Simple Linear Regression

Definition. ***Method of Least Squares*** *is a method for determining parameters in curve-fitting problems, where we want to predict dependent variable Y from X. Consider fitting simpliest model to data*

$$Y = \beta_0 + \beta_1 X$$

Denote ith residual as

$$e_i = y_i - \beta_0 - \beta_1 x_i = y_i - \hat{y}_i$$

We want to minimize e_i as small as possible. The ***least squares estimators*** *of β_0 and β_1 are the minimizers of the* ***residual sum of squares***

$$RSS(\beta_0, \beta_1) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

In other words we choose a linear fit $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ such that

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Proposition. *The least squares estimators of β_0 and β_1 are given by*

$$\begin{cases} \hat{\beta}_1 = \dfrac{S_{XY}}{S_X^2} \\ \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \end{cases}$$

where $S_{XY} = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ is the ***sample covariance*** *of X and Y and $S_X^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ is the* ***sample variance*** *of X.*

Lemma. ***some useful properties in proving previous proposition and facilitates computation***

1. $$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

2. $$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - 2n\overline{xy} + n\overline{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

Definition. ***The normal equations*** *Denote $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \cdots, n$ the residue is hence $e_i = y_i - \hat{y}_i$. The residuals of the least square fit satisfy*

1.

$$0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i$$

2.

$$0 = \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = \sum_{i=1}^{n} x_i e_i$$

which is derived from the process of finding least squared estimators, when we are taking first order partial with respect to β_0 *and* β_1

Definition. ***Standard Statistical Model*** *stipulates that the observed value of* y *is a linear function of* x *plus random noise*

$$y(x) = \beta_0 + \beta_1 x + \epsilon(x)$$

where x *is not a random variable, and* $y(x)$ *is a random through inclusion of random noise* $\epsilon(x)$ *where we assume*

1. *by Normal equation*

$$\mathbb{E}(\epsilon(x)) = 0 \text{ for all } x$$

2. *Noise at different* x *are uncorrelated and variance around the regression line is same for all values of* x *(homoscedasticity)*

$$Cov(\epsilon(x), \epsilon(x')) = \begin{cases} \sigma^2 & x = x' \\ 0 & x \neq x' \end{cases}$$

Hence the model can be denoted as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, n$$

with

$$\mathbb{E}[\epsilon_i] = 0 \text{ for all } i \qquad Var(\epsilon_i) = \sigma^2 \text{ for all } i \qquad \mathbb{E}[\epsilon_i \epsilon_j] = 0 \text{ for } i \neq j$$
$$\text{hence} \quad Var[y_i] = 0 \qquad Cov(y_i, y_j) = 0 \text{ for } i \neq j$$

Definition. ***LS estimator as linear estimator*** *Let* $y_1, \cdots, y_n \sim f_\theta$*. Any estimator of* θ *of the form*

$$\hat{\theta} = \sum_{i=1}^{n} c_i y_i$$

is called a linear estimator (linear combination of observations)

1. $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are linear estimators.*

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x}) y_i}{\sum_j (x_j - \overline{x})^2} = \sum_i a_i y_i \qquad \textit{where} \qquad a_i = \frac{x_i - \overline{x}}{\sum_j (x_j - \overline{x})^2}$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_i y_i - \overline{x} \frac{\sum_i (x_i - x) y_i}{\sum_j (x_j - \overline{x})^2} = \sum_i b_i y_i \qquad \textit{where} \qquad b_i = \frac{1}{n} - \frac{\overline{x}(x_i - x)}{\sum_j (x_j - \overline{x})^2}$$

2. *In fact, they are also unbiased estimators, i.e.* $\mathbb{E}[\hat{\beta}_0] = \beta_0$ *and* $\mathbb{E}[\hat{\beta}_1] = \beta_1$

3.

$$Var[\hat{\beta}_1] = Var\left[\sum_i a_i y_i\right] = \sigma^2 \sum_i a_i^2 == \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2}$$

$$Var[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}\right] \qquad \textit{and} \qquad Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \overline{x}}{\sum_i (x_i - \overline{x})^2}$$

4. *Unbiased estimator of noise variance is given by*

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Definition. ***The Gauss-Markov Theorem*** *Under standard model assumptions, no linear unbiased estimator of* β_0 *(*β_1*) has a smaller variance than the least squares estimator* $\hat{\beta}_0$ *(*$\hat{\beta}_1$*).*

Remark. This shows that least square estimators are the **best linear unbiased estimator (BLUE)**

Definition. ***Correlation Coefficient*** *The correlation coefficient of random variables* X *and* Y *is*

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X *and* σ_Y *are the standard deviations of* X *and* Y*, respectively.*

1. $|\rho_{XY}| \leq 1$*, where equality holds iff* X *and* Y *are perfect linear function of one another. In other words,* $|\rho_{XY}|$ *is a measure of linear relationship between* X *and* Y

2. ***Sample Correlation Coefficient*** *is defined to be*

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

which shares the above property and happens to be a consistent estimator of ρ_{XY}

Definition. ***Explained Variation*** *Variation in value of Y is the* ***Total Sum of Squares***

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Now

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad \text{(by Normal Equation)} \\ TSS &= RSS + ESS\end{aligned}$$

the ***Proportion of explained variance*** *is defined to be*

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \qquad \text{and} \qquad R^2 = r_{XY}^2$$

R^2 is an indication of good linear fit

Statistical Inference under Gaussian Noise

Definition. *A linear model with following assumptions*

1. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ *where* $i = 1, \cdots, n$
2. $\mathbb{E}[\epsilon_i] = 0$ *for* $i = 1, \cdots, n$
3. $Var(\epsilon_i) = \sigma^2$ $i = 1, \cdots, n$ *(homoscedastic) and* $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ *for* $i \neq j$ *(uncorrelated)*
4. *distribution of ϵ_i is normal for $i = 1, \cdots, n$*
5. *Uncorrelated normal random variable is independent.*

allows for ***statistical inference****, i.e. hypothesis testing, calculate confidence interval etc. An unbiased estimator of noise variance σ^2 is given by*

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \qquad \text{and} \qquad \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2$$

Since ϵ_i is normal, we have

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear estimators, i.e. of the form $\hat{\beta} = \sum_i c_i y_i$, then we derive

Regression coefficients

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2})$$

where σ^2 is variance of error. Under normality

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_i (x_i - \overline{x})^2}}} \sim \mathcal{N}(0, 1)$$

Now we replace unknown σ^2 with unbiased estimator $S^2 = \dfrac{1}{n-2} \sum_i e_i^2$ we have

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{\sum_i (x_i - \overline{x})^2}}} \sim t_{n-2}$$

We can then do hypothesis tests on the slope coefficient $\hat{\beta}_1$

Definition. ***Hypothesis tests on the slope β_1 to evaluate correlation*** *Testing $\mathcal{H}_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, then, by*

$$\mathcal{T} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \overset{H_0}{\sim} t_{n-2} \qquad \textit{where} \qquad s_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2}\sum_i e_i^2}{\sum_i (x_i - \overline{x})^2}}$$

and a $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} \frac{S}{\sqrt{\sum_i (x_i - \overline{x})^2}}$$

Definition. ***Confidence interval for mean response*** *Want to estimate mean response*

$$\mu(x_0) := \mathbb{E}[y(x_0)] = \beta_0 + \beta_1 x_0$$

The prediction at x_0 may be used as an estimator

$$\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

we have

$$\mathbb{E}[\hat{y}(x_0)] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \mathbb{E}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0 = \mu(x_0)$$
$$Var[\hat{y}(x_0)] = Var[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right\}$$

Since least square estimators are linear estimators we can write

$$\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \sum_i b_i y_i + x_0 \sum_i a_i y_i = \sum_i (b_i + x_0 a_i) y_i$$

hence $\hat{y}(x_0)$ *has normal distribution so then,*

$$\hat{y}(x_0) \sim \mathcal{N}\left(\mu(x_0), \sigma^2 \left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right\}\right)$$

Now we replace σ^2 *with* $S^2 = \dfrac{1}{n-2}\sum_i e_i^2$ *would result in* t *distribution*

$$\hat{y}(x_0) \pm t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

is a $100(1-\alpha)\%$ *confidence interval for mean response* $\mathbb{E}[y(x_0)]$, *where* $\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$. *As* x_0 *vary, we have a confidence band centered at the least squares fit, that get narrower as* x_0 *draws closer to* $\overline{x}$

Definition. ***Least Square Estimators under standard normal model are MLEs*** *Under additional assumption that random noise is Gaussian, the least squares estimators* $\hat{\beta}_0^{LS}$ *and* $\hat{\beta}_1^{LS}$ *are maxmum likelihood estimators of* β_0 *and* β_1, *respectively.*

Diagnostic Plot

Definition. *Linear regressio model relies heavily on assumptions about random errors* ϵ_i. *the residual* e_i *should be*

1. *normal*
2. *independent*
3. *homoscedasticitic*
4. *distribution of* ***standardized residuals*** $\dfrac{e_i}{S} \sim \mathcal{N}(0,1)$

Two plots are given

1. ***Residuals vs Fitted Value Plot*** *plot of* e_i *vs.* $\hat{y}_i$.
   (a) *Symmetry about 0, with homogeneity of the noise variance (homoscedastic), and no trends or pattern implies a good fit for linear models*
   (b) *Streaks of positive/negative residual indicates observation is correlated, violating the independence assumption*

(c) *The trend resembles an upward or downward curve indicates model misspecification. The assumption of linearity is violated*

(d) *Increasing variance along the dependent $\hat{y}_i$ axis violates homoscedasticity assumption*

2. ***Quantile-Quantile Plot*** *A plot for comparing two probability distributions by plotting their quantiles against each other. In evaluating good fit for linear model we plot sample quantiles of standardized residues vs. theoretical quantiles of standard normal distribution.*

   (a) *If points approximately lie on the line $y = x$, then the distribution in comparison are similar, i.e.* $\dfrac{e_i}{S} \sim \mathcal{N}(0, 1)$.

   (b) *If lower quantiles are too small and upper quantiles are too large - a heavy-tailed noise. Perhaps assuming t distribution, thus violating the normality assumption*