The Elements of Statistical Learning
WS 2019/2020

Jilles Vreeken and Tobias Marschall
Exercise Sheet #2: Regression 1

Yashaswini Mysuru Udaya Kumar – 2581353
Pramod Ramesh Rao – 2581261

Assignment - 2
Exercise group – We

Problem 1

Prove Gauss Markov Theorem

We will prove Gauss Markov Theorem using the Matrix form.

Given :- $y = X\beta + \epsilon$.

The normal form of least square estimator $\hat{\beta}$ is :-

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Suppose we have another unbiased linear estimator $\tilde{\beta}$ such that $\tilde{\beta} = Ay$ where $A = (X^T X)^{-1} X^T + D$. D is some matrix

$$\therefore \tilde{\beta} = Ay$$

$$= \left((X^T X)^{-1} X^T + D\right) y.$$

$$= (X^T X)^{-1} X^T y + Dy.$$

Substituting the value of y in the above equation

$$= (X^T X)^{-1} X^T (X\beta + \epsilon) + D(X\beta + \epsilon)$$

$$= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon + DX\beta + D\epsilon$$

Taking Expectation on $\tilde{\beta}$ we get.

$$E[\tilde{\beta}] = E\left[\underbrace{(X^T X)^{-1} X^T X}_{\text{Identity matrix}}\beta\right] + E\left[(X^T X)^{-1} X^T \epsilon\right] + E[DX\beta] + E[D\epsilon]$$

($E[(X^T X)^{-1} X^T \epsilon] \to 0$, $E[D\epsilon] \to 0$)

Since we have $E(\epsilon) = 0$

$$\therefore E[\tilde{\beta}] = I\beta + DX\beta$$

$$= (I + DX)\beta$$

Here $\tilde{\beta}$ can be unbiased only if $DX = 0$

We now have to find an expression for variance of $\tilde{\beta}$ i.e $Var(\tilde{\beta})$

we need to make use of few general matrix results

→ Var of any matrixes $AX = Var(AX) = A\,Var(X)\,A^T$

→ $(AB)^T = B^T A^T$

→ $(A^{-1})^T = (A^T)^{-1}$

We have $Var(\tilde{\beta}) = var((X^TX)^{-1}X^T + D)y)$

$\therefore$ Let $C = (X^TX)^{-1}X^T + D$, then

$$Var(\tilde{\beta}) = Var(Cy)$$

$$= C\,Var(y)\,C^T \qquad [Var(y) = \sigma^2]$$

$$= \sigma^2 CC^T$$

Substituting the value of C we get:

$$= \sigma^2 ((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T)$$

$$= \sigma^2 [(X^TX)^{-1}X^T \cdot X(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1} + DD^T]$$

because $DX = 0$ and $(DX)^T = X^TD^T = 0$

$$= \sigma^2 [(X^TX)^{-1}X^T \times (X^TX)^{-1} + DD^T]$$

$$= \sigma^2 [I\,(X^TX)^{-1} + DD^T]$$

(I: Identity matrix)

$$= \sigma^2 [I\,(X^TX)^{-1}] + \sigma^2[DD^T]$$

$\sigma^2[I\,(X^TX)^{-1}]$: Variance of Least square estimator.

$$= \sigma^2(\hat{\beta}) + \sigma(DD^T]$$

Any matrix times transpose of that matrix makes it a positive Semidefinite matrix.

$V^TAV \geq 0$ where V is a vector and A is semidefinite matrix

. It is semi definite when the above scalar product is ≥ 0

$$= \sigma^2(\hat{\beta}) + \sigma(DD^T)$$

Hence, this is a scalar analog which is ≥ 0

Thus we have $Var(\tilde{\beta}) \geq \sigma^2(\hat{\beta})$

$$= Var(\tilde{\beta}) \geq Var(\hat{\beta})$$

Hence, we have proved that the least square estimator for parameter β is the best linear unbiased estimater possible. as it has a lower variance.

Problem 3

① (a) Given Set of observations

No of features or p = 1 is X

X is uniformly distributed in [0, 1]

Predict response for a test observation with X = 0.6 using observation in range [0.55, 0.65]

i e using observations within 10% of range X closest to test observation

→ Fraction of available obs. used to make Prediction $= \frac{0.65 - 0.55}{1 - 0}$

[diff. given obs. range / diff. range of X]

$= \frac{0.1}{1} = 10\%$

(b) $p = 2$, X_1 and X_2

(X_1, X_2) uniformly distributed on $[0,1] \times [0,1]$

Predict response for test obs. using 10% of range of X_1 and X_2.

$\Rightarrow$ 10% of X_1 × 10% of X_2

$\frac{10}{100} * \frac{10}{100} = 0.1 \times 0.1 = 0.01 = 1\%$

(c) $p = m$, $X_1, X_2, X_3 \ldots X_m$

Predict response for test obs. using 10% of range of range of $X_1, X_2 \ldots X_m$

$\Rightarrow$ 10% of X_1 * 10% of X_2 * ..... * 10% of X_m.

$\therefore \left(\frac{10}{100}\right)^m = (0.1)^m$ or $(10\%)^m$.

Problem 3

(2) Given : N data points uniformly distributed in a p dimensional unit ball centered at the origin.

→ Unit ball - a ball with radius $r = 1$

- Consider a nearest neighbor estimate at the origin & Show that the median distance from origin to the closest data point is:-

$$d(p, N) = \left((1 - 1/2)^{1/N}\right)^{1/p}$$

- We have the median distance as $d(p, N)$, firstly let us consider the number of data points to be 1. $\therefore N = 1$

Then, we have $d(p, 1)$

- We are given the volume of a p-dimensional sphere by

$V(r, p) = G_1(p) r^p$ where $G_1(p)$ is probability constant.

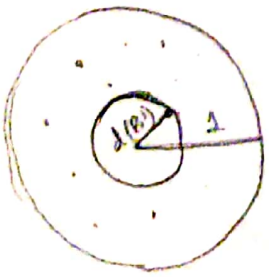
[volume of a p-dimensional ball of radius r is proportional to r^p]

- Let d be the distance to any point such that $P(d \leq d(p, 1)) = 1/2$

Because, if we want to know the distance of origin to its nearest neighbour it will have different values for different data sets. Hence performing many experiments with randomly distributed data points inside a sphere and then find the median of all obtained values.

In this case if we take median as d, then half of the time the nearest neighbour will be farther away than d and half of the time its nearer than d. Hence the probability of this happening is $1/2$. (50:50)

$\therefore P(d \leq d(p, 1)) = 1/2.$



Representing probability in terms of volume

$$\frac{V(d(p,1))}{V(1)} = \frac{G_1(p)\,(d(p,1))^p}{G_1(p)\,1^p}$$

$$= (d(p,1))^p$$

$$\Rightarrow (d(p,1)) = (1/2)^{1/p}$$

→ This is when $N = 1$, we will now consider N data points.

Generally, let a be any value b/w 0 to 1 i.e $[0,1] \rightarrow 0 \le a \le 1$

$$P(d \le a) = a^p$$

Let d be distance to any data point inside the sphere of radius a

$$P(d \le a) = \frac{V(a^p) q(p)}{V(1) q(p)} = \frac{a^p}{1^p} = a^p \quad \text{— (A)}$$

Now consider the case with N datapoint $x_1, x_2, x_3 \ldots\ldots x_N$
The distance to the closest point is :-

$$d = \min\left(\|x_1\|, \|x_2\|, \|x_3\|, \ldots \|x_N\|\right)$$

Thus, like the previous case, we have.

$$\left(d \le d(p,N)\right) = 1/2.$$

As discussed previous, the probability that the nearest neighbour will be nearear and farther away, to add up to 1

i.e $P(d \le d(p,N) + P(d > d(p,N)) = 1$

$\therefore P(d \le d(p,N) = 1 - P(d > d(p,N))$

So, $1 - P(d > d(p,N)) = 1/2$

$\rightarrow P(d > d(p,N)) = 1/2$

Since we have N data points, considering each point

$\rightarrow P(\|x_1\| > d(p,N)) P(\|x_2\| > d(p,N)) (P\|x_3\| > d(p,N)) \ldots P(\|x_N\| > d(p,N) = 1/2$

$\rightarrow \prod_{i=1}^{N} \left(1 - P(\|x_i\| \le d(p,N))\right) = 1/2$

Since x_i is independent and identically distributed variables with same PD and Considering $P(\|x_i\| \le d(p,N)))$ and solving like eqn (A), we get

$\rightarrow (1 - d(p,N)^p)^N = 1/2$

$\rightarrow (1 - d(p,N)^p) = (1/2)^{1/N}$

$\rightarrow d(p,N)^p = 1 - (1/2)^{1/N}$

$\rightarrow d(p,N) = (1 - (1/2)^{1/N})^{1/p}.$

Hence, we obtain $\boxed{d(p,N) = \left(1 - (1/2)^{1/N}\right)^{1/p}}$

For the nearest neighbour method, using high dimensional space will not yield good response or prediction. Hence knn doesnt serve well in high dimensional spaces. The methods that involve capturing a fixed neighbours around the points give high variance for the fit..

1. Create scatterplots between all the variables. Is the relationship between those variables linear? Describe the connection between the variables. (Exclude the name variable, which is qualitative.)

   Solution.
   Refer Section 1 of the code.

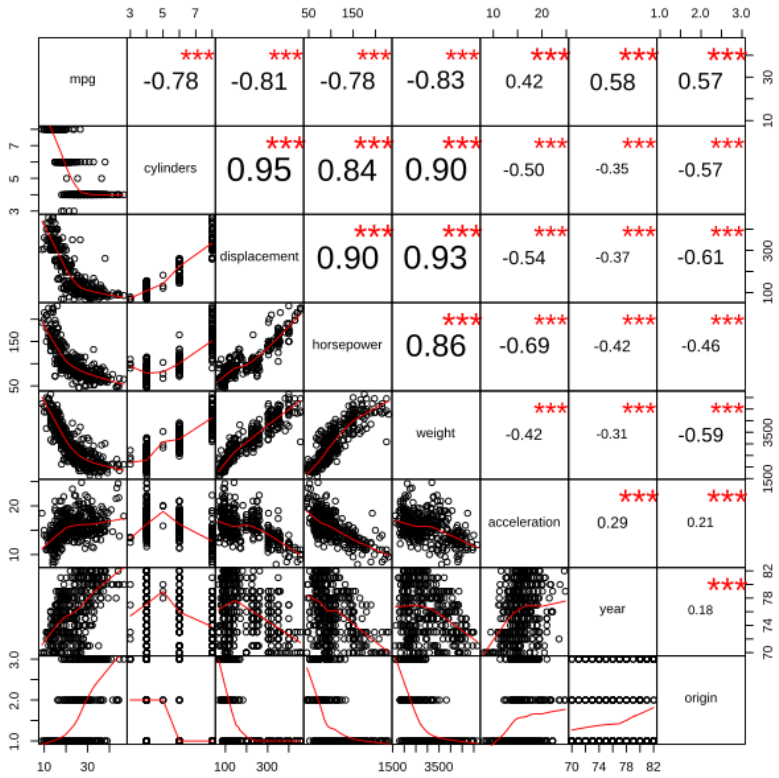


Figure 1: Scatter Plot between all the variables

From the **Figure**1, we can draw the following connections:

- The response variable, "mpg" is positively correlated with the variables acceleration, year and origin. While it is negatively correlated with cylinders, displacement, horsepower and weight.
- The variable "cylinders" is strongly correlated with displacement, horsepower and weight and is negatively correlated with acceleration, origin and year.
- The variable "displacement" is strongly correlated to horsepower and weight and is correlated to acceleration, origin and year in a negative manner.
- The variable "horsepower" is positively correlated to weight and has a negative correlation with acceleration, year and origin.
- The variables, "acceleration", "year" and "origin" are positively correlated with each other, but the correlation is relatively weak.

■

2. Detect the variables in the scatterplots that appear to be most highly correlated and anti-correlated,respectively. Justify your choice using the cor()function.

   Solution.
   The upper triangle of the **Figure**1 shows the correlation values among the different variables.

   It is clear that the variables, displacement, cylinders, horsepower and weight are strongly correlated. At the same time, these variables have a strong anti-correlation with the response variable mpg.

   Refer to the section 2 of the code. ■

3. Perform simple linear regression with mpg as the response using the variables cylinders, displacement,horsepower and year, respectively, as features. Which predictors appear to have a statistically significant relationship to the outcome and how good are the resulting models (measured using R^2)?

   Solution.
   Refer Section 3 of the code.

   The statistical analysis and the RSS value for each estimate is summarised in the following tables : **Table 1**, **Table 2**, **Table 3**, **Table 4** as follows:

Table 1: Cylinders vs mpg

	Coeff	SE	t-stat	p-value
intercept	42.9155	0.8349	51.40	<2e-16
cylinders	-3.5581	0.1457	-24.43	<2e-16

Table 2: Displacement vs mpg

	Coeff	SE	t-stat	p-value
intercept	35.12064	0.49443	71.03	<2e-16
displacement	-0.06005	0.00224	-26.81	<2e-16

Table 3: Horsepower vs mpg

	Coeff	SE	t-stat	p-value
intercept	39.935861	0.717	55.66	<2e-16
horsepower	-0.157845	0.0064	-24.49	<2e-16

Table 4: year vs mpg

	Coeff	SE	t-stat	p-value
intercept	-70.01167	6.645	-10.54	<2e-16
year	1.23004	0.087	14.08	<2e-16

- R-squared$_{\text{cylinder}}$: 0.6047
- R-squared$_{\text{displacement}}$: 0.6482
- R-squared$_{\text{horsepower}}$: 0.6059
- R-squared$_{\text{year}}$: 0.337

We can see that all the p-values for the predictors are less than 0.05. Therefore, all the predictors are statistically significant.

By comparing the R^2 values, we can see that model using displacement as the predictor is highest, hence simple linear regression model with displacement is the best. Further, the model using year has the lowest R^2 value hence it is the worst performing model. ■

4. Use the **lm()**function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the **summary()** function to print the results. Compare the full model to those generated in 3) in terms of their model fit. What can you observe in the different models concerning the significance of the relationship between response and individual predictors? What does the sign of the coefficient tell you about the relationship between the predictor and the response?

Solution.
Refer Section 4 of the code.

The statistical analysis of the multiple linear regression model along with its RSS value is given in the **Table5**

Table 5: Multiple Linear Regression model

	Coeff	SE	t-stat	p-value)
Intercept	-17.218435	4.644294	-3.707	**0.00024**
cylinders	-0.493376	0.323282	-1.526	0.1278
displacement	0.019896	0.007515	2.647	**0.00844**
horsepower	-0.016951	0.013787	-1.23	0.21963
weight	-0.006474	0.000652	-9.929	**<2e-16**
acceleration	0.080576	0.098845	0.815	0.41548
year	0.750773	0.050973	14.729	**<2e-16**
origin	1.426141	0.278136	5.127	**4.67E-07**

- R-squared$_{\text{multiple}}$: 0.8215

Comparing the RSS values of the multiple regression model and the individual simple regression model, it is clear that the best RSS value in simple linear model was using displacement feature, given by R-squared$_{\text{displacement}} = 0.6482$ while the RSS value mutiple regression model is 0.8215. Hence, we can say that the multiple linear model fits better than the simple regression models. ■

5. Use the **plot()** function to produce diagnostic plots of the linear regression fit. Does the residual plot suggest any non-linearity in the data? Does the residual plot suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

Solution.

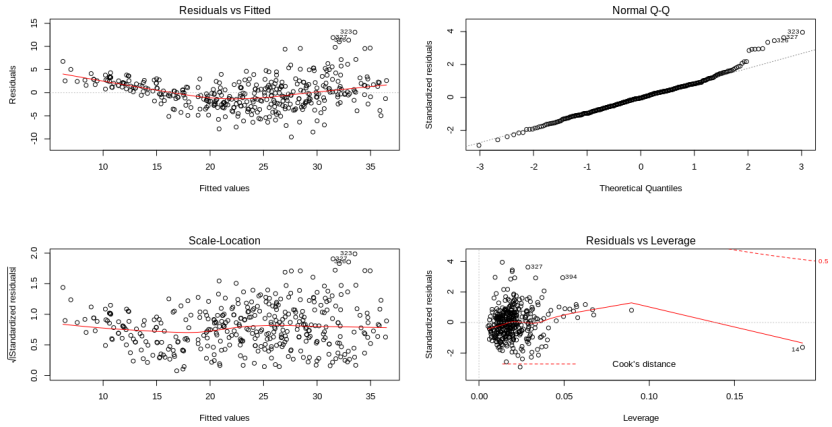


Figure 2: Scatter Plot between all the variables

From the **Figure**1, we can draw the following connections:

- The Residual vs Fitted plot is not horizontal across the zero, hence the presence of a U-shape is indicative of existence of non-linear relationship in the data.
- From the Residual vs Leverage plot, there seem to be two-three points beyond 3 standard deviations. If we consider 3 standard deviation as the threshold then there are few points around point **327** that are outliers.
- There are no points beyond the Cook's distance line, hence there are no points with unusually high leverage.

■

6. The statistical summary for the different interaction models and non linear models are given below.

Table 6: Cylinder-Weight Interaction Model

	Coeff	SE	T-stat	P-value
Intercept	65.3864559	3.7333137	17.514	<2e-16
cylinders	-4.209795	0.7238315	-5.816	1.26e-08
weight	-0.0128348	0.0013628	-9.418	<2e-16
cylinders:weight	0.0010979	0.0002101	5.226	2.83e-07

Table 7: Weight-Year Interaction Model

	Coeff	SE	T-stat	P-value
Intercept	-1.11E+02	1.30e+01	-8.531	3.30e-16
weight	2.76E-02	4.41e-03	6.242	1.14e-09
year	2.04e+00	1.72e-01	11.876	<2e-16
weight:year	-4.58e-04	5.91e-05	-7.752	8.02e-14

Table 8: Year-Cylinders Interaction Model

	Coeff	SE	T-stat	P-value
Intercept	-61.61775	15.10277	-4.08	5.47e-05
year	1.34054	0.19909	6.733	5.99e-11
cylinders	5.51044	2.73705	2.013	0.04478
year:cylinders	-0.1135	0.03647	-3.112	0.00199

Table 9: Sqrt(Displacement) Model

	Coeff	SE	T-stat	P-value
Intercept	47.11839	0.86246	54.63	<2e-16
sqrt(displacement)	-1.75878	0.06186	-28.43	<2e-16

Table 10: Log(displacement) Model

	Coeff	SE	T-stat	P-value
Intercept	85.6906	2.1422	40	<2e-16
log(displacement)	-12.1385	0.4155	-29.21	<2e-16

Table 11: Displacement-Squared Model

	Coeff	SE	T-stat	P-value
Intercept	35.12064	0.49443	71.03	<2e-16
Displacementˆ2	-0.06005	0.00224	-26.81	<2e-16

7. Generate three linear models that are based on all pairwise interaction terms (X1X2) for cylinders, weight, and year as well as on the non-linear transformations log (X); pX; X2 for the displacement variable (one per linear model). Comment on your findings.

   Solution.
   The R^2 values for the interaction models are as follows:

   - R-squared$_{\text{cylinder-weight}}$: 0.7174
   - R-squared$_{\text{weight-year}}$: 0.8339
   - R-squared$_{\text{year-cylinder}}$: 0.722

   Hence, based on the p-values it can be inferred from the **Table 6, 7 and 8** that the interaction models have all the predictors statistically significant. In addition, the weight-year combination has a better fit compared to the multiple regression model.

   The R^2 values for the non-linear models with displacement are as follows:

   - R-squared$_{\text{sqrt(displacement)}}$: 0.6746
   - R-squared$_{\text{log(displacement)}}$: 0.6863
   - R-squared$_{\text{displacement-squared}}$: 0.6482,

   Again,from the **Table 9, 10 and 11**, it is evident that the non linear models have statistically significant predictors but the multiple regression model fits far better than the non-linear model based on displacement. Interestingly, the non-linear model especially log and square fit better than the simple linear model with displacement. Also, the line of fit between the simple linear model and squared model for displacement. ■

Problem 2

2) By definition R^2 is given by

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \sum_{i=1}^{n} \frac{(y_i - \hat{y})^2}{(y_i - \bar{y})^2}$$

Also correlation is given by,

$$\rho(\hat{y}, y) = \frac{\frac{1}{n} \sum_{i=1}^{n} (\hat{y} - \bar{y})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2\right)\left(\frac{1}{n}\sum (y_i - \bar{y})\right)^2}}$$

where $\hat{y}_i$ - predicted value
$\bar{y}_i$ - mean
y_i - actual observation

If we make an assumption that $\hat{y}_i$ is given by a predictor of best fit and hence has a optimum, This assumption ensures we need not scale or shift the regression line. Mathematically it can be written as . The mean square loss $L(\hat{y}, y)$ between the prediction & observations based on the above assumptions can be given by

$$MSE(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad -(1)$$

for any α, β (1) can be expressed as an linear combination α, β & $\hat{y}$ & the (1) becomes.

$$MSE(\hat{y}, y) \leq \frac{1}{n} \sum_{i=1}^{n} (\alpha \hat{y}_i + \beta - y_i)^2 \quad -(2)$$

~~For simplicity if we let $\alpha = 0, \beta = 0$~~

~~Let $\hat{y}_i$ & y_i exist of a point~~, consider a specific observation y_i & it respective prediction $\hat{y}_i$. We can express (2) as a function of α, β

$$f(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (\alpha \hat{y}_i + \beta - y_i)^2 \quad -(3)$$

③ is min when α, β are (1,0) respectively, since we assumed we have obtained the line of best fit at ①

$\therefore$ w.k.t at optimal point $\frac{\partial f}{\partial \alpha} = 0$ & $\frac{\partial f}{\partial \beta} = 0$.

$$\therefore \left.\frac{\partial f}{\partial \beta}\right|_{1,0} = \frac{1}{n}\sum_{i=1}^{n} 2(\alpha \hat{y}_i + \beta - y_i)\Big|_{1,0} = 0$$

$$= \sum_{i=1}^{n} (\hat{y}_i - y_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i \Rightarrow \bar{\hat{y}}_i = \bar{y}_i \quad -④$$

$$\left.\frac{\partial f}{\partial \alpha}\right|_{1,0} = \frac{1}{n}\sum_{i=1}^{n} 2(\alpha \hat{y}_i + \beta - y_i)\hat{y}_i\Big|_{1,0} = 0$$

$$= \sum_{i=1}^{n} (\hat{y} - y_i)\hat{y}_i = 0$$

$$(\hat{y}_i)^2 = \hat{y}_i \cdot y_i \quad -⑤$$

Consider equation for R^2

$$1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

if we shift the co-ordinate system by $\bar{y}$ distance such that it becomes the reference, then R^2 becomes.

$$1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i)^2} = 1 - \frac{y \cdot y - 2y \cdot \hat{y} + \hat{y} \cdot \hat{y}}{y \cdot y}$$

$\because y\hat{y} = \hat{y}_i \cdot \hat{y}_i$

$$= 1 - \frac{yy - \hat{y}\hat{y}}{yy} = \frac{\hat{y}\hat{y}}{yy} \quad -⑥$$

Consider the correlation equation.

$$\rho(\hat{y}, y) = \frac{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i)\cdot(y_i)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)}} \quad \text{(shift by } \bar{y}\text{)}$$

$$= \frac{\hat{y}\cdot y}{\sqrt{(\hat{y}\cdot\hat{y})(y\cdot y)}} = \frac{\hat{y}\,\hat{y}}{\sqrt{(\hat{y}\,\hat{y})(y\cdot y)}} \quad \text{(from (5))}$$

$$\rho(\hat{y}, y) = \sqrt{\frac{\hat{y}\cdot\hat{y}}{y\cdot y}} \quad — \quad (4)$$

clearly, $\rho^2(\hat{y}, \hat{y}) = R^2$.

If we consider $ax + b = \hat{y}$ is the line of best fit. for the given observation. ~~X~~ Y. The loss is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)$$

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

the above equation is similar to (3)

∴ we will get $x_i \cdot x_i = y_i x_i$ & $\bar{x}_\ell = \bar{y}_\ell$

∴ shifty the co-ordinate system by $\bar{x}_\ell$ wrt

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

~~$y_i y_i + ax_i ax_i + b^2 = 2abx_i + y_i =$~~

~~$(y_i)^2$~~

$$1 - \frac{y_i \cdot y_i + (ax_i + b)^2 - 2(ax_i + b) y_i}{y_i y_i}$$

$$1 - \frac{y_i y_i + a^2 x_i x_i + b^2 + 2 a x_i b - 2 a x_i y_i - 2 b y_i}{y_i y_i}$$

$$= \frac{a^2 x_i \cdot x_i + 2(a x_i b - y_i b)}{y_i y_i} = \frac{a^2 x_i^2 + b^2 - 2abx_i}{y_i^2}$$

$$f(x,y) = \frac{y_n \sum \hat{y}_i y_i}{\sqrt{(y_n \sum \hat{y}_i \cdot)^2 (y_n \sum y^*_i)^2}}$$

$$= \frac{\hat{y}_i \cdot y}{\sqrt{\hat{y}_i \hat{y}_i}}$$

$$= \frac{(ax+b)(y)}{\sqrt{(ax+b)^2 y)}} = \sqrt{\frac{(ax+b)^2}{y^2}}$$