L 2. Simple Linear Regression

- Fitted value and residuals
- Estimate of σ^2
- Distribution of $\hat{\beta}_1$
- Sum of Squares Phenomenal

Recall:

Simple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

Least-square estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize $S(\beta_0, \beta_1) = \Sigma (y_i - \beta_0 - \beta_1 x_i)^2$

then the **fitted value of the i^{th} observation** is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

For any given value $X = x$, we would estimate $E(y) = \beta_0 + \beta_1 x$

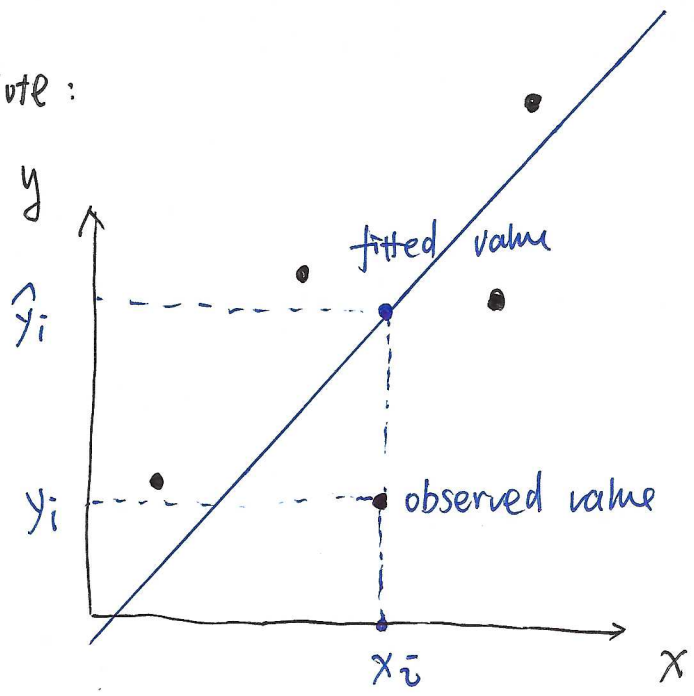
(doesn't have to be in the data)

by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \leftarrow$$

Best prediction of y given $X = x$ using the existing data

Note:



← fitted line using the data
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

points are the given data

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ predicts $E(y|_{x=x}) = b_0 + b_1 x$
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ predicts the value of y at $X = x$

What's the difference? The estimated value is the same, but the estimation error (variance) is different. will discuss in L3.

1. Fitted Value and Residuals

Now we have $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

and fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Then ε_i can be estimated by **residuals**

$$\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$$

Can you see e_i is unbiased estimate of ε_i?

Properties of $\hat{y}_i$ and e_i

(1) The fitted residuals must sum to zero

$$\sum e_i = \sum (y_i - \hat{y}_i)$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = n(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}) = 0$$

— from LSE

Consequently

(2) $\sum y_i = \sum \hat{y}_i$

(3) $\sum e_i^2$ is at minimum, there's no better fitted line to make it smaller — from LSE

(4) $\sum e_i x_i = 0$ — from LSE

(5) $\sum e_i \hat{y}_i = 0$

II. Estimate of σ^2

Recall we have some unfinished business: unknown σ^2

If we use e_i to estimate ε_i, then we may use it to estimate σ^2.

— recall in stats, estimate of σ^2 always starts with χ^2 distribution.

Sum of Squared Errors $SSE = \sum_{i=1}^{n} e_i^2 = \Sigma (y_i - \hat{y}_i)^2$

where $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, since H's somehow the sum of squares of normal distributions, we grab the result:

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

The proof is lengthy and not required, but I will provide it in the materials for the ones who're interested.

Therefore, $E\left(\frac{SSE}{\sigma^2}\right) = n-2 \iff E\left(\frac{SSE}{n-2}\right) = \sigma^2$

Mean Squared Errors

$$MSE = \frac{SSE}{n-2} = \frac{\Sigma e_i^2}{n-2}$$

Is the unbiased estimator of σ^2

III. Distribution of $\hat{\beta}_1$

Recall the facts we have known about $\hat{\beta}_1$:

$E(\hat{\beta}_1) = \beta_1$, $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, where σ^2 can be estimated by MSE.

- When σ^2 is known (magic!)

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

  so $\hat{\beta}_1 = \sum k_i y_i \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

- When σ^2 is unknown (common!), and we estimate it by MSE, then the studentized distribution becomes:

$$\frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} \sim t_{(n-2)}$$

Hint: $S(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$

Why do we like the distribution of $\hat{\beta}_1$?

It can give us the **t-test for the significance of X**

- **Idea**: $y = \beta_0 + \beta_1 x + \varepsilon$

  If X is a "significant" factor to y, when x changes, y should change with it significantly. β_1 is the slope for this change.

- **Hypothesis**: $H_0: \beta_1 = 0$ v.s. $H_1: \beta_1 \neq 0$

  When H_1 is true, β_1 is significantly different from 0, therefore a change in x cause significant change in y — X is significant to Y.

- **Test statistic**: $$t_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}$$

- **Decision rule**: When $|t_0| \geq t_{\alpha/2}$, $df = n-2$

  or p-value $\leq \alpha$ $\Rightarrow$ Reject H_0 $\Rightarrow$ X is significant to y.

IV. Sum of Squares Phenomenal — ANOVA and R^2

- ANOVA: testing the hypothesis related to equality of more than one parameters
  - More useful in Multiple Linear Regression
  - Just to develop the basic math here
  - In SLR, it's also testing $H_0: \beta_1 = 0$

(a) Goal: The variation in y can be explained by the variation in regression line and the variation in the error. Or.

Total Sum of Square in y = Regression sum of squares + Error sum of squares

$$SST = SSR + SSE$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Proof: $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$

$$= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

Note that

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = \sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i - \sum_{i=1}^{n}(y_i-\hat{y}_i)\bar{y}$$

$$= \sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i$$

$$= 0 - 0$$

$$= 0$$

so

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

or

$$SST = SSR + SSE$$

(b) F test for $H_0: \beta_1 = 0$

Use distributions

	sum of squares	d.f.	mean squares
Regression	SSR	1	MSR = SSR/1
Error	SSE	n-2	MSE = SSE/(n-2)
Total	SST	n-1	

we know that $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$

and $\frac{SSR}{\sigma^{*2}} \sim \chi^2(1)$

Ideas

- If the regression line can explain "more" of the variation and the error explains "less" ~~variable~~ variation, then it means including x in the line is significant
- This comparison can be done using the ratio $\frac{MSR}{MSE}$

since $\frac{\chi^2(df_1)/df_1}{\chi^2(df_2)/df_2}$ gives a F distribution. so we can do test!

- To test $H_0: \beta_1 = 0$ v.s. $H_1: \beta_1 \neq 0$,

think how it affects MSR and MSE:

we know that $E(MSE) = \sigma^2$

we can prove that $E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

Hint: Show $MSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

then $E(MSR) = E(\hat{\beta}_1^2) \Sigma (X_i - \bar{X})^2$

- When $\beta_1 = 0$: $E(MSR) = E(MSE) = \sigma^2$

- When $\beta_1 \neq 0$: $E(MSR) > E(MSE)$

So we are expecting $\frac{MSR}{MSE} >> 1$.

- The test. statistic $F = \frac{MSR}{MSE} \underset{\text{under } H_0}{\sim} F(1, n-2)$

- Decision Rule: Reject H_0 when $F > F_{1-\alpha}(df_1 = 1, df_2 = n-2)$

or when p-value $< \alpha$.

□

- **R Squared**

A similiar idea can be measured to see the

goodness of fit of the model

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- This is known as the "coefficient of determination"
- This measures how much variation in y (stated by $SST = \Sigma (y_i - \bar{y})^2$), is explained by the variation in regression (stated by $SSR = \Sigma(\hat{y}_i - \bar{y})^2$) and how much unexplainable part is contained in the variation of error (stated by $SSE = \Sigma(\hat{y}_i - y_i)^2$).

- It can be proved that

$$R^2 = r_{xy}^2 = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

the square of the sample correlation between x and y.

- $0 \leq R^2 \leq 1$, closer to 1 indicates better fit.

Summary and Take-aways of this lecture.

Fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Residual $e_i = y_i - \hat{y}_i$

Fitted line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

→

$E(\hat{y}_i) = \beta_0 + \beta_1 x_i = E(y_i)$

$Var(\hat{y}_i)$ will be discussed later

$E(e_i) = \varepsilon_i$

$E(\bar{y}) = \beta_0 + \beta_1 \bar{x}$

sum of squared errors

$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$

$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$

mean squared errors

$MSE = \frac{SSE}{n-2}$

$E(MSE) = \sigma^2$

→

SSE measures the variation in errors;

MSE is the unbiased estimate of σ^2

When σ^2 is known:

$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

When σ^2 is unknown

$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{(n-2)}$

→

Use the t distribution of $\hat{\beta}_1$ to test:

$H_0: \beta_1 = 0$ v.s. $H_1: \beta_1 \neq 0$

for the significance of X to Y.

ANOVA Test for $H_0: \beta_1 = 0$ v.s. $\beta_1 \neq 0$

	s.s.	d.f.	M.S	F stats
Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2$	1	$MSR = SSR/1$	$\frac{MSR}{MSE}$
Error	$SSE = \sum(\hat{y}_i - y_i)^2$	$n-2$	$MSE = \frac{SSE}{(n-2)}$	
Total	$SST = \sum(y_i - \bar{y})^2$	$n-1$		

Reject H_0 if $F_{stat} > F_{1-\alpha, df_1=1, df_2=n-2}$

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, closer to 1 better fit

- ANOVA is essential comparing two models:

  $H_0: \beta_1 = 0 \iff y = \beta_0 + \varepsilon$

  $H_1: \beta_1 \neq 0 \iff y = \beta_0 + \beta_1 x + \varepsilon$

  If reject H_0, the model including x factor is significantly better than the null model.

- In the simple regression case, $F_{stat} = \frac{\hat{\beta}_1^2 \sum(x_i - \bar{x})^2}{MSE} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}\right)^2$

  $= t_{stat}^2$

  the test result is often the same.