INDIAN STATISTICAL INSTITUTE
Saptarshi Sinha(bs2031)

A note to Professor

Sir, since you told us that you are going to write a suitable reference book on this course and our assignments are going to help, I have tried my best to present this assignment as it is a part of a book. Since I have no experience in writing book, I request you to guide me by correcting the mistakes I made here as a good author and give me advice so that I can present you with a better manuscript next time.

Acknowledgement

In writing this assignment I want to acknowledge Prof. Arnab Chakraborty for providing a neat sketch of the proof as well as visuals in his web page, and MAN-MADE from stackexchange for providing me a marvelous proof of the result, "trace of an idempotent matrix is its rank".

1 Estimating σ^2

So far we have seen results regarding the $\hat{\vec{\beta}}$ in the Gauss-Markov setup;

$$\vec{Y}_{nx1} \;=\; X_{nxp}\vec{\beta}_{px1} + \vec{\epsilon}_{nx1}$$

where $\vec{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ and $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, $\forall i$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$.

Though we have an estimate of the β, wich is $\hat{\beta}$, we don't know the value of the σ that we have introduced in order to make this a "Statistical Model". So, our next task is to find the estimate of this variance. A very used unbiased estimator is:

$$\hat{\sigma^2} \;=\; \frac{\left\|\vec{y} - X\hat{\vec{\beta}}\right\|^2}{n - rank(X)}$$

The proof that this estimator is indeed unbiased requires some knowledge of linear algebra. But before going through rigor, let's appreciate this idea intuitively.

1.1 Geometric Interpretation

We will again return to that picture,

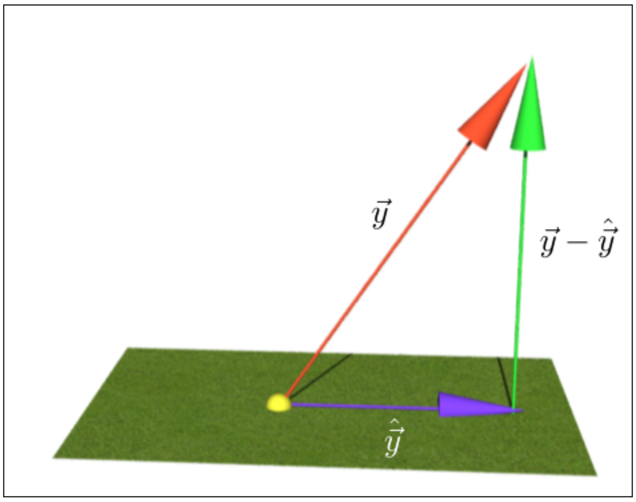

Figure 1: Pictorial diagram of a linear model

As we have seen that $X\hat{\vec{\beta}}$ provides the "best" approximation to $\vec{Y}$, as it minimizes the norm of their difference. Observe that, if $\sigma^2 = 0$ in the model, then the vector $(\vec{Y} - X\vec{\beta})$ is always $\vec{0}$ and if σ^2 is large, the the length of $(\vec{Y} - X\vec{\beta})$ can vary to a great extent. So intuitively, $\left\|\vec{Y} - X\hat{\vec{\beta}}\right\|$ may be somewhat related to the variance of $\vec{\epsilon}$. The denominator can be thought of as (n — number of independent estimable parameters), kind of an analogy to that of the $(n-1)$ in the denominator of sample variance. To know in much greater details about the choice of this denominator, visit Prof. Arnab Chakraborty's page.

1.2 Proof of the result

Now comes rigor. But before going through the actual proof, we need to have some prerequisites. One of them is the **Projection Map**.

1.2.1 Projection map

We have seen that $X\hat{\vec{\beta}}$ gives the "best approximation" to $\vec{Y}$. In linear algebraic terms, it means that $X\hat{\vec{\beta}}$ is the orthogonal projection of $\vec{Y}$ onto the column space of X, i.e. $\mathcal{C}(X)$ (since this idea is generalised- "perpendicular distance between a point and a line/plane is always the minimum distance between them in 3D space").

Let $\vec{y} \in \mathrm{R}^n$ and we want to break $\vec{y}$ into two orthogonal vectors where one of them will lie in $\mathcal{C}(X)$, i.e.

$$\vec{y} \;=\; \vec{y_1} + \vec{y_2} \;\;, \ni \vec{y_1} \in \mathcal{C}(X) \text{ and } \vec{y_2}^T\vec{y_1} \;=\; 0$$

We wish to define a map $\Phi_X \; : \mathrm{R}^n \to \mathcal{C}(X)$[1] which can project any vector onto $\mathcal{C}(X)$ orthogonally. Notice that this kind of map is linear and hence we can associate a matrix P_X with Φ_X. It is for the readers to check explicitly the following properties of P_X:

1. If $\vec{v} \in \mathcal{C}(X)$, then $P_X\vec{v} \;=\; \vec{v}$ (because it is natural to think this way as $\vec{v}$ itself makes the distance between it and $\mathcal{C}(X)$ equals 0)
2. P_X is symmetric and idempotent. ($\because P_X(P_X\vec{y}) \;=\; P_X\vec{y}$, $\forall\vec{y} \in \mathrm{R}^n$)
3. $\mathcal{C}(X) \;=\; \mathcal{C}(P_X)$ (use the idea of 1. to show)

Therefore, we can write $X\hat{\vec{\beta}} = P_X\vec{y}$

1.2.2 Proof

To show $\hat{\sigma^2}$ is unbiased, we have to show $\mathrm{E}(\hat{\sigma^2}) \;=\; \sigma^2$. We proceed with;

$$\begin{aligned}
\left\|\vec{y} - X\hat{\vec{\beta}}\right\|^2 &= <\vec{y} - X\hat{\vec{\beta}}, \vec{y} - X\hat{\vec{\beta}}> \\
&= (\vec{y} - X\hat{\vec{\beta}})^T(\vec{y} - X\hat{\vec{\beta}}) && \mid \text{Under the regular inner product} \mid \\
&= (\vec{y} - P_X\vec{y})^T(\vec{y} - P_X\vec{y}) && \mid \text{From the previous section} \mid \\
&= \vec{y}^T.\vec{y} - \vec{y}^TP_X\vec{y} - (P_X\vec{y})^T.\vec{y}^T + (P_X\vec{y})^T.P_X\vec{y} \\
&= \vec{y}^T.\vec{y} - \vec{y}^TP_X\vec{y} - \vec{y}^T(P_X)^T\vec{y}^T + \vec{y}^T(P_X)^TP_X\vec{y} \\
&= \vec{y}^T.\vec{y} - \vec{y}^TP_X\vec{y} - \vec{y}^TP_X\vec{y}^T + \vec{y}^TP_X(P_X\vec{y}) && \mid \text{Use symmetry of } P_X \mid \\
&= \vec{y}^T\vec{y} - \vec{y}^TP_X\vec{y} - \vec{y}^TP_X\vec{y} + \vec{y}^TP_X\vec{y} && \mid \text{By property 1 of } P_X \mid \\
&= \vec{y}^T\vec{y} - \vec{y}^TP_X\vec{y} \;=\; \vec{y}^T(I - P_X)\vec{y} \\
&= \mathrm{Tr}(\vec{y}^T(I - P_X)\vec{y}) && \mid \mathrm{Tr}() \text{ i.e. Trace of a scalar is the scalar itself} \mid \\
&= \mathrm{Tr}((I - P_X)\vec{y}\vec{y}^T) && \mid \text{For matrices } A, B, C;\ \mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) \mid
\end{aligned}$$

[1]The map is defined after X is given, though Φ_X does not depend on X

[2]A generalised result is "Trace of product of matrices is invariant under **cyclic permutations**". It may not hold for any permutations. For example, let A, B, C be three matrices. Then $\mathrm{Tr}(ABC)$= $\mathrm{Tr}(CAB)$= $\mathrm{Tr}(BCA)$. But $\mathrm{Tr}(ABC)$ may not always be equal to $\mathrm{Tr}(ACB)$.

Now, treating $\vec{y}$ as a random vector and taking expectations on both sides yield:

$$
\begin{aligned}
\mathrm{E}\left(\left\|\vec{Y} - X\hat{\vec{\beta}}\right\|^2\right) &= \mathrm{E}\left(\mathrm{Tr}((I-P_X)\vec{Y}\vec{Y}^T)\right) \\
&= \mathrm{Tr}\left(\mathrm{E}((I-P_X)\vec{Y}\vec{Y}^T)\right) && \big| \text{ Trace is just a summation } \big| \\
&= \mathrm{Tr}\left((I-P_X)\mathrm{E}(\vec{Y}\vec{Y}^T)\right) && \big| \ \mathrm{E}(A\vec{Y}) = A\mathrm{E}(\vec{Y}), \text{ where } A \text{ is not a Random variable } \big|
\end{aligned}
$$

Since,

$$
\begin{aligned}
\mathrm{E}(\vec{Y}\vec{Y}^T) &= \mathrm{E}\left(\left(X\vec{\beta}+\vec{\epsilon}\right)\left(X\vec{\beta}+\vec{\epsilon}\right)^T\right) = \mathrm{E}\left(X\vec{\beta}\vec{\beta}^T X^T + X\vec{\beta}\vec{\epsilon}^T + \vec{\epsilon}\vec{\beta}^T X^T + \vec{\epsilon}\vec{\epsilon}^T\right) \\
&= \mathrm{E}\left(X\vec{\beta}\vec{\beta}^T X^T\right) + \mathrm{E}\left(X\vec{\beta}\vec{\epsilon}^T\right) + \mathrm{E}\left(\vec{\epsilon}\vec{\beta}^T X^T\right) + \mathrm{E}\left(\vec{\epsilon}\vec{\epsilon}^T\right) \\
&= X\vec{\beta}\vec{\beta}^T X^T + X\vec{\beta}\mathrm{E}\left(\vec{\epsilon}^T\right) + \mathrm{E}\left(\vec{\epsilon}\right)\vec{\beta}^T X^T + \mathrm{E}\left(\vec{\epsilon}\vec{\epsilon}^T\right)
\end{aligned}
$$

$$
\left| \begin{array}{c} \text{If } A \text{ is not a Random variable, but just a matrix then:} \\ \mathrm{E}(A\vec{Y}) = A\mathrm{E}(\vec{Y}) \ , \ \mathrm{E}(\vec{Y}^T A) = \mathrm{E}(\vec{Y}^T)A \ , \ \mathrm{E}\left(A\right) = A \end{array} \right|
$$

$$
= X\vec{\beta}\vec{\beta}^T X^T + \sigma^2 I \qquad \left| \begin{array}{c} \text{In Gauss-Markov setup, } E(\vec{\epsilon}) = \vec{0}, \ \mathrm{Var}(\epsilon_i) = \sigma^2, \forall \text{ i} \\ \text{and } \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \ \forall i \neq j \end{array} \right|
$$

$$
\begin{aligned}
\therefore \ \mathrm{E}\left(\left\|\vec{Y} - X\hat{\vec{\beta}}\right\|^2\right) &= \mathrm{Tr}\left((I-P_X)(X\vec{\beta}\vec{\beta}^T X^T + \sigma^2 I)\right) = \mathrm{Tr}\left(X\vec{\beta}\vec{\beta}^T X^T - P_X X\vec{\beta}\vec{\beta}^T X^T + \sigma^2(I-P_X)\right) \\
&= \mathrm{Tr}\left(X\vec{\beta}\vec{\beta}^T X^T - X\vec{\beta}\vec{\beta}^T X^T + \sigma^2(I-P_X)\right) && \left| \begin{array}{c} P_X X\vec{\beta}\vec{\beta}^T X^T = P_X(X(\vec{\beta}\vec{\beta}^T X^T)) \\ \text{use property 1 of } P_X \end{array} \right| \\
&= \mathrm{Tr}\left(\sigma^2(I-P_X)\right) = \sigma^2(\mathrm{Tr}(I) - \mathrm{Tr}(P_X)) && \big| \text{ Trace is a linear function } \big| \\
&= \sigma^2(n - \mathrm{rank}(X)) && \big| \ \mathrm{Tr}(P_X) = \mathrm{rank}(P_X) = \mathrm{rank}(X) \text{ from property 3 } \big| \qquad \blacksquare
\end{aligned}
$$

Linear Algebra Corner

The result, "trace of an **idempotent matrix** equals it's rank" generally uses arguments regarding Eigen values. But there is another way using rank factorization. Let P_{nxn} be an(non-null) idempotent matrix of rank $r > 0$, otherwise it is trivial. Then by rank-factorisation, $\exists \ B_{nxr}, C_{rxn}$, with B being left invertible and C being right invertible, $\ni$

$$
P = BC \quad \Longrightarrow \quad P^2 = BCBC
$$

$$
\therefore \ BCBC = BC \qquad |\because \ P^2 = P|
$$

$$
\Longrightarrow \ CB = I_{rxr} \qquad |\text{B and C have left and right inverses respectively}|
$$

$$
\text{Hence,} \qquad \mathrm{Tr}(P) = \mathrm{Tr}(BC) = \mathrm{Tr}(CB) = \mathrm{Tr}(I_{rxr}) = \text{r}
$$