L2. Simple Linear Regression

- Fitted values and residuals
- Estimate σ^2
- Distribution of $\hat{\beta}_1$
- Sum of squares phenomenal

1. Fitted values and residuals

$\boxed{\text{Recall}}$ SLR model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

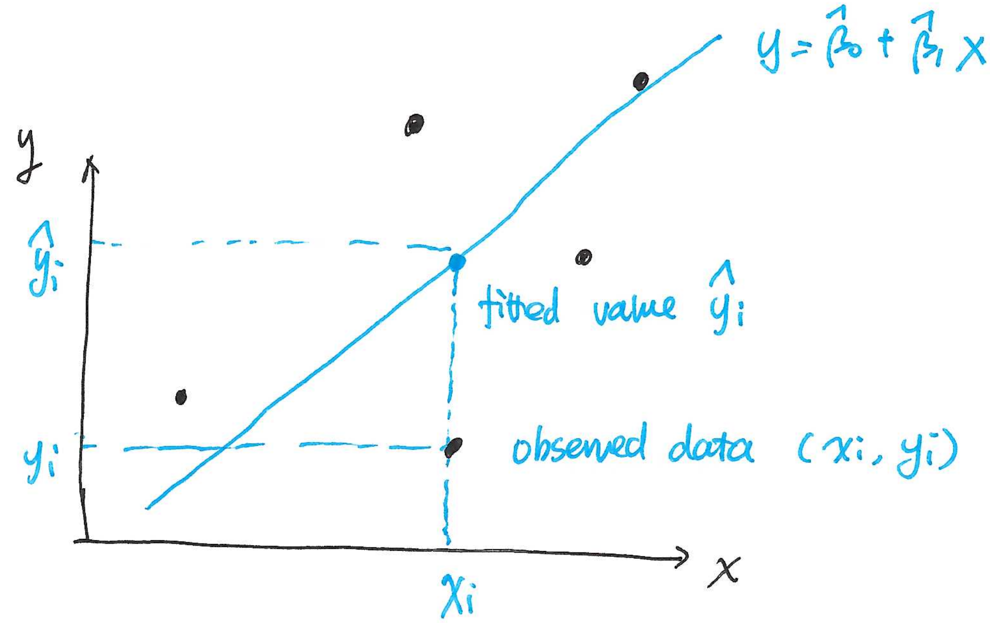
Find OLSE $\hat{\beta}_0$ and $\hat{\beta}_1$

Then the $\boxed{\text{fitted value of the } i\text{th observation}}$ is

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

the fitted regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{or} \quad y = \hat{\beta}_0 + \hat{\beta}_1 x$$



- $\hat{y}_i$ predicts the value of Y at $X = x_i$

- $\hat{y}_i$ estimates $E(Y)$ at $X = x_i$

Difference? The point estimate value is the same,

but the estimation error (variance) is different. (L3)

The estimated prediction error is denoted by $\boxed{\text{residuals}}$

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

Difference between ε_i and e_i?

$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ where β_0 and β_1 are the true unknown parameters.

$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

- ε_i random $\overset{iid}{\sim} N(0, \sigma^2)$

- e_i random $\sim N\left(0,\ \sigma^2\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_j-\bar{x})^2}\right]\right)$

  - why it's a r.v.?
  - why normal distribution?
  - why $E(e_i)=0$, $Var(e_i)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_j-\bar{x})^2}\right]$

$$Var(e_i) = Var\left[(\hat{\beta}_0+\hat{\beta}_1 x_i)-(\beta_0+\beta_1 x_i+\varepsilon_i)\right]$$

$$= Var\left[(\bar{y}-\hat{\beta}_1\bar{x})+\hat{\beta}_1 x_i-(\beta_0+\beta_1 x_i)-\varepsilon_i\right]$$

$$= Var(\bar{y}) + (x_i-\bar{x})^2\, Var(\hat{\beta}_1) + Var(\varepsilon_i)$$

$$= \frac{\sigma^2}{n} + (x_i-\bar{x})^2\frac{\sigma^2}{S_{xx}} + \sigma^2$$

$$= \sigma^2\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right]$$

Hint:

all the covariances are 0. For e.x.

$$cov(\bar{y}, \hat{\beta}_1) = cov\left(\sum\frac{y_i}{n}, \sum k_j y_j\right)$$

$$= \sum cov\left(\frac{1}{n}y_i, k_j y_j\right)$$

$$= \sum\frac{k_j}{n}cov(y_i, y_j)$$

$$= \sum\frac{k_i}{n}\sigma^2$$

$$= \frac{\sigma^2}{n}\sum k_i = 0$$

Properties of $\hat{y}_i$ and e_i : good practice to prove them

(i) $\sum_{i=1}^{n} e_i = 0$

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

(iii) $\sum e_i^2$ is at mininum, no better fitted line to make it smaller

(iv) $\sum e_i x_i = 0$

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

II. Estimate σ^2

Recall: The variance of all random components : ε_i, Y_i, $\hat{\beta}_0$, $\hat{\beta}_1$ and etc are related to the given data and σ^2

But σ^2 is normally unknown.

Recall: How did we estimate σ^2 of a given sample from normal distribution? What distribution is involved with the estimator?

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

where $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, since it'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 materals for the ones who're interested.

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

$\boxed{\text{Mean Squared Errors}}$

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

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

- When σ^2 is unknown (common!), and we esitimate 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 $\boxed{\text{t- test for the significance of X}}$

- 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 signicantly different from 0, therefore a change in x cause significant change in y — X is significant to Y.

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

- 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

$$\boxed{\begin{array}{ccccc} SST & = & SSR & + & SSE \\ \sum_{i=1}^{n}(y_i-\bar{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 & + & \sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \end{array}}$$

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

$$SST \quad = \quad SSR \quad + \quad SSE$$

(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 $\beta_1 = 0$ or not would affect 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 find $E(\hat{\beta}_1)^2$

- when $H_0: \beta_1 = 0$ is true: $E(MSE) = E(MSR) = \sigma^2$

- when $H_1: \beta_1 \neq 0$ is true: $E(MSR) > E(MSE)$

Recall: t test for mean: $H_0: \mu = 0$ v.s. $H_1: \mu \neq 0$

when H_0 is true, $E(\bar{X}) = 0$

when H_1 is true, $E(\bar{X}) \neq 0$

For a given sample, define test stat $t_{stat} = \dfrac{\bar{X}}{se(\bar{X})}$

if $|t_{stat}| \geq t_{\alpha/2}$, then the sample mean is far away from 0 enough to make a conclusion that $E(\bar{X}) \neq 0$, with a type I error α.

Similarly, for a given data set, calculate MSE and MSR.

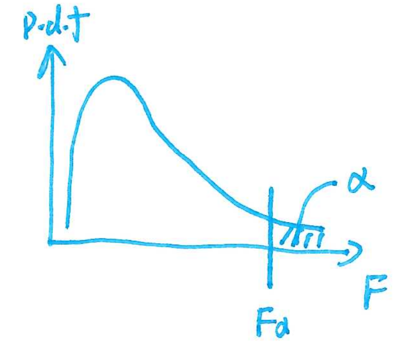
define

- test stat $F_{stat} = \frac{MSR}{MSE}$

- based on sample distribution, $F_{stat} \underset{H_0 \text{ is true}}{\sim} F(1, n-2)$

- decision rule: reject H_0 if $F_{stat} > F_{\alpha, df_1=1, df_2=n-2}$

  or when p-value $\le \alpha$



- conclusion if rejected:

  β_1 is significantly different from 0,

  X is a significant variable to be kept in the model.

More generally, we can write the test as:

$H_0: \; y_i = \beta_0 + \varepsilon_i$ — reduced model

$H_1: \; y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ — full model

To test which model interpret the data better.

and re-write $F_{stat} = \dfrac{(SST_{Full} - SSE_{Full}) / (1)}{SSE_{Full} \; / \; (n-2)}$

$$= \frac{(SSE_{Reduced} - SSE_{Full}) / (df_{Reduced} - df_{Full})}{SSE_{Full} \, / \, df_{Full}}$$

- $SST_{Full} = SST_{Reduced} = SSE_{Reduced}$
- $df_{Full} = n-2$ for SSE_{Full}

  $df_{Reduce} = n-1$ for $SSE_{Reduced}$

- Since $MSR_{Full} = SSE_{Reduced} - SSE_{Full} > 0$

  $\boxed{SSE_{Reduced} > SSE_{Full}}$: when adding more predictors, you always get a reduction in SSE

- when H_0 is true, adding X to the model does not help reducing the variation caused by the error significantly, then the reduced model makes more sense : less calculation burden

  with similar information

- When H_1 is true: adding the predictor X substantially reduce the variation in error. so it makes sense to choose the full model.

- More discussion on the d.f. of SSE:

d.f. of SSE has an impact on the precise estimation of the parameters and tests: $Var(\hat{\beta}_1)$ is estimated by

$$\frac{SSE/df}{S_{xx}}$$

when df in SSE drops (more predictors in the model), it increases the sample variance of $\hat{\beta}_1$.

we are always finding a balance of underfitting (SSE is high) and overfitting (df is low)

- In SLR, ANOVA and t test are identical.
- We will have more discussions in MLR.

- R Squared

R square is used for model selection purpose, will discuss more in MLR

A similiar idea can be measured to see the $\boxed{\text{goodness of fit}}$ 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.

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

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

$\rightarrow$

$\hat{y}_i$ gives the estimate of $Y|x_i$

and $E(Y)|x_i$

$Var(\hat{y}_i)$ needs to be discussed separately.

$E(e_i) = 0$, $Var(e_i) = \sigma^2 [1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}]$

Sum of squared errors

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

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

$MSE = \frac{SSE}{n-2}$ is an unbiased estimate of σ^2

t-test on $H_0: \beta_1 = 0$

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

or $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$ when σ^2 unknown

$\rightarrow$

$t_{stat} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}$

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

or p-value $\leq \alpha$

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

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

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, $F_{stat} = \frac{\hat{\beta}_1^2 \Sigma(x_i - \bar{x})^2}{MSE} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}\right)^2$

$= t_{stat}^2$

the test result is often the same.