Section 3D. Coefficient Uncertainty

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Coefficients Accuracy

- Assume that we observe data generated by an additive linear model of the form

$$\mathbf{x}_i \sim f_X \text{ and } \mathbf{y} = M_X\beta + \varepsilon, \text{ where } \varepsilon_i \sim f_\varepsilon$$

- We can estimate the linear coefficients as $\widehat{\beta} = \left(M_X^\intercal M_X\right)^{-1} M_X^\intercal \mathbf{y}$. Note that, since the datapoints $(\mathbf{x}_i, y_i)$ are random, *the coefficients* $\widehat{\beta}_i$ *are random variables themselves!*
- **Mean analysis**: One can prove that $\mathbb{E}\left[\widehat{\beta}_i\right] = \beta_i$ for all i (unbiased estimator)
- **Covariance analysis**: Assuming that we are given a data matrix M_X, the covariance matrix of $\widehat{\beta}$ satisfies (without proof)

$$\mathrm{Cov}\left[\widehat{\beta}|M_X\right] = \sigma^2\left(M_X{}^\intercal M_X\right)^{-1}$$

Coefficients Accuracy: Univariate Case

- For the particular case $p = 1$, the linear regression takes the form $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$
- The diagonal elements of the covariance matrix are the variances of the estimated coefficients. These variances are:

$$\mathrm{SD}\left(\widehat{\beta}_1\right)^2 = \mathrm{Var}\left[\widehat{\beta}_1 \middle| \{x_i\}_{i=1}^N\right] = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \overline{x})^2}$$

$$\mathrm{SD}\left(\widehat{\beta}_0\right)^2 = \mathrm{Var}\left[\widehat{\beta}_0 \middle| \{x_i\}_{i=1}^N\right] = \sigma^2 \left(\frac{1}{N} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$$

with $\sigma^2 = \mathrm{Var}(\varepsilon)$, $\overline{y} = \frac{1}{N}\sum_{i=1}^N y_i$ and $\overline{x} = \frac{1}{N}\sum_{i=1}^N x_i$. In this course, we use $\mathrm{SD}(\cdot)$ to denote the Standard Deviation of a r.v.

Coefficients Accuracy: Numerical Example

Numerical validation: Let us consider an unrealistic, but illustrative situation

- Consider 1,000 independent realizations of a training dataset, $\mathcal{D}_{\text{Tr}}^k$ for $k = 1, 2, \ldots, 1000$. Notice that, in practice, we will only have access to a single realization of a training dataset!
- Each dataset contains $N = 100$ sample points $\mathcal{D}_{\text{Tr}}^k = \left\{ \left(\mathbf{x}_1^k, y_1^k\right), \ldots, \left(\mathbf{x}_{100}^k, y_{100}^k\right) \right\}$
- For each $\mathcal{D}_{\text{Tr}}^k$, we compute the corresponding estimates $\widehat{\beta}_0^k$ and $\widehat{\beta}_1^k$ for $k = 1, 2, \ldots, 1000$

Coefficients Accuracy: Numerical Example

In this example, $\beta_1 = \mathbb{E}\left[\widehat{\beta}_1\right] = 0.5$ and $\mathrm{Var}\left[\widehat{\beta}_1 \middle| \{x_i\}_{i=1}^N\right] = 0.05^2$.

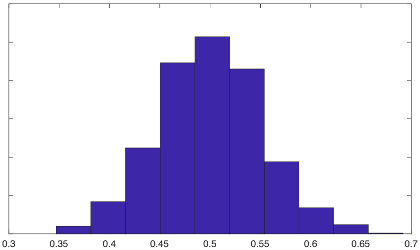

Figure: Histogram of the values of $\widehat{\beta}_1^k$. We have used a linear model $Y = 1 + 0.5X + \varepsilon$, where $X \sim \mathcal{N}(0, 1)$ and $\mathrm{Var}(\varepsilon) = \sigma^2 = 1$.