LM

STAT 331 - Applied Linear Models

Prof. Glen McGee

Contents

1. Review and Simple Linear Regression

1.1 Simple Linear Regression

Definition 1.1.1 Let X, Y be a continuous random variable that models two populations with pdf's $f_X, f_Y : y \mapsto [0,1]$ and support A for Y.

1. The **population mean** of Y is

$$E(Y) = \int_A y f(y) dy.$$

2. The **sample mean** of Y given a sample $y_1, \ldots, y_n$ is

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

3. The **population variance** of Y is

$$\text{Var}(Y) = E((Y - E(Y))^2).$$

4. The **sample variance** of Y given a sample $y_1, \ldots, y_n$ is

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

5. The **population covariance** of X and Y is

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))).$$

6. The **sample covariance** for respective samples from the populations modeled by X and Y, $(x_1,y_1),\ldots,(x_n,y_n)$, is

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

7. The **population correlation** between X and Y is

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

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ and $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$.

8. The **sample correlation** for observations $(x_1,y_1),\ldots,(x_n,y_n)$ is

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

where s_{xy} is the sample covariance, s_x^2 is the sample variance of $x_1,\ldots,x_n$, and s_y^2 is the sample variance of $y_1,\ldots,y_n$.

Proposition 1.1.2

1. Suppose X,Y are rv's, then

$$\mathrm{Var}(Y) = E(Y^2) - E(Y)^2$$

and

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y).$$

2. Suppose $Y_1,\ldots,Y_m$ are rv's and a_i,b_i, $1\le i\le m$ are constants, then

$$E\left(\sum_{i=1}^{m} a_iY_i + b_i\right) = \sum_{i=1}^{m} a_iE(Y_i) + \sum_{i=1}^{m} b_i.$$

3. Suppose Y is a rv and $a,b\in\mathbb{R}$, then

$$\mathrm{Var}(aY+b) = a^2\,\mathrm{Var}(Y).$$

4. Suppose X,Y are rv's and $a,b,c,d\in\mathbb{R}$, then

$$\mathrm{Cov}(aY+c,bX+d) = ab\cdot\mathrm{Cov}(X,Y).$$

5. Suppose X,Y,U,V are rv's, then

$$\mathrm{Cov}(U+V,X+Y) = \mathrm{Cov}(U,X)+\mathrm{Cov}(X,Y)+\mathrm{Cov}(V,X)+\mathrm{Cov}(V,Y).$$

6. Suppose X,Y are rv's, then

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

and

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y).$$

Proof. 1. STAT330(S) Thm. 1.7.6 and STAT330(S) Thm. 2.6.8.
2. STAT330(S) Thm. 1.7.4.
3. STAT330(S) Thm. 1.7.6.
4. For convenience write $\text{Cov}(aY+c, bX+d)$ as $\text{Cov}(Z)$. By (1),

$$\text{Cov}(Z) = E((aY+c)(bX+d)) - E(aY+c)E(bX+d),$$

which, by (2), yields

$$\begin{aligned}\text{Cov}(Z) &= E(abXY+cbX+adY+cd) - (aE(Y)+c)(bE(X)+d)\\ &= abE(XY)+bcE(X)+adE(Y)+cd-(aE(Y)+c)(bE(X)+d)\\ &= abE(XY)-abE(X)E(Y)\\ &= ab\,\text{Cov}(X,Y)\end{aligned}$$

where the last line comes from (1).
5. By (1),

$$\begin{aligned}&\text{Cov}(U+V, X+Y)\\ =&E((U+v)(X+Y)) - E(U+V)E(X+Y)\\ =&E(UX+UY+VX+VY) - (E(U)+E(V))(E(X)+E(Y))\\ =&E(UX)-E(U)E(X)+(E(UY)-E(U)E(Y))+(E(VX)-E(V)E(X))+(E(VY)-E(V)E(Y))\\ =&\text{Cov}(U,X)+\text{Cov}(U,Y)+\text{Cov}(V,X)+\text{Cov}(V,Y)\end{aligned}$$

where the last line again comes from (1).
6. STAT330(S) Thm. 2.6.10.

■

Definition 1.1.3 The **Gamma function** is

$$\begin{aligned}\Gamma : (0,\infty) &\to \mathbb{R}\\ \alpha &\mapsto \int_0^\infty y^{\alpha-1}e^{-y}dy\end{aligned}$$

Proposition 1.1.4 Let $\alpha > 1$, $n \in \mathbb{N}$, then
1. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$.
2. $\Gamma(n) = (n-1)!$.
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proof. STAT330(S) Proposition 1.4.6. ■

Definition 1.1.5 1. Let $\mu, \sigma \in \mathbb{R}$, Z be a continuous rv with pdf

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right), z \in \mathbb{R},$$

then Z is said to follow a **normal distribution**, written as $Z \sim N(\mu, \sigma^2)$.
2. Let $k \in \mathbb{N}$ and X be a continuous rv with pdf

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}, x > 0,$$

then X is said to follow a **chi-squared distribution with k degrees of freedom**, written

as $X \sim \chi^2_k$ or sometimes $X \sim \chi^2(k)$.

3. Let $k \in \mathbb{N}$, and suppose Y is a continuous rv with pdf

$$f(y) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{y^2}{k}\right)^{-\frac{k+1}{2}}, y \in \mathbb{R},$$

then Y is said to follow a **student t-distribution with k degrees of freedom**, and we write $Y \sim t_k$ or sometimes $Y \sim t(k)$.

Proposition 1.1.6

1. If $Z \sim N(\mu, \sigma^2)$, then

$$E(Z) = \mu, \mathrm{Var}(Z) = \sigma^2.$$

2. If $X \sim \chi^2_k$ for some $k \in \mathbb{N}$, then

$$E(X) = k, \mathrm{Var}(X) = 2k.$$

3. If $Y \sim t_k$ for some $k \in \mathbb{N}$, then

$$E(Y) = 0 \text{ if } k = 2,3,\ldots$$

$$\mathrm{Var}(Y) = \frac{k}{k-2} \text{ if } k = 3,4,\ldots$$

and otherwise $E(Y)$ and $\mathrm{Var}(Y)$ do not exist.

Proof. We omit the proof. ■

Proposition 1.1.7

1. If $Z_i \sim N(\mu_i, \sigma_i^2)$, $\alpha_i \in \mathbb{R}$, $1 \le i \le n$, then

$$U := \sum_{i=1}^n \alpha_i Z_i \sim N\left(\sum_{i=1}^n \alpha_i \mu_i, \sum_{i=1}^n \alpha_i^2 \sigma_i^2\right).$$

2. If $Z_i \sim N(0,1)$ independently, $1 \le i \le n$, then

$$X := \sum_{i=1}^n Z_i^2 \sim \chi^2_n.$$

3. If $Z \sim N(0,1)$ and $X \sim \chi^2_k$ independently for some $k \in \mathbb{N}$, ten

$$\frac{Z}{\sqrt{\frac{X}{k}}} \sim t_k.$$

Proof.

1. STAT330(S) Thm. 3.3.2.
2. STAT330(S) Proposition. 3.3.5.
3. STAT330(S) Lemma. 3.3.9.

■

Remark 1.1.8 In regression, we are interested in predicting an observed random variable y given an observed random variable x. To quantify any relationships between x and y, we may use simple linear regression.

We use the convention that y is the variable we are interested in predicting, and x is the "given" observed random variable.

We usually call y to be the **independent variable** and x to be the **dependent variable**.

Definition 1.1.9 Let $Y := (Y_1, \dots, Y_n)$ be a random vector and $x := (x_1, \dots, x_n)$ be an observed random vector that represents dependent variables. A **simple linear regression** is a group of n functions

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, 1 \leq i \leq n$$

for some unobserved $\beta_0, \beta_1 \in \mathbb{R}$ and $\varepsilon_i \sim N(0, \sigma^2)$ (called **error terms**) independently (with respect to i) for some unobserved σ^2.

In matrix form, a simple linear regression is written as

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

Remark 1.1.10 As a consequence of the above definition, the random observations Y_i has the distribution

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

independently. This in turn yields

$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i.$$

The underlying assumptions of this model are

1. linearity between X and Y
2. independent distribution of the error terms
3. normality of the error terms distribution
4. (homoskedasticity) equal variance of Y_i's with respect to i.

Theorem 1.1.11 Let X, Y be rv's and assume a simple linear regression model

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

then the maximum likelihood estimates for β_0, β_1, and σ^2 are

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

where (x_i, y_i), $1 \leq i \leq n$, is an observed sample of size n.

Proof. We omit the proof. ■

Definition 1.1.12 Let X, Y be rv's, (x_i, y_i), $1 \leq i \leq n$ be observations, and assume a simple linear regression model $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Let $\hat{\beta}_0$, $\hat{\beta}_1$ be estimates for β_0, β_1, then the **fitted values** of the model are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, 1 \leq i \leq n$$

and the **residuals** of the model are

$$E_i := Y_i - \hat{y}_i, 1 \leq i \leq n.$$

Proposition 1.1.13 Let X,Y be ev's with observations (x_i,y_i), $1 \le i \le n$ and assume a simple linear regression model $Y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$, then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} E_i^2 \sim \chi^2_{n-2},$$

where E_i's are residuals.

Proof. We omit the proof. ■

Corollary 1.1.14 With the same setup as above, the estimator

$$\tilde{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} E_i^2$$

for σ^2 is unbiased.

Proof. We have, by Proposition 1.1.12 and the expected value of the χ^2 distribution,

$$E(\tilde{\sigma}^2) = \sigma^2 E\left(\frac{1}{\sigma^2(n-2)}\sum_{i=1}^{n} E_i^2\right) = \frac{\sigma^2}{n-2}E\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} E_i^2\right) = \frac{\sigma^2}{n-2}(n-2) = \sigma^2.$$

This completes the proof. ■

Definition 1.1.15 Let X,Y be rv's with observations (x_i,y_i), $1 \le i \le n$, modelled by a simple linear regression. The **least squared error(LSE) estimates** for parameters β_0, β_1 are $\hat{\beta}_0$, $\hat{\beta}_1$ such that they minimise the **sum squared error (SSE)**

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

Theorem 1.1.16 In a simple linear regression, the LSE estimators are equivalent to the maximum likelihood estimators for the parameters β_0 and β_1.

Proof. For convenience, all $\sum$ are with respcet to i from $i=1$ to n.

Setting

$$\frac{\partial S(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_0} = 0 = \frac{\partial S(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_1}$$

simultaneously yields

$$\begin{cases} 0 &= \sum 2(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))(-1) \\ 0 &= \sum 2(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))(-x_i) \end{cases}$$

$$\begin{cases} 0 &= \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum x_i \\ 0 &= \sum(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \end{cases}$$

$$\begin{cases} \hat{\beta}_0 &= \frac{1}{n}\sum y_i - \hat{\beta}_1\frac{1}{n}\sum x_i \\ 0 &= \sum y_i x_i - \hat{\beta}_0 n\bar{x} - \hat{\beta}_1\sum x_i^2 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ 0 &= \sum y_i x_i - (y - \hat{\beta}_1 \bar{x})\bar{x} - \hat{\beta}_1 \sum x_i^2 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum y_i x_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \end{cases}$$

Immediately we get equivalence of $\hat{\beta}_0$ with Thm. 1.1.11. For $\hat{\beta}_1$, first we observe that

$$\sum \bar{y}(x_i - \bar{x}) = \bar{y}(\sum x_i - n\bar{x}) = \bar{y} \cdot 0 = 0$$

and similarly $\sum \bar{x}(x_i - \bar{x}) = 0$.

Now the numerator of $\hat{\beta}_1$ is

$$\sum y_i x_i - n\bar{x}\bar{y} = \sum y_i x_i - n\bar{x}\frac{1}{n}\sum y_i = \sum y_i(x_i - \bar{x}).$$

Because $\sum \bar{y}(x_i - \bar{x}) = 0$, we can add this term to the numerator for free to get the numerator to be

$$\sum (y_i - \bar{y})(x_i - \bar{x}) = (n-1)s_{xy}.$$

Similarly, for the denominator, we have

$$\begin{aligned} &\sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - n\bar{x}\frac{1}{n}\sum x_i = \sum x_i(x_i - \bar{x}) = \sum x_i(x_i - \bar{x}) - 0 \\ &= \sum x_i(x_i - \bar{x}) - \sum \bar{x}(x_i - \bar{x}) = \sum (x_i - \bar{x})^2 = (n-1)s_{xx} \end{aligned}$$

Thus

$$\hat{\beta}_1 = \frac{(n-1)s_{xy}}{(n-1)s_{xx}} = \frac{s_{xy}}{s_{xx}},$$

which corresponds to Thm. 1.1.11 as well. ■

Remark 1.1.17 The MLE is the LSE estimators in a simple linear regression model may not agree if the normality assumption in Remark 1.1.10 is violated.

However we do continue to hold this assumption going forward unless otherwise stated, and when we write $\hat{\beta}_0$ and $\hat{\beta}_1$, we just mean the LSE/MLE estimate.

Proposition 1.1.18 Let X, Y be rv's and assume a simple linear regression model

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Define, for $1 \leq i \leq n$,

$$w_i = \frac{X_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

then

$$\tilde{\beta}_1 \sim N\left(\sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i), \sigma^2 \sum_{i=1}^n w_i^2\right).$$

Proof. From Thm. 1.1.16 we get

$$\tilde{\beta}_1 = \frac{S_{XY}}{s_{xx}} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2|} = \frac{\sum_{i=1}^n Y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i Y_i.$$

Now each $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, so via Proposition 1.1.7 we get

$$\tilde{\beta}_1 \sim N\left(\sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i), \sigma^2 \sum_{i=1}^n w_i^2\right).$$

■

Corollary 1.1.19 With the same setup as above, the estimator

$$\tilde{\beta}_1 = \frac{S_{XY}}{s_{xx}}$$

is unbiased.

Proof. By Proposition 1.1.18 we immeidately get

$$E(\tilde{\beta}_1) = \sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i).$$

Substituting in w_i yields

$$\begin{aligned}
E(\tilde{\beta}_1) &= \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}(\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \sum_{i=1}^n \frac{(x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\beta_0}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n \frac{(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\beta_0}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot 0 + \beta_1 \\
&= \beta_1
\end{aligned}$$

as required. ■

Corollary 1.1.20 With the same setup as Proposition 1.1.18,

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{s_{xx}} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Proof. Again, directly from Proposition 1.1.18 we have

$$\mathrm{Var}(\tilde{\beta}_1) = \sigma^2 \sum_{i=1}^n w_i^2 = \sigma^2 \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{s_{xx}}.$$

as required. ■

Proposition 1.1.21 With the same setup as Proposition 1.1.18,

$$\tilde{\beta}_0 \sim N\left(\beta_0, \sigma\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right),$$

in particular, $\tilde{\beta}_0$ is an unbiased estimator of β_0.

Proof. From the proof of Thm. 1.1.16, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Now each

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), 1 \leq i \leq n,$$

which means

$$\bar{Y} \sim N\left(\beta_0 + \beta_1 \bar{x}, \frac{\sigma^2}{n}\right).$$

Thus, along with the fact that $E(\tilde{\beta}_1) = \beta_1$, we have

$$E(\tilde{\beta}_0) = E(\bar{Y}) - \bar{x}E(\tilde{\beta}_1) = \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0.$$

Next,

$$\mathrm{Var}(\tilde{\beta}_0) = \mathrm{Var}(\bar{Y} - \tilde{\beta}_1 \bar{x}) = \mathrm{Var}(\bar{Y}) + \bar{x}^2 \mathrm{Var}(\tilde{\beta}_1) + 2(-\bar{x})\mathrm{Cov}(\bar{Y}, \tilde{\beta}_1),$$

where

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}, \tilde{\beta}_1) &= \mathrm{Cov}\left(\frac{1}{n}\sum_i Y_i, \frac{\sum_i Y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right) \\
&= \frac{1}{n\sum_i (x_i - \bar{x})^2} \mathrm{Cov}\left(\sum_i Y_i, \sum_j Y_j (x_j - \bar{x})^2\right) \\
&= \frac{1}{n\sum_i (x_i - \bar{x})^2} \sum_j (x_j - \bar{x}) \sum_i \mathrm{Cov}(Y_i, Y_j) \\
&= \frac{1}{n\sum_i (x_i - \bar{x})^2} \sum_j (x_j - \bar{x})\sigma^2 \\
&= 0
\end{aligned}$$

and $\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n}$, $\mathrm{Var}(\tilde{\beta}_1) = \sigma^2 \sum_i w_i^2$. So

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}_0) &= \frac{\sigma^2}{n} + \bar{x}^2 \sigma^2 \sum_i w_i^2 \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \sigma^2 \sum_i \frac{(x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \sigma^2 \frac{1}{\sum_i (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)
\end{aligned}$$

Finally, $\tilde{\beta}_0$ follows a normal distribution because it is a linear combination of normal distributions, thus

$$\tilde{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right)$$

as required. ■

Proposition 1.1.22 Let X, Y be rv's and assume a simple linear regression model

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), 1 \leq i \leq n.$$

Fix $q \in [0,1]$.

1. If σ^2, the population variance, is known, then the $100q\%$ confidence interval of the parameter β_1 is

$$\left[\hat{\beta}_1 - a\frac{\sigma}{\sqrt{s_{xx}}}, \hat{\beta}_1 + a\frac{\sigma}{\sqrt{s_{xx}}}\right]$$

where $a = \Phi^{-1}\left(1 - \frac{1}{2}(1-q)\right)$, Φ is the distribution function of the standard normal distribution.

2. If the population variance is unknown, then the $100q\%$ confidence interval of β_1 is

$$\left[\hat{\beta}_1 - b\frac{\hat{\sigma}}{\sqrt{s_{xx}}}, \hat{\beta}_1 + b\frac{\hat{\sigma}}{\sqrt{s_{xx}}}\right]$$

where $b = \Psi_{n-2}^{-1}\left(1 - \frac{1}{2}(1-q)\right)$, n is the sample size, Ψ_{n-2} is the distribution function of the student t-distribution with $n-2$ degrees of freedom, and

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

is the sample estimator of σ^2.

Proof. We know that

$$\tilde{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

and so $\frac{\tilde{\beta}_1 - \beta_1}{\sigma/\sqrt{s_{xx}}} \sim N(0,1)$.

Meanwhile, we know that, from Proposition 1.1.13,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2 \sim \chi^2_{n-2}.$$

Write

$$\frac{\tilde{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{xx}}} = \frac{\tilde{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}/\sqrt{s_{xx}}} = \frac{\sigma}{\sigma}\cdot\frac{(\tilde{\beta}_1 - \beta_1)/(1/\sqrt{s_{xx}})}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}} = \frac{(\tilde{\beta}_1 - \beta_1)/(\sigma/\sqrt{s_{xx}})}{\sqrt{\frac{1}{\sigma^2(n-2)}\sum_{i=1}^{n} e_i^2}}.$$

Denote $Z := (\tilde{\beta}_1 - \beta_1)/(\sigma/\sqrt{s_{xx}})$ and $V := \frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2$, we have $Z \sim N(0,1)$ and $V \sim \chi^2_{n-2}$. By Proposition 1.1.7(3), we have

$$\frac{\tilde{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{xx}}} = \frac{Z}{\sqrt{V/(n-2)}} \sim t_{n-2}.$$

We now have everything we need for the actual proof concerning the confidence intervals.

1. Set $a = \Phi^{-1}(1 - \frac{1}{2}(1-q))$. By the symmetry of the normal distribution, we get

$$\Pr\left(-a \leq \frac{\tilde{\beta}_1 - \beta_1}{\sigma/\sqrt{s_{xx}}} \leq a\right) = q$$

or equivalently

$$q = \Pr\left(\tilde{\beta}_1 - a\frac{\sigma}{\sqrt{s_{xx}}} \leq \beta_1 \leq \tilde{\beta}_1 + a\frac{\sigma}{\sqrt{s_{xx}}}\right).$$

By the definition of a confidence interval,

$$\left[\hat{\beta}_1 - a\frac{\sigma}{\sqrt{s_{xx}}}, \hat{\beta}_1 + a\frac{\sigma}{\sqrt{s_{xx}}}\right]$$

is a $100q\%$ confidence interval of β_1.

2. Set $b = \Psi_{n-2}^{-1}\left(1 - \frac{1}{2}(1-q)\right)$. By the symmetry of the student t-distribution,

$$q = \Pr\left(-b \leq \frac{\tilde{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{xx}}} \leq b\right)$$

or equivalently

$$q = \Pr\left(\tilde{\beta}_1 - b\frac{\sigma}{\sqrt{s_{xx}}} \leq \beta_1 \leq \tilde{\beta}_1 + b\frac{\sigma}{\sqrt{s_{xx}}}\right).$$

By the definition of a confidence interval,

$$\left[\hat{\beta}_1 - b\frac{\sigma}{\sqrt{s_{xx}}}, \hat{\beta}_1 + b\frac{\sigma}{\sqrt{s_{xx}}}\right]$$

is a $100q\%$ confidence interval of β_1.

■

1.2 Hypothesis Testing in Simple Linear Regression

Remark 1.2.1 Let X, Y be rv's and suppose (x_i, y_i), $1 \leq i \leq n$ are sample observations. Assume a simple linear regression model

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

and suppose $\hat{\beta}_0$, $\hat{\beta}_1$ are LSE based on the observations.

Let $\theta_0 \in \mathbb{R}$. We are interested in testing the simple hypothesis

$$H_0 : \beta_1 = \theta_0,$$

The "evidence" for the hypothesis test is quantified by a probability: the probability of observing $\hat{\beta}_1$ and $\hat{\sigma}$ if H_0 is true. The lower the probability is , the stronger the evidence against H_0 is.

The convention in linear regression hypothesis testing is to test $H_0 : \beta_1 = 0$, i.e. no linear relationship between X and Y.

We already know that

$$T := \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{xx}}} \sim t_{n-2}.$$

The probability of observing (x_i, y_i), $1 \leq i \leq n$, or something even more extreme, given $\beta_1 = 0$ is

$$\Pr\left(|T| > \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{s_{xx}}}\right|\right) = 2\Pr\left(T \geq \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{s_{xx}}}\right|\right), T \sim t_{n-2}.$$

1.3 Prediction of a Simple Linear Regression Model

Definition 1.3.1 Let X, Y be rv's with observations (x_i, y_i), $1 \leq i \leq n$. Assume a simple linear regression model

$$Y_i = \beta_0 + \beta_1 + \varepsilon_i, 1 \leq i \leq n$$

and let $\hat{\beta}_0$, $\hat{\beta}_1$ be LSEs for β_0 and β_1.

The **mean response** of this model given an arbitrary input x_0 is

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The **predicted value** of Y given a new observation x_{new} based on this model is

$$y_{new} = \beta_0 + \beta_1 x_{new}.$$

Proposition 1.3.2 With the same setup as Def. 1.3.1, the mean response LSE

$$\tilde{\mu}_0 = \tilde{\beta}_0 + \tilde{\beta}_1 x_0$$

follows the distribution

$$N\left(\mu_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right) = N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

Proof. By the linearity of expectation, Corollary 1.1.19, and Proposition 1.1.20, we have

$$E(\tilde{\mu}_0) = E(\tilde{\beta}_0) + x_0 E(\tilde{\beta}_1) = \beta_0 + \beta_1 x_0 = \mu_0.$$

On the other hand

$$\mathrm{Var}(\tilde{\mu}_0) = \mathrm{Var}(\tilde{\beta}_0 + \tilde{\beta}_1 x_0) = \mathrm{Var}(\bar{Y} - \tilde{\beta}_1 \bar{x} + \tilde{\beta}_1 x_0) = \mathrm{Var}(\bar{Y} + \tilde{\beta}_1 (x_0 - \bar{x})).$$

Writing is out yields

$$\mathrm{Var}(\tilde{\mu}_0) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i + \left(\sum_{i=1}^n \frac{x_i - \bar{x}}{s_{xx}} y_i\right)(x_0 - \bar{x})\right) = \mathrm{Var}\left(\sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{s_{xx}}\right) Y_i\right).$$

By model assumption, each $\mathrm{Var}(Y_i) = \sigma^2$. Moreover the Y_i's are assumed to be independent, so by Proposition 1.1.7,

$$\begin{aligned}
\mathrm{Var}(\tilde{\mu}_0) &= \sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{s_{xx}}\right)^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{2}{n}\left(\frac{(x_i - \bar{x})(x_0 - \bar{x})}{s_{xx}}\right) + \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{s_{xx}^2}\right) \\
&= \sigma^2 \left(\frac{n}{n^2} + \frac{2(x_0 - \bar{x})}{s_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{(x_0 - \bar{x})^2}{s_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2\right) \\
&= \sigma^2 \left(\frac{1}{n} + 0 + \frac{(x_0 - \bar{x})^2}{s_{xx}^2} s_{xx}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)
\end{aligned}$$

where $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Finally, the above shows that $\tilde{\mu}_0$ can be written as a linear combination of Y_i's, all of which are normally distributed, so

$$\tilde{\mu}_0 \sim N\left(\mu_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$$

as required. ■

Proposition 1.3.3 With the same setup as Def. 1.3.1, $\alpha \in [0,1)$, and known parameter σ^2, the $100\alpha\%$ confidence interval of μ_0 is

$$\left[\hat{\mu}_0 - a\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}, \hat{\mu}_0 + a\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}\right]$$

where $a = \Phi^{-1}\left(\frac{1}{2}(1+\alpha)\right)$, Φ being the distribution function of the standard normal distribution.

Proof. The arguments are analogous to Proposition 1.1.22(1). ■

Proposition 1.3.4 With the same setup as Def. 1.3.1, $\alpha \in [0,1)$, and unknown population variance σ^2, the $100\alpha\%$ confidence interval of μ_0 is

$$\left[\hat{\mu}_0 - q_\alpha\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}, \hat{\mu}_0 + q_\alpha\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}\right]$$

where

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

is the variance estimate, and $q_\alpha = \Psi_{n-2}^{-1}\left(1 - \frac{1}{2}(1-\alpha)\right)$, Φ_{n-2} being the distribution function of the t_{n-2} distribution.

Proof. From Proposition 1.3.2,

$$\frac{\tilde{\mu}_0 - \mu_0}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \sim N(0,1).$$

Now by Proposition 1.1.13, $\frac{1}{\sigma^2}\sum_{i=1}^{n} E_i^2 \sim \chi^2_{n-2}$, and so

$$\begin{aligned}
\frac{\tilde{\mu}_0 - \mu_0}{\tilde{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} &= \frac{\tilde{\mu}_0 - \mu_0}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n} E_i^2}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \cdot \frac{\sigma}{\sigma} \\
&= \frac{\tilde{\mu}_0 - \mu_0}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \cdot \frac{\sigma}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n} E_i^2}} \\
&= Z \cdot \sqrt{\frac{(n-2)\sigma^2}{\sum_{i=1}^{n} E_i^2}}, Z = \frac{\tilde{\mu}_0 - \mu_0}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \sim N(0,1) \\
&= \frac{Z}{\sqrt{\frac{\sum_{i=1}^{n} E_i^2}{\sigma^2}/(n-2)}} \sim t_{n-2} \text{ by Proposition 1.1.7(3).}
\end{aligned}$$

By the symmetry of the t-distribution, we look for q_α such that

$$\begin{aligned}\alpha &= \Pr\left(-q_\alpha \le \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \le q_\alpha\right) \\ &= \Pr\left(\hat{\mu}_0 - q_\alpha\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}} \le \mu_0 \le \hat{\mu}_0 + q_\alpha\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}\right).\end{aligned}$$

By this we get

$$\alpha = 1 - 2\Pr(T \ge q_\alpha) = 1 - 2(1 - \Pr(T \le q_\alpha)), T \sim t_{n-2},$$

which yields

$$\Pr(T \le q_\alpha) = 1 - \frac{1}{2}(1-\alpha), T \sim t_{n-2},$$

completing the proof. ■

Proposition 1.3.5 With the same setup as Def. 1.3.1, we have

$$\tilde{y}_{new} - y_{new} = N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{s_{xx}}\right)\right).$$

Proof. By the linearity of expectation and the model assumption,

$$\begin{aligned}E(\tilde{y}_{new} - y_{new}) &= E(\tilde{\beta}_0 + \tilde{\beta}_1 x_{new} - \beta_0 - \beta_1 x_{new} - \varepsilon_{new}) \\ &= E(\tilde{\beta}_0) + E(\tilde{\beta}_1 x_{new}) - \beta_0 - \beta_1 x_{new} - E(\varepsilon_{new}) \\ &= \beta_0 + \beta_1 x_{new} - \beta_0 - \beta_1 x_{new} - 0 \\ &= 0.\end{aligned}$$

Now observe that $\tilde{y}_{new} = \tilde{\beta}_0 + \tilde{\beta}_1 x_{new}$ is a linear combination of y_i's, $1 \le i \le n$, as pointed out in the proof of Proposition 1.3.2, so $\tilde{y}_{new} \perp y_{new}$, and so

$$\begin{aligned}\mathrm{Var}(\tilde{y}_{new} - y_{new}) &= \mathrm{Var}(\tilde{\beta}_0 + \tilde{\beta}_1 x_{new}) - \mathrm{Var}(y_{new}) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{s_{xx}}\right) + \sigma^2 \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{s_{xx}}\right).\end{aligned}$$

Thus $\tilde{y}_{new} - y_{new} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{s_{xx}}\right)\right)$ as required. ■

Proposition 1.3.6 With the same setup as Def. 1.3.1, $\alpha \in [0,1)$, then

1. If the population variance σ^2 is known, then the $100\alpha\%$ confidence interval of y_{new} given x_{new} is

$$\left[\hat{y}_{new} - q_\alpha\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{s_{xx}}}, \hat{y}_{new} + q_\alpha\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{s_{xx}}}\right]$$

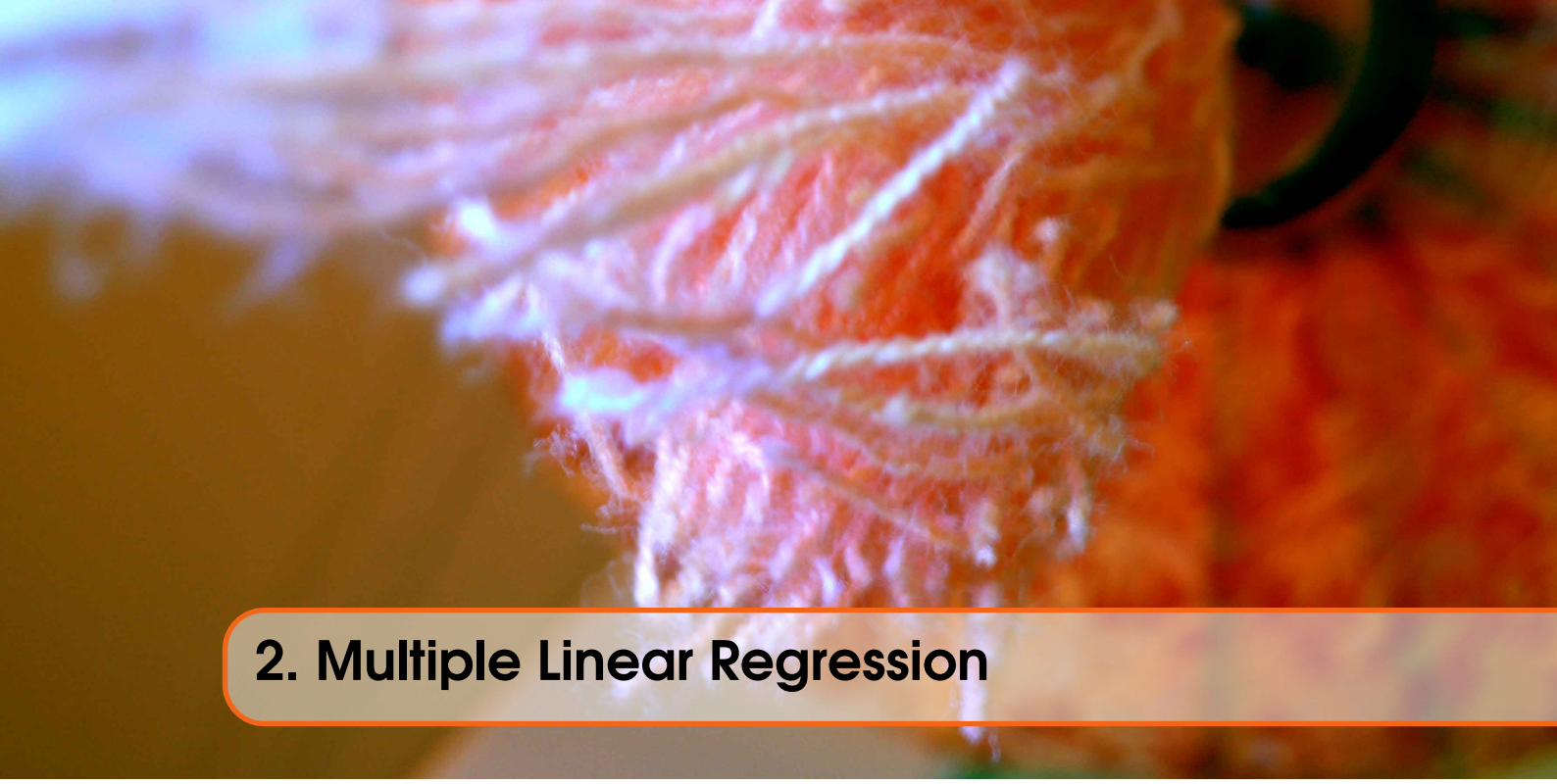
where $q_\alpha = \Phi^{-1}\left(1 - \frac{1}{2}(1-\alpha)\right)$, Φ being the distribution function of the standard normal distribution.

2. If the population variance σ^2 is unknown, then the $100\alpha\%$ confidence interval of y_{new} given x_{new} is

$$\left[\hat{y}_{new} - q_\alpha \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{new}-\bar{x})^2}{s_{xx}}}, \hat{y}_{new} + q_\alpha \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{new}-\bar{x})^2}{s_{xx}}}\right]$$

where $\hat{\sigma} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$, and $q_\alpha = \Psi_{n-2}^{-1}\left(1-\frac{1}{n}(1-\alpha)\right)$, Ψ_{n-2} being the distribution function of the t_{n-2} distribution.

Proof. Analogous to Proposition 1.3.4, 1.3.5. ■



2. Multiple Linear Regression

2.1 Pre-Requisites and Definitions

Proposition 2.1.1 — Matrix algebra. Let $C \in \mathbb{R}_{m\times n}$, $A, B \in \mathbb{R}_{n\times n}$ for some $m, n \in \mathbb{N}$, then

1. $(AB)^T = B^T A^T$.
2. If B is not singular, i.e. $\det(B) \neq 0$, then $B^{-1}B = I \in \mathbb{R}_{n\times n}$.
3. If AB is not singular, then $(AB)^{-1} = B^{-1}A^{-1}$.
4. If A^T is not singular, then $(A^T)^{-1} = (A^{-1})^T$.
5. Define the trace function

$$\begin{aligned} \operatorname{tr} : \mathbb{R}^{n\times n} &\to \mathbb{R} \\ [a_{ij}]_{ij} &\mapsto \sum_{i=1}^{n} a_{ii} \end{aligned}$$

then (5.1) $\operatorname{tr}(A+\alpha B) = \operatorname{tr}(A) + \alpha \operatorname{tr}(B)$ for all $\alpha \in \mathbb{R}$, (5.2) $\operatorname{tr}(A) = \operatorname{tr}(A^T)$, (5.3) $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

Proof. We omit the proof. ■

Proposition 2.1.2 — Calculus with matrices. Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a function and $z = f(y)$, then

1.

$$\frac{\partial z}{\partial y} = \begin{bmatrix} \frac{\partial z}{\partial y_1} & \frac{\partial z}{\partial y_2} & \dots & \frac{\partial z}{\partial y_n} \end{bmatrix}$$

where $y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$, each $y_i \in \mathbb{R}$, $1 \leq i \leq n$.

2. If $f(y) = a^T y$ for some $a \in \mathbb{R}^n$, then

$$\frac{\partial f}{\partial y} = a.$$

3. If $f(y) = y^T A y$ for some matrix $A \in \mathbb{R}_{n\times n}$, then

$$\frac{\partial f}{\partial y} = Ay + A^T y.$$

In particular, if A is symmetric, then $\frac{\partial f}{\partial y} = 2Ay = 2A^T y$.

Proof. We omit the proof. ■

Definition 2.1.3 Let $Y \in \mathbb{R}^n$ be a random vector. The **expectation** of Y is

$$E(Y) = (E(Y_1) \dots E(Y_n))^T$$

where $Y_1, \dots, Y_n$ are random variables.

The **variance-covariance matrix** of Y is

$$\mathbb{R}_{n\times n} \ni V := \mathrm{Var}(Y) = \begin{bmatrix} \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1, Y_2) & \dots & \mathrm{Cov}(Y_1, Y_n) \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Var}(Y_2) & \dots & \mathrm{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(Y_n, Y_1) & \mathrm{Cov}(Y_n, Y_2) & \dots & \mathrm{Var}(Y_n) \end{bmatrix}$$

which is derived from the definition

$$\mathrm{Var}(Y) = E((Y-\mu)(Y-\mu)^T), Y, \mu \in \mathbb{R}^n$$

where $\mu = (E(Y_1) \dots E(Y_n))^T$.

Proposition 2.1.4 Let $Y \in \mathbb{R}^n$ be a random vector, then

1. $\mathrm{Var}(Y) = E(YY^T) - E(Y)E(Y)^T$.
2. $\mathrm{Var}(aY) = a\,\mathrm{Var}(Y)a^T$ for constant $a \in \mathbb{R}^n$.
3. $\mathrm{Var}(Y+b) = \mathrm{Var}(Y)$ for all $b \in \mathbb{R}^n$.

Proof. 1. From definition

$$\begin{aligned} \mathrm{Var}(Y) &= E((Y-\mu)(Y-\mu)^T) \\ &= E(YY^T - \mu Y^T - Y\mu^T + \mu\mu^T) \\ &= E(YY^T) - \mu E(Y^T) - E(Y)\mu^T + \mu\mu^T \text{ where } E(Y) = \mu \\ &= E(YY^T) - E(Y)E(Y^T). \end{aligned}$$

2. From definition

$$\begin{aligned} \mathrm{Var}(aY) &= E((aY - E(aY))(aY - E(aY))^T) \\ &= E((aY - aE(Y))(aY - aE(Y))^T) \\ &= E(a(Y - E(Y))(Y - E(Y))^T a^T) \\ &= aE((Y - E(Y))(Y - E(Y))^T)a^T \\ &= a\,\mathrm{Var}(Y)a^T. \end{aligned}$$

3. From definition

$$\begin{aligned}\mathrm{Var}(Y+b) &= E((Y+b-E(Y+b))(Y+b-E(Y+b))^T)\\ &= E((Y+b-E(Y)-b)(Y+b-E(Y)-b)^T)\\ &= E((Y-E(Y))(Y-E(Y))^T)\\ &= \mathrm{Var}(Y).\end{aligned}$$

■

Corollary 2.1.5 Let $Y \in \mathbb{R}^n$ be a random vector, then $\mathrm{Var}(Y)$ is semi-positive definite, i.e. for all $a \in \mathbb{R}^n$, each term of the matrix $a\,\mathrm{Var}(Y)a^T$ is nonnegative.

Proof. From Proposition 2.1.4(2), $a\,\mathrm{Var}(Y)a^T = \mathrm{Var}(aY)$ for all $a \in \mathbb{R}$, and all terms of a variance-covariance matrix, by the definition of variance and covariance in the random variable case, must be nonnegative. ■

Proposition 2.1.6 Suppose $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is a random vector, $a_1, \ldots, a_n \in \mathbb{R}$, $c_1, \ldots, c_n \in \mathbb{R}$, $b_1, \ldots, b_n \in \mathbb{R}$, and $d_1, \ldots, d_n \in \mathbb{R}$ are constants, then

$$\mathbb{R} \ni Z := \sum_{i=1}^{n} a_i Y_i + c_i, U := \sum_{i=1}^{n} b_i Y_i + d_i$$

are random variables with

$$\begin{aligned}E(Z) &= \sum_{i=1}^{n} a_i E(Y_i) + c_i\\ E(U) &= \sum_{i=1}^{n} b_i E(Y_i) + d_i\\ \mathrm{Cov}(Z,U) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(Y_i, Y_j).\end{aligned}$$

In matrix notation, we have

$$\begin{aligned}\mathbb{R} \ni Z &= a^T Y + c, U = b^T Y + d\\ E(Z) &= a^T \mu + c \in \mathbb{R}\\ \mathrm{Cov}(Z,U) &= a^T\,\mathrm{Var}(Y) b \in \mathbb{R}\end{aligned}$$

where $\mu = (E(Y_1), \ldots, E(Y_n))^T \in \mathbb{R}^n$, $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$, etc.

Proof. This follows directly from the definitions of expectation, variance, and matrix arithmetic. ■

Proposition 2.1.7 Let $Y \in \mathbb{R}^n$ be a random vector and $A \in \mathbb{R}_{k\times n}$ be a matrix, then $Z := (Z_1, \ldots, Z_k)^T$, defined by each

$$Z_i = \sum_{j=1}^{n} a_{ij} Y_j,$$

is a random vector in $\mathbb{R}^k$, with

$$\mathbb{R}^k \ni E(Z) = AE(Y),$$

$$\mathbb{R}^k \ni \text{Var}(Z) = A\,\text{Var}(Y)A^T.$$

In matrix notation, we have

$$\mathbb{R}^k \ni Z = AY.$$

Proof. We realise that

$$Z_i = \sum_{j=1}^{n} a_{ij}Y_j$$

for $i = 1,\dots,k$, so $Z = AY$. The expectation result follows from Proposition 2.1.6. For variance, we use its definition

$$\begin{aligned}\text{Var}(Z) = \text{Var}(AY) &= E((AY - E(AY))(AY - E(AY))^T) \\ &= E(A(Y - E(Y))(Y - E(Y))^T A^T) \\ &= AE((Y - E(Y))(Y - E(Y))^T)A^T \\ &= A\,\text{Var}(Y)A^T\end{aligned}$$

as required. ■

Definition 2.1.8 Suppose $Y \in \mathbb{R}^n$ is a random vector, $\mu, y \in \mathbb{R}^n$, and $\Sigma \in \mathbb{R}_{n\times n}$ is a semi-positive definite symmetric matrix that is non-singular, then Y follows a **multivariate normal distribution** with population mean μ and **population variance-covariance matrix** Σ, written as

$$Y \sim \text{MVN}(\mu, \Sigma),$$

if Y has the pdf

$$f : y \mapsto \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}$$

where $|\Sigma| = \det(\Sigma)$.

We have

$$[\Sigma]_{ij} = \text{Cov}(Y_i, Y_j), 1 \le i \le j \le n.$$

Proposition 2.1.9 Suppose $Z_1,\dots,Z_n \sim N(0,1)$ independently, $A \in \mathbb{R}_{n\times n}$ is a matrix, and $\mu \in \mathbb{R}^n$, then $Y = AZ + \mu$ ($Z := (Z_1,\dots,Z_n)$) has distribution

$$Y \sim \text{MVN}(\mu, \Sigma)$$

where $\Sigma = AA^T$ and $\mu = E(Y)$.

Proof. We omit the proof. ■

Proposition 2.1.10 Suppose $Y \in \mathbb{R}^n$ is a random vector and $Y \sim \text{MVN}(\mu, \Sigma)$, then

1. If $C \in \mathbb{R}_{n\times n}$ and $d \in \mathbb{R}^n$ are constants, the random vector $U := CY + d$ has

$$U \sim \text{MVN}(C\mu + d, C\Sigma C^T).$$

2. If $\tilde{Y} := (Y_1, \ldots, Y_m)^T$ for some $m \leq n$, then

$$\tilde{Y} \sim \text{MVN}\left(\begin{bmatrix}\mu_1\\ \vdots \\ \mu_m\end{bmatrix}, \begin{bmatrix}\text{Var}(Y_1) & \ldots & \text{Cov}(Y_1, Y_m)\\ \vdots & \ddots & \vdots \\ \text{Cov}(Y_m, Y_1) & \ldots & \text{Var}(Y_m)\end{bmatrix}\right).$$

Proof. We omit the proof. ■

Proposition 2.1.11 Suppose $Y \in \mathbb{R}^n$ is a random vector with

$$Y \sim \text{MVN}(\mu, \Sigma)$$

for some $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}_{n\times n}$, and let $0 < p < n$, $p \in \mathbb{N}$. Define

1.
$$Y_a = \begin{bmatrix}Y_1\\ \vdots \\ Y_p\end{bmatrix}, Y_b = \begin{bmatrix}Y_{p+1}\\ \vdots \\ Y_n\end{bmatrix}$$

2.
$$\mu_a = \begin{bmatrix}\mu_1\\ \vdots \\ \mu_p\end{bmatrix}, \mu_b = \begin{bmatrix}\mu_{p+1}\\ \vdots \\ \mu_n\end{bmatrix}$$

3. $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$ such that

$$\Sigma_{11} \in \mathbb{R}_{p\times p}, \Sigma_{12} \in \mathbb{R}_{p\times(n-p)}$$
$$\Sigma_{21} \in \mathbb{R}_{(n-p)\times p}, \Sigma_{22} \in \mathbb{R}_{(n-p)\times(n-p)}$$

and

$$\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

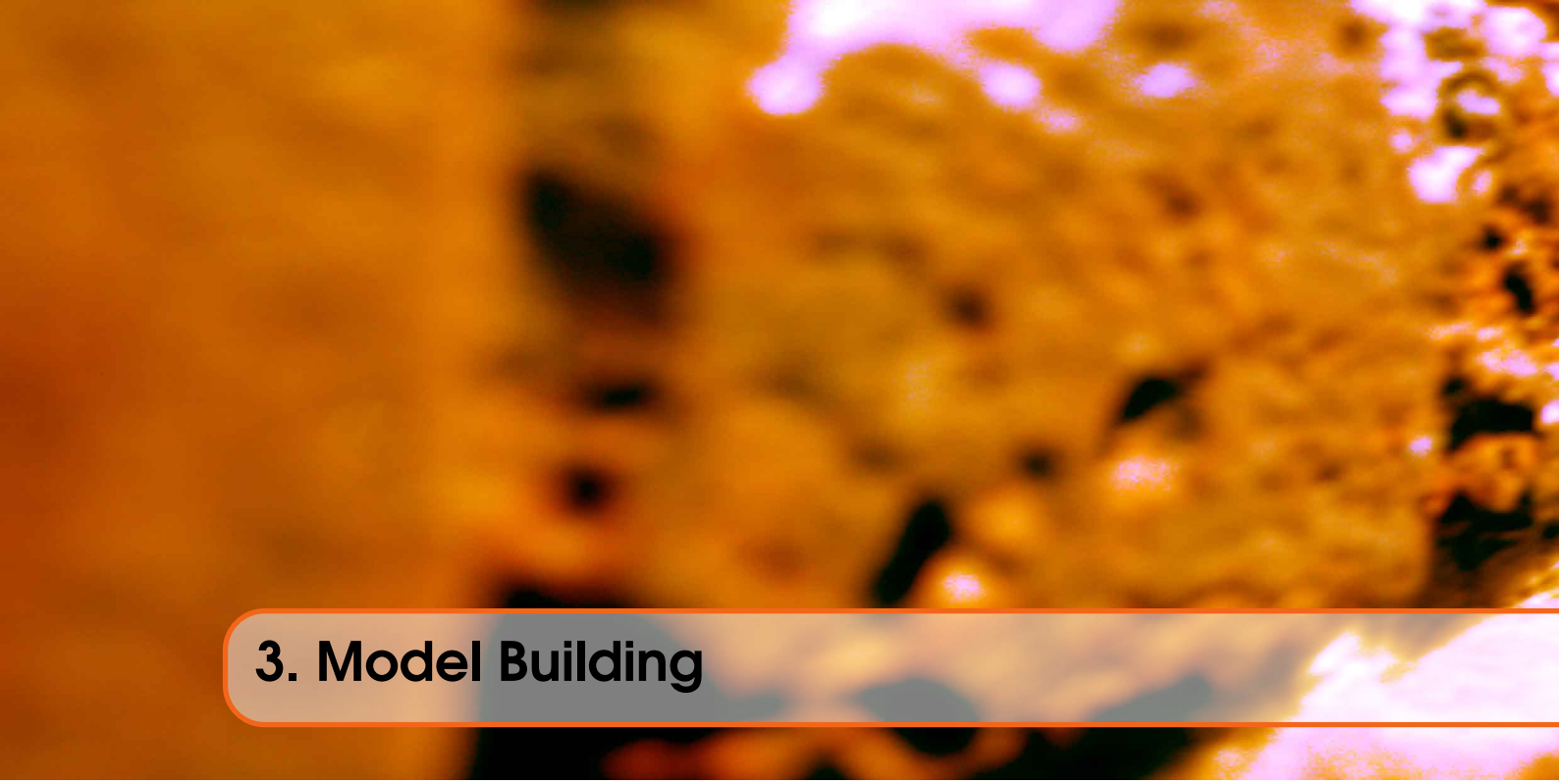
then the conditional distribution of Y_a given $Y_b = b$ for some constant $b \in \mathbb{R}^{n-p}$ is

$$Y_a | Y_b = b \sim \text{MVN}(\bar{\mu}, \bar{\Sigma})$$

where

$$\bar{\mu} = \mu_a + \Sigma_{12}\Sigma_{22}^{-1}(b - \mu_b)$$
$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof. We omit the proof. ■



3. Model Building

4. Model Diagnostics

Index