MULTIVARIATE STATISTICS, Lesson 8-9.
Multivariate Regression and ANalysis Of VAriance

- **Linear regression**: given the expectation vector and covariance matrix of the random vector $(Y, X_1, \dots, X_p)^T$, we want to approximate Y (target variable) with a linear combination of the predictor variable $\mathbf{X} = (X_1, \dots, X_p)^T$ in such a way that

$$\mathbb{E}(Y - \ell(\mathbf{X}))^2 = \mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b))^2 = \mathbb{E}(Y - \mathbf{a}^T \mathbf{X} - b)^2 \to \textit{minimum}$$

in the parameters $\mathbf{a} = (a_1, \dots, a_p)^T$ and b.

1. *Theoretical solution*: $\mathbf{a} = \mathbf{C}^{-1}\mathbf{d}$, $b = \mathbb{E}(Y) - \mathbf{a}^T \mathbb{E}(\mathbf{X})$,where $\mathbf{C} = \mathbb{D}^2(\mathbf{X}) > 0$ and $\mathbf{d}$ contains the cross-covariances between Y and X_i's. If the expectations are 0's, $\ell(\mathbf{X})$ and the error term $\varepsilon := Y - \ell(\mathbf{X})$ are uncorrelated, hence

$$\mathbb{D}^2(Y) = \mathbb{D}^2 l(\mathbf{X}) + \mathbb{D}^2 \varepsilon = r_{Y\mathbf{X}}^2 \cdot \mathbb{D}^2(Y) + (1 - r_{Y\mathbf{X}}^2) \cdot \mathbb{D}^2(Y),$$

where $r_{Y\mathbf{X}} = corr(Y, \ell(\mathbf{X}))$ is the **multiple correlation** between Y and $\mathbf{X}$. If $Y_1 = \ell_1(\mathbf{X}) + \varepsilon_1$ and $Y_2 = \ell_2(\mathbf{X}) + \varepsilon_2$, then the **partial correlation** between Y_1 and Y_2 after eliminating the effect of $\mathbf{X}$ is

$$r_{Y_1 Y_2 | \mathbf{X}} = corr(\varepsilon_1, \varepsilon_2) = \frac{corr(Y_1, Y_2) - corr(Y_1, X) \cdot corr(Y_2, X)}{\sqrt{(1 - corr^2(Y_1, X)) \cdot (1 - corr^2(Y_2, X))}},$$

where the last formula is applicable only in the $p = 1$ case.

2. *Linear model*: $\mathbf{Y} = \mathbf{X}\mathbf{a} + \underline{\varepsilon}$, where measurements $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ are random variables due to the independent, zero expectation, homoscedastic measurement errors contained in $\underline{\varepsilon} := (\varepsilon_1, \dots, \varepsilon_n)^T$ (the errors have the same variance σ^2). The errors are added to a linear combination of given measurement points of the $n \times p$ matrix $\mathbf{X}$. The parameter vector $\mathbf{a} \in \mathbb{R}^r$ minimizing $\|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2$ is obtained by solving the **Gauss normal equations**: $\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{Y}$. The solution is: $\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, if $\text{rang}\,(\mathbf{X}) = p < n$, otherwise we need generalized inverse: $\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{Y}$.

3. *Proclaim*: in the linear model, if $\text{rang}\,(\mathbf{X}) = p < n$ and $\underline{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, then $\hat{\mathbf{a}} \sim \mathcal{N}_p(\mathbf{a}, \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1})$. In this case, $\hat{\mathbf{a}}$ is also ML-estimate of $\mathbf{a}$.

4. **Gauss –Markov Theorem**: for any other unbiased linear estimate $\tilde{\mathbf{a}}$ of $\mathbf{a}$: $\mathbb{D}^2(\hat{\mathbf{a}}) \le \mathbb{D}^2(\tilde{\mathbf{a}})$, that is $\hat{\mathbf{a}}$ is **BLUE** (Best, Linear, Unbiased) Estimate of $\mathbf{a}$.

5. *Testing the significance of the regression.* Under $H_0 : \mathbf{a} = \mathbf{0}$

$$F = \frac{SSR}{SSE} \cdot \frac{n-p}{p} = \frac{SST - SSE}{SSE} \cdot \frac{n-p}{p} = \frac{R^2}{1 - R^2} \cdot \frac{n-p}{p} \sim \mathcal{F}(p, n-p),$$

where SSR, SSE, and R are sample estimates of $\mathbb{D}^2(\ell(\mathbf{X}))$, $\mathbb{D}^2(\varepsilon)$ and $r_{Y\mathbf{X}}$.

- **ANOVA models**:

1. *One-way ANOVA*: $X_{ij} = m + a_i + \varepsilon_{ij} \qquad (j = 1, \dots, n_i;\ i = 1, \dots, k)$
2. *Two-way ANOVA without interaction*: 1 observation/cell

$$X_{ij} = m + a_i + b_j + \varepsilon_{ij}, \qquad (i = 1, \dots, k;\ j = 1, \dots, p)$$

3. *Two-way ANOVA with interaction*: n observations/cell

$$X_{ijl} = m + a_i + b_j + c_{ij} + \varepsilon_{ijl}, \qquad (i = 1, \dots, k;\ j = 1, \dots, p;\ l = 1, \dots, n)$$

4. *Three-way ANOVA without interaction: latin squares.*