Least Squares Estimator

CON 2012: Consumer Big Data Analysis I

Instructor: Tae-Young Pak

Spring 2022

How can we estimate the regression model?

$\Rightarrow$ Find a line that <u>best</u> explains the data

$\Rightarrow$ Find a line that <u>best</u> explains the association between x and y

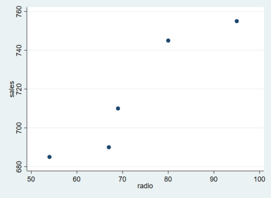
Example: sales and advertising expenditures

- Data on radio advertising expenditures (x) and sales (y)

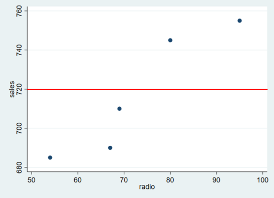
x (=in \$1000)	y (=unit of sales)
54	685
67	690
69	710
80	745
95	755

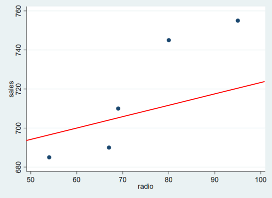
- (x_i, y_i) for person $i = 1, \cdots, n$
- $n = 5$
- $(x_1, y_1) = (54, 685)$, $(x_2, y_2) = (67, 690)$, $(x_3, y_3) = (69, 710)$, $(x_4, y_4) = (80, 745)$, $(x_5, y_5) = (95, 755)$

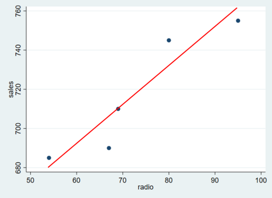
Example: sales and advertising expenditures



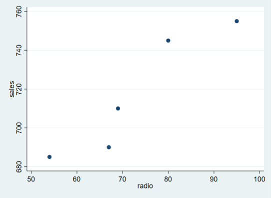
Three different fits to the data: which one looks the best? how can we define “best”?...



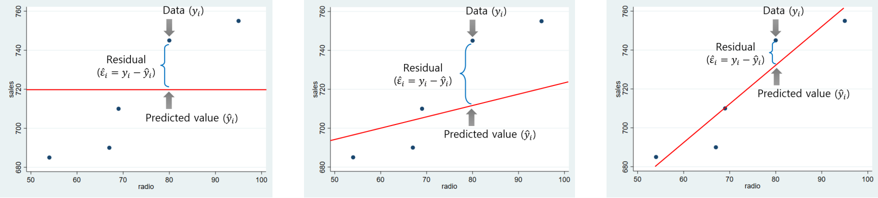




Example: sales and advertising expenditures



Three different fits of the data: which one looks the best? how can we define "best"?...



"The line fits the data well" is equivalent to, saying

= "the data points are close to the line"

= "distance from each data point to the line is short"

= "the sum of distance from each data point to the line is minimized"

sum of distances:

regression (a):
$$Q = (685 - 720) + (690 - 720) + (710 - 720) + (745 - 720) + (755 - 720) = -15$$

regression (b):
$$Q = (685 - 695) + (690 - 702) + (710 - 704) + (745 - 712) + (755 - 721) = 51$$

regression (c):
$$Q = (685 - 680) + (690 - 707) + (710 - 710) + (745 - 730) + (755 - 762) = -4$$

"The line fits the data well" is equivalent to, saying

= "the data points are close to the line"

= "distance from each data point to the line is short"

= "the sum of distance from each data point to the line is minimized"

sum of absolute distances:

regression (a): $Q = |685-720| + |690-720| + |710-720| + |745-720| + |755-720| = 135$

regression (b): $Q = |685-695| + |690-702| + |710-704| + |745-712| + |755-721| = 95$

regression (c): $Q = |685-680| + |690-707| + |710-710| + |745-730| + |755-762| = 44$

"The line fits the data well" is equivalent to, saying

= "the data points are close to the line"

= "distance from each data point to the line is short"

= "the sum of distance from each data point to the line is minimized"

sum of squared distances:

regression (a):
$Q = (685 - 720)^2 + (690 - 720)^2 + (710 - 720)^2 + (745 - 720)^2 + (755 - 720)^2 = 4075$

regression (b):
$Q = (685 - 695)^2 + (690 - 702)^2 + (710 - 704)^2 + (745 - 712)^2 + (755 - 721)^2 = 2525$

regression (c):
$Q = (685 - 680)^2 + (690 - 707)^2 + (710 - 710)^2 + (745 - 730)^2 + (755 - 762)^2 = 588$

$\rightarrow$ regression (c) has the lowest sum of squared distances

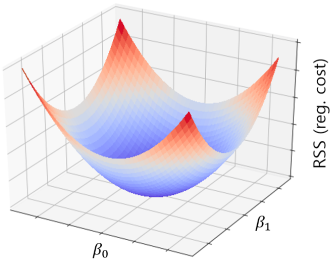
$\rightarrow$ regression (c) is better than other two!!

Residual sum of squares (RSS)

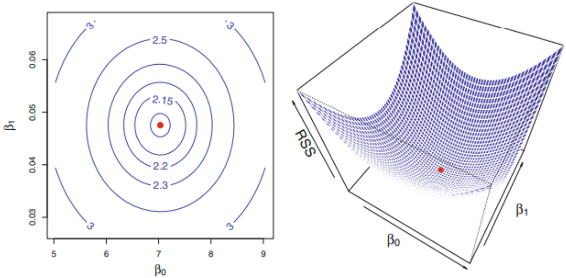
- RSS $= \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$
- often called "**cost function**" or "**loss function**"
- shows the sum of deviations from each data point to the fitted regression line
- an indicator of how well the regression line explains data; measure of "model fit"
- the lower the better fit
- unit dependent ($\Rightarrow$ later we will define a unit independent measure of model fit)

Least Squares (LS) criterion:

- Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimzes sum of squared residuals (RSS)
- Minimize, RSS $= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$
- Two approaches for minimization
  - ◇ Analytic approach: check the first order condition and second order condition to identify $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizing the RSS
  - ◇ Numerical approach: gradient descent ($\Rightarrow$ slide down the curve until the RSS no longer changes)



Analytic solution:



Derivation: take a partial derivative of RSS with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{cases} \dfrac{\partial RSS}{\partial \hat{\beta}_0} = \displaystyle\sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \overset{\text{set}}{=} 0 & (1) \\ \dfrac{\partial RSS}{\partial \hat{\beta}_1} = \displaystyle\sum_{i=1}^{n} -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \overset{\text{set}}{=} 0 & (2) \end{cases}$$

From equation (1),

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i \tag{1.1}$$

From equation (2),

$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \tag{2.1}$$

Equations (1) and (2) or (1.1) and (2.1) are called "**normal equations**"

Re-arranging (2.1) for $\hat{\beta}_0$ leads to

$$\hat{\beta}_0 \sum x_i = \sum x_i y_i - \hat{\beta}_1 \sum x_i^2$$

$$\hat{\beta}_0 = \frac{\sum x_i y_i - \hat{\beta}_1 \sum x_i^2}{\sum x_i} \tag{2.2}$$

If we plug (2.2) into (1.1)

$$\sum y_i = \frac{n\left(\sum x_i y_i - \hat{\beta}_1 \sum x_i^2\right)}{\sum x_i} + \hat{\beta}_1 \sum x_i$$

$$\sum x_i \sum y_i = n\left(\sum x_i y_i - \hat{\beta}_1 \sum x_i^2\right) + \hat{\beta}_1 \left(\sum x_i\right)^2$$

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\left(n \sum x_i^2 - \left(\sum x_i\right)^2\right)}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\left(\sum x_i^2 - n\bar{x}^2\right)}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum \left(x_i - \bar{x}\right)^2}$$

$$= \frac{cov(x, y)}{var(x)}$$

Dividing (1.1) by n leads to

$$\frac{\sum y_i}{n} = \hat{\beta}_0 + \hat{\beta}_1 \frac{\sum x_i}{n}$$
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least Squares estimator (for simple linear regression)

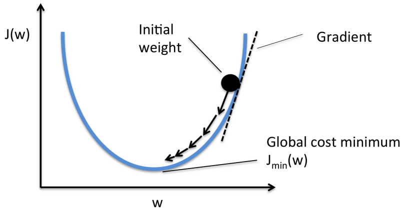
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\left(\sum x_i^2 - n\bar{x}^2\right)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum \left(x_i - \bar{x}\right)^2} = \frac{cov(x,y)}{var(x)}$$

Numerical properties of the LS estimator

- The regression line passes through the sample means of x and y
- Gives the lowest RSS among all possible linear regression fits

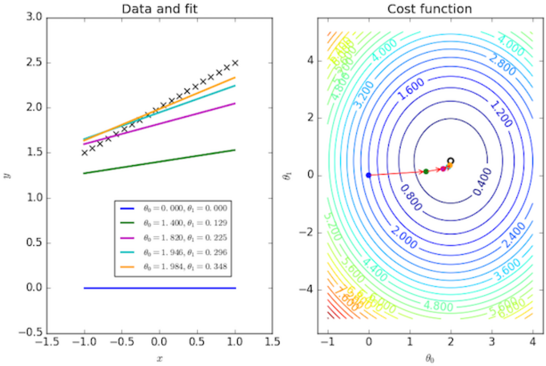
Gradient descent (or, steepest descent): concept

- An optimization algorithm used to find the values of parameters (coefficients) that minimize a cost function (e.g., RSS)
- Best used when the parameters cannot be calculated analytically (e.g. using linear algebra)



- Gradually reduce model error until the stopping condition is met
- Take a big step when the absolute value of gradient is high and take a small step when the absolute value of gradient is low

Gradient descent:



Gradient descent: algorithm

a. Define the loss function and its first derivatives (i.e., gradient)

b. Pick random values for the parameters ($\Rightarrow$ initialization)

c. Plug the parameter vaues into the derivatives

d. Calculate step size:

   $\Rightarrow$ **step size** $=$ **slope** $\times$ **learning rate**

e. Update parameters:

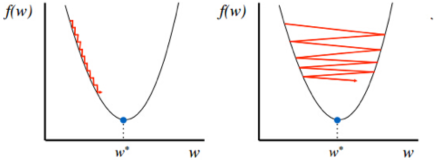
   $\Rightarrow$ **new parameter** $=$ **old parameter** $-$ **step size**

f. Repeat steps d. through e. until the stopping condition is met

   Stopping condition: $\Rightarrow$ $|$step size$| < \nu$, $|\triangle$parameter$| < \delta$, or no. of iterations $< \lambda$

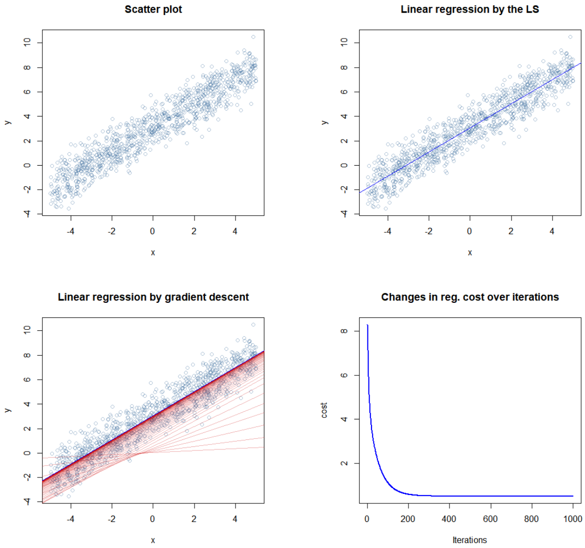
Learning rate

- A step size in gradient descent
- Usually in range of 0 to 1
  - Too high $\rightarrow$ minimum is not reached
  - Too low $\rightarrow$ more iteration is needed to get to a minimum



Too small: converge very slowly
Too big: overshoot and even diverge

Analytic solution vs. Gradient descent:



Analytic solution vs. Gradient descent:

- Gradient descent is preferred over analytic solution when
  - ⋄ explanatory variables include lots of zeros (sparse data)
  - ⋄ data is exceptionally large
  - ⋄ cost function takes a very complex form
- Otherwise, analytic solution is better than gradient descent because
  - ⋄ it is much quicker in calculation
  - ⋄ it always leads to the minimum of regression cost

Numerical example:

x (=in \$1000)	y (=unit of sales)
54	685
67	690
69	710
80	745
95	755

Numerical example:

x	y	$x \cdot y$	x^2
54	685	36990	2916
67	690	46230	4489
69	710	48990	4761
80	745	59600	6400
95	755	71725	9025
$\Sigma = 365$	3585	263535	27591

$$\Rightarrow \bar{x} = 73, \bar{y} = 717$$

$$\Rightarrow \therefore \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\left(\sum x_i^2 - n\bar{x}^2\right)} = \frac{263535 - (5 \cdot 73 \cdot 717)}{27591 - (5 \cdot 73^2)} \approx 1.93$$

$$\Rightarrow \therefore \hat{\beta}_0 \approx 717 - (1.93 \cdot 73) = 575.78$$

$$\Rightarrow \therefore \hat{y} = 575.78 + 1.93 \cdot x$$

Estimated regression line and RSS:

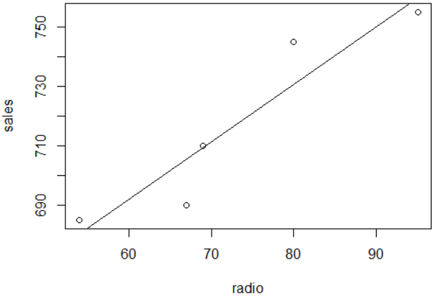
$$\hat{y} = 575.78 + 1.93 \cdot x$$
$$\text{RSS} = 489.94$$

R results:

```
Call:
lm(formula = sales ~ radio, data = prac)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 575.7844    30.8647  18.655 0.000336 ***
radio         1.9345     0.4155   4.656 0.018693 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> plot(prac$radio, prac$sales, xlab="radio", ylab="sales")
> abline(res.lm)
>
> prac$sales_hat <- res.lm$coefficients[1] + (res.lm$coefficients[2]
> rss <- sum((prac$sales - prac$sales_hat)^2)
> rss
[1] 489.9366
>
> ### assumed model fit, sales_hat = 570 + 2*radio
> prac$sales_hat <- 570 + (2.0 * prac$radio)
> rss <- sum((prac$sales - prac$sales_hat)^2)
> rss
[1] 499
>
> ### assumed model fit, sales_hat = 580 + 1.85*radio
> prac$sales_hat <- 580 + (1.85 * prac$radio)
> rss <- sum((prac$sales - prac$sales_hat)^2)
> rss
[1] 515.6975
>
> ### assumed model fit, sales_hat = 600 + 1.7*radio
> prac$sales_hat <- 600 + (1.7 * prac$radio)
> rss <- sum((prac$sales - prac$sales_hat)^2)
> rss
[1] 793.99
```



R code:

```
setwd("E:/Spring 2022/CON 2012/notes/week5_least squares/")
prac <- read.table("advertising.txt", header=TRUE)

res.lm <- lm(sales ~ radio, data = prac)
summary(res.lm)

plot(prac$radio, prac$sales, xlab="radio", ylab="sales")
abline(res.lm)

prac$sales_hat <- res.lm$coefficients[1] + (res.lm$coefficients[2] * prac$radio)
rss <- sum((prac$sales - prac$sales_hat)^2)
rss

### assumed model fit, sales_hat = 570 + 2*radio
prac$sales_hat <- 570 + (2.0 * prac$radio)
rss <- sum((prac$sales - prac$sales_hat)^2)
rss

### assumed model fit, sales_hat = 580 + 1.85*radio
prac$sales_hat <- 580 + (1.85 * prac$radio)
rss <- sum((prac$sales - prac$sales_hat)^2)
rss

### assumed model fit, sales_hat = 600 + 1.7*radio
prac$sales_hat <- 600 + (1.7 * prac$radio)
rss <- sum((prac$sales - prac$sales_hat)^2)
rss
```

Goodness of fit

- How well the regression model fits (explains) the data; in other words, how much variation in y is explained by the model
- indicates explanatory power of the model
- an important indicator of model's capacity to explain y and to predict y
- determines the thickness of prediction interval

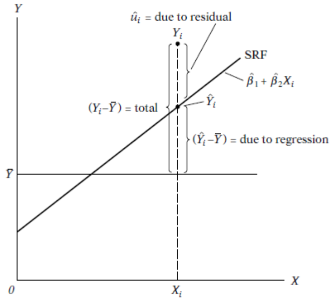
Goodness of fit

- TSS (total sum of squares) $= \Sigma(y_i - \bar{y})^2$

  $\rightarrow$ total variation of the actual y values about their sample mean
- RSS (residual sum of squares) $= \Sigma(y_i - \hat{y}_i)^2$

  $\rightarrow$ residual or unexplained variation of the y values about the SRL
- ESS (explained sum of squares) $= \Sigma(\hat{y}_i - \bar{y})^2$

  $\rightarrow$ explained variation of the estimated y values about their mean
- TSS = RSS + ESS



Goodness of fit

$$\begin{aligned} \mathsf{TSS} &= \mathsf{RSS} + \mathsf{ESS} \\ \rightarrow 1 &= \frac{\mathsf{RSS}}{\mathsf{TSS}} + \frac{\mathsf{ESS}}{\mathsf{TSS}} = \frac{\mathsf{RSS}}{\mathsf{TSS}} + R^2 \\ \rightarrow R^2 &= \frac{\mathsf{ESS}}{\mathsf{TSS}} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}} = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} \end{aligned}$$

R^2 (R-squared):

- measures the percentage of the total variation in y explained by the regression model
- the higher the better for prediction (in most cases.... not always though...)
- unit independent
- $0 \leq R^2 \leq 1$
  - $R^2 = 1$: perfect fit $\rightarrow y_i = \hat{y}_i$ for all i
  - $R^2 = 0$: zero explanatory power; the RHS variables are not explaining any variation in y

Numerical example:

x (=in \$1000)	y (=unit of sales)
54	685
67	690
69	710
80	745
95	755

$$\hat{y} = 575.78 + 1.93 \cdot x$$

TSS $= (685 - 717)^2 + (690 - 717)^2 + (710 - 717)^2 + (745 - 717)^2 + (755 - 717)^2 = 4030$

RSS $= \cdots = 489.94$

ESS $= 4030 - 489.94 = 3540.06$

$\rightarrow R^2 = \frac{\text{ESS}}{\text{TSS}} = 0.878$

$\rightarrow$ The estimated regression model explains about 87.8% of the variation in y

Little more about interpretation of R^2

- No agreed-upon threshold for high R^2; some argue an R^2 of ≥ 0.5 for cross sectional data
- R^2 is a measure of the model's capacity to predict y
- R^2 is less important for getting an unbiased estimate of β (i.e., inference)

Hypothesis testing: testing the significance of regression coefficients

- Test for whether β is significantly different from zero or not
- $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$
  - $\beta = 0 \Rightarrow$ there is no association between x and y
  - $\beta \neq 0 \Rightarrow$ there is a significant association between x and y
- Significance level: $\alpha = 5\%$ (often $\alpha = 1\%$ or 0.1%)
- Test statistic:

$$t = \frac{\hat{\beta} - 0}{se(\hat{\beta})} = \frac{\hat{\beta}}{se(\hat{\beta})} \sim t_{(\frac{\alpha}{2}, n-k)}$$

- Decision rule: rejects the null hypothesis when $|t| > t_{(\frac{\alpha}{2}, n-k)}$

Little more on $se(\hat{\beta})$

- Higher RSS leads to higher $se(\hat{\beta})$
- Higher R^2 leads to lower $se(\hat{\beta})$
- More data points (greater number of observations) leads to lower $se(\hat{\beta})$

R results:

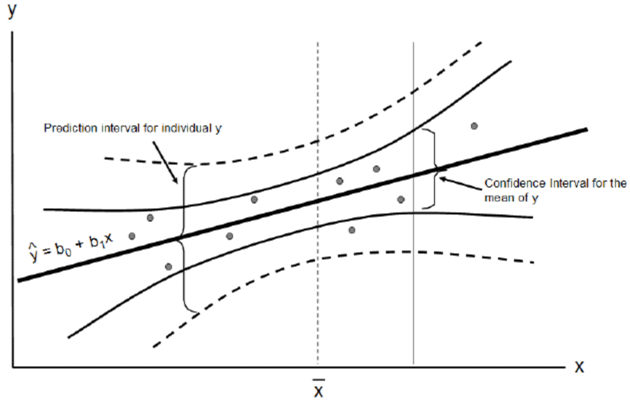
```
Call:
lm(formula = sales ~ radio, data = prac)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 575.7844    30.8647  18.655 0.000336 ***
radio         1.9345     0.4155   4.656 0.018693 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.78 on 3 degrees of freedom
Multiple R-squared:  0.8784,Adjusted R-squared:  0.8379
F-statistic: 21.68 on 1 and 3 DF,  p-value: 0.01869
```

- Null hypothesis?
- Test results?

Confidence interval for the mean of y, prediction interval for individual y



Things to note:

- The intervals show possible ranges of the mean of y and individual y
- The thickness of the intervals are inversely related to the model fit (i.e., R^2)
- The narrower the better for prediction accuracy...
- Both intervals get thicker as x deviates from $\bar{x}$; y values around the boundaries are poorly predicted.
- It's always difficult to predict y around the upper and lower bound of x (extreme values...)