22nd March'2024

Section 6.4: Straight-Line Models.

(A.K.A, Linear Regression Models)

Recall our assumptions from Sec 6.1:

$$Y_i \sim N(\mu_i, \delta^2), \quad i = 1, 2, \dots, n, \text{ all independent.}$$

where, $\mu_i = \alpha + \beta x_i$ is constant for given x_i

Note: Constant Variance !!

$$\therefore Y_i = \alpha + \beta x_i + \varepsilon_i, \text{ where } \varepsilon_i \overset{iid}{\sim} N(0, \delta^2)$$

Error terms are where Y gets its randomness.

$$\varepsilon_i = \underbrace{y_i}_{\text{Obs}} - \underbrace{(\alpha + \beta x_i)}_{\text{Expected}}$$

Obs − Expected.

We want to assess data in Ordered pair form:

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

x_i = predictor/explanatory variable (independent)

y_i = response variable (dependent)

Let's look at Example 6.4.1 in the Lecture Notes...

⇒ We want to relate the monthly Co-op Salary (y_i) to the # of work terms (x_i)

(1) Explain the relationship between y and x.

(2) Predict y given some x.

Look at Fig 6.6 : Boxplots of Salary / work term

↳ What do we notice about this graph ?

- WT 7 has no whiskers ⟹ means no outliers.
- Median increasing with WT #.
- Box plots overlap as WT increases.
- The fitted line passes through all the boxes.
- Box size fluctuates.
- Linear Model seems appropriate.
- Variation is large for earlier W.T's

$$Y_i \sim N(\alpha + \beta x_i, \delta^2) \quad \text{independent.}$$

Using likelihoods, we can find $(\hat{\alpha}, \hat{\beta}, \hat{\delta}^2)$, our joint MLE.

$$l(\alpha, \beta, \delta^2) = -n\ln(\delta) - \frac{1}{2\delta^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

Now, $\frac{\partial l}{\partial \alpha}, \frac{\partial l}{\partial \beta}, \frac{\partial l}{\partial \delta} \longrightarrow$ we can find that.

- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
- $\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} = \frac{S_{xy}}{S_{xx}}$ } related to correlation $\rho = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$
- $\hat{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

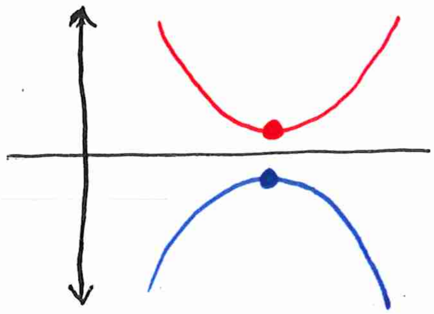
$$\Rightarrow \hat{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}$$

$\hat{\varepsilon}_i$ is called "the residuals".

So, $E(\hat{\delta}^2) \neq \delta^2 \Rightarrow$ it is a biased estimate.

$\hat{\alpha}, \hat{\beta}$ are also called Least Square Estimates, why?

- $(\hat{\alpha}, \hat{\beta})$ maximize $\quad - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$
- $(\hat{\alpha}, \hat{\beta})$ also minimize $\quad \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$



where, $y_i - \alpha - \beta x_i = \varepsilon_i$
"Error term"

However, the L.S.E for σ^2 is different:

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$
$$= \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$$

☆ From $\left.\frac{\partial l}{\partial \alpha}\right|_{\hat{\alpha}, \hat{\beta}, \hat{\sigma}}$ and $\left.\frac{\partial l}{\partial \beta}\right|_{\hat{\alpha}, \hat{\beta}, \hat{\sigma}}$, we find that $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

In R: the function for fitting a linear Model is

lm( ) ⟶ "Linear model"

Look at pg 33 of Chapter 6 Lecture Notes.

Returning to Example 6.4.1, the fitted model from R is given below.

```
> Sal.lm<-lm(SalMonth~WTNumN, data=salarynz)
> summary(Sal.lm)
```

$\widehat{E(y)} = \hat{\alpha} + \hat{\beta}X$

```
Call:
```

Function call:

```
lm(formula = SalMonth ~ WTNumN, data = salarynz)

Residuals:
    Min      1Q  Median      3Q     Max
-2960.8  -522.6  -157.4   406.4  4136.2
```

5 number summary of residuals

t obs

p-value

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2887.40      56.26   51.33   <2e-16 ***
WTNumN        234.99      21.06   11.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

α $\hat{\alpha}$

β $\hat{\beta}$

Ho: $\alpha = 0$ given β in the model

Ho: $\beta = 0$ given α in the model

```
Residual standard error: 874.7 on 1149 degrees of freedom
Multiple R-squared:  0.09779,Adjusted R-squared:  0.097
F-statistic: 124.5 on 1 and 1149 DF,  p-value: < 2.2e-16
```

Std error is a part of CI calculation.

Figure 6.7: R Output: Linear regression for salary data

- The estimated relationship between monthly salary and work term number is:

$$\text{Salary} = 2887.40 + 234.99 \times \text{Work Term number}.$$

- The estimate of σ is $s =$"Residual standard error" $= 874.7$ on 1149 degrees of freedom.
- We estimate that monthly salary increases by \$234.99 for each additional work term.
- The intercept estimate is the estimated monthly salary for zero work terms, but this is not meaningful here. Instead, we could quote the estimated monthly salary for work term 1, \$2887.40 + \$234.99.