Ve406 Lecture 4

Jing Liu

UM-SJTU Joint Institute

May 23, 2018

- Consider simple linear regression, which has the regression/link function

$$\mathbb{E}\left[Y_i \mid X_i = x_i\right] = \beta_0 + \beta_1 x_i$$

- Recall we concluded that sample estimate of β_0 and β_1 are reasonable

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{c_{xy}}{s_x^2}$$

- This is a result of first minimising with respect to β_0 and β_1 of the following

$$\mathrm{MSE}(\beta_0, \beta_1) = \mathbb{E}\left[\left(Y - (\beta_0 + \beta_1 X)\right)^2\right]$$

then using the unbiased sample values for the population parameters

$$\beta_0 = \mathbb{E}\left[Y\right] - \beta_1 \mathbb{E}\left[X\right] \qquad \text{where} \quad \beta_1 = \frac{\mathrm{Cov}\left[X, Y\right]}{\mathrm{Var}\left[X\right]}$$

- However, if we step back, and consider what exact the following is

$$\widehat{\mathrm{MSE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (b_0 + b_1 x_i)\right)^2$$

and if our samples (x_i, y_i) are all independent for any fixed (b_0, b_1), then

$$\widehat{\mathrm{MSE}} \to \mathrm{MSE} \qquad \text{when} \quad n \to \infty$$

since the law of large numbers is applicable.

- Thus it is also natural to consider minimise $\widehat{\mathrm{MSE}}$ with respect to b_0 and b_1

$$\frac{\partial \widehat{\mathrm{MSE}}}{\partial b_0} = 0 \qquad \text{and} \qquad \frac{\partial \widehat{\mathrm{MSE}}}{\partial b_1} = 0$$

which lead us to the normal equations for least-squares estimation (LSE)

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - (b_0 + b_1 x_i)\right) = 0 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (b_0 + b_1 x_i)\right)(x_i) = 0$$

- Using the sample mean notation,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(y_i-(b_0+b_1x_i)) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}(y_i-(b_0+b_1x_i))(x_i) &= 0 \end{aligned} \implies \begin{aligned} \bar{y}_n - b_0 - b_1\bar{x}_n &= 0 \\ \overline{(xy)}_n - b_0\bar{x}_n - b_1\overline{x_n^2} &= 0 \end{aligned}$$

- The first equation gives

$$b_0 = \bar{y}_n - b_1\bar{x}_n$$

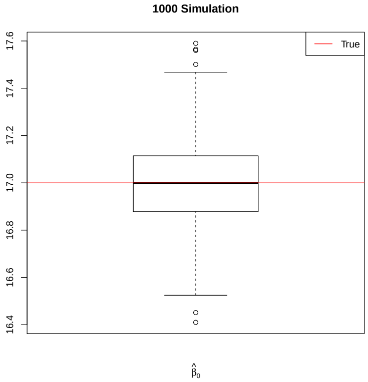
- Substituting into the second equation, we have

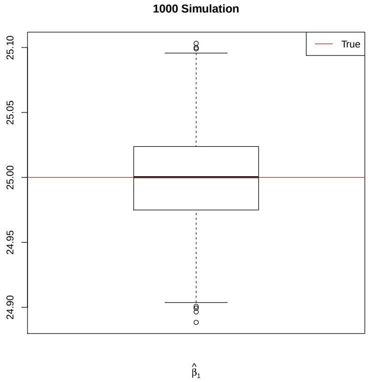
$$\overline{(xy)}_n - \bar{x}_n\bar{y}_n + b_1(\bar{x}_n)^2 - b_1\overline{x_n^2} = 0 \implies c_{xy} - b_1 s_x^2 = 0 \implies b_1 = \frac{c_{xy}}{s_x^2}$$

- Hence we obtain the same estimates by using LSE, a.k.a. OLS.

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad \text{and} \qquad b_1 = \hat{\beta}_1 = \frac{c_{xy}}{s_x^2}$$

Q: We know they are consistent, but are they unbiased?





- They seem to be unbiased, but let us show why they are unbiased and the role of our assumptions play in their unbiasedness.

```
> n = 100          # Sample size
> num = 1000       # Number of repetition
>
> beta0 = 17       # True intercept
> beta1 = 25       # True slope
>
> b.df = data.frame(b0 = double(), b1 = double())
>
> for (i in 1:num){
+   x.vec = rchisq( n, 4)
+   m = beta0 + beta1 * x.vec
+   y.vec = rnorm(n, mean = m, sd = 1)
+   b.df[i,"b1"] = cov(x.vec, y.vec) / var(x.vec)
+   xbar = mean(x.vec)
+   ybar = mean(y.vec)
+   b.df[i,"b0"] = ybar -  b.df[i,"b1"] * xbar
+
+ }
```

```
> # Intercept
> boxplot(b.df[,"b0"],
+         xlab = expression(hat(beta)[0]),
+         main = "1000 Simulation")
>
> abline(h = beta0, col = 2)
>
> legend("topright", "True", lty = 1, col = 2)
>
> # Slope
> boxplot(b.df[,"b1"],
+         xlab = expression(hat(beta)[1]),
+         main = "1000 Simulation")
>
> abline(h = beta1, col = 2)
>
> legend("topright", "True", lty = 1, col = 2)
```

- Notice, of course, the estimators will have the same properties

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad b_1 = \hat{\beta}_1 = \frac{c_{xy}}{s_x^2}$$

whether we treat them as LSE or not. Starting with the slope, we have

$$\hat{\beta}_1 = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)\left(\beta_0 + \beta_1 x_i + e_i - \overline{\beta_0 + \beta_1 x_i + e_i}\right)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

- Simplifying the last expression, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(\beta_0 + \beta_1 x_i + e_i - \beta_0 - \beta_1 \bar{x}_i - \bar{e}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(\beta_1 x_i - \beta_1 \bar{x}_n + e_i - \bar{e}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(e_i - \bar{e}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) e_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

Q: Where did the missing term go?

- Now consider the conditional expectation

$$\mathbb{E}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right] = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)\,\mathbb{E}\left[\varepsilon_i \mid X_1, X_2, \ldots, X_n\right]}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

- Invoking assumption 2,

$$\mathbb{E}\left[\varepsilon_i \mid X_1, X_2, \ldots, X_n\right] = 0$$

we see the slope estimator is unbiased both conditionally and unconditionally

$$\begin{aligned}
&\mathbb{E}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right] = \beta_1 \\
&\qquad \Longrightarrow \mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right]\right] = \mathbb{E}\left[\beta_1\right] = \beta_1
\end{aligned}$$

- For the intercept,

$$
\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0 \mid X_1, X_2, \ldots, X_n\right] &= \mathbb{E}\left[\bar{Y}_n - \hat{\beta}_1 \bar{X}_n \mid X_1, X_2, \ldots, X_n\right] \\
&= \mathbb{E}\left[\beta_0 + \beta_1 \bar{X}_n + \bar{\varepsilon}_n - \hat{\beta}_1 \bar{X}_n \mid X_1, X_2, \ldots, X_n\right] \\
&= \beta_0 + \mathbb{E}\left[\bar{\varepsilon}_n \mid X_1, X_2, \ldots, X_n\right] \\
&= \beta_0 + \mathbb{E}\left[\varepsilon_i \mid X_1, X_2, \ldots, X_n\right] \\
&= \beta_0
\end{aligned}
$$

Q: What is the standard error of $\hat{\beta}_1$?

$$
\mathrm{Var}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right] = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \, \mathrm{Var}\left[\varepsilon_i \mid X_1, X_2, \ldots, X_n\right]}{\left(\sum_{i=1}^{n} (x_i - \bar{x}_n)^2\right)^2}
$$

- Putting in terms of sample variance, we have

$$\mathrm{Var}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right] = \frac{\sigma^2}{(n-1)s_x^2}$$

- Thus the conditional standard error of $\hat{\beta}_1$ is

$$\mathrm{SE}\left(\hat{\beta}_1\right) = \frac{\sigma}{\sqrt{(n-1)s_x^2}}$$

- To obtain the conditional standard error, we use the total variance formula

$$\mathrm{Var}\left[Z\right] = \mathbb{E}\Big[\mathrm{Var}\left[Z \mid W\right]\Big] + \mathrm{Var}\Big[\mathbb{E}\left[Z \mid W\right]\Big]$$

- Since β_1 is a constant, the variance of the estimator is given by

$$\begin{aligned}\mathrm{Var}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\mathrm{Var}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right]\right] \\ + \mathrm{Var}\left[\mathbb{E}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right]\right] = \frac{\sigma^2}{n-1}\mathbb{E}\left[\frac{1}{s_x^2}\right]\end{aligned}$$

- Suppose we make some stronger assumptions about the error ε.

1. The conditional mean of the response is linear in terms of β_0, β_1, x_i

$$\mathbb{E}\left[Y_i \mid X_i = x_i\right] = \beta_0 + \beta_1 x_i$$

2. The errors have zero mean and constant variance

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{and} \quad \operatorname{Var}\left[\varepsilon_i \mid X_i\right] = \sigma^2 \qquad \text{where} \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

3. The errors are independent of X_i, and of each other.
4. The errors follow the normal distribution of $N\left(0, \sigma^2\right)$.

- With this set of stronger assumptions, we can consider MLE

$$\begin{aligned} f_{\varepsilon_i}(e_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(e_i\right)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right) \end{aligned}$$

- The independency of ε_i means the likelihood function is given by,

$$\mathcal{L}(b_0, b_1, s^2 \mid y_i, x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(y_i - (b_0 + b_1 x_i))^2}{2s^2}\right)$$

- The MLE of β_0, β_1 and σ^2 are values of b_0, b_1 and s^2 that maximised

$$\mathcal{L}(b_0, b_1, s^2 \mid y_i, x_i)$$

- This is equivalent to finding b_0, b_1 and s^2 that maximise the log-likelihood,

$$\begin{aligned} \ell(b_0, b_1, s^2 \mid y_i, x_i) &= \sum_{i=1}^{n} \left(-\frac{1}{2}\ln(2\pi) - \ln s - \frac{(y_i - b_0 - b_1 x_i)^2}{2s^2}\right) \\ &= -\frac{n}{2}\ln(2\pi) - n\ln s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

  which is more convenient to work with.

- Setting the first derivatives of ℓ with respect to b_0, b_1, and s to 0, we have

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)(x_i) = 0$$

$$-\frac{n}{s} + \frac{1}{s^3}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = 0$$

- Notice the first two equations are essentially the same as those from LSE, so we will have the same estimators, and all have been said about them hold.
- From the third equation, we have

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \widehat{\mathrm{MSE}}$$