Section 15

TA: Yasi Zhang

May 23, 2022

1 Linear Regression

Given pairs of observations $\{(X_i, Y_i)\}_{i=1}^n$ where $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^p, X_{i1} = 1$, we fit the data to a linear model :

$$Y_i = X_i'\theta + \varepsilon_i = \theta_1 + X_{i2}\theta_2 + \cdots + X_{ip}\theta_p + \varepsilon_i.$$

Remark 1 *Our aim is to find the relationship between Y_i and X_i. They might have a perfect linear relationship, or they might not. We could fit the data to other models, e.g. $Y_i = \theta_1 + X_{i2}^2\theta_2 + \cdots + X_{ip}^p\theta_p + \varepsilon_i$.. Finally, we could compare these different models and pick the best one as our final model.*

After assuming the form of linear model, we need to find the best estimated θ. Pay attention that we never know how large the true parameter θ is. What we can do is to estimate it. The model is assumed for population, but we only have a sample of X_i and Y_i of size n.

We need a criterion to find the best $\hat{\theta}$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^n (Y_i - X_i'\theta)^2$$

By FOC:

$$\hat{\theta} = (\sum_{i=1}^n X_i X_i')^{-1} \sum_{i=1}^n X_i Y_i$$

Remark 2 *You could use other criterions like MAE. But we are used to using MSE as the criterion.*

Linear Regression Example

Yasi Zhang

2022/5/23

```r
X = 1:3
Y = c(2.1, 2.9,4.1)
print(paste('(', 'X',',', 'Y',')'))
```

```
## [1] "( X , Y )"
```

```r
for(i in 1:3){
    print(paste('(', X[i],',', Y[i],')'))
}
```

X_{i2} Y

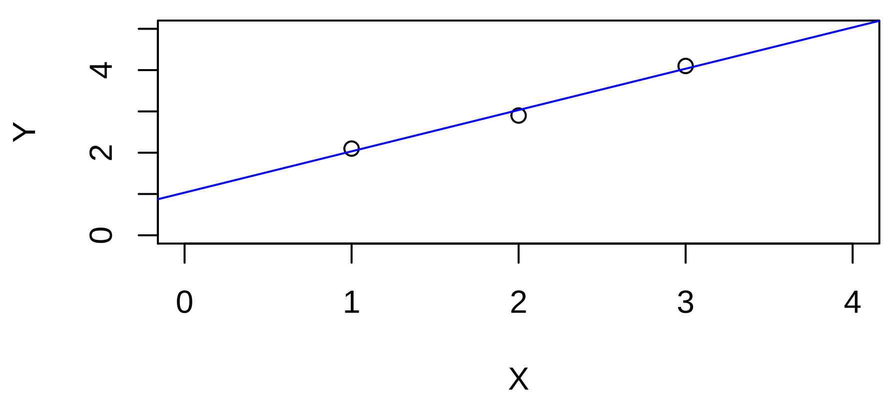
```
## [1] "( 1 , 2.1 )"
## [1] "( 2 , 2.9 )"
## [1] "( 3 , 4.1 )"
```

$X_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $X_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ $X_3 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$

$Y_1 = 2.1$ $Y_2 = 2.9$ $Y_3 = 4.1$

Assume $Y_i = \theta_1 + \theta_2 X_{i2} + \varepsilon_i$, derive the best estimated θ.

```r
model = lm(Y~X)
plot(X,Y, xlim = c(0,4), ylim=c(0,5))
abline(model, col='blue')
```



$$\hat{\theta} = (\sum_{i=1}^{n} X_i X_i')^{-1} \sum_{i=1}^{n} X_i Y_i$$

```r
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##        1        2        3
##  0.06667 -0.13333  0.06667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0333     0.2494   4.143   0.1508
## X             1.0000     0.1155   8.660   0.0732 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1633 on 1 degrees of freedom
## Multiple R-squared:  0.9868, Adjusted R-squared:  0.9737
## F-statistic:    75 on 1 and 1 DF,  p-value: 0.07319
```

$$\hat{\theta} = \left[\begin{pmatrix}1\\1\end{pmatrix}(1\ \ 1) + \begin{pmatrix}1\\2\end{pmatrix}(1\ \ 2) + \begin{pmatrix}1\\3\end{pmatrix}(1\ \ 3)\right]^{-1}\left[\begin{pmatrix}1\\1\end{pmatrix}2.1 + \begin{pmatrix}1\\2\end{pmatrix}2.9 + \begin{pmatrix}1\\3\end{pmatrix}4.1\right]$$

$$= \left[\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix} + \begin{pmatrix}1 & 2\\2 & 4\end{pmatrix} + \begin{pmatrix}1 & 3\\3 & 9\end{pmatrix}\right]^{-1}\begin{pmatrix}2.1+2.9+4.1\\2.1+5.8+12.3\end{pmatrix}$$

$$= \begin{pmatrix}3 & 6\\6 & 14\end{pmatrix}^{-1}\begin{pmatrix}9.1\\20.2\end{pmatrix}$$

$$= \frac{1}{\begin{vmatrix}3 & 6\\6 & 14\end{vmatrix}}\begin{pmatrix}14 & -6\\-6 & 3\end{pmatrix}\begin{pmatrix}9.1\\20.2\end{pmatrix}$$

$$= \begin{pmatrix}\frac{7}{3} & -1\\-1 & \frac{1}{2}\end{pmatrix}\begin{pmatrix}9.1\\20.2\end{pmatrix} = \begin{pmatrix}1.033\\1\end{pmatrix}$$

2 Method of Moment Estimator

- Compute population moments $E_\theta(X_i^k)$, $k = 1, 2, \cdots$, under the PMF/PDF model $f(x, \theta)$:

$$\begin{aligned} M_k(\theta) &= E_\theta(X_i^k) \\ &= \begin{cases} \int_{-\infty}^{\infty} x^k f(x,\theta)dx & \text{if } X \text{ is a CRV,} \\ \\ \sum_{x\in\Omega_X} x^k f(x,\theta) & \text{if } X \text{ is a DRV;} \end{cases} \end{aligned}$$

- Compute the sample moments from random sample $\mathbf{X}^n = (X_1, \cdots, X_n)$:

$$\hat{m}_k = n^{-1}\sum_{i=1}^{n} X_i^k, \ k = 1, 2, \cdots;$$

- Match the sample moments and the population moments by choosing some parameter value $\hat{\theta}_n$. In general, if θ is a $p \times 1$ parameter vector, we need p equations:

$$\begin{cases} \hat{m}_1 = M_1\left(\hat{\theta}_n\right), \\ \hat{m}_2 = M_2\left(\hat{\theta}_n\right), \\ \quad \cdots \\ \hat{m}_p = M_p\left(\hat{\theta}_n\right). \end{cases}$$

  Solving for these p equations will yield an MME $\hat{\theta}_n = \hat{\theta}(\mathbf{X}^n)$.

We now provide some examples.

2.1 Practice

Suppose $\{X_i\}_{i=1}^n \sim iid\ U(a, b)$. Find MME for a and b.

3 Maximum Likelihood Estimator

We now summarize the procedure of MLE:

- Find the log-likelihood function, $\ln \hat{L}(\theta|\mathbf{X}^n)$. For an IID random sample with population PMF/PDF $f(x,\theta)$, we have $\ln \hat{L}(\theta|\mathbf{X}^n) = \sum_{i=1}^{n} \ln f(X_i, \theta)$.
- Solve for the first order conditions (FOC) and find $\hat{\theta}_n$.
- Check the second order conditions (SOC) to ensure $\hat{\theta}_n$ is a global maximizer or at least a local maximizer.

3.1 Practice

1. Suppose $\{X_i\}_{i=1}^n \sim iid\ Poi(\lambda)$.

a. Given $X_i = 169$, find the MLE for λ.

b. Given $\{X_i\}_{i=1}^n$, find the MLE for λ.

2. Suppose $\{X_i\}_{i=1}^n \sim iid\ U(a,b)$. Find MLE for a and b.

4 Optimization Problem

4.1 Unconstrained Problem

$$\min_x f(x)$$

FOC: Find the x_* that lets $f'(x_*) = 0$.

Remark 3 *Our target is to find a global minimum or maximum. FOC only ensures x_* is a local maximum or minumun.*

After finding the x_*, think about how the function $f(x)$ looks and check x_* is a local maximum or minimum, or calculate its second-order derivative.

Some examples: $f(x) = x^2 - x, f(x) = x^3, f(x) = sin(x)$

4.2 Constrained Problem

The **standard form** of a continuous optimization problem is

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g(x) = 0 \end{aligned}$$

where f : $\mathbb{R}^n \to \mathbb{R}$ is the objective function to be minimized over the n-variable vector x, $g(x) = 0$ is called equality constraint.

A maximization problem can be treated by negating the objective function.

We introduce a new variable λ called a Lagrange multiplier and study the Lagrange function defined by

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x),$$

By FOC,

$$\nabla_x \mathcal{L} = 0 \tag{1}$$

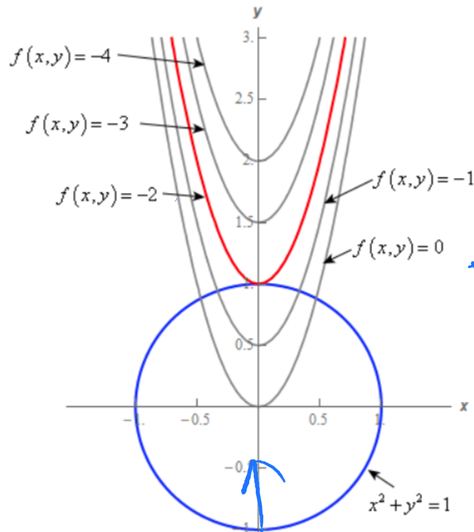
$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \tag{2}$$

4.3 Example

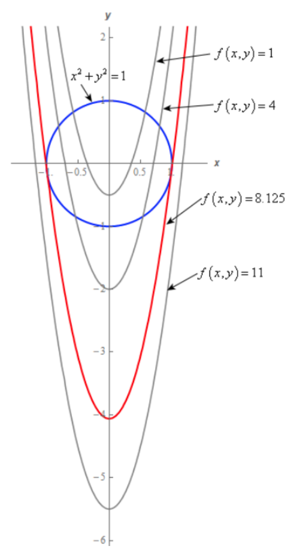
$$\min_{x,y} \;or\; \max_{x,y} \; 8x^2 - 2y$$
$$s.t.\; x^2 + y^2 = 1$$

$\mathcal{L}(x,y,\lambda)$
$= 8x^2 - 2y + \lambda(x^2+y^2-1)$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 16x + 2\lambda x = 0 \\ \frac{\partial \mathcal{L}}{\partial y} = -2 + 2\lambda y = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = x^2+y^2-1 = 0 \end{cases}$$



$\nabla f(x,y)$ paralle to $\nabla g(x,y)$



① $x = 0$
$y = \pm 1$

② $x \neq 0$
$\lambda = -8$
$y = \frac{-1}{8}$
$x^2 = 1 - y^2$
$= \frac{63}{64}$
$x = \pm \frac{3\sqrt{7}}{8}$

$(0 \;\; 1) \Rightarrow -2$
$(0 \;\; -1) \Rightarrow 2$
$(\frac{3\sqrt{7}}{8}, -\frac{1}{8}) \Rightarrow 8{,}125$
$(-\frac{3\sqrt{7}}{8}, -\frac{1}{8}) \Rightarrow 8.125$

4.4 Practice

Example 213 (8.15). Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an independent but not identically distributed random sample, with $E(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma_i^2 < \infty, i = 1, \cdots, n$. Find a uniformly best linear unbiased estimator of μ within the class of estimators

$$\Gamma = \left\{ \hat{\mu}_n : \mathbb{R}^n \to \mathbb{R} \mid \hat{\mu}_n = \sum_{i=1}^{n} c_i X_i, \ (c_1, \cdots, c_n) \in \mathbf{R}^n \right\},$$

where $\sum_{i=1}^n c_i = 1$.

5 Take-Home Practice

5. (a) Consider i.i.d. Pois(λ) r.v.s $X_1, X_2, \ldots$. The MGF of X_j is $M(t) = e^{\lambda(e^t - 1)}$. Find the MGF $M_n(t)$ of the sample mean $\bar{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$
(b) Find the limit of $M_n(t)$ as $n \to \infty$. (You can do this with almost no calculation using a relevant theorem; or you can use (a) and the fact that $e^x \approx 1+x$ if x is very small.

Consider the MGF for the standardized Poisson

$$M_X(t) = e^{(\lambda e^{t/\sqrt{\lambda}} - \lambda - t\sqrt{\lambda})}$$

Prove

$$\lim_{\lambda \to \infty} M_X(t) = e^{\frac{t^2}{2}}$$

①

$$\begin{cases} EX = \dfrac{a+b}{2} \\ Var(X) = \dfrac{(a-b)^2}{12} \end{cases} \qquad \begin{cases} \hat{E}(X) = \dfrac{X_1+\cdots+X_n}{n} \triangleq \hat{m}_1 \\ \hat{E}(X^2) = \dfrac{X_1^2+\cdots+X_n^2}{n} \triangleq \hat{m}_2 \end{cases}$$

$$\Rightarrow \begin{cases} EX = \dfrac{a+b}{2} \\ EX^2 = Var X + (EX)^2 = \dfrac{(a-b)^2}{12} + \left(\dfrac{a+b}{2}\right)^2 \end{cases}$$

$$\begin{cases} \dfrac{\hat{a}+\hat{b}}{2} = \hat{m}_1 \\ \dfrac{(\hat{a}-\hat{b})^2}{12} + \left(\dfrac{\hat{a}+\hat{b}}{2}\right)^2 = \hat{m}_2 \end{cases}$$

$$\begin{cases} \hat{a}+\hat{b} = 2\hat{m}_1 \\ (\hat{a}-\hat{b})^2 = 12(\hat{m}_2 - \hat{m}_1^2) \end{cases}$$

$$[\hat{a}-(2\hat{m}_1-\hat{a})]^2 = 12(\hat{m}_2-\hat{m}_1^2)$$

$$4(\hat{a}-\hat{m}_1)^2 = 12(\hat{m}_2-\hat{m}_1^2) \qquad (\hat{a} < \hat{m}_1 < \hat{b})$$

$$\begin{cases} \hat{a} = \hat{m}_1 - \sqrt{3(\hat{m}_2-\hat{m}_1^2)} \\ \hat{b} = \hat{m}_1 + \sqrt{3(\hat{m}_2-\hat{m}_1^2)} \end{cases}$$

or $$\begin{cases} \hat{a} = \hat{\mu} - \sqrt{3\hat{\sigma}^2} \\ \hat{b} = \hat{\mu} + \sqrt{3\hat{\sigma}^2} \end{cases}$$

Remark: We can see that there may exist some sample points outside $(\hat{a}, \hat{b})$

3.1

① $X_1 \sim Poi(\lambda)$

$$\mathcal{L}(\lambda) = f(X = x_1; \lambda)$$

$$= e^{-\lambda} \frac{\lambda^{x_1}}{x!}$$

$$\ln \mathcal{L}(\lambda) = -\lambda + x_1 \ln \lambda - \ln(x!)$$

$$\frac{\partial \ln \mathcal{L}(\lambda)}{\partial \lambda} = -1 + \frac{x_1}{\lambda} = 0$$

$$\hat{\lambda} = x_1 = 169.$$

② $\{X_i\}_{i=1}^{n}$

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} f(X = x_i; \lambda)$$

$$= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

$$\ln \mathcal{L}(\lambda) = \sum_{i=1}^{n} \left[-\lambda + x_i \ln \lambda - \ln(x_i!) \right]$$

$$= -n\lambda + \left(\sum_{i=1}^{n} x_i \right) \ln \lambda + C_0$$

$$\frac{\partial \ln \mathcal{L}(\lambda)}{\partial \lambda} = -n + \left(\sum_{i=1}^{n} x_i \right) \frac{1}{\lambda} = 0$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x_n}$$

MLE for $U(a,b)$

$$\mathcal{L}(a,b) = \prod_{i=1}^{n} f(x_i; a,b)$$

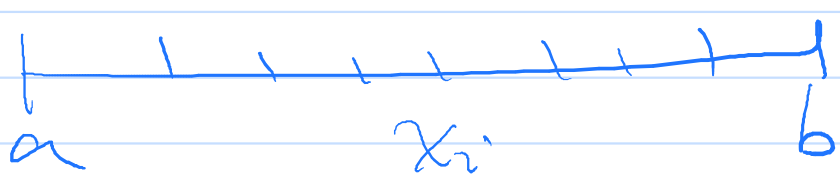
$$f(x_i; a,b) = \begin{cases} \frac{1}{b-a} & a<x_i<b \\ 0 & \text{o.w.} \end{cases}$$

If any $x_i \notin (a,b)$, then $\mathcal{L}(a,b)=0$,

which will never be the maximum

Assert: $a < x_i < b$, $\forall i$

$$\mathcal{L}(a,b) = \left(\frac{1}{b-a}\right)^n, \quad a<x_i<b, \ \forall i$$



$$\text{Max } \mathcal{L} \iff \text{Min } (b-a).$$

$$\begin{cases} \hat{b} = \max\limits_i x_i \\ \hat{a} = \min\limits_i x_i \end{cases}$$

Remark: All $\{x_i\}$'s are inside $(\hat{a}, \hat{b})$

$E X_i = \mu, \quad Var X_i = \sigma_i^2$

$$\min \quad \frac{1}{n^2} \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

$$\text{s.t.} \quad \sum_{i=1}^{n} c_i = 1$$

$$\mathcal{L}(\vec{c}, \lambda) = \frac{1}{n^2} \sum_{i=1}^{n} c_i^2 \sigma_i^2 + \lambda \left(\sum_{i=1}^{n} c_i - 1 \right)$$

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial c_i} = \dfrac{1}{n^2} \sigma_i^2 (2 c_i) + \lambda = 0, \ \forall i \\ \dfrac{\partial \mathcal{L}}{\partial \lambda} = \sum_{i=1}^{n} c_i - 1 = 0 \end{cases}$$

$$\Rightarrow \quad c_i = \frac{-\lambda n^2}{2 \sigma_i^2} = \alpha \frac{1}{\sigma_i^2}, \ \forall i$$

$$\Rightarrow \quad \sum_{i=1}^{n} \alpha \frac{1}{\sigma_i^2} = 1$$

$$\alpha = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

$$\Rightarrow \quad c_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

Remark: $\sigma_i \uparrow \quad c_i \downarrow$

Think About why it's a min point rather than a max point.

Take-Home

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

$$\begin{aligned}
\gamma(t) &= E e^{t\bar{X}_n} \\
&= E e^{\frac{1}{n}t(X_1 + \cdots + X_n)} \\
&= \prod_{i=1}^{n} E e^{\frac{1}{n}t X_i} \\
&= \prod_{i=1}^{n} e^{\lambda(e^{\frac{1}{n}t} - 1)} \\
&= e^{n\lambda(e^{\frac{1}{n}t} - 1)}
\end{aligned}$$

$$\begin{aligned}
&\lim_{n\to\infty} e^{n\lambda(e^{\frac{1}{n}t} - 1)} \\
&= \lim_{n\to\infty} e^{n\lambda(1 + \frac{t}{n} + o(\frac{1}{n}) - 1)} \qquad e^x = 1 + x + o(x) \\
&= \lim_{n\to\infty} e^{\lambda t + o(1)} \qquad o(\tfrac{1}{n}) \times n = o(1) \\
&= \lim_{n\to\infty} e^{\lambda t + o(1)}. \\
&= e^{\lambda t}
\end{aligned}$$

$$\begin{aligned}
&\lim_{\lambda\to 0} e^{\lambda e^{t/\sqrt{\lambda}} - \lambda - t\sqrt{\lambda}} \\
&= \lim_{\lambda\to 0} e^{\lambda(1 + \frac{t}{\sqrt{\lambda}} + \frac{1}{2}\frac{t^2}{\lambda} + o(\frac{1}{\lambda})) - \lambda - t\sqrt{\lambda}} \\
&= \lim_{\lambda\to 0} e^{\frac{1}{2}t^2 + o(1)} \\
&= e^{\frac{1}{2}t^2}
\end{aligned}$$