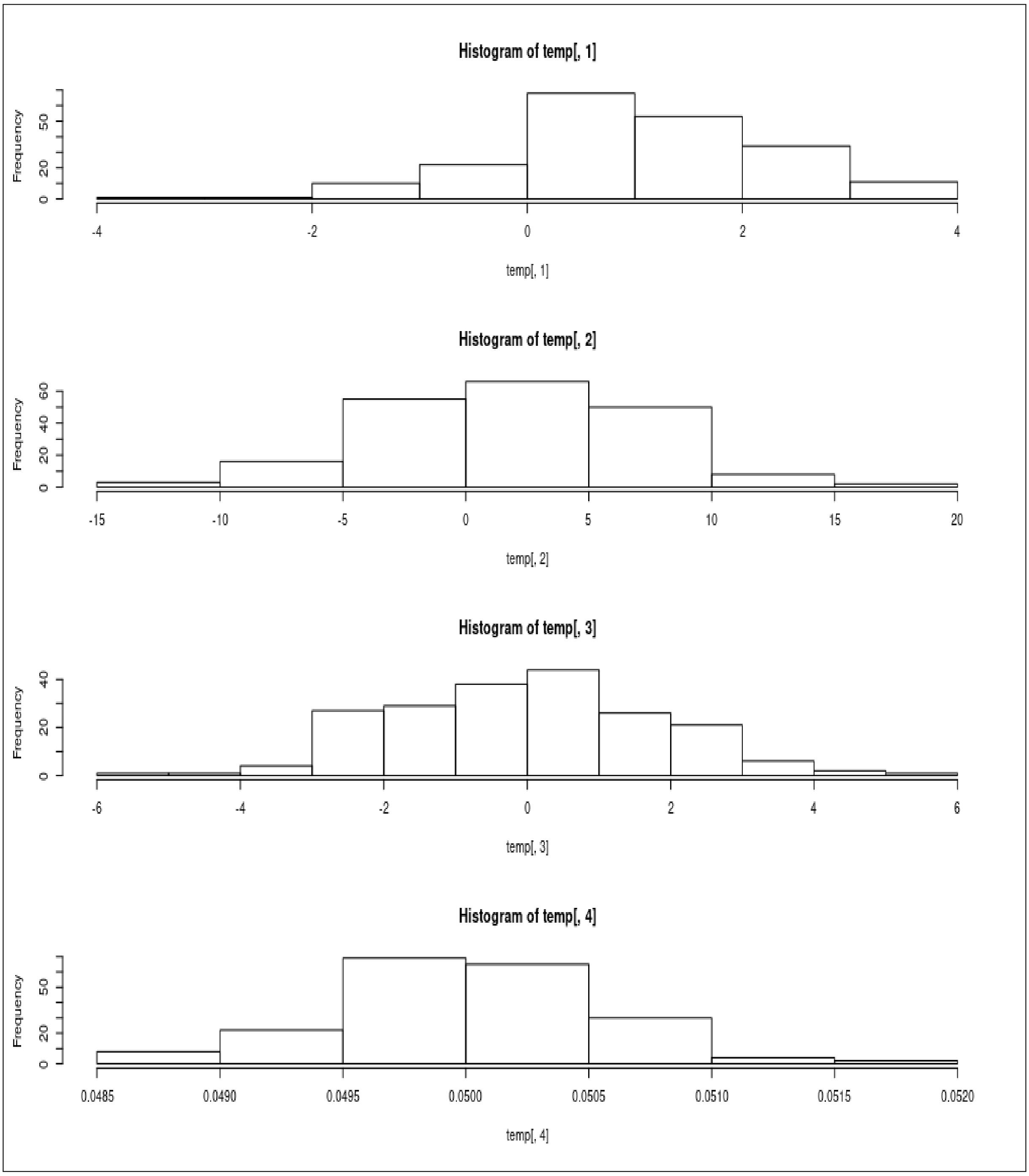

Figure 1: Histogram plots of: Intercept, X1, X2, X3

The variance-covariance matrix is,

$$\begin{array}{ccccc} & \text{Intercept} & X1 & X2 & X3 \\ \text{Intercept} & 1.5818705983 & 1.308630785 & -4.570597e-01 & 5.043144e-04 \\ X1 & 1.3086307853 & 30.031308179 & -1.002442e+01 & 2.083890e-04 \\ X2 & -0.4570596725 & -10.024421925 & 3.346509e+00 & -7.972711e-05 \\ X3 & 0.0005043144 & 0.000208389 & -7.972711e-05 & 3.020110e-07 \end{array}$$

It is clear from both the histogram and the matrix, that the estimators for the coefficients of X1, X2 is much less reliable than that of X3, since their variance is much more compared to that of X3. That's the reason why we saw weird estimates for coefficients of X1,X2 across samples. In some samples we got $\hat{x}1=$ **9.89755532,** $\hat{x}2=$ **-2.6264126464**, while in other samples, we got $\hat{x}1=$ **-4.20037679,** $\hat{x}2=$ **2.0379043834**. In short, X2 was 3 times X1, but in almost all the samples, this relation broke and also the estimates were faulty.

In conclusion, because X1 and X2 were linearly related, they gave **"bad estimates"**, whereas, since X3 had no such linear relation, it gave reliable estimates. Thus we demonstrated the dangers of multicollinearity.