0.1 APPENDIX

0.2 Statistics Basics

To understand subject of Linear Regression Through the Origin, it is better to revise some basic probability theory and statistic notions and theorems. This section also sets a commonground for the most of the mathemtical notation that is going to be used throughout the paper.

Definition (Data) Let $(x_1,\ldots,x_n)$, where $x_i \in S$ for $i = 1,\ldots,n$. The set S is typically $\mathbb{R}$, $\mathbb{R}^d$, or it can be any abstract set. However, for our purposes, S (the sample space) will usually be $\mathbb{R}$.

Definition (Sample) In statistics, our data are often modeled by a vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of i.i.d. (independent, identically distributed) random variables, called the sample (of which size is n), where the random variables X_i take values in $\mathbb{Z}$ or $\mathbb{R}$. The common distribution of the X_i is called the parent distribution, and we say that the sample is from that parent distribution.

Definition (Model) A statistical model is a family $\{P_\theta \mid \theta \in \Theta\}$ of distributions on the sample space. When $\Theta \subset \mathbb{R}^d$, we say that we have a parametric model, and we call Θ the parameter set (space).

Definition (p-th Quantile of Data) If $p \in (0,1)$, then a p-th quantile (or a p-th percentile) of the data $(x_1,\ldots,x_n)$ is a p-th quantile of the corresponding empirical distribution function $\hat{F}_n$.

Definition (Sample mean) Let $(X_1,\ldots,X_n)$ be a sample. Then the random variable

$$\bar{X} = X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is called the sample mean.

Definition (Estimator) An estimator is a statistic (a function of the sample data) used to estimate an unknown parameter in a statistical model. An estimator for the parameter θ, denoted as $\hat{\theta}$, is any measurable function of the random variables $X_1, X_2, \ldots, X_n$.

Definition (Unbiased Estimator) If $\hat{\theta}$ is an estimator of θ, then we can define the quantity $Bias(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$. The estimator $\hat{\theta}$ is called unbiased if its bias is 0.

Definition (MSE of an Estimator) Let us have the model $\{P_\theta \mid \theta \in \Theta\}$ and let us have the sample $(X_1, \ldots, X_n)$ from it. The mean square error (or the quadratic risk) of an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ for the parameter θ is defined by

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta((\hat{\theta} - \theta)^2)$$

when θ is the true parameter.

Steiner's identity: $\mathbb{E}((X-a)^2) = \mathrm{Var}(X) + (a - \mathbb{E}(X))^2$

Interpretation in the context of mean square error (MSE):

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + (\mathrm{Bias}_\theta(\hat{\theta}))^2$$

Definition (Sufficiency) Let the model be $\{P_\theta \mid \theta \in \Theta\}$ and $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample from it. The statistic T is called *sufficient* for the parameter θ (or, for the model $\{P_\theta \mid \theta \in \Theta\}$) if the conditional distribution $P_\theta(\mathbf{X} \in \cdot \mid T = t)$ does not depend on θ.

Theorem (Neyman-Fisher Factorization Theorem) If the model is $\{p(x|\theta) \mid \theta \in \Theta\}$ where $p(x|\theta)$ is a probability mass/density function and $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample from it, then the statistic T is *sufficient* for the parameter θ if and only if we can find nonnegative functions g and h such that

$$p_{\mathbf{X}}(x|\theta) = g(T(x), \theta)h(x).$$

Definition (Likelihood) Let $\{p(x, \theta), \theta \in \Theta\}$ be a model. If the observed value of X

is x, we say that $p(x|\theta)$ is the *likelihood* of θ: $L(\theta) = p(x|\theta)$. Thus, we are considering the mass/density as a function of θ, for a fixed x. If $x = (x_1, \ldots, x_n)$ is a realization of the sample $\mathbf{X} = (X_1, \ldots, X_n)$, then $p(x|\theta)$ is the product of the marginals,

$$L(\theta) = p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

Theorem (Rao-Blackwell) Let $\{P_\theta \mid \theta \in \Theta\}$ be a model and $(X_1, \ldots, X_n)$ be a sample. Let $\hat{\theta}$ be an estimator of θ with $\mathrm{Var}_\theta(\hat{\theta})$ finite for each θ. If T is a sufficient statistic for θ, then $\theta^* = \mathbb{E}_\theta(\hat{\theta}|T)$ is a statistic, and we have for all θ that

$$\mathrm{MSE}_\theta(\theta^*) \leq \mathrm{MSE}_\theta(\hat{\theta}) \quad (1)$$

and the inequality is strict unless $\hat{\theta}$ is a function of T with probability 1.

0.3 Simple Linear Regression

In simple linear regression we have a random sample $((x_1,y_1),...,(x_n,y_n))$ of observed points. And our goal is to find a linear function $y = \beta_0 + \beta_1 x$ that describes the relationship between two variables in our sample as accurately as possible. The accuracy in this paper will be measured using least squares method. Namely,

If x_i is a predictor variable from our sample, then we say that $\hat{y}_i = \beta_0 + \beta_1 x_i$ is a fitted value for response variable y_i. We then can model the relationship between those two using the following formula:

$$y_i = \hat{y}_i + \varepsilon_i$$

Then the objective is to minimize the sum of squares of error terms $\varepsilon_i,\ i = 1,...,n$.

and our job is to find a model that describes the relationship between two two variables x and y

Consider the model function:

$$y = \beta_0 + \beta_1 x,$$

which describes a line with slope β_1 and y-intercept β_0. In general, such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; we call the unobserved deviations from the above equation the errors. Suppose we observe n data pairs and call them $\{(x_i,y_i), i = 1,\ldots,n\}$. We can describe the underlying relationship between y_i and x_i involving this error term ε_i by:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

This relationship between the true (but unobserved) underlying parameters β_0 and β_1

and the data points is called a linear regression model.

The goal is to find estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 which would provide the "best" fit in some sense for the data points. As mentioned in the introduction, in this paper the "best" fit will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals $\hat{\varepsilon}_i$ (differences between actual and predicted values of the dependent variable y), each of which is given by, for any candidate parameter values β_0 and β_1:

$$\hat{\varepsilon}_i = y_i - \beta_0 - \beta_1 x_i.$$

In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ solve the following minimization problem:

$$\text{Find } \min_{\beta_0,\beta_1} Q(\beta_0,\beta_1), \quad \text{for } Q(\beta_0,\beta_1) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

We assume that the response variable is normally distributed as follows:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \ldots, n$$

such that the Y_i are independent and β_0, β_1, and σ are unknown parameters.

If $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, and we let $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$, then it is easy to see that

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$