STAT24510 (W25): Statistical Theory and Methods IIa

Lecturer: Mei Wang
Notes by: Aden Chen

Tuesday 4th March, 2025

Contents

This set of notes is mostly unreadable.

1 Introduction

The goal of statistics is often to estimate a (population) parameter θ. From data, we may obtain point estimates $\hat{\theta}$ that depends on data, and construct confidence intervals to quantify the uncertainty of the estimate, with which we can conduct hypothesis testing.

In this course we will start from confidence intervals and hypothesis testing, and then move on to linear models. This course will be less structured; intend, it will be more like a collection of methods.

1.1 Examples of Construction: Review of Wald and Wilson

1.1.1 Pivotal method

Example 1.1 (Pivotal method, Normal, Known σ^2). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 5^2)$, $i = 1, \ldots, n$. Our goal is to construct a $1 - \alpha$ CI for μ. We have the MLE estimator $\hat{\mu} = \overline{X} = n^{-1} \sum_{i=1}^{n} X_i$.

Note that

$$\frac{\overline{X} - \mathrm{E}[X]}{\sqrt{\mathrm{Var}(X)}} = \frac{\overline{X} - \mu}{\sqrt{5^2/n}} \sim \mathcal{N}(0, 1).$$

In particular, note that the left side is a function of data and parameters, while the right side is free of parameters.

We thus have

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sqrt{5^2/n}} \leq z_{1-\alpha/2} \right) = 0.95,$$

using which we can construct the CI of μ: With $I \coloneqq \left[\overline{X} - z_{1-\alpha/2}\sqrt{5^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{5^2/n} \right]$, we have $\mathbb{P}(\mu \in I) = 1 - \alpha$.

Notice that we obtain a probability statement of random interval containing a fixed quantity from a probability statement of a fixed interval containing a random quantity.

Example 1.2 (Pivotal method, Normal, Known μ). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(4, \sigma^2)$, $i = 1, \ldots, n$. The goal: CI for σ^2, i.e., to find random variables L and U such that $\mathbb{P}\left(L \leq \sigma^2 \leq U\right) = 1 - \alpha$.

Note that

$$Y_i \coloneqq \frac{X_i - 4}{\sigma} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

and thus

$$T_n \coloneqq \sum Y_i^2 = \sum \left(\frac{X_i - 4}{\sigma} \right)^2 \sim \chi^2_n.$$

Again, we obtained a function of data and parameters that follows a known distribution. From

$$\mathbb{P}\left(\chi^2_{n,\alpha/2} \leq T_n \leq \chi^2_{n,1-\alpha/2} \right) = 1 - \alpha$$

we may again obtain the CI for σ^2,

$$\left[\frac{\sum (X_i - 4)^2}{\chi^2_{n,1-\alpha/2}}, \frac{\sum (X_i - 4)^2}{\chi^2_{n,\alpha/2}} \right].$$

Example 1.3 (Pivot failing). Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. The goal: CI for p. The MLE for p is $\hat{p} = \overline{X}$, thus we may be temped to try

$$T_n \coloneqq \frac{\hat{p} - p}{\sqrt{\text{Var}(\hat{p})}},$$

but the distribution of T_n depends on p. The method of pivots fail.

1.1.2 Asymptotic CI

I.e., when we have large sample size n.

Example 1.4 (Wald CI). Let X_i be iid with mean μ and variance σ^2. From the CLT we have

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2),$$

that is,

$$\lim_{n\to\infty} \mathbb{P}(T_n \le x) = P(Z \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t.$$

Thus we have

$$\mathbb{P}\left(z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \le z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

When σ^2 is known, we may derive an approximate CI for μ:

$$\mathbb{P}\left(\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n} \le \mu \le \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n}\right) \approx 1 - \alpha.$$

When σ^2 is unknown: If there exists random variables $U_n \to_{\text{p}} \sigma^2$ (that is, $\lim \mathbb{P}\left(U_n = \sigma^2\right) = 1$), then

$$T_n \coloneqq \frac{\overline{X} - \mu}{\sqrt{U_n/n}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{U_n/\sigma^2}}$$

where $(\overline{X} - \mu)/\sqrt{\sigma^2/n} \xrightarrow{\mathscr{D}} \mathcal{N}(0,1)$ and $\sqrt{U_n/\sigma^2} \to_{\text{p}} 1$, and thus by Slutsky's theorem we have

$$T_n \xrightarrow{\mathscr{D}} \mathcal{N}(0,1),$$

using which we can again construct an approximate CI. Note that we used asymptotic approximation multiple times. This is called the Wald confidence interval.

Example 1.5 (Wald CI). Let $X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. The goal: asymptotic CI for λ. Note that we have the MLE of λ, $\hat{\lambda} = \overline{X}$, with $\text{E}[\hat{\lambda}] = \lambda$ and $\text{Var}[\hat{\lambda}] = \lambda/n$. We then have

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \approx Z \sim \mathcal{N}(0,1).$$

We approximate a second time: $\frac{\overline{X}-\lambda}{\sqrt{\hat{\lambda}/n}} \approx Z$, from which we obtain the Wald CI for λ:

$$\left[\hat{\lambda} - z_{1-\alpha/2}\sqrt{\hat{\lambda}/n}, \hat{\lambda} + z_{1-\alpha/2}\sqrt{\hat{\lambda}/n}\right].$$

Example 1.6 (Wilson's method). Assume the same setup as above. Again, we use

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \approx Z \sim \mathcal{N}(0, 1),$$

which gives

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \leq z_{1-\alpha/2}\right) = \mathbb{P}\left(\left(\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}}\right)^2 \leq z_{1-\alpha/2}^2\right) \approx 1 - \alpha.$$

Solving for λ in the middle expression gives the Wilson CI. We used one fewer approximation.

2 Confidence Intervals

2.1 Constructing CI

- Exact (i.e., coverage probability is exactly $1-\alpha$) confidence intervals: **pivot method**:
  1. Find statistic $T_n = T(X, \theta)$ whose distribution is known and independent of θ. Such a statistic is called a pivot.
  2. Using knowledge on the distribution, find c_L and c_U such that $\mathbb{P}(c_L \le T_n \le c_R) = 1-\alpha$.
  3. Convert the probability statement of the pivot to a probability statement of the parameter.
- Approximation methods, or asymptotic (sample size n is large) methods.
  - **Wald confidence interval** (the default CI produced by standard software): more than one approximations.
  - **Wilson's confidence interval /** score method **/** duality method: using CLT once.
    - Wilson's CI is better in the sense that the actual coverage is closer to the desired coverage.
  - **Variance stabilization transformation** (VST) method.

Example 2.1 (Pivot, Normal). Assume $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$. Goal: $(1-\alpha)$ CI for μ.

1. Recall that
$$T_n = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}.^{1}$$
2. We have
$$\mathbb{P}\left(t_{n-1,\alpha/2} \le \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \le t_{n-1,1-\alpha/2} \right) = 1-\alpha.$$
3. Rearrangement gives
$$\mathbb{P}\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \right) = 1-\alpha.$$

Example 2.2 (Wald, Bernoulli). Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, so that $\sum X_i \sim \text{Binomial}(n, p)$. Goal: $(1-\alpha)$ CI for p. Note that $\hat{p} = \overline{X}$ is the MLE of p. CLT gives
$$\frac{\hat{p} - p}{\sqrt{\text{Var}(\hat{p})}} \approx \mathcal{N}(0, 1).$$

[1] t distribution has fatter tails than $\mathcal{N}(0, 1)$.

From Slutsky's theorem we have a **second approximation**

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}.$$

This second approximation is characteristic of Wald's method. Using this second approximation we have

$$\frac{\overline{X} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx \mathcal{N}(0,1),$$

with which we can construct the desired CI:

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}\right].$$

Example 2.3 (Wilson's CI; Bernoulli). We assume the same Bernoulli setup. Note that we obtained using just the CLT that

$$\mathbb{P}\left(z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha.$$

We rewrite the middle expression as

$$n\,(\hat{p} - p)^2 \leq z^2_{1-\frac{\alpha}{2}} p(1-p).$$

Solving a quadratic equation gives the desired Wilson's CI.

Remark 2.4.

- Note that the Wald CI is centered at $\hat{p}$, but Wilson's is not.
- Wilson's CI will always be contained in $[0,1]$; the lower bound of Wald might be negative.

2.2 Variance Stabilization Transformation (VST) Method

Example 2.5 (VST, Poisson). Let $X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Note that

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} \approx \mathcal{N}(0,1).$$

and

$$\sqrt{n}(\hat{\lambda} - \lambda) \approx \mathcal{N}(0, \lambda).$$

Goal: find a transformation (usually smooth) g such that

$$\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \approx \mathcal{N}(0,1).$$[2]

[2] Or just a normal distribution with fixed variance.

Tool: delta method (Taylor expansion for random variables). By Taylor expansion,

$$g(\hat{\lambda}) = g(\lambda) + g'(\lambda)(\hat{\lambda} - \lambda) + \frac{g''(\lambda)}{2}(\hat{\lambda} - \lambda)^2 + \dots.$$

We thus have the approximation

$$g(\hat{\lambda}) \approx g(\lambda) + g'(\lambda)(\hat{\lambda} - \lambda) + O(n^{-1}),$$

where the last term follows from the fact that $(\hat{\lambda} - \lambda)/\sqrt{\lambda/n} \approx \mathcal{N}(0, 1)$. Then,

$$\mathrm{E}[g(\hat{\lambda})] \approx g(\lambda) + g'(\lambda)\,\mathrm{E}[\hat{\lambda} - \lambda].$$

Since $\hat{\lambda}$ is unbiased, we have

$$\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \approx g'(\lambda)\sqrt{n}(\hat{\lambda} - \lambda),$$

giving

$$\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \approx \mathcal{N}(0, [g'(\lambda)]^2\lambda),$$

where we used the approximation

$$\mathrm{Var}(g(\hat{\lambda})) = \mathrm{E}[(g(\hat{\lambda}) - \mathrm{E}[g(\hat{\lambda})])^2] \approx [g'(\lambda)]^2\,\mathrm{Var}(\hat{\lambda}).$$

To obtain $\sqrt{n}(g(\hat{\lambda}) - \lambda) \approx \mathcal{N}(0, 1)$, we need only $g'(\lambda)^2 = 1/\lambda$. $g(\lambda) = 2\sqrt{\lambda}$ will do.

That is, we have

$$\sqrt{n}(2\sqrt{\hat{\lambda}} - 2\sqrt{\lambda}) \approx \mathcal{N}(0, 1).$$

Using this we can obtain a CI for λ. Note that the left endpoint of the CI for $\sqrt{\lambda}$ may be negative, so in such cases when obtaining the CI for λ we need to use 0 instead.

2.3 Delta Method

Consider an unbiased estimator $\hat{\theta}$. $\mathrm{E}[\hat{\theta}] = \theta$. Assume that the variance is a function of the mean, $\mathrm{Var}[\hat{\theta}] = v(\theta)$.[3] Goal: find a function $g(t)$ such that $\mathrm{Var}[g(\hat{\theta})]$ is constant. Taylor gives

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta).$$

Note that we have

$$\mathrm{E}[g(\hat{\theta})] \approx g(\theta) + g'(\theta)\,\mathrm{E}[\hat{\theta} - \theta] = g(\theta).$$

and

$$\mathrm{Var}[g(\hat{\theta})] = \mathrm{E}\left[\left(g(\hat{\theta}) - g(\theta)\right)^2\right] \approx \mathrm{E}\left[g'(\theta)^2\left(\hat{\theta} - \theta\right)^2\right] = g'(\theta)^2\,\mathrm{Var}[\hat{\theta}],$$

which we want to be constant, say 1. Thus we seek g such that

$$g'(\theta) = \frac{1}{\sqrt{v(\theta)}}.$$

A choice is of course $g(\theta) = \int 1/\sqrt{v(\theta)}\,\mathrm{d}\theta$.

[3] This is a common case.

Example 2.6 (Delta Method, Exponential, distribution of MLE). Consider $X_i \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$. Recall that the MLE of λ is $\hat{\lambda} = 1/\overline{X}$. By the CLT we have for large n that

$$\frac{\overline{X} - \mathrm{E}\, X}{\sqrt{\mathrm{Var}\, \overline{X}}} \approx \mathcal{N}(0, 1),$$

That is,

$$\frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{n\lambda^2}}} = \sqrt{n}\left(\overline{X} - \frac{1}{\lambda}\right) \approx \mathcal{N}(0, \lambda^{-2}).$$

Goal: find an approximate distribution of $\hat{\lambda} = 1/\overline{X}$ using Delta Method. Using $g(t) = t^{-1}$ we have

$$\sqrt{n}\left(g(\hat{\lambda}) - g(\lambda)\right) \approx \mathcal{N}(0, [g'(\lambda)]^2 \,\mathrm{Var}(\hat{\lambda})) = \mathcal{N}(0, \lambda^2).$$

(Note that Fisher's theorem gives another way to derive the above.)

Example 2.7 (VST, Exponential, CI). Consider the same setup as above. We can then derive an approximate CI for λ using VST. Recall that from Delta Method, we have

$$\sqrt{n}\left(g(\hat{\lambda}) - g(\lambda)\right) \approx \mathcal{N}(0, [g'(\lambda)]^2 \,\mathrm{Var}(\hat{\lambda})).$$

Thus we seek g such that $g'(\lambda)^2\lambda^2$ is a constant, say 1, or $g'(\lambda) = 1/\lambda$. It is easy to see that $g = \log$ is such an option. Then,

$$\sqrt{n}\left(\log\hat{\lambda} - \log\lambda\right) \approx \mathcal{N}(0, 1),$$

using which we can obtain an approximate CI for $\log\hat{\lambda}$ and then $\hat{\lambda}$:

$$\left[\hat{\lambda}\exp\left(-\frac{z_{1-\alpha/2}}{\sqrt{n}}\right), \hat{\lambda}\exp\left(\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)\right]$$

Theorem 2.8 (Delta Method)**.** *If X_n is such that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2)$$

and g is continuously differentiable, then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2[g'(\theta)]^2).$$

Remark 2.9. Intuition: We can write

$$\sqrt{n}(g(X_n) - g(\theta)) = g'(\tilde{\theta}_n)\sqrt{n}(X_n - \theta), \quad \tilde{\theta}_n \in (x_n, \theta).$$

We know that $g'(\tilde{\theta}_n) \to_{\mathrm{p}} g'(\theta)$ and $\sqrt{n}(X_n - \theta) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2)$, so Slutsky's gives the desired result.

3 Change of Variable / Variable Transformation

3.1 The Discrete Case

Let X and $Y = g(X)$ be discrete. Then,

$$\mathbb{P}(Y = k) = \sum_{X:g(X)=k} \mathbb{P}(X = k).$$

3.2 The Continuous, 1D Case

Consider a continuous random variable X and transformation $Y = g(X)$ with g smooth with derivative vanishing nowhere, that is, $g' \neq 0$. These conditions guarantee that Y is also continuous. Suppose further that g is smooth and $g' \neq 0$. We have then that

$$f_Y(y) = f_X(x)\frac{1}{|g'(x)|} = f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

Example 3.1. Let $g' > 0$. Note that

$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Thus

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \leq y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right)\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y).$$

Similarly, if $g' < 0$, then

$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

Then

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \leq y) = -f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) = f_X(g^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

In general, if g is piecewise monotone, then

$$f_Y(y) = \sum_{\{x:g(x)=y\}} f_X(x)\frac{1}{|g'(x)|}.$$

Example 3.2. $Y = X^2$.

$$f_Y(y) = \sum_{\{x:x^2=y\}} f_X(x)\frac{1}{|2x|} = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{\left|-2\sqrt{y}\right|}.$$

3.3 The Continuous, 2D Case

Suppose X and Y are continuous with densities f_X and f_Y and joint density $f_{X,Y}$. Consider

$$\begin{cases} U = g_1(X, Y) \\ V = g_2(X, Y) \end{cases}$$

Suppose g_i's are smooth. Question: what is the joint density of U and V?

$$f_{U,V}(u, v) = f_{X,Y}(x, y)|J|,$$

where

$$\begin{aligned} J = \det \frac{\partial(x, y)}{\partial(u, v)} &= \det \begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix} \\ &= \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \end{aligned}$$

and we assume $J \neq 0$.

We have also

$$f_{XY}(x, y) = f_{UV}(u, v)|J|^{-1},$$

where

$$J^{-1} = \det \frac{\partial(u, v)}{\partial(x, y)} = \frac{1}{J}.$$

Example 3.3 (2D Change of Variables, Quotient). Let $U = X/Y$ with $Y \neq 0$ and $V = Y$. We have then that

$$\begin{cases} X = UV \\ Y = V \end{cases}.$$

Then

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{bmatrix} v & u \\ 0 & 1 \end{bmatrix}$$

so $J = v$. Alternatively,

$$\det \frac{\partial(u, v)}{\partial(x, y)} = \det \begin{bmatrix} \frac{1}{y} & -\frac{x}{y^2} \\ 0 & 1 \end{bmatrix} = \frac{1}{y} = J^{-1}.$$

This gives

$$f_{UV}(u, v) = f_{XY}(uv, v)|v|.$$

If we assume further that $X \perp\!\!\!\perp Y$, then

$$f_{UV}(u, v) = f_X(uv) f_Y(v)|v|.$$

4 Binomial and Poisson Distributions

Consider Binomial(n, p_n) with $np_n \to \lambda$ (thus $p_n \to 0$). We have then that

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.\end{aligned}$$

Since k is fixed, as $n \to \infty$, and recalling that $(1 + X/n)^n \to e^X$, we have

$$\mathbb{P}(X = k) \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

That is, if $np \to \lambda$, we have Binomial$(n, p) \to$ Poisson(λ). In this sense, the Poisson distribution is the limit of the binomial distribution, and we can thus use it as an approximation for the binomial distribution.

But note that for large n and np, since a Binomial is the sum of Bernoulli's, the CLT can be used to approximate the binomial distribution as $\mathcal{N}(np, np(1-p))$. Similarly, the Poisson distribution can be approximated as $\mathcal{N}(\lambda, \lambda)$.

Example 4.1. Let $X \sim$ Binomial(n, p). The CLT gives

$$\mathbb{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \le x\right) \approx \Phi(x)$$

and thus

$$\begin{aligned}\mathbb{P}(X = k) = \mathbb{P}(k-1 < X \le k) &= \mathbb{P}\left(\frac{k-1-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} \le \frac{k-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-1-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

It turns out that the **continuity correction** almost always gives a better approximation:

$$\begin{aligned}\mathbb{P}(X = k) = \mathbb{P}(k-0.5 < X \le k+0.5) &= \mathbb{P}\left(\frac{k-0.5-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-0.5-np}{\sqrt{np(1-p)}}\right).\end{aligned}$$

5 Correlation

5.1 Bivariate Normal

Let (X, Y) be bivariate normal. That is, X and Y have joint density

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right).$$

If $X \perp\!\!\!\perp Y$, then $\rho = 0$ and the joint density simplifies to $f_X(x) f_Y(y)$.

5.2 Bivariate Normal, the Standard Case

The standard bivariate normal (U, V) is

$$f_{UV}(u,v) = \frac{1}{2\pi(1-\rho^2)} \exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\rho uv)\right).$$

Note that

$$u^2 + v^2 - 2\pi uv = (v-\rho u)^2 + (1-\rho^2)u^2.$$

Thus we have

$$\begin{aligned} f_V(v) = \int_{-\infty}^{\infty} f_{UV}(u,v)\,\mathrm{d}v &= \frac{1}{2\pi(1-\rho^2)} \int_{-\infty}^{\infty} \exp\left(-\frac{(1-\rho^2)u^2}{2(1-\rho^2)}\right) \exp\left(-\frac{(v-\rho v)^2}{2(1-\rho^2)}\right) \mathrm{d}v \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(v-\rho u)^2}{2(1-\rho^2)}\right) \mathrm{d}v. \end{aligned}$$

So the marginal distributions are also normal. One can verity similarly that $\mathrm{E}[UV] = \rho$, giving

$$\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V) = \rho$$

and

$$\mathrm{Corr}(U,V) = \frac{\mathrm{Cov}(U,V)}{\mathrm{Var}(U)\,\mathrm{Var}(V)} = \rho.$$

Note that

$$\begin{aligned} f_{U|V}(u|v) = \frac{f_{UV}(u,v)}{f_V(v)} &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\rho uv)\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v^2}{2}\right)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\pi uv-(1-\rho^2)v^2)\right) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right). \end{aligned}$$

Thus,

$$U|V = v \sim \mathcal{N}(\rho v, (1-\rho^2))$$

and similarly

$$V|U = u \sim \mathcal{N}(\rho u, (1 - \rho^2)).$$

Thus the conditional expectation functions are

$$U = \rho V$$

and

$$V = \rho U.$$

Remark 5.1. This demonstrates regression toward the mean, since $\rho \leq 1$.

5.3 Bivariate Normal, the General Case

We may write

$$\begin{cases} X = \mu_X + \sigma_X U \sim \mathcal{N}(\mu_x, \sigma_X^2) \\ Y = \mu_Y + \sigma_Y V \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \end{cases}.$$

We have

$$\text{Cov}(X, Y) = \text{Cov}(\mu_X + \sigma_X U, \mu_Y + \sigma_Y V) = \sigma_X \sigma_Y \text{Cov}(U, V)$$

and thus

$$\text{Corr}(X, Y) = \rho.$$

From

$$\frac{X - \mu_X}{\sigma_X} | \frac{Y - \mu_Y}{\sigma_Y} \sim \mathcal{N}\left(\rho \cdot \frac{y - \mu_Y}{\sigma_Y}, 1 - \rho^2\right)$$

we know

$$X|Y = y \sim \mathcal{N}\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho^2)\right).$$

Similarly,

$$Y|X = x \sim \mathcal{N}\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho^2)\right).$$

Note that the variance is smaller.

5.4 Bivariate Normal Data

Suppose (x_i, y_i) are iid bivariate normal with

$$(x_i, y_i) \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{\sigma_1 \sigma_2} \\ \rho_{\sigma_1 \sigma_2} & \sigma_2^2 \end{bmatrix}\right).$$

Then $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and $\text{Cov}(X_i, Y_i) = \rho$.

5.5 MLE

We have

$$\begin{aligned}
&L(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) \\
&= \prod\prod f(x_i, y_i) \\
&= \frac{1}{(2\pi)^2(1-\rho^2)^{n/2}} \\
&\quad \exp\left(\frac{1}{2(1-\rho^2)} \sum \left(\left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 + \left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 - 2\pi\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(\frac{y_i-\mu_Y}{\sigma_Y}\right)\right)\right).
\end{aligned}$$

We have[4]

$$\hat{\mu}_X = \overline{X}, \quad \hat{\mu}_X = \overline{Y}, \quad \hat{\sigma}_X = n^{-1}\sum (X_i - \overline{X})^2, \quad \hat{\sigma}_Y = n^{-1}\sum (Y_i - \overline{Y})^2$$

and

$$R \coloneqq \hat{\rho} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}.$$

Definition 5.2. R is called the **Pearson correlation coefficient**.

5.6 Distribution of the Pearson Correlation Coefficient

It is easy to see that $\operatorname{supp} R = [-1, 1]$, but its distribution is quite complicated. We know however that

$$\mathrm{E}[\hat{\rho}] = \rho - \frac{\rho(1-\rho^2)}{2n} + \dots$$

and

$$\operatorname{Var}[\hat{\rho}] = \frac{(1-\rho^2)^2}{n} + \dots.$$

We have approximately that

$$\operatorname{Var}[\hat{\rho}] \propto (1-\rho^2)^2 = v(\rho).$$

By Fisher we know that for large n we have

$$\hat{\rho} \sim \mathcal{N}\left(\rho, \frac{(1-\rho^2)^2}{n}\right).$$

To implement VST, we seek g such that $g'(\rho) = 1/(1-\rho^2)$. A choice is

$$g(\rho) = \int \frac{1}{(1+\rho)(1-\rho)}\, \mathrm{d}\rho = \frac{1}{2}\int \frac{1}{1+\rho} + \frac{1}{1-\rho}\, \mathrm{d}\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho}.$$

[4] These are *not* unbiased estimators. For the unbiased estimators, we need to divide by $n-1$ instead of n.

This is called the **Fisher transformation**. Note that it is monotone (and thus the construction of CIs is somewhat easy).

We have that

$$\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} \sim \mathcal{N}\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right).$$

It turns out that $1/(n-3)$ is better than $1/n$ for $n > 5$.

Using this we can construct a CI for $g(\rho)$

$$\left[\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} - z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}, \frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} + z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}\right] = [L, U]$$

and for then ρ: Note that

$$T = \frac{1}{2}\log\frac{1+R}{1-R} \iff e^{2T} = \frac{1+R}{1-R} \iff R = \frac{e^{2T}-1}{e^{2T}+1} = \tanh T.$$

So a CI for ρ is

$$\left[\frac{e^{2L}-1}{e^{2L}+1}, \frac{e^{2U}-1}{e^{2U}+1}\right].$$

Remark 5.3. This is the standard way to construct a CI for the correlation coefficient in software packages, regardless of the distribution of the original data. In our derivation, however, we assumed bivariate normality. For small n, be mindful of the many layers of approximation!

5.6.1 The Matrix Notation

For the standard bivariate normal, with

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

we may write $[U, V]^\intercal \sim \mathcal{N}([0, 0]^\intercal, \Sigma)$, and then

$$f(u, v) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\exp\left(\frac{1}{2}[uv]^\intercal\Sigma^{-1}[uv]\right),$$

where

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

For the more general case, with $X \sim \mathcal{N}(\mu, \Sigma)$, we have

$$f(x) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\exp\left(\frac{1}{2}[x-\mu]^\intercal\Sigma^{-1}[x-\mu]\right),$$

5.6.2 Properties

Proposition 5.4. *For bivariate normally distributed variables, we have the variables are uncorrelated if and only if they are independence.*

Proposition 5.5. *If $X \sim \mathcal{N}_2(\mu, \Sigma)$ if and only if $aX + bY$ is of (univariate) normal distribution for any $a, b \in \mathbb{R}$.*

Example 5.6 (Non-example). Let $X \sim \mathcal{N}(0,1)$ and

$$Y = \begin{cases} X & \text{with probability } 1/2 \\ -X & \text{with probability } 1/2 \end{cases}.$$

Then $Y \sim \mathcal{N}(0,1)$, but $[XY]^\intercal$ is not of bivariate normal distribution. A way to see this is that the linear combination $X + Y$ has point mass at 0 with probability $1/2$.

Proposition 5.7. *If $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then*

$$\begin{bmatrix} X \\ Y \end{bmatrix} = A \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} a_{11}Z_1 + a_{12}Z_2 \\ a_{21}Z_1 + a_{22}Z_2 \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, AA^\intercal \right)$$

is bivariate normal.

Proposition 5.8. *If $\left[X_1 X_2\right]^\intercal \sim \mathcal{N}_2(\mu, \Sigma)$ and $[Y_1 Y_2]^\intercal = A[X_1 X_2]^\intercal$ is a linear transformation of $[X_1 X_2]^\intercal$ such that A^{-1} exists, then $[Y_1 Y_2]^\intercal \sim \mathcal{N}(A\mu, A\Sigma A')$. Similarly, $A[X_1 X_2]^\intercal + [b_1 b_2]^\intercal \sim \mathcal{N}_2([b_1 b_2]^\intercal + A\mu, A\Sigma A')$.*

5.7 Theorem of Sampling Distributions

5.7.1 The χ and t Distributions

χ^2 distribution. Let $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then

$$Z_1^2 + \cdots + Z_n^2 \sim \chi_k^2.$$

It turns out that $\chi_k^2 \sim \text{Gamma}(k/2, 1/2)$. If $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi_n^2$ with Z and W independent, then

$$\frac{Z}{\sqrt{W/k}} \sim t_k.$$

We have

$$f_{t_k}(t) = C_k \frac{1}{(1 + t^2/k)^{\frac{k+1}{2}}},$$

where C_k is a constant.

Remark 5.9.

- Note that as $k \to \infty$,
$$(1 + t^2/k)^{\frac{k+1}{2}} \longrightarrow e^{t^2/2}.$$

- $t_1 = Z_1/Z_2$ with $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ is also called the **Cauchy distribution**. It has density

$$f_1(t) = \frac{1}{\pi} \cdot \frac{1}{1+t^2}.$$

Note however that it does not have an expected value, since

$$\frac{1}{\pi} \int_0^\infty \frac{t}{1+t^2} \, \mathrm{d}t \longrightarrow \infty,$$

and similarly the left side also diverges. The variance also does not exist. The law of large number does not apply!

Theorem 5.10. *Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$. Then we have*

(i) $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

(ii) $\frac{1}{\sigma^2} \sum (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}$.

(iii) $\sum (X_i - \overline{X}_n)^2 \perp\!\!\!\perp \overline{X}_n$.

(iv) $\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$, *where* $S^2 = \frac{1}{n-1} \sum X_i$ *is the sample variance.*

These are exact distributions, not approximations.

Proof.

(i) Follows from linearity of the normal distribution.

(ii) & (iii) We prove by induction. Let $Z_i \coloneqq (X_i - \mu)/\sigma$. Then $\overline{Z}_n = n^{-1} \sum Z_i = (\overline{X} - \mu)/\sigma$. Thus we have

$$\frac{1}{\sigma^2} \sum (X_i - \overline{X}_n)^2 = \sum \left(\frac{X_i - \overline{X}_n}{\sigma} \right)^2 = \sum (Z_i - \overline{Z})^2.$$

We need thus only show that $\sum (Z_i - \overline{Z})^2 \sim \chi^2_{n-1}$ and is independent of $\overline{Z}_n$.

For $n = 2$, we have

$$(Z_1 - \overline{Z}_2)^2 + (Z_2 - \overline{Z}_2)^2 = \left(Z_1 - \frac{Z_1 + Z_2}{2} \right)^2 + \left(Z_2 - \frac{Z_1 + Z_2}{2} \right)^2 = \frac{(Z_1 - Z_2)^2}{2}.$$

Recalling that $(Z_1 - Z_2)/\sqrt{2} \sim \mathcal{N}(0,1)$, we have $(Z_1 - Z_2)^2/2 \sim \chi^2_1$.

Next, recall that for bivariate normal (X, Y) we have that $X \perp\!\!\!\perp Y$ if and only if $\rho = 0$. Thus, noting that $[Z_1 - Z_2 \; Z_1 + Z_2]^\intercal$ is bivariate normal and

$$\mathrm{Cov}(Z_1 - Z_2, Z_1 + Z_2) = \mathrm{Var}(Z_1) - \mathrm{Var}(Z_2) = 0,$$

we know that $Z_1 - Z_2 \perp\!\!\!\perp Z_1 + Z_2$ and thus $(Z_1 + Z_2)^2/2 \perp\!\!\!\perp (Z_1 + Z_2)/2 = \overline{Z}_2$.

For the general case, suppose that $\sum_{i=1}^{n}(Z_i - \overline{Z})n)^2$ is of χ^2_{n-1} and is independent of $\overline{Z}_n$. Note that

$$\begin{aligned}\sum_{i=1}^{n}(Z_i - \overline{Z}_{n+1})^2 &= \sum_{i=1}^{n}(Z_i - \overline{Z}_n + \overline{Z}_n - \overline{Z}_{n+1})^2 + (Z_{n+1} - \overline{Z}_{n+1}) \\ &= \sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2 + 2\sum_{i=1}^{n}(Z_1 - \overline{Z}_n)(\overline{Z}_n - \overline{Z}_{n+1}) \\ &\quad + \sum_{i=1}^{n}(\overline{Z}_n - \overline{Z}_{n+1})^2 + (Z_{n+1} - \overline{Z}_{n+1})^2.\end{aligned}$$

Note that the first term is χ^2_{n-1} and the second term is 0. Thus we need only show that

$$J \coloneqq n(\overline{Z}_n - \overline{Z}_{n-1})^2 + (Z_{n+1} - \overline{Z}_{n+1})^2$$

is of χ^2_1 distribution and is independent of the first term.

Observe now that

$$\overline{Z}_{n+1} = \frac{n}{n+1}\overline{Z}_n + \frac{Z_{n+1}}{n+1},$$

and thus

$$\begin{aligned}J &= n\left(\overline{Z}_n - \frac{n}{n+1}\overline{Z}_n - \frac{1}{n+1}Z_{n+1}\right)^2 + \left(Z_{n+1} - \frac{n}{n+1}\overline{Z}_n - \frac{1}{n+1}Z_{n+1}\right) \\ &= \frac{n}{(n+1)^2}(\overline{Z}_n - Z_{n+1})^2 + \frac{n^2}{(n+1)^2}(Z_{n+1} - \overline{Z}_n)^2 \\ &= \frac{n+n^2}{(n+1)^2}(\overline{Z}_n - Z_{n-1})^2 = \frac{n}{n+1}(\overline{Z}_n - Z_{n+1})^2.\end{aligned}$$

Since $\overline{Z} - Z_{n+1} \sim \mathcal{N}(0, 1/n+1)$, we know that J is of χ^2_1 distribution. For claim (ii), it remains to show that $J = n/(n+1)\cdot(\overline{Z}_n - Z_{n+1})^2$ is independent to $\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$. Since $Z_{n+1}, \overline{Z}_n$, and $\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$ are pairwise independent, we obtain (ii). To complete the proof, note that a similar argument as before shows (iii).

(iv) We have

$$\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum(X_i - \overline{X})^2}{\sigma^2}/(n-1)}} = \frac{Z}{\sqrt{W/n-1}},$$

where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_n$.

□

5.8 Two-Sample

Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ and we want to estimate $\mu_1 - \mu_2$. A choice of an estimator is $\hat{\mu}_1 - \hat{\mu}_2 = \overline{X} - \overline{Y}$, but how can we obtain an CI?

Case 1: A "paired" sample. Suppose $\text{Cov}(X_1, Y_1) > 0$. We can use

$$\text{Var}(\overline{X} - \overline{Y}) = \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) - 2\,\text{Cov}(\overline{X}, \overline{Y}) < \text{Var}(\overline{X}) + \text{Var}(\overline{Y})$$

to obtain a narrower CI.

Define $D_i \coloneqq X_i - Y_i$ to reduce the problem to that of a one sample problem. We have in particular that

$$D_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right),$$

where $\rho \coloneqq \text{Corr}(X_i, Y_i)$.

If σ is unknown, we may use $\hat{\text{Var}}[\overline{D}] = S_D^2/n$ to obtain

$$\frac{\overline{D}_i - (\mu_1 - \mu_2)}{\sqrt{S_D^2/n}} \sim t_{n-1}.$$

Case 2: $\text{Cov}(X_i, Y_i) = 0$ and $\sigma_1 = \sigma_2$. Note that the normality assumption gives $X_i \perp\!\!\!\perp Y_i$. Then, $\text{Var}[\overline{X} - \overline{Y}] = \text{Var}[\overline{X}] + \text{Var}[\overline{Y}] = \sigma^2/n + \sigma^2/m$. We may use the sample variances to estimate σ^2:

$$S_X^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2, \quad S_Y^2 = \frac{1}{m-1}\sum (Y_i - \overline{Y})^2,$$
$$S_{\text{pool}}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

We use the last term for higher power and narrower CI. We use

$$\hat{\text{Var}}[\overline{X} - \overline{Y}] = S_{\text{pool}}^2\left(\frac{1}{n} + \frac{1}{m}\right)$$

to obtain

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_{\text{pool}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

and

$$\frac{S_{\text{pool}^2}}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Case 3: $X_i \perp\!\!\!\perp Y_i$ and $\sigma_1 \neq \sigma_2$. We have $\text{Var}[\overline{X} - \overline{Y}] = \sigma_1^2/n + \sigma_2^2/m$. Then, approximately, we have

$$T \coloneqq \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0, 1).$$

It turns out that for moderate sized n and m, a better approximation, called the **Welch–Satterthwaite** approximation is

$$T \sim t_\nu,$$

where ν is a function of sample variances and sample sizes. This is what is implemented in software packages.

5.8.1 Hypothesis Testing for Two-Sample

We define the likelihood ratio as

$$LR \coloneqq \frac{\max_{H_0} \text{likelihood function}}{\max_{H_0 \cup H_a} \text{likelihood function}} \in (0, 1].$$

Note that is H_0 is true, LR is likely to be large (close to 1).

Remark 5.11. To maximize $f \in C^2$, we impose:

- (FOC) $\nabla f(x) = 0$
- (SOC) $H(x) = \left[\partial^2 f / \partial x_i \partial x_j\right]$ to be negative definite.

Definition 5.12. Let $A = [a_{ij}]_{p \times p}$ be a symmetric matrix. We say A is **negative definite** if one of the following equivalent statements hold:

- All eigenvalues of A are negative.
- Let M_k be the determinant of the kth leading principal minor. If $M_k < 0$ when k is odd, and $M_k > 0$ when k is even.

Proposition 5.13. *Under H_0, we have the following approximate asymptotic distribution of LR:*

$$-2 \log LR \approx \chi^2_d,$$

where $d = \dim(H_a \cup H_0) - \dim H_0$.

Example 5.14. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma^2)$ and $X_i \perp\!\!\!\perp Y_i$. Consider the hypothesis $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$. Under H_0 we have $\mu_1 = \mu_2 = \mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}$, so $\dim H_0 = 2$. Similarly we have $\dim(H_a \cup H_0) = 3$. Thus $d = 1$.

We have

$$\begin{aligned} L(\mu_1, \mu_2, \sigma^2) &= \prod f_X(x_i) f_Y(y_i) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu_1)^2 - \frac{1}{2\sigma^2} \sum (y_i - \mu_2)^2\right). \end{aligned}$$

The log-likelihood function is

$$\ell(\mu_1, \mu_2, \sigma^2) = -\frac{n+m}{2} \log(2\pi) - \frac{n+m}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu_1)^2 - \frac{1}{2\sigma^2} \sum (y_i - \mu_2)^2.$$

Under H_0 we have $\mu_1 = \mu_2 = \mu_0$, and for large n,

$$-2 \log LR \approx \chi_1^2.$$

Further, under H_0,

$$\frac{\partial \log L}{\partial \mu_0} = -\frac{1}{\sigma^2} \sum (x_i - \mu_0) - \frac{1}{\sigma^2} \sum (y_i - \mu_0).$$

Setting $\partial \log L / \partial \mu_0 = 0$, we get

$$\hat{\mu}_0 = \frac{\sum x_i + \sum y_i}{n + m}$$

for any σ^2. Thus

$$\max_{H_0} L(\mu_1, \mu_2, \sigma^2) = L(\hat{\mu}_0).$$

Next, we have

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n+m}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum (x_i - \mu_1)^2 + \frac{1}{2(\sigma^2)^2} \sum (y_i - \mu_2)^2.$$

Setting $\partial \log L / \partial \sigma^2 = 0$ at $\mu_1 = \mu_2 = \hat{\mu}_0$, we get

$$\hat{\sigma}_0^2 = \frac{1}{n+m} \left(\sum (x_i - \hat{\mu}_0)^2 + \sum (y_i - \hat{\mu}_0)^2 \right).$$

We have then that

$$\begin{aligned} \max_{H_0} L(\mu_1, \mu_2, \sigma^2) &= L(\hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0^2) \\ &= \left(\frac{1}{\sqrt{\pi \hat{\sigma}_0^2}} \right)^{n+m} \exp\left(-\frac{\sum (x_i - \hat{\mu}_0)^2 + \sum (y_j - \hat{\mu}_0)^2}{2\hat{\sigma}_0^2} \right) \\ &= \left(\frac{1}{\sqrt{\pi \hat{\sigma}_0^2}} \right)^{n+m} \exp\left(-\frac{n+m}{2} \right). \end{aligned}$$

Under $H_0 \cup H_a$: we have

$$\frac{\partial \log L}{\partial \mu_1} = \frac{1}{\sigma^2} \sum (x_i - \mu_1), \quad \frac{\partial \log L}{\partial \mu_2} = \frac{1}{\sigma^2} \sum (y_i - \mu_2).$$

Setting the above terms to zero gives

$$\hat{\mu}_1 = \overline{X}, \quad \hat{\mu}_2 = \overline{Y}.$$

Again, setting $\partial \log L / \partial \sigma^2 = 0$ at $\mu_1 = \hat{\mu}_1$ and $\mu_2 = \hat{\mu}_2$, we get

$$\hat{\sigma}^2 = \frac{1}{n+m} \left(\sum (x_i - \hat{\mu}_1)^2 + \sum (y_i - \hat{\mu}_2)^2 \right).$$

We have then that

$$\begin{aligned}\max_{H_0 \cup H_a} L(\mu_1, \mu_2, \sigma^2) &= L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) \\ &= \left(\frac{1}{\sqrt{\pi\hat{\sigma}^2}}\right)^{n+m} \exp\left(-\frac{\sum(x_i - \hat{\mu}_1)^2 + \sum(y_j - \hat{\mu}_2)^2}{2\hat{\sigma}^2}\right) \\ &= \left(\frac{1}{\sqrt{\pi\hat{\sigma}^2}}\right)^{n+m} \exp\left(-\frac{n+m}{2}\right).\end{aligned}$$

Then,

$$LR = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n+m}{2}} = \left(\frac{\sum(x_i - \overline{x})^2 + \sum(y_i - \overline{y})^2}{\sum(x_i - \hat{\mu}_0)^2 + \sum(y_i - \hat{\mu}_0)^2}\right)^{\frac{n+m}{2}}.$$

6 ANOVA (Analysis of Variance)

This chapter is very wrong

For $i = 1, \ldots, g$ with $g > 2$ suppose $Y_{i1}, \ldots, Y_{in} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. Consider $H_0 : \mu_1 = \cdots = \mu_g$ and H_a: $\mu_i \neq \mu_j$ for some $i \neq j$. (Consider for example the outcomes after different treatment levels i.)

Note that there are two types of variation: within group and across group. Denote the mean of the whole sample as $\overline{y} = y_{\cdot\cdot}$. We may decompose the variance accordingly:

$$\begin{aligned}\sum_{i=1}^{g}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}\right)^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n}\left[\left(y_{ij}-\overline{y}_i\right)-\left(\overline{y}_i-\overline{y}\right)\right]^2 \\ &= \sum_{i=1}^{g}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2+\sum_{i=1}^{g} n_i\left(\overline{y}_i-\overline{y}\right)^2,\end{aligned}$$

where the first term is within group variation (the noise), and the second term is the across group variation (the treatment effect, the signal). Equivalently, we write this decomposition as

$$SS_{\text{total}} = SS_{\text{residual}} + SS_{\text{treatment}},$$

where SS stands for sum of squares.

6.0.1 ANOVA Table

Source	SS	df	MS	Var Ratio (F)
Treatment	SS_{trt}	$g-1$	$SS_{\text{trt}}/(g-1)$	$MS_{\text{trt}}/MS_{\text{e}}$
Residual (error)	SS_{e}	$n-g$	$SS_{\text{e}}/(n-g)$	
Total	SS_{total}	$n-1$		

6.0.2 F distribution

If $W_1 \sim \chi^2_{k_1}$ and $W_2 \sim \chi^2_{k_2}$ with $W_1 \perp\!\!\!\perp W_2$, then

$$\frac{W_1/k_1}{W_2/k_2} \sim F_{k_1,k_2}.$$

Note that supp $F \subset \mathbb{R}_+$.

Proposition 6.1. *We have under H_0 that*

$$\frac{MS_{trt}}{MS_e} \sim F_{g-1,n-g}.$$

Perhaps some intuition:
Note that

$$\begin{aligned}\mathrm{E}[SS_{\text{e}}] &= \mathrm{E}\left[\sum_i (n_i-1)\sum\frac{(Y_{ij}-\overline{Y}_i)^2}{n_i-1}\right] \\ &= \sum_i (n_i-1)\,\mathrm{E}[S_i^2] = \sum_i (n_i-1)\sigma^2 = (n-g)\sigma^2.\end{aligned}$$

Recalling that

$$\frac{1}{\sigma^2}\sum_j (Y_{ij} - \overline{Y}_i)^2 \sim \chi^2_{n_i - 1},$$

we have

$$\frac{1}{\sigma^2} SS_{\text{e}} \sim \chi^2_{n-g}.$$

Next,

$$\begin{aligned}\mathrm{E}[SS_{\text{trt}}] = \mathrm{E}\left[\sum n_i (\overline{Y}_i - \overline{Y})^2\right] &= \sum n_i \,\mathrm{E}\left[(\overline{Y}_i - \overline{Y})^2\right] \\ &= \sum n_i \left[\mathrm{Var}[\overline{Y}_i - \overline{Y}] + [\mathrm{E}(\overline{Y}_i - \overline{Y})]^2\right].\end{aligned}$$

Now note that

$$\mathrm{E}[\overline{Y}_i - \overline{Y}] = \mathrm{E}\left[\frac{1}{n_i}\sum_j Y_{ij}\right] - \mathrm{E}\left[\frac{1}{n}\sum_i\sum_j Y_{ij}\right] = \mu_i - \frac{1}{n}\sum_i n_i\mu_i =: \mu_i - \mu.$$

Noting that $\mathrm{Cov}(Y_{ij}, Y_{kl} = 0)$ for $i \neq k$ and $\mathrm{Cov}(Y_{ij}, Y_{ik}) = 0$ for $j \neq k$, we have

$$\mathrm{Cov}(\overline{Y}_i, \overline{Y}) = \mathrm{Cov}\left(\frac{1}{n}\sum_i Y_{ij}, \frac{1}{n}\sum_i\sum_j Y_{ij}\right) = \frac{1}{n}\frac{1}{n_i}\sum_i \mathrm{Var}(Y_{ij}) = \frac{\sigma^2}{n}.$$

We have thus that

$$\mathrm{Var}(\overline{Y}_i - \overline{Y}) = \mathrm{Var}(\overline{Y}_i) + \mathrm{Var}(\overline{Y}) - 2\,\mathrm{Cov}(\overline{Y}_i, \overline{Y}) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} = \frac{\sigma^2}{n_1} - \frac{\sigma^2}{n}.$$

Plugging back,

$$\mathrm{E}[SS_{\text{trt}}] = \sum n_i \left[\left(\frac{\sigma^2}{n_i} - \frac{\sigma^2}{n}\right) + (\mu_i - \mu)^2\right] = (g-1)\sigma^2 + \sum_i n_i(\mu_i - \mu)^2.$$

Then,

$$\mathrm{E}[MS_{\text{trt}}] = \frac{(g-1)\sigma^2}{g-1} + \sum_i n_i(\mu_i - \mu)^2 = \sigma^2 + \frac{1}{g-1}\left(\sum_i (\mu_i - \mu)^2\right).$$

Under H_0, there holds $\mathrm{E}[MS_{\text{trt}}] = \sigma^2$. Moreover,

$$\frac{\sum\sum(Y_{ij} - \overline{Y})^2}{\sigma^2} = \frac{\sum\sum(Y_{ij} - \overline{Y})^2}{\sigma^2} + \frac{\sum n_i(\overline{Y}_i - \overline{Y})}{\sigma^2} \sim \chi^2_{n-1}.$$

6.0.3 Relation to Linear Models

We may rewrite the assumptions for one-way ANOVA as: $Y_{ij} = \mu_i + \epsilon_{ij}$ with $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. (Alternatively, this may be viewed as an assumption of the data generation process.)

Proposition 6.2. *The MLE of μ_i is $\overline{y}_i$.*

Another way to parameterize the model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. We impose constraint on the parameters of the form $\sum \alpha_i = 0$, or $\alpha_0 = 0$, or $\alpha_g = 0$.

These are constraints that we can freely impose without loss of generality; they may be viewed as a way to define μ. But we have always that $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$.

Example 6.3. Use $\partial \log L / \partial \alpha_i = 0$ for $i = 0, \ldots, g-1$ and use constraint $\sum \alpha_i = 0$ to derive MLE for α_i and μ. It turns out that $\hat{\mu} = \sum \overline{y}_i / g$.

7 Two Way ANOVA

This chapter is very wrong

	SS	df		SS	df
trt_A		$I-1$	trt_B		$J-1$
Res		$n-I$	Res		$n-J$

Consider y_{ijk}, where $i = 1, \dots, I$ is treatment level of A, $j = 1, \dots, J$ is treatment level of B, and $k = 1, \dots, K_{ij}$ is the replicate. Consider first the case where $K_{ij} = K$. We denote the means as:

- $\overline{y}_{i\cdot\cdot} \coloneqq \frac{1}{IK} \sum_{j-1}^{J} \sum_{k}^{K} y_{ijk}$.
- $\overline{y}_{\cdot j\cdot} \coloneqq \frac{1}{JK} \sum_{i-1}^{I} \sum_{k}^{K} y_{ijk}$.

Source	SS	df	MS	F
Treatment A	$SS_A = JK \sum_i (\overline{y}_{i\cdot\cdot} - \overline{y})^2$	$I-1$	$SS_A/(I-1)$	MS_A/MS_e
Treatment B	$SS_B = IK \sum_j (\overline{y}_{\cdot j\cdot} - \overline{y})^2$	$J-1$	$SS_B/(J-1)$	MS_B/MS_e
A × B	SS_{AB}	$(I-1)(J-1)$		MS_{AB}/MS_e
Residue	$SS_e = \sum_{ijk} (y_{ijk} - \overline{y}_{ij\cdot})^2$	df_e		
Total	$SS_{\text{total}} = \sum_{ijk} (y_{ijk} - \overline{y})^2$	$IJK-1$		

When there is no interactions, we have the partition

$$SS_{\text{total}} = \sum_{ijk} (y_{ijk} - \overline{y})^2 = SS_A + SS_B + SS_e.$$

7.1 Relation to Linear Models

We may write the model as

$$y_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2)$$

or

$$y_{ijk} \overset{\text{iid}}{\sim} \mu_{ij} + \epsilon_{ijk}$$

with $\epsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$. That is,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

A typical constraint is $\sum \alpha_i = \sum \beta_j = \sum_j \gamma_{1j} = \sum_i \gamma_{i1} = 0$.

Proposition 7.1. *We have the MLE*

$$\hat{\mu} = \overline{y}, \quad \hat{\alpha}_i = \overline{y}_{i\cdot\cdot} - \overline{y}, \quad \hat{\beta}_j = \overline{y}_{\cdot j\cdot} - \overline{y},$$
$$\hat{\gamma}_{ij} = \overline{y}_{ij\cdot} - (\hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = \overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}.$$

7.2 Hypothesis testing

- $H_A : \alpha_i = 0, \forall i$. We have $\frac{MS_A}{MS_e} \sim F_{I-1,df_e}$.
- $H_B : \beta_i = 0, \forall j$. We have $\frac{MS_B}{MS_e} \sim F_{J-1,df_e}$.
- $H_{AB} : \gamma_{ij} = 0, \forall i, j$. We have $\frac{MS_{AB}}{MS_e} \sim F_{df_{AB},df_e}$.

Example 7.2. Consider the case of no replicates.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

$y_{ij} = \hat{\mu}_{ijk}$	B-1	B-2	B-3	B-4	row sum	row mean
A-1	3	2	1	2	8	2
A-2	8	7	8	5	28	7
A-3	7	6	9	2	24	6
col sum	18	15	18	9	60	
col mean	6	5	6	3		

We have then that $\hat{\mu} = 60/12 = 5$. We subtract the mean.

$\hat{\mu}_{ijk} - \hat{\mu}$	B-1	B-2	B-3	B-4	row sum	$\hat{\alpha}_i$
A-1	-2	-3	-4	-3	12	-3
A-2	3	2	3	0	8	2
A-3	2	1	4	-3	4	1
col sum	3	0	3	-6		
$\hat{\beta}_j$	1	0	1	-2		

We then subtract the row and column means to obtain the interactions (or residual if no interaction in model).

$\hat{\mu}_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$	B-1	B-2	B-3	B-4
A-1	0	0	-2	2
A-2	0	0	0	0
A-3	0	0	2	-2

We have the following ANOVA table:

Source	SS	df
A	SS_A	$I - 1$
B	SS_B	$J - 1$
Res	SS_E	$IJK - 1$

Definition 7.3 (Nested Models)**.** Consider models

$$M_1 : \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \cdots + \beta_{p+q} x_{pq} + \epsilon$$

and

$$M_2 : \beta_0 + \beta_{p+1} x_{p+1} + \cdots + \beta_{p+q} x_{pq} + \epsilon.$$

We say M_2 is nested in M_1 and write $M_2 \subset M_1$.

Proposition 7.4. *Let $M_{small} \subset M_{large}$ and denote the residual sum of squares SS_e as RSS. We have then that $\mathrm{RSS}_{large} \leq \mathrm{RSS}_{small}$. Moreover, we have*

$$\mathrm{RSS}_{small} = \mathrm{RSS}_{large} + \sum (\textit{fitted value}_{small} - \textit{fitted value}_{large})^2.$$

8 Comparison of Nested Models, Sequential ANOVA

Suppose

$$Y_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2), \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad k = 1, \dots, K_{ij}.$$

Consider the model $\mu_{ij} = \mu + \alpha_i + \beta_j$ with constraints $\sum \alpha_i = \sum \beta_j = 0$.

Proposition 8.1. *We have the following MLE:*

- $\hat{\mu} = \overline{y}$
- $\hat{\alpha}_i = \overline{y}_{i\cdot} - \overline{y}$
- $\hat{\beta}_j = \overline{y}_{\cdot j} - \overline{y}$

If $K_{ij} \equiv K$, by writting

$$y_{ijk} - \overline{y} = y_{ijk} - \overline{y} - (\overline{y}_{i\cdot} - \overline{y}) - (\overline{y}_{\cdot j} - \overline{y}) + \overline{y} - \overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}$$

and recalling

$$(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc,$$

we have

$$\begin{aligned} \sum_{ijk} (y_{ijk} - \overline{y})^2 &= \sum_{ijk} (y_{ijk} - \overline{y}_i - \overline{y}_j + \overline{y})^2 + \sum_{ijk} (\overline{y}_i - \overline{y})^2 + \sum_{ijk} (\overline{y}_j - \overline{y})^2 \\ &= SS_e + SS_A + SS_B, \end{aligned}$$

where the remaining three terms are 0 by the constraints and $K_{ij} \equiv K$.

8.1 Residual Sum of Squares

Example 8.2. Consider the two sample case with

$$H_0 : \mu_1 = \mu_2, \quad H_a : \mu_1 \neq \mu_2.$$

Consider the models

$$\begin{aligned} M_0 &: X_i = \mu_0 + \epsilon_i, \quad Y_j = \mu_0 + \epsilon_j, \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \\ M_1 &: X_i = \mu_1 + \epsilon_i, \quad Y_j = \mu_2 + \epsilon_j, \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \end{aligned}$$

Note that M_0 corresponds to the null hypothesis, and M_1 includes both the null and the alternative hypotheses.

We have

$$\hat{\mu}_0 = \frac{\sum X_i + \sum Y_j}{n + m}, \quad \hat{\mu}_1 = \overline{X}, \quad \hat{\mu}_2 = \overline{Y}.$$

The residual sum of squares are

$$RSS_0 = \sum (X_i - \hat{\mu}_0)^2 + \sum (Y_j - \hat{\mu}_0)^2, \quad RSS_1 = \sum (X_i - \hat{\mu}_1)^2 + \sum (Y_j - \hat{\mu}_2)^2.$$

We have always that $RSS_0 \geq RSS_1$.

We write

$$\sum (X_i - \hat{\mu}_0)^2 = \sum (X_i - \hat{\mu}_1 + \hat{\mu}_1 - \hat{\mu}_0)^2 = \sum (X_i - \hat{\mu}_1)^2 + \sum (\hat{\mu}_1 - \hat{\mu}_0)^2$$

and similarly for Y_j. Then,

$$\begin{aligned} RSS_0 &= RSS_1 + \sum (\hat{\mu}_1 - \hat{\mu}_0)^2 + \sum (\hat{\mu}_2 - \hat{\mu}_0)^2 \\ &= RSS_1 + \sum (\text{fitted value}_1 - \text{fitted value}_0)^2. \end{aligned}$$

Example 8.3. In one way ANOVA we have

$$\sum (y_{ij} - \overline{y})^2 = \sum (\overline{y}_i - \overline{y})^2 + \sum (y_{ij} - \overline{y}_i)^2.$$

That is,

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{residual}}.$$

Consider

$$H_0 : \mu_1 = \cdots = \mu_g, \quad H_a : \mu_i \neq \mu_j \text{ for some } i \neq j.$$

The corresponding models are

$$\begin{aligned} M_0 &: Y_{ij} = \mu_0 + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \\ M_1 &: Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \end{aligned}$$

We have then that $\hat{\mu}_0 = \overline{y}$ and $\hat{\mu}_i = \overline{y}_i$.

$$SS_{\text{total}} = RSS_0, \quad SS_{\text{treatment}} = \sum (\text{fitted value}_1 - \text{fitted value}_0)^2, \quad SS_{\text{residual}} = RSS_1.$$

8.2 Sequential ANOVA

Consider models

$$\begin{aligned} M_0 &: \quad Y_{ijk} = \mu + \epsilon_{ijk} \\ M_1 &: \quad Y_{ijk} = \mu + \alpha_i + \epsilon_{ij} \\ M_1^* &: \quad Y_{ijk} = \mu + \beta_j + \epsilon_{ij} \\ M_2 &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \end{aligned} \qquad M_3 : \quad Y_{ijk} \quad = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

Source	SS	df
A	$SS_A = \sum (\overline{y}_i - \overline{y})^2$	
Residual	$SS_E = \sum (y - \overline{y}_i)$	

Source	SS	df
B	$SS_B = \sum (\overline{y}_j - \overline{y})^2$	
Residual	SS_E	

Source	SS	df
A	SS_A	
B	SS_B	
Residual	SS_E	

Source	SS	df
A	SS_A	
B	SS_B	
A × B	SS_{AB}	
Residual	SS_E	

In general, consider a string of models

$$M_0 \subset M_1 \subset \cdots \subset M_n.$$

The parameter space gets larger:

$$\Theta_0 \subset \Theta_1 \subset \cdots \subset \Theta_n.$$

Sequential ANOVA SS columns: we are effectively partitioning the residual sum of squares as follows

$$\begin{aligned} RSS_0 &= RSS_0 - RSS_1 \\ &\quad + RSS_1 - RSS_2 \\ &\quad \vdots \\ &\quad + RSS_{n-1} - RSS_n \\ &\quad = RSS_n. \end{aligned}$$

Remark 8.4. The effect in interested should be conducted in the last step.

We have then that for $H_1 : \alpha_o = i \equiv 0$,

$$\frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_n/df_n} \sim F_{df_0 - df_1, df_n}$$

and so on.

8.3 Design Matrix

Consider the model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

We may write this as

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1J} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{IJ} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{bmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Note in particular that we have $\mathbf{Y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times(g+1)}$, $\boldsymbol{\beta} \in \mathbb{R}^{(g+1)\times 1}$, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$. Note that the rank of X is g, since one column can be written as a linear combination of the others.

To make the design matrix full rank, we need parameter constrains such as $\sum \alpha_i = 0$. Under this constraint, we have $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$. We may then write the design matrix as

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1J} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{IJ} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & -1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & -1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{I-1} \end{bmatrix}.$$

The design matrix is now full rank.

9 Simple Linear Regression

Recall that with (x_i, y_i) bivariate normal we have the following MLE:

$$\hat{\mu}_x = \overline{x}, \quad \hat{\mu}_y = \overline{y}, \quad \hat{\sigma}_x^2 = \frac{1}{n}\sum (x_i - \overline{x})^2, \quad \hat{\sigma}_y^2 = \frac{1}{n}\sum (y_i - \overline{y})^2$$

and

$$\hat{\rho} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}.$$

We have also

$$Y|X = x \sim \mathcal{N}\left(\mu_y + \frac{\sigma_Y}{\sigma_x}\rho(x - \mu_x), \sigma_y^2(1 - \rho^2)\right).$$

We consider now an alternative model which focuses on the distribution of one variable conditioning on a specific value of the other variables. Suppose we have data $(x_1, y_1), \ldots, (x_n, y_n)$. Writing

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x,$$

we have comparing to the previous results for bivariate normal variables that

$$\beta_0 = \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x, \quad \beta_1 = \rho\frac{\sigma_y}{\sigma_x}.$$

Consider thus the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2).$$

Definition 9.1 (Least Square Estimator)**.** The least square estimators of $\{\beta_0, \beta_1\}$ are the values that minimize

$$\sum (y_i - (\beta_0 + \beta_1 x))^2 \coloneqq \mathrm{RSS}\,.$$

That is,

$$\{\hat{\beta}_0, \hat{\beta}_1\} \coloneqq \arg\min \mathrm{RSS}\,.$$

Note that

$$\frac{\partial\,\mathrm{RSS}}{\partial\beta_0} = -2\sum (y_i - \beta_0 - \beta_1 x)\,, \quad \frac{\partial\,\mathrm{RSS}}{\partial\beta_1} = -2\sum x_i\,(y_i - \beta_0 - \beta_1 x)\,x.$$

Setting the above to 0 gives

$$n^{-1}\sum y_i - \beta_0 - \beta_1 n^{-1}\sum x_i = 0, \quad n^{-1}\sum x_i y_i - n^{-1}\sum x_i\beta_0 - n^{-1}\sum \beta_1 x_i^2 = 0$$

and thus

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum x_i y_i - (\frac{1}{n}\sum x_i)(\frac{1}{n}\sum y_i)}{\frac{1}{n}\sum x_i^2 - (\frac{1}{n}\sum x_i)^2} = \frac{\text{sample covariance of } x_i \text{ and } y_i}{\text{sample variance of } x_i}.$$

9.1 Properties of least square estimators

Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are random. Think of the x_i's the formula for $\hat{\beta}_1$ as fixed, and only y_i's are random:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum x_i y_i - (\frac{1}{n}\sum x_i)(\frac{1}{n}\sum y_i)}{\frac{1}{n}\sum x_i^2 - (\frac{1}{n}\sum x_i)^2}.$$

We may rewrite the above as

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \overline{x}\sum y_i}{\sum x_i^2 - \overline{x}\sum x_i} = \frac{\sum (x_i - \overline{x}) y_i}{\sum (x_i - \overline{x}) x_i} = \frac{\sum (x_i - \overline{x}) y_i}{\sum (x_i - \overline{x})^2} = \sum c_i y_i,$$

a linear combination of the y_i's.

Alternatively, we can write

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2},$$

the familiar covariance over variance form.

From the linear combination form, we have

$$\mathrm{E}[\hat{\beta}_1] = \mathrm{E}\left[\sum c_i Y_i\right] = \sum c_i \,\mathrm{E}[Y_i] = \sum c_i(\beta_0 + \beta_1 x) = \left(\sum c_i\right)\beta_0 + \left(\sum c_i x_i\right)\beta_0.$$

Recalling that

$$c_i = \frac{x_i - \overline{x}}{\sum (x_i - \overline{x})^2},$$

we see that

$$\sum c_i = 0, \quad \sum c_i x_i = \sum \frac{(x_i - \overline{x})(x_i - \overline{x})}{\sum (x_i - \overline{x})^2} = 1.$$

Therefore $\mathrm{E}[\hat{\beta}_1] = \hat{\beta}_1$. Moreover, we have

$$\mathrm{E}[\hat{\beta}_0] = \mathrm{E}[\overline{Y} - \hat{\beta}_1 \overline{x}] = \mathrm{E}[\overline{Y}] - \overline{x}\,\mathrm{E}[\hat{\beta}_1] = (\beta_0 + \beta_1 \overline{x}) - \overline{x}\beta_1 = \beta_0.$$

That is, we have:

Proposition 9.2. *The least square estimators are unbiased.*

We consider next the variance. We have

$$\begin{aligned}\mathrm{Var}[\hat{\beta}_1] = \mathrm{Var}[\sum c_i Y_i] &= \sum c_i \,\mathrm{Var}[Y_i] \\ &= \left(\sum c_i^2\right)\sigma^2 = \sigma^2 \frac{1}{\sum (x_i - \overline{x})^2} = \frac{\sigma^2}{n}\frac{1}{n^{-1}\sum (x_i - \overline{x})^2},\end{aligned}$$

where we used the fact that

$$\sum c_i^2 = \sum \frac{(x_i - \overline{x})^2}{(x_i - \overline{x})^4} = \frac{1}{\sum (x_i - \overline{x})^2}.$$

Remark 9.3. The variance is large if either the noise σ^2 is large or x_i's are concentrated ($n^{-1}\sum(x_i - \overline{x})^2$ is small).

For $\hat{\beta}_0$, we can write

$$\begin{aligned}\operatorname{Var}[\hat{\beta}_0] = \operatorname{Var}[\overline{Y} - \hat{\beta}_1\overline{x}] &= \operatorname{Var}[\overline{Y}] + \overline{x}^2 \operatorname{Var}[\hat{\beta}_1] - 2\overline{x}\operatorname{Cov}(\overline{Y}, \hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{\sum(x_i - \overline{x})^2},\end{aligned}$$

where the last inequality comes from

$$\operatorname{Cov}(\overline{y}, \hat{\beta}_1) = \frac{1}{n}\operatorname{Cov}\left(\sum Y_i, \sum c_i Y_i\right) = \frac{1}{n}\sum c_i \operatorname{Cov}(Y_i, Y_i) = 0.$$

$$\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{Cov}(\overline{Y} - \hat{\beta}_1\overline{x}, \hat{\beta}_1) = \operatorname{Cov}(\overline{Y}, \hat{\beta}_0) - \overline{x}\operatorname{Var}(\hat{\beta}_1) = -\frac{\overline{x}\sigma^2}{\sum(x_i - \overline{x})^2}$$

9.2 Distributional results

We now consider the model

$$\begin{aligned}& Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \\ \iff & Y_i = \beta_0 + \beta_1 x_i + \sigma^2 Z_i, \quad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \\ \iff & Y_i \text{ independent }, \quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).\end{aligned}$$

Proposition 9.4. *Under the assumption above, the MLE of β_0 and β_1 are the least square estimators.*

With the added bivariate normality assumption, we have $[\hat{\beta}_0\hat{\beta}_1]^\intercal$ is bivariate normal as a linear combination of the data. Specifically, with the results above, we have

$$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}, \sigma^2\begin{bmatrix}\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i-\overline{x})^2} & -\frac{\overline{x}}{\sum(x_i-\overline{x})^2} \\ -\frac{\overline{x}}{\sum(x_i-\overline{x})^2} & \frac{1}{\sum(x_i-\overline{x})^2}\end{bmatrix}\right)$$

9.2.1 Prediction

Consider a new observation of x, say x_0. We have

$$\mathrm{E}[Y_0] = \beta_0 + \beta_1 x_0, \quad \hat{y}_0 \coloneqq \mathrm{E}[\hat{Y}_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

and

$$\begin{aligned}\operatorname{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= \frac{\sigma^2}{n} + \frac{\sigma^2\overline{x}^2}{\sum(x_i - \overline{x})^2} + \frac{\sigma^2 x_0^2}{\sum(x_i - \overline{x})^2} + \frac{-2x_0\overline{x}\sigma^2}{\sum(x_i - \overline{x})^2} \\ &= \frac{\sigma^2}{n} + \sigma^2\frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2}.\end{aligned}$$

We have

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}}} \sim \mathcal{N}(0,1).$$

We may then contract a CI at $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n} + \sigma^2 \frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}},$$

We use the estimator

$$\hat{\sigma}^2 \coloneqq \frac{\mathrm{RSS}}{n-2}.$$

Remark 9.5. The CI is wider when x_0 is far from the center $\overline{x}$.

It turns out that

$$\frac{\mathrm{RSS}}{\hat{\sigma}^2} \sim \chi^2_{n-2},$$

where the degree of freedom is $n-2$ because we are estimating two parameters, and

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}}} \cdot \frac{1}{\sqrt{\frac{\mathrm{RSS}}{\sigma^2}/(n-2)}} \sim t_{n-2},$$

since the first term is $\mathcal{N}(0,1)$ and the second is $1/\sqrt{\chi^2_{n-2}/(n-2)}$. We then have the CI at $(x_0, \hat{\mathrm{E}}(Y_0)) = (x_0, \hat{\beta}_0 + \hat{\beta}_1 x_0)$ for $\beta_0 + \beta_1 x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,1-\frac{\alpha}{2}} \sqrt{\frac{\mathrm{RSS}}{n-2}\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$$

Consider now

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

Consider

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

We have the LSE/MLE for $\{b_0, b_1\}$

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

and the MLE for σ^2

$$\hat{\sigma}^2_{ML} = \frac{\mathrm{RSS}}{\sigma^2}.$$

Recall that we can write $\hat{\beta}_0$ and $\hat{\beta}_1$ as linear combinations of Y_i. They are thus jointly normal.

Note that the regression line always pass through $(\overline{x}, \overline{y})$.

10 Multiple Linear Regression Model

Consider

$$(x_{i1}, x_{i2}, \ldots, x_{ir}, y_i), \quad i = 1, \ldots, n.$$

Model assumption:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

or equivalently

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}, \sigma^2).$$

Think of $x_{i1}, \ldots, x_{ir}$ as fixed.

We have

$$\begin{aligned} L(\beta, \sigma^2; x, y) &= \prod_{i=1}^{n} f(y_i; \beta, x, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_r x_{ir})^2}{2\sigma^2}\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}))^2\right). \end{aligned}$$

We write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Note that $\mathbf{y}, \boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ where $p = r + 1$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$. Using this notation we can write

$$\sum (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}))^2 = \boldsymbol{\epsilon}^2 = \boldsymbol{\epsilon}^\intercal \boldsymbol{\epsilon} = \langle \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \rangle = \|\boldsymbol{\epsilon}\|^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

The log likelihood is then

$$\ell = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

We see thus that the MLE is attained when $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|$ is minimized. The MLE coincides with the least square estimator. Specifically, we have

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Noting that

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\| = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^\intercal \mathbf{Y} - 2\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{Y} + \boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta},$$

we have

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2} \left(-2\mathbf{X}^\intercal \mathbf{Y} + 2\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta}\right) = -\frac{1}{\sigma^2} \left(\mathbf{X}^\intercal \mathbf{Y} - \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta}\right).$$

Setting it to 0 gives $\mathbf{X}^\intercal = \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta}$. Thus if $\mathbf{X}$ has rank $p \le n$, we have

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y}.$$

That is, we have

Proposition 10.1. *The MLE and the least square estimator coincide, and are given by*

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y.$$

Proof. Note that

$$\|Y - X\beta\| = (Y - X\beta)^\intercal (Y - X\beta) = Y^\intercal Y - 2\beta^\intercal X^\intercal Y + \beta^\intercal X^\intercal X\beta,$$

Taking the derivative with respect to β gives

$$\frac{\partial \|Y - X\beta\|}{\partial \beta} = -2X^\intercal Y + 2X^\intercal X\beta,$$

where the second term comes from noting that

$$\begin{aligned}
\frac{\partial}{\partial x_k}(x^\intercal Ax) &= \frac{\partial}{\partial x_k}\left(\sum^2 a_{ii}x_i^2 + \sum_{i\neq j} a_{ij}x_i x_j\right) \\
&= 2a_{kk}x_k + \sum_{i\neq k} a_{ik}x_i + \sum_{j\neq k} a_{kj}x_j \\
&= \sum_i a_{ik}x_i + \sum_j a_{kj}x_j \\
&= k\text{th entry of } Ax + k\text{th entry of } A^\intercal x.
\end{aligned}$$

Thus $\partial(x^\intercal Ax)/\partial x_k = Ax + A^\intercal x$ and when A is symmetric, we have $\partial(x^\intercal Ax)/\partial x = 2Ax$. □

Then

$$\mathrm{E}[\boldsymbol{\beta}] = (\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \,\mathrm{E}[\mathbf{Y}] = (\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal (\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}.$$

The least square estimator is unbiased.

The variance is

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \mathrm{Var}\left[(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{Y}\right] = (\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \,\mathrm{Var}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1},$$

where we recall

$$\mathrm{Cov}(AY, BY) = A\,\mathrm{Cov}(Y, Y)B^\intercal.$$

Since $\mathbf{X}$ and $\boldsymbol{\beta}$ are fixed, we have

$$\mathrm{Var}[Y] = \mathrm{Var}[\boldsymbol{\epsilon}] = \sigma^2 I_n.$$

Then,

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\intercal \mathbf{X})^{-1}.$$

So $\hat{\boldsymbol{\beta}}$ is normally distributed with

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal \mathbf{X})^{-1}).$$

The RSS is

$$\mathrm{RSS} = \left\|Y - X\hat{\beta}\right\|^2$$

with

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p},$$

where, recall, $p = r - 1$.

We have the following

Proposition 10.2.

- $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal \mathbf{X})^{-1})$.
- $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$.
- $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \text{RSS}$.

10.1 Projection Matrix

We write

$$X = \begin{bmatrix} 1 & X_1 & \dots & X_{p-1} \end{bmatrix}.$$

Then,

$$X\beta = \beta_0 1 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

is a linear combination of the columns of X. In particular, $X\beta$ is in the column space of X, and the least squares condition dictates that it is the orthogonal projection of Y onto the column space of X.

$$X\hat{\beta} = X(X^\intercal X)^{-1} X^\intercal Y = P_X Y =: \hat{Y},$$

where P_X or H denotes the projection matrix.

Definition 10.3. P is a **projection matrix** if $P = P^\intercal$ and $P = P^2$.

Example 10.4.

- I_n, the identity matrix.
- $H = P_X \coloneqq X(X^\intercal X)^{-1} X^\intercal$ is a projection matrix onto the column space of X.
- Let v be a vector with norm $\|v\| = v^\intercal v = 1$. Then $P = vv^\intercal$ is a projection matrix onto the line spanned by v (a "rank-1" matrix). This is a special case of the previous example.
- Let $P \in \mathbb{R}^{n \times n}$ be projection matrix. Then $I_n - P$ is also a projection matrix.

Proposition 10.5. *The only two possible eigenvalues of a projection matrix P is* 1 *and* 0.

Proof. Let $v_1, \ldots, v_n$ be a set of orthonormal eigenvectors of P with $Pv_i = \lambda_i v_i$ for $i = 1, \ldots, n$ and $\lambda_1 \geq \cdots \geq \lambda_n$. Note that since P is symmetric, v_i's exist and $\lambda_i \in \mathbb{R}$ are real. We have then that

$$P \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & \ldots & \lambda_n v_n \end{bmatrix} = \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}.$$

Denote the last two terms by V and Λ. Since $v_1, \ldots, v_n$ is orthonormal, we have $V^\intercal V = I_n$. Then from

$$PV = V\Lambda$$

we have by multiplying each side by $V^\intercal$ that

$$P = V\Lambda V^\intercal.$$

Since P is a orthogonal projection, we have

$$P^2 = V\Lambda V^\intercal V\Lambda V^\intercal = V\Lambda^2 V^\intercal = P = V\Lambda V^\intercal.$$

In particular,

$$V\Lambda^2 V^\intercal = V\Lambda V^\intercal.$$

Multiplying both sides by $V^\intercal$ on the left and V on the right gives

$$\Lambda^2 = \Lambda,$$

which gives the desired result. □

Note that we obtained $P = V\Lambda V^\intercal$, which gives

Corollary 10.6. *The rank of P is equal to the number of eigenvalues equal to* 1. *Thus,* $\operatorname{rank} P = \sum \lambda_i$.

Corollary 10.7. $\operatorname{tr} P = \operatorname{rank} P$

Proof.

$$\operatorname{tr} P = \operatorname{tr}(V\Lambda V^\intercal) = \operatorname{tr}(V^\intercal V\Lambda) = \operatorname{tr}\Lambda = \operatorname{rank} P.$$

□

Note that we used the following:

Proposition 10.8.

- $\operatorname{tr}(AB) = \operatorname{tr}(BA)$,
- $\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B$.

An application:

Proposition 10.9. *Let P be a projection matrix and $Z \in \mathcal{N}(0, I_n)$. We have*

$$\|PZ\|^2 \sim \chi^2_k,$$

where $k \coloneqq \operatorname{rank} P$.

Proof. Let V and Λ be as in the previous proof (note that λ_i's are in decreasing order). Then

$$\begin{aligned}\|PZ\|^2 = (PZ)^\top (PZ) &= (V\Lambda V^\top Z)^\top (V\Lambda V^\top Z) \\ &= Z^\top V \Lambda^2 V^\top Z \\ &= Z^\top V \Lambda V^\top Z = Y^\top \Lambda Y,\end{aligned}$$

where we define $Y \coloneqq Z^\top V$.

Note that Y is a linear combination of Z's and thus of multi-normal distribution. In particular we have

$$\mathrm{E}[Y] = V^\top \mathrm{E}[Z] = V^\top 0 = 0$$

and

$$\mathrm{Var}[Y] = \mathrm{Cov}(V^\top Z) = V^\top \mathrm{Var}[Z] V = V^\top I_n V = I_n.$$

Then,

$$\|PZ\|^2 = Y^\top \Lambda Y = Y_i^1 + \cdots + Y_i^k \sim \chi^2_k.$$

□

Note that we used the following:

Proposition 10.10. *Let A and B be matrices and X and Y be random vectors. Then,* $\mathrm{Cov}(AX, BY) = A \,\mathrm{Cov}(X, Y) B^\top$.

10.2 Statistical Properties of the Least Squares Estimators

Recall that for $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and the hypothesis $H_0 : \mu = \mu_0$, we have under null that

- $\frac{\overline{Y} - \mu_0}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1)$,
- $\frac{\sum (Y_i - \overline{Y})^2}{\sigma} \sim \chi^2_{n-1}$,
- $S^2 \perp\!\!\!\perp \hat{\mu}$,

which gives

$$\frac{\overline{Y} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1},$$

using which we can construct a CI for μ. This is a simple case of the following multiple regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{i(p-1)} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

or

$$Y = X\beta + \epsilon, \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_n).$$

Recall that we obtained the MLE/LSE of β as $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$. Thus,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^\intercal X)^{-1}\right).$$

Definition 10.11. The standard error is defined by

$$\begin{aligned} \operatorname{se} \hat{\beta}_k &= \sqrt{\operatorname{Var}(\hat{\beta}_k)} \\ &= \text{the } (k+1)\text{th diagonal entry of the covariance matrix } \sigma^2 (X^\intercal X)^{-1} \end{aligned}$$

We will prove the following:

Proposition 10.12. *We have the following:*

- $\frac{\text{RSS}}{\sigma^2} = \frac{\|Y - X\hat{\beta}\|}{\sigma^2} = \chi^2_{n-p}$.
- $\text{RSS} \perp\!\!\!\perp \hat{\beta}$.

This similarly will enable us to do all kinds of t tests.

Proof.

- We have
$$\text{RSS} = b\left\|Y - X\hat{\beta}\right\|^2 = \|Y - HY\|^2 = \|(I - H)Y\|^2.$$
Note that
$$\begin{aligned} (I - H)Y &= (I - H)(X\beta + \epsilon) \\ &= X\beta - H\beta + (I - H)\epsilon = (I - H)\epsilon. \end{aligned}$$
Plugging back in, we obtain
$$\text{RSS} = \sigma^2 \|(I - H)\epsilon\|^2,$$
where, note, $I - H$ is a projection matrix. Since $\epsilon =: \sigma Z \sim \mathcal{N}(0, \sigma^2 I)$. We have
$$\operatorname{rank}(I - H) = \operatorname{tr}(I - H) = \operatorname{tr}(I) - \operatorname{tr}(H) = n - p,$$
where we used $\operatorname{tr}(P) = \operatorname{tr}(V \Lambda V^{-1}) = \operatorname{tr}(\Lambda V^\intercal V) = \operatorname{tr}(\Lambda) = p$. Using the result from the previous subsection, we have
$$\frac{\text{RSS}}{\sigma^2} = \frac{\|(I - H)Z\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

- In light of normality, we need only check

$$\mathrm{Cov}(Y - X\hat{\beta}, \hat{\beta}) = 0.$$

We have

$$\begin{aligned}
\mathrm{Cov}(Y - X\hat{\beta}, \hat{\beta}) &= \mathrm{Cov}(Y - X(X^\intercal X)^{-1}X^\intercal Y, (X^\intercal X)^{-1}X^\intercal Y) \\
&= \mathrm{Cov}\left([I - X(X^\intercal X)^{-1}X^\intercal]Y, (X^\intercal X)^{-1}X^\intercal Y\right) \\
&= [I - X(X^\intercal X)^{-1}X^\intercal]\,\mathrm{Cov}(Y, Y)[(X^\intercal X)^{-1}X^\intercal]^\intercal \\
&= [I - X(X^\intercal X)^{-1}X^\intercal]\left(\sigma^2 I_n\right)X(X^\intercal X)^{-1} \\
&= \sigma^2[I - X(X^\intercal X)^{-1}X^\intercal]X(X^\intercal X)^{-1} = 0.
\end{aligned}$$

□

Note that from the first result we can obtain an unbiased estimator for σ^2:

$$\mathrm{E}\left(\frac{\mathrm{RSS}}{n-p}\right) = \sigma^2.$$

We summarize the results we obtained so far as follows:

Proposition 10.13. *Consider the linear model*

$$Y = \mathcal{N}_n(X\beta, \sigma^2 I_n).$$

We have

- $$\hat{\beta} = (X^\intercal X)^{-1}X^\intercal Y \sim \mathcal{N}_p(\beta, \sigma^2(X^\intercal X)^{-1}).$$
- $$\frac{\mathrm{RSS}}{\sigma^2} = \frac{\left\|Y - X\hat{\beta}\right\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$
- $$\hat{\beta} \perp\!\!\!\perp \mathrm{RSS} \quad (\textit{thus } MSS \perp\!\!\!\perp \mathrm{RSS}).$$
- *Under the null hypothesis that $\beta_k = 0$, we have*

$$\frac{\hat{\beta}_k - 0}{\mathrm{se}\,\hat{\beta}_k} \sim t_{n-p},$$

where $\mathrm{se}\,\hat{\beta}_k = \sqrt{\mathrm{Var}(\hat{\beta})}$ *and* $\mathrm{Var}[\hat{\beta}_k]$ *is the ith diagonal entry of* $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^\intercal X)^{-1}$.

10.3 Mode Comparison

Consider the following two models:

$$\begin{aligned} M_0: &\qquad Y_i = \beta_0 + \epsilon_i, &&\qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2) \\ M_1: &\qquad Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, &&\qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \end{aligned}$$

We can rewrite the models as

$$\begin{aligned} M_0: &\qquad Y_i \sim \mathcal{N}(\beta_0 \mathbb{1}_n, \sigma_0^2 I_n), \\ M_1: &\qquad Y_i \sim \mathcal{N}(X\beta, \sigma^2 I_n), \end{aligned}$$

where $\mathbb{1}_n$ is a vector of 1's. The model estimates are

$$\begin{aligned} M_0: &\qquad \hat{Y}_i = \overline{Y} = J_n Y, \\ M_1: &\qquad \hat{Y}_i = X\hat{\beta} = P_X Y, \end{aligned}$$

where $J_n \coloneqq n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal$ is easily checked to be a projection matrix of rank 1 (it's column space is just a line). We have for both models that

Proposition 10.14.

$$\left\| Y - \overline{Y} \right\| = \left\| Y - \hat{Y} \right\|^2 + \left\| \hat{Y} - \overline{Y} \right\|^2.$$

Or,

$$SS_{total} = SS_e + SS_{model}.$$

Since the column space of $\mathbb{1}_n$ is clearly a subspace of the column space of X, we have

$$\left\| Y - \hat{Y}_0 \right\|^2 \leq \left\| Y - \overline{Y}_1 \right\|^2.$$

We have moreover that

Proposition 10.15.

$$\|Y - J_n Y\|^2 = \|Y - HY\|^2 + \|HY - J_n Y\|^2.$$

Proof. We have

$$\begin{aligned} \|Y - J_n Y\|^2 &= \|Y - HY + HY - J_n Y\|^2 \\ &= \|Y - HY\|^2 + \|HY - J_n Y\|^2 + 2(Y - HY)^\intercal (HY - J_n Y). \end{aligned}$$

Since

$$\begin{aligned} (Y - HY)^\intercal (HY - J_n Y) &= Y^\intercal (I_n - H)(H - J_n)Y \\ &= Y^\intercal (H - J_n - H^2 + HJ_n)Y. \end{aligned}$$

Since the column space of $\mathbb{1}_n$ is a subspace of the column space of X, we know that $HJ_n = J_n$, thus completing the proof. Alternatively, from $HX = X$ and the fact that the first column of X is $\mathbb{1}_n$, we know that $H\mathbb{1}_n = \mathbb{1}_n$, from which we have similarly that $HJ_n = J_n$. □

Remark 10.16. Remember the picture!

We summarize the results as follows:

Proposition 10.17. *We have*

- $SS_{total} = SS_e + SS_{model}$.
- $SS_e/\sigma^2 = \mathrm{RSS}/\sigma^2 \sim \chi^2_{n-p}$.
- $SS_e \perp\!\!\!\perp SS_{model}$.
- *Under the null hypothesis that M_0 is true, we have*

$$\frac{SS_{total}}{\sigma^2} \sim \chi^2_{n-p}, \quad \frac{SS_{model}}{\sigma^2} \sim \chi^2_{p-1}.$$

Thus,

$$\frac{(SS_{total} - SS_e)/[(n-1)-(n-p)]}{SS_e/(n-p)} = \frac{SS_{model}/(p-1)}{SS_e/(n-p)} \sim F_{p-1,n-p}.$$

More generally,

$$\frac{\left(SS_{small} - SS_{large}\right)/[(n-p_s)-(n-p_l)]}{SS_{large}/(n-p_l)} \sim F_{p_s-p_l,n-p_l}.$$

10.4 LR and ANOVA

Example 10.18. Consider the case $Y_i = \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$. We have $X^\intercal X = \|X\|$ and so $P_X = X(X^\intercal X)^{-1}X^\intercal = XX^\intercal/\|X\|$.

Example 10.19. Consider the case $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, $i = 1, \ldots, n$. We have $X^\intercal X = \|X\|$ and so $P_X = X(X^\intercal X)^{-1}X^\intercal = XX^\intercal/\|X\|$. Thus

$$X^\intercal X = \begin{bmatrix} x_1^\intercal x_1 & x_1^\intercal x_2 \\ x_2^\intercal x_1 & x_2^\intercal x_2 \end{bmatrix}.$$

If in addition $x_1^\intercal x_2 = 0$, then

$$x^\intercal x = \begin{bmatrix} \|x_1\|^2 & 0 \\ 0 & \|x_2\|^2 \end{bmatrix}$$

and thus

$$P_X = X(X^\intercal X)^{-1}X^\intercal = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{bmatrix} \frac{1}{\|x_1\|^2} & 0 \\ 0 & \frac{1}{\|x_2\|^2} \end{bmatrix} \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \end{bmatrix} = \begin{bmatrix} \frac{x_1 x_1^\intercal}{\|x_1\|^2} & 0 \\ 0 & \frac{x_2 x_2^\intercal}{\|x_2\|^2} \end{bmatrix}$$

This shows that $P_{x_1} \perp P_{x_2}$, which gives $P_X = P_{x_1} + P_{x_2}$. We have

$$P_X Y = P_{x_1} Y + P_{x_2} Y.$$

From $Y = P_X Y + (I - P_X)Y$ we have

$$\|Y\|^2 = \left\|P_{x_1} Y\right\|^2 + \left\|P_{x_2} Y\right\|^2 + \|(I - P_X)Y\|^2,$$

the two way ANOVA partition. We see that the order does not matter for sequential ANOVA if $x_1^\intercal x_2 = 0$.

Example 10.20. Consider

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}.$$

We may write it as

$$Y_i = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,$$

where x_{1i} is a dummy for receiving treatment A at level i, and x_{2i} a dummy for B at level i. Consider the usual constraints $\sum \alpha_i = \sum \beta_j = 0$. The usual adjustment of the design matrix is needed to obtain a full rank matrix. Writing $X = \begin{bmatrix} X_0 & X_A & X_B & X_{AB} \end{bmatrix}$ we see that $X_A \perp X_B$ if the design is balanced.

10.5 Q-Q Plot

$$\text{RSS} = SS_e = \sum \hat{\epsilon}_i^2 = \left\|Y - X\hat{\beta}\right\|^2.$$

$\hat{\epsilon}_i$'s should be close to iid $\mathcal{N}(0, \hat{\sigma}^2)$. To test this hypothesis, we may use a Q-Q plot (quantile-quantile plot).

Suppose we have data $X_1, \dots, X_n$ (e.g., the residuals of a linear model). Order the X_i's as $X_{(1)} \le \cdots \le X_{(n)}$. Are X_i's approximately iid of some normal distribution $\mathcal{N}(\mu, \sigma^2)$. Pick $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and order them as $Z_{(1)} \le \cdots \le Z_{(n)}$. We plot $(Z_{(i)}, X_{(i)})$ for $i = 1, \dots, n$.

Claim 10.21. If the X_i's are approximately iid of some normal distribution, the points should approximately lie on the line $X = \mu + \sigma Z$.

In practice, to make the process deterministic, we often plot $\left(\mathrm{E}[Z_{(i)}], X_{(i)}\right)$, with

$$\mathrm{E}[Z_{(i)}] = \Phi^{-1}(\Phi(\mathrm{E}[Z_{(i)}])) \approx \Phi^{-1}\left(\mathrm{E}[\Phi(Z_{(i)})]\right) = \Phi\left(\frac{i}{n+1}\right).$$

The last equality comes from the discussion below.

10.5.1 Order Statistics

Let $X_1, \dots, X_n$ be iid with CDF F and density f, both continuous. Order them as $X_{(1)} \le \cdots \le X_{(n)}$.

Example 10.22. We have

$$F_{(1)}(x) = \mathbb{P}\left(X_{(1)} \le x\right) = 1 - \mathbb{P}\left(X_{(1)} \ge x\right) = 1 - \mathbb{P}\left(X_i > x, \forall i\right) = 1 - (1 - F(x))^n.$$

$$F_{(n)}(x) = \mathbb{P}\left(X_{(n)} \le x\right) = \mathbb{P}\left(X_i < x, \forall i\right) = F(x)^n.$$

More generally,

$$F_{(k)}(x) = \mathbb{P}\left(X_{(k)} \le x\right) = \mathbb{P}\left(\text{as least } k \text{ of } X_i \le x\right)$$
$$= \sum_{i=k}^{n} \binom{n}{i} F(x)^i \left(1 - F(x)\right)^{n-i}.$$

Hence,

$$\begin{aligned} f_{(k)} = F'_{(k)}(x) &= \sum_{i=k}^{n} \frac{n!}{(n-i)!i!} iF(x)^{i-1} f(x) \left(1 - F(x)\right)^{n-i} \\ &\quad - \sum_{i=k}^{n} \frac{n!}{(n-i)!i!} F(x)^i (n-i) \left(1 - F(x)\right)^{n-i-1} f(x) \\ &= nf(x) \left[\sum_{i=k}^{n} \frac{(n-1)!}{(n-i)!(i-1)!} F(x)^{i-1} \left(1 - F(x)\right)^{n-i} + \sum_{i=k}^{n} \frac{(n-1)!}{i!(n-i-1)!} F(x)^i \left(1 - F(x)\right)^{n-i-1} \right] \\ &= nf(x) \frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!} F(x)^{k-1} \left(1 - F(x)\right)^{(n-1)-(k-1)} \\ &= nf(x) \binom{n-1}{k-1} F(x)^{k-1} \left(1 - F(x)\right)^{n-k}. \end{aligned}$$

That is, we have

Proposition 10.23. *Let $X_1, \ldots, X_n$ be iid with CDF F and density f, both continuous. Order them as $X_{(1)} \le \cdots \le X_{(n)}$. We have then that*

$$F_{(k)} = \sum_{i=k}^{n} \binom{n}{i} F(x)^i \left(1 - F(x)\right)^{n-i}$$

and

$$f_{(k)} = nf(x) \binom{n-1}{k-1} F(x)^{k-1} \left(1 - F(x)\right)^{n-k}.$$

Remark 10.24 (Sidenote). Let $W = \Phi(Z)$ and $Z \sim \mathcal{N}(0,1)$. We have

$$\mathbb{P}\left(W \le w\right) = \mathbb{P}\left(\Phi(Z) \le w\right) = \mathbb{P}\left(Z \le \Phi^{-1}(w)\right) = \Phi(\Phi^{-1}(w)) = w$$

and thus

$$W \sim \text{Uniform}[0,1].$$

This is true in fact for all random variables. That is, we have

Proposition 10.25. *Let X be a random variable with distribution F. Then*

$$F(X) \sim \text{Uniform}[0,1], \quad X = F^{-1}(U),$$

where

$$F^{-1}(u) \coloneqq \inf\left\{y : f(y) \ge u\right\}.$$

The latter result is useful for simulation.

We can now apply this to the discussion of Q-Q plots above: For $U \sim \mathcal{N}(0,1)$, we have

$$f_{(k)}(x) = n f(x) \binom{n-1}{k-1} F(x)^{k-1} (1 - F(x))^{n-k} \sim \text{Beta}(k, n-k+1).$$

Thus it has mean

$$\mathrm{E}[f_{(k)}] = \frac{k}{k + (n-k+1)} = \frac{k}{n+1}.$$

11 Simple Regression (A)

Proposition 11.1. *The least square estimators are*

$$\hat{\beta}_1 = \frac{\overline{(xy)} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

We can also write

$$\overline{\beta}_1 = \sum c_i Y_i, \quad c_i \coloneqq \frac{x_i - \overline{x}}{\sum (x_i - \overline{x})^2}.$$

Note that $\sum c_i = 0$, $\sum c_i x_i = 1$, $\sum c_i^2 = 1/[\sum (x_i - \overline{x})^2]$. We can use these to calculate the mean and variance of the least square estimators.

Thus we have:

Proposition 11.2. *The least square estimators are unbiased.*