Lecture 9

Last class : Categorical variables (variables that can take values that fall into several categories). For >2 categories,

→ Treat them as numerical (if appropriate)

→ Converted into indicator/dummy variables

- In general, k categories will require $k-1$ indicator variables.

ANOVA

- Consider the regression analysis from new perspective called ANalysis Of VAriance (ANOVA)
- How well does a regression model fit the response variable?
- Idea of ANOVA is to partition the variability in the responses. In particular, the variability in the response is measured by Total Sums of Squares (SS(tot)) :

$$SS(tot) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$: clearly related to sample var. of $y_1, \cdots, y_n$.

(Sample variance is $SS(tot)/n-1$).

→ Measures the deviation of the responses from the sample mean.

→ Greater the SS(tot), greater the variation.

ANOVA decompose SS(tot) as follows:

$$SS(tot) = SS(res) + SS(reg)$$

(1) $SS(res) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} e_i^2$ (residual sums of squares).

→ Measures the deviation of responses from fitted values.

(2) SS(reg) : regression sums of squares

$$SS(reg) = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 \qquad \left(\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \right)$$

→ Measures the deviation of fitted values from sample mean.

* SS(reg) : variability that is explained by our model.

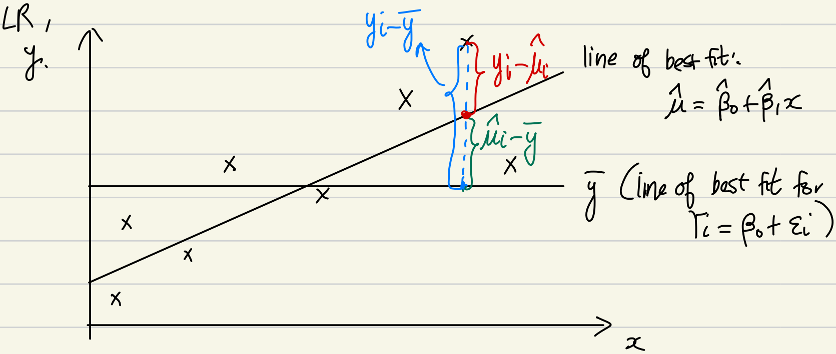
SS(res) : variability that is not explained by our model.

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

Decomposition of $y_i-\bar{y}$:

$$y_i-\bar{y} = y_i-\hat{\mu}_i + \hat{\mu}_i-\bar{y}$$

In SLR,



Note 1. When regression model fits the data well, the observations will be close to fitted values $\Rightarrow$ $y_i-\hat{\mu}_i$ may be smaller, $\hat{\mu}_i-\bar{y}$ may be larger.

2. Line given by $\bar{y}$ is obtained using regression where it's assumed that $\beta_1=0$ (no relationship between explanatory var. and response)

Mathematically, we can show the partition of SS(tot) as follows:

$$\begin{aligned}SS(tot) = \sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{\mu}_i+\hat{\mu}_i-\bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2 + \sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2 + 2\sum_{i=1}^{n}\overbrace{(y_i-\hat{\mu}_i)}^{e_i}(\hat{\mu}_i-\bar{y}) \\ &= \underbrace{\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2} + \underbrace{\sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2} + 2\underbrace{\sum_{i=1}^{n}e_i\hat{\mu}_i}_{=0} - 2\bar{y}\underbrace{\sum_{i=1}^{n}e_i}_{=0} \\ &= SS(res) + SS(reg).\end{aligned}$$

$$\therefore \underbrace{SS(tot)}_{\text{Measure of deviation of responses from sample mean}} = \underbrace{SS(res)}_{\text{Measure of deviation of responses from fitted values}} + \underbrace{SS(reg)}_{\text{Measure of deviation of fitted values from sample mean.}}$$

Breakdown of degrees of freedom (df):

- $SS(tot) = \sum_{i=1}^{n}(y_i-\bar{y})^2$ has $n-1$ df.
  - → 1 df lost due to estimation of sample mean.
  - → denominator of sample variance of response.
- $SS(res) = \sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$ has $n-p-1$ df
  - → $p+1$ df lost due to estimation of $\vec{\beta}$ for estimating fitted values.
  - → denominator of $\hat{\sigma}^2$
- $SS(reg) = \sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2$ has p df.
  - → df of SS(res) + df of SS(reg) equals df of SS(tot)
    - ⇒ df of $SS(tot) = n-1$, $(n-1)-(n-p-1)=p$
  - → p is the number of explanatory variables.

Summaries in ANOVA table

Source	SS	df	Mean Squares (→ MS)	F
Regression	SS(reg)	p	$MS(reg) = SS(reg)/p$	$MS(reg)/MS(res)$
Residual	SS(res)	$n-p-1$	$MS(res) = SS(res)/n-p-1$	
Total	SS(tot)	$n-1$		

Mean Squares:

- $MS(reg) = SS(reg)/p$
- $MS(res) = SS(res)/n-p-1 = \hat{\sigma}^2$

F-test: (next class)

- Used to test for the significance of our regression model.

Coefficient of Determination (R^2):

$$R^2 = SS(reg)/SS(tot)$$

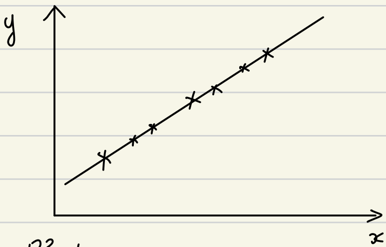
→ R^2 is the proportion of total variation in the response variable that's explained by regression model.

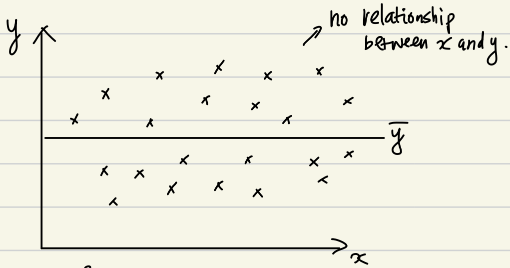
→ Bigger the R^2 value, the better fit. (fitted values are closer to y_i's.)

→ R^2 will be in $[0,1]$.

When $R^2=1$, then this implies a perfect fit s.t. $\hat{\mu}_i = y_i \;, \forall i$
and that $SS(res)=0$ and $SS(reg) = SS(tot)$ (generally not the case)

In SLR,





$R^2=1$

$SS(reg) = SS(tot)$, $SS(res)=0$

$\hat{\mu}_i = y_i \,, \forall i$

$R^2=0$

$SS(res) = SS(tot)$, $SS(reg) = 0$ → $\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 = 0$

$\hat{\mu}_i = \bar{y} \,, \forall i$ $(\hat{\beta}_0 = \bar{y}, \hat{\beta}_1 = 0)$.

Note: In SLR, R^2 is equal to squares of sample correlation between x and y (or r)

$\therefore R^2 = r^2$

Recall: $r = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \Rightarrow r^2 = \dfrac{S_{xy}^2}{S_{xx}S_{yy}}$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \rightsquigarrow \sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \quad (\text{def}^n \text{ of } \hat{\beta}_0)$$

$$= \sum_{i=1}^{n}\left[\frac{S_{xy}}{S_{xx}}(x_i - \bar{x})\right]^2 \quad (\text{def}^n \text{ of } \hat{\beta}_1)$$

$$= \frac{S_{xy}^2}{S_{xx}}$$

$$\therefore R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \Rightarrow R^2 = r^2 \quad \blacksquare$$

Numerical Example Lung Function Data ($n = 655$; age, height, gender, smoke).

Source	SS	df	MS	F
Reg	369.88	4	92.47	462
Res	130.67	650	0.20 $\nearrow \hat{\sigma}^2$	
Total	500.55	654		

$R^2 = SS(reg) / SS(tot)$

$= 0.74.$

$\therefore$ regression model with age, height, gender and smoking status explains 74% of variation in the response.