GU 5205/GR 4205: Linear Regression Models	Fall 2020

Lecture 2 — September 14, 2020

Prof. Xiaofei Shi *Scribe: Aiden Kenny, Zhanghao Zhang*

Regression analysis

- *Regression*: a statistical method used to study the dependence between variables in the presence of noise.
- *Linear regression*: a statistical method used to study *linear* dependence between variables in the presence of noise.
- e.g. Here is a linear function
$$Y = 104 - 14X.$$
Because the intercept is negative, we say there is a *negative* linear relationship.

(Simple) Linear Regression Procedure

1. *Estimate your model*
    - We have two random variables: a predictor X and a response Y.
    - We assume a *linear relationship*: $Y = \beta_0 + \beta_1 X + \varepsilon$.
    - We want to find β_0, β_1 to minimize $\mathbb{E}\left[(Y - \beta_0 - \beta_1 X)^2\right]$.
2. *Estimate your model*
    - We have n distinct observations of each predictor and response, $(x_1, y_1), \dots, (x_n, y_n)$.
    - We use this observed data to minimize
    $$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$
    - This is a function of β_0, β_1, so we find the values that *minimize* Q.
3. *Understand your model*
    - The simple linear regression model has several useful properties (homework 1).
    - Can use your model to make predictions and inferences.

Theoretical Example: Bivariate Noraml Distrubution

Here we have $(X, Y) \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$, where $X \sim \mathrm{N}(0,1)$, $Y \sim \mathrm{N}(0,1)$, and $\rho = \mathrm{Cov}(X, Y)$.

We want to find a distribution of ε (i.e. $\varepsilon \sim \mathrm{N}(?, ?)$) such that ϵ and X are independent and $Y = X\rho + \varepsilon$. Note then that $\varepsilon = Y + X\rho$.

Because of this, we can show that:

1. $\mathbb{E}[\varepsilon] = 0$. Since $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, we have $\mathbb{E}[\varepsilon] = \mathbb{E}[Y + X\rho] = \mathbb{E}[Y] + \rho\mathbb{E}[X] = 0$.
2. $\text{Var}[\varepsilon] = 1 - \rho^2$. Proof is similar.

We can generalize this to the case where $X \sim \text{N}(\mu_X, \sigma_X)$ and $Y \sim \text{N}(\mu_Y, \sigma_Y)$, so they are no longer standard normal. Then $(X, Y) \sim \text{N}\left(\begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}, \begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right)$.

Useful trick: *normalization*! If $\tilde{X} = \frac{X - \mu_X}{\sigma_X}$ and $\tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}$, then both $\tilde{X}$ and $\tilde{Y}$ are standard normal, and $\text{Cov}(\tilde{C}, \tilde{Y}) = \rho$.

We've essentially reduced more general case back to simple case, so $\tilde{Y} = \tilde{X}\rho + \tilde{\varepsilon}$. Writing this in terms of X and Y gives us $Y = (X - \mu_X)\,\rho\frac{\sigma_Y}{\sigma_X} + \tilde{\varepsilon}\sigma_Y + \mu_Y$.

Define $\varepsilon = \tilde{\varepsilon}\sigma_Y$. If we compare this result to the linear regression model $\beta_0 + \beta_1 X + \varepsilon$, we have

- $\beta_0 = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$
- $\beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$ (and so $\beta_0 = \mu_Y - \beta_1\mu_X$)

UMVUL for bivariate normal

From before, we had $\beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$ and $\beta_0 = \mu_Y - \beta_1\mu_X$. This is known as the **ground truth**.

We can get **estimates** for these parameters as $\hat{\beta}_1 = \frac{\text{Cov}(X,Y)}{\text{Var}[X]}$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$. These estimators have several properties:

1. *Uniformity*: converges to the ground truth, i.e. $\lim_{n\to\infty}\hat{\beta}_1 - \beta_1$ and $\lim_{n\to\infty}\hat{\beta}_0 - \beta_0$.
2. *Unbiased*: $\mathbb{E}[\hat{\beta}_1|X] = \beta_1$ and $\mathbb{E}[\hat{\beta}_0|X] = \beta_0$.
3. *Linear*.
4. *Minimum Variance*

Adding Data

$$\beta_1 = \rho\frac{\sigma_Y}{\sigma_X} \to \hat{\beta}_1 = \frac{Cov(X,Y)}{Var|X|}$$
$$\beta_0 = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X \to \hat{\beta}_0 = Y - X\hat{\beta}_1$$

1. Recall the estimation for μ_X, μ_Y
   Using data: $\hat{\mu}_X = \bar{x}, \hat{\mu}_Y = \bar{y}$
2. And the estimation for σ_X, σ_Y
   Using data: $\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2, \sigma_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$
   In matrix form: $\hat{\sigma}_X^2 = \frac{1}{n-1}||x - \bar{x}1_n||^2, \hat{\sigma}_Y^2 = \frac{1}{n-1}||y - \bar{y}1_n||^2$

3. As well as for $Cov(X,Y) \rightarrow \hat{Cov}(X,Y) = \frac{1}{n-1}(x - \bar{x}1_n)^T(y - \bar{y}1_n)$
   $\rho = Cov(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \rightarrow \hat{\rho} = \frac{\hat{Cov}(X,Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$

Q: I see that your unbiased estimator use n - 1 instead of n, is that because it is a sample instead of the population?
When data is big ($n > 40$), then the difference is negligible.

More general case...

1. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be samples from the same model
2. If the SLR model holds, we write $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$
3. Here, ϵ_i satisfies $E[\epsilon_i] = 0$ and $E[\epsilon_i\epsilon_j] = \sigma^2\delta_{ij}$
4. Observations: predictor: $x_1, x_2, \ldots, x_n$ response: $y_1, y_2, \ldots, y_n$
5. Preference: $Q = \sum_{i=1}^n (y_i - \beta_0 - x_i\beta_1)^2$
6. Model parameters: $\beta_0, \beta_1(, \sigma^2)$

General Methodology

1. Preference + data $\rightarrow$ Q = Q(model paramters; data)
2. Estimation of model parameters $\leftrightarrow$ Minimizing Q wrt model parameters $\rightarrow$ Taking partial derivatives of Q wrt model parameters and sent them to 0!

$$
\begin{aligned}
Q =& Q(\beta_0, \beta_1 | (x_1, \ldots, x_n), (y_1, \ldots, y_n)) \\
=& \sum_{i=1}^n (y_i - \beta_0 - x_i\beta_1)^2 \\
=& \sum_{i=1}^n (y_i^2 + \beta_0^2 + x_i^2\beta_1^2 - 2\beta_0 y_i - 2x_i y_i \beta_1 + 2x_i\beta_0\beta_1) \\
=& (\sum_{i=1}^n y_i^2) + n\beta_0^2 + (\sum_{i=1}^n x_i^2)\beta_1^2 - 2\beta_0(\sum_{i=1}^n y_i) - 2(\sum_{i=1}^n x_i y_i)\beta_1 + 2(\sum_{i=1}^n x_i)\beta_0\beta_1
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_0} =& 2(\sum_{i=1}^n x_1)\beta_1 + 2n\beta_0 - 2(\sum_{i=1}^n y_i) = 0 \\
\frac{\partial Q}{\partial \beta_1} =& 2(\sum_{i=1}^n x_i)\beta_0 + 2(\sum_{i=1}^n x_i^2)\beta_1 - 2(\sum_{i=1}^n x_i y_i) = 0
\end{aligned}
$$

$$\rightarrow \begin{cases} 2nb_0 - 2i_n^T y + 21_n^T x b_1 = 0 \\ 2x^T x b_1 + 21_n^T x b_0 - 2x^T y = 0 \rightarrow 2(x - \bar{x}1_n)^T (x - \bar{x}1_n) b_1 - 2(x - \bar{x}1_n)^T (y - \bar{y}1_n) = 0 \end{cases}$$

$$\rightarrow \begin{cases} b_0 = \bar{y} - \bar{x} b_1 \\ b_1 = \frac{(x-\bar{x}1_n)^T (y-\bar{y}1_n)}{||x-\bar{x}1_n||^2} \end{cases}$$

Prediction and residual

1. Prediction: $\hat{y}_i = b_0 + x_i b_1$
2. Residual: $e_i = y_i - \hat{y}_i = y_i - b_0 - x_i b_1$
3. Residuals can be viewed as the estimation of unobservable error terms

$$\hat{\epsilon}_i = e_i = y_i - \hat{y}_i = y_i - b_0 - x_i b_1$$

4. Estimation of $\hat{\sigma}^2 = \mathbf{MSE} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{||y-\hat{y}||^2}{n-2}$

Recall Q doesn't have σ^2, where $\sigma^2 = E[\epsilon_i^2]$ is level of noise.