

Introduction to linear regression

(Week 05 lecture notes)

Wei Q. Deng

Department of Statistical Sciences

August 8th 2018

Topics covered in this lecture

- Distinguish between a functional relationship and a statistical relationship.
- Understand the least squares (LS) method.
- Know how to derive and obtain the LS estimates b_1 , b_0 .
- Recognize the difference between a population regression line and the estimated regression line.
- Know the Gauss-Markov conditions for simple linear regression.
- Know the normal error regression model
- Interpret the intercept b_0 and slope b_1 of an estimated regression equation.
- Understand the coefficient of determination R^2 and how to interpret it

What is regression?

- Regression means "going back"
- Linear regression/linear models: a procedure to analyze data
- Historically, *Francis Galton* (1822-1911) invented the term and concepts of regression and correlation.
 - He predicted child's height from fathers height
 - Sons of the tallest fathers tended to be taller than average, but shorter than their fathers.
 - Sons of the shortest fathers tended to be shorter than average, but taller than their fathers.
 - He was deeply concerned about "regression to mediocrity".
 - A brief history of Linear Regression and more about Galton, <http://www.amstat.org/publications/jse/v9n3/stanton.html>
- Regression analysis is a statistical approach to summarize and study the **relationships** between variables in a data set.

Response and predictor variables

- One variable, denoted Y , is regarded as the **response (or outcome, or dependent)** variable
 - whose behaviour that we want to study under the impact of the other variables
 - sometimes unavailable, so we would like to be able to **predict**
- The other variable, denoted X , is regarded as the **predictor (or explanatory, or independent)** variable.
 - variables that gives combinations of conditions under which we want to examine the response variable (e.g. drug treatments, experimental conditions)
 - variables that provides information (e.g. covariates including demographical information, educational background, etc.)

Types of relationships

Relationship between Y and X

- Functional (or deterministic) Relationships
 - $Y = f(X)$, where $f()$ is some function. eg. Circumference $= \pi \times$ diameter.
 - the relationship is deterministic through the function, for each possible X there is only one Y .
- Statistical Relationships
 - $Y = f(X) + \epsilon$, where ϵ is the random error term. eg. a simple linear regression model.
 - the relationship is probabilistic (contains random noise) due to the random error, for each possible X , there is a unique Y (if ϵ is continuous, then the probability of obtaining two exactly the same value is 0)

What does the data look like?

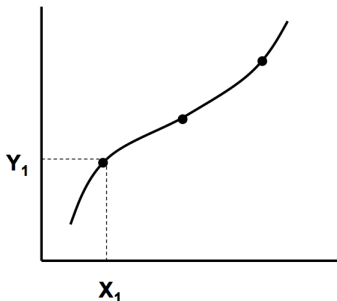
| i | X | Y |
|-----|-----|-------|
| 1 | 0 | 6.95 |
| 2 | 1 | 5.22 |
| 3 | 2 | 6.46 |
| 4 | 3 | 7.03 |
| 5 | 4 | 9.71 |
| 6 | 5 | 9.67 |
| 7 | 6 | 10.69 |
| 8 | 7 | 13.85 |
| 9 | 8 | 13.21 |
| 10 | 9 | 14.82 |

There are $n = 10$ observations. The third observation is the pair $(x_3, y_3) = (2, 6.46)$. In a real dataset, usually you don't have the index i column as given in the table.

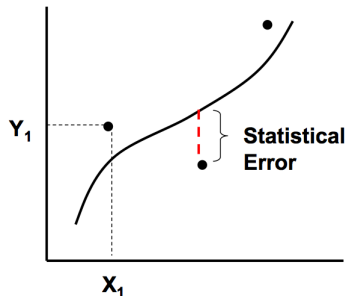
Types of relationships

- Scatter plots of data pair (Y_i, X_i)

Functional Relationship



Statistical Relationship



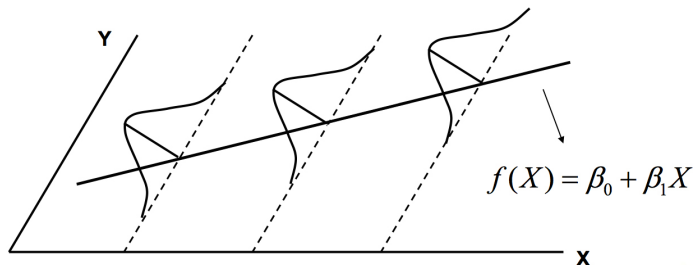
- For each of these relationships, the equation, $Y = f(X)$, describes the relationship between the two variables.
- We are not interested in the functional relationship (deterministic) in this course.
- Instead, we are interested in **statistical relationships**, in which the relationships between the variables is not perfect.

Regression Models

- **Regression model** describes the statistical relationship between the response variable Y and one or more predictor variable(s)
 - The response variable Y has a tendency to vary with the predictor variable X in a systematic fashion.
 - The (data) pairs are **scattered** around the regression curve.
- Regression model assumes a distribution for Y at each level of X .
- When the relationship between Y and X is linear, we call it **linear regression**.
 - In linear regression model, if it concerns study of only one predictor, then we have **simple linear regression model**.
 - In contrast, we have **multiple linear regression**.

A simple linear regression

- It concerns about the statistical relationship between Y and a single predictor X .
- The regression curve is a straight line.



The relationship is linear if it is linear in the model parameters (β_0, β_1) and nonlinear, if it is not linear in parameters.

A simple linear regression

- A formal statistical model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (1)$$

- Y_i is the value of response variable in the i^{th} observation (random but observable).
- X_i is the predictor in the i^{th} observation (a known constant).
- β_0 is the intercept of the regression line (model parameter: assume constant but unknown).
- β_1 is the slope of the regression line (model parameter: assume constant but unknown).
- ϵ_i is the error term (random and unobservable)

A Simple Linear Regression: summary

| R/C | Known | Unknown |
|----------|-------|------------------------------|
| Random | Y | ϵ |
| Constant | X | $\beta_0, \beta_1, \sigma^2$ |

- The parameters associated with a simple linear regression model are $\beta_0, \beta_1, \sigma^2$.
- We need to find estimators for all three of them in order to make use of the model.

Example 1: hourly wage (Y) and years of education (X)

Variables

- Y: hourly wage(pound)
- X: years of education

Parameter interpretation

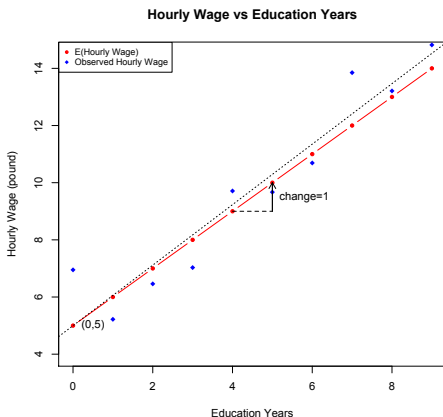
- β_0 : Y-intercept, the starting salary or the baseline salary
- β_1 : slope, the hourly wage increase per one more year of education

Example 1: hourly wage (Y) and years of education (X)

| EducYrs | $E(Y)=E(HWage_T)$ | $Y=HWage_O$ |
|---------|-------------------|-------------|
| 0 | 5 | 6.95 |
| 1 | 6 | 5.22 |
| 2 | 7 | 6.46 |
| 3 | 8 | 7.03 |
| 4 | 9 | 9.71 |
| 5 | 10 | 9.67 |
| 6 | 11 | 10.69 |
| 7 | 12 | 13.85 |
| 8 | 13 | 13.21 |
| 9 | 14 | 14.82 |

- EducYrs (X): years of education;
- $HWage_T$ (true $E(Y)$): the true expected hourly wage (in pounds).
- $HWage_O$ (observed Y): the observed hourly wage (in pounds)

Example 1: hourly wage (Y) and education years (X)



The observed Y goes up and down around the population regression line. In real world, we don't observed the true error term (ϵ), instead we have data (EducYrs , HWage_O). We aim to reveal the true relationship between Y and X using the data we observed. That is, how to use observed data to estimate β_0, β_1 ?

True vs Estimated model

Assume we have a data set of size $n : (Y_i, X_i), i = 1, \dots, n$.

True regression model (or population regression model)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = f(X) + \epsilon_i, \quad f(X) = \beta_0 + \beta_1 X_i$$

Estimated regression model (or sample regression model)

$$\hat{Y}_i = b_0 + b_1 X_i = \hat{f}(X), \quad \hat{f}(X) = b_0 + b_1 X_i$$

- Point estimators of β_0, β_1 are denoted by b_0, b_1 respectively.
- An estimate of Y_i (for a given X_i) is denoted by \hat{Y}_i .
- An estimate of ϵ_i (for given X_i) is denoted by e_i

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

This implies that

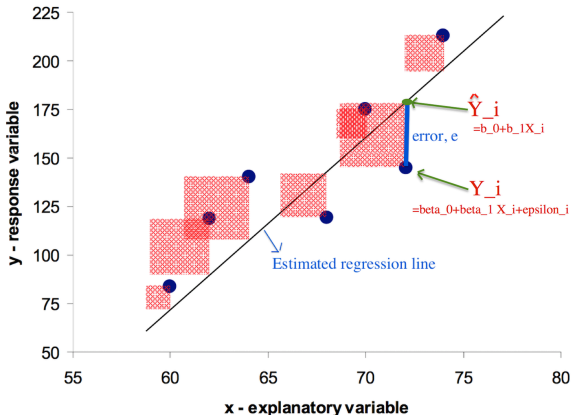
$$Y_i = \hat{Y}_i + e_i = (b_0 + b_1 X_i) + e_i$$

True vs Estimated model

- Difference between $\hat{Y}_i = b_0 + b_1X_i$ and $Y_i = \beta_0 + \beta_1X_i + \epsilon_i$.
- Note that we never observed ϵ_i , but we could estimate it by e_i .

$$Y_i = \hat{Y}_i + e_i = \hat{f}(X) + \text{estimated error}_i,$$

where $e_i = Y_i - \hat{Y}_i$.



Estimation by Least Squares method

Gauss-Markov Assumptions

- **Gauss-Markov Assumptions:**

1. Dependent variable is linear in parameter and can be written as :
$$Y = \beta_0 + \beta_1 X + \epsilon$$
2. $E(\epsilon_i) = 0$. ϵ_i is R.V. with mean 0.
3. $V(\epsilon_i) = \sigma^2$, this homoskedasticity implies that the model uncertainty is identical across observations.
4. $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. ϵ_i and ϵ_j are uncorrelated:

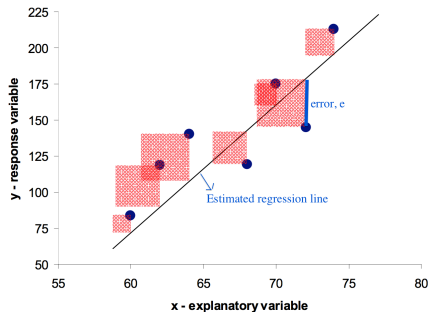
- X is assumed to be constant, ie, X is uncorrelated with the error term ($Cov(X_i, \epsilon_i) = 0$).
- $cov(\epsilon_i, \epsilon_j) = 0$ does not guarantee ϵ_i and ϵ_j are independent. But if they are independent, their covariance must be 0.

- **Above assumptions imply:**

- $E(Y_i|X_i) = \mu_i = \beta_0 + \beta_1 X_i$, that is $f(X) = \beta_0 + \beta_1 X$
- $V(Y_i|X_i) = V(\mu_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$
- $Cov(Y_i, Y_j|X_i) = E\{(Y_i - \mu_i)(Y_j - \mu_j)\} = E(\epsilon_i \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$

We often drop $|X$ notation in above because X is non-random.

Least Square Method



- The equation of the estimated model (or best fitting line) is:
 $\hat{Y}_i = b_0 + b_1 X_i$
- We need to find the values b_0, b_1 that make the sum of the squared prediction error the smallest it can be. That is, find b_0 and b_1 that minimize the objective function Q .

$$Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

Least Square Estimates b_0, b_1

$$Q = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

Minimizing Q gives

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x} \quad (2)$$

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (3)$$

where

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Substituting b_0 in the estimated model, it can be rewritten as

$$\hat{Y}_i = b_0 + b_1 X_i = \bar{Y} + b_1 (X_i - \bar{X}),$$

this also implies

$$Y_i = \bar{Y} + b_1 (X_i - \bar{X}) + e_i$$

i.e. The estimated regression line always goes through the point data point (\bar{X}, \bar{Y}) .

Proof

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) = 0 \quad (4)$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) X_i = 0 \quad (5)$$

These lead to the **Normal equations:**

$$\sum_{i=1}^n Y_i = nb_0 + b_1 \sum_{i=1}^n X_i$$

$$\sum_{i=1}^n X_i Y_i = b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2$$

The normal equations can be solved simultaneously for b_0 and b_1 given in equation (2) and (3) respectively.

proof (not tested)

The Hessian matrix which is the matrix of second order partial derivatives in this case is given as

$$H = \begin{pmatrix} \frac{\partial Q}{\partial \beta_0^2} & \frac{\partial Q}{\partial \beta_0 \beta_1} \\ \frac{\partial Q}{\partial \beta_0 \beta_1} & \frac{\partial Q}{\partial \beta_1^2} \end{pmatrix} = 2 \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$$

- The 2 by 2 matrix H is **positive definite** if its determinant and the element in the first row and column of H are positive.
- The determinant of H is given by $|H| = 4n \sum (x_i - \bar{x})^2 > 0$ given $x \neq c$ (some constant).
- So H is positive definite for any (β_0, β_1) , therefore Q has a global minimum at (b_0, b_1) .

Equivalent formula for b_1

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}} \quad (6)$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{S_{xx}} \quad (7)$$

$$= \sum_{i=1}^n \frac{X_i - \bar{X}}{S_{xx}} y_i = \sum_{i=1}^n k_i Y_i \quad (8)$$

$$= \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{S_{xx}} \quad (9)$$

$$= \sum_{i=1}^n k_i Y_i \quad (10)$$

where

$$k_i = \frac{X_i - \bar{X}}{S_{xx}} = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

suggesting that b_1 is a linear combination of Y_i (assume constant X) and hence is a linear estimator.

Equivalent formula for b_0

$$\begin{aligned} b_0 &= \bar{Y} - b_1 \bar{X} = \sum_{i=1}^n \frac{1}{n} Y_i - \bar{X} \sum_{i=1}^n k_i Y_i \\ &= \sum_{i=1}^n \left(\frac{1}{n} - k_i \bar{X} \right) Y_i \\ &= \sum_{i=1}^n w_i Y_i \end{aligned}$$

where

$$w_i = \frac{1}{n} - k_i \bar{X},$$

suggesting that b_0 is also a linear combination of Y_i and hence is a linear estimator.

Estimation of error terms variance σ^2

- Error sum of squares (SSE) or residual sum of square (RSS)

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

- SSE has $n-2$ degrees of freedom associated with it. Two degrees of freedom are lost because both β_0 and β_1 had to be estimated in obtaining estimated means \hat{Y}_i
- In LS method, the error term variance $\sigma^2 = V(\epsilon_i)$ for all i , is estimated by the error mean square (MSE)

$$s^2 = \text{MSE} = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{(Y_i - \hat{Y}_i)^2}{n-2}$$

Example 2: Estimation (by hand)

- Annual salary (Y) and years of service (X)

| | x_i | y_i | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|-----|-------|-------|-----------------|-----------------|---------------------|---------------------|----------------------------------|
| i=1 | 3 | 34 | -5 | -4 | 25 | 16 | 20 |
| i=2 | 6 | 34 | -2 | -4 | 4 | 16 | 8 |
| i=3 | 10 | 38 | 2 | 0 | 4 | 0 | 0 |
| i=4 | 8 | 37 | 9 | -1 | 0 | 1 | 0 |
| i=5 | 13 | 47 | 5 | 9 | 25 | 81 | 45 |
| Sum | 40 | 190 | 0 | 0 | 58 | 114 | 73 |

Above calculation gives $\bar{X} = 40/5 = 8$ and $\bar{Y} = 190/5 = 38$.

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{73}{58} = 1.258621$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 38 - 1.258621 \times 8 = 27.931$$

Example 2: Estimation (by hand)

- Find $\hat{Y}_i = 27.931 + 1.25862X_i$
 - $\hat{Y}_i = c(31.70686, 35.48272, 40.51720, 37.99996, 44.29306)$
- Find $e_i = Y_i - \hat{Y}_i$
 - $e_i = c(2.29314, -1.48272, -2.51720, -0.99996, 2.70694)$
- Estimate σ^2 by MSE: $s^2 = \hat{\sigma}^2 = \sum e_i^2 / (n - 2) = 7.373563$
 - $\hat{\sigma} = \sqrt{7.373563} = 2.715431$

Inference on model parameters β_1 and β_0

Motivations

- Recall the assumptions:
 1. Dependent variable is linear in parameter and can be written as :
$$Y = \beta_0 + \beta_1 X + \epsilon$$
 2. $E(\epsilon_i) = 0$. ϵ_i is R.V. with mean 0.
 3. $V(\epsilon_i) = \sigma^2$, this homoskedasticity implies that the model uncertainty is identical across observations.
 4. $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. ϵ_i and ϵ_j are uncorrelated:
- In other words, the least squares estimators hold no matter what the distribution of *epsilon*.
- However, if we want to make inference on β_0 or β_1 , we need to make distributional assumptions.

Normal error regression models

Similar to the previous model, the true regression model (or population regression model)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = f(X) + \epsilon_i, \quad f(X) = \beta_0 + \beta_1 X_i$$

The only difference is that for $i = 1, \dots, n$:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In other words,

- The errors are IID normal with mean 0 and **unknown** variance σ^2
- $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ are independent.

Extended to multiple predictors (multiple input variables)

- the output is an **outcome** variable Y
- the inputs are the **predictor** variables X_1, \dots, X_p
- the relationship is described by a linear regression model of the form

$$E(Y|X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

and a model on $Y|X_1, \dots, X_p$ or the distribution of Y can be described by

$$Y \sim \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

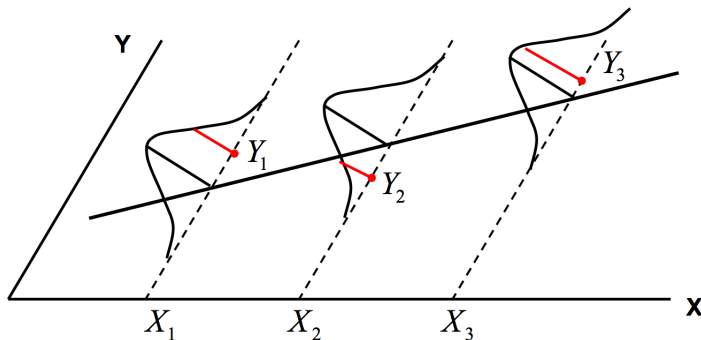
Normal error regression models

This formulation enables

- estimation of β_0 , β_1 and σ^2 through maximum likelihood method
- and
- inference on these parameters via a known sampling distribution

Maximum Likelihood Estimation for Regression

- Similarly, we find the values of parameters $\theta = (\beta_0, \beta_1, \sigma^2)$ in the normal error model that maximize the relevant likelihood function.



- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, the density function at y_i is

$$f(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}$$

MLE for Regression

- Likelihood function

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f(y_i; \theta) = \left\{ \frac{1}{2\pi\sigma^2} \right\}^{n/2} \exp\left\{ -\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}$$

- Log likelihood function

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

Partial differentiation of $\ell(\beta_0, \beta_1, \sigma^2)$ yields

$$\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = \frac{1}{\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i) \quad (11)$$

$$\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum x_i (y_i - \beta_0 - \beta_1 x_i) \quad (12)$$

$$\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (y_i - \beta_0 - \beta_1 x_i)^2 \quad (13)$$

MLE vs LSE for Regression

Set partial derivatives equal to zero and solve them to obtain the MLE of $\beta_0, \beta_1, \sigma^2$.

| Parameters | MLE | Same as LSE? |
|------------|--|---|
| β_0 | $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ | Same. $\hat{\beta}_0 = b_0$ |
| β_1 | $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ | Same. $\hat{\beta}_1 = b_1$ |
| σ^2 | $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}$ | Different. $\hat{\sigma}^2 = \frac{n-2}{n} MSE$ |

Note that the MLE of σ^2 is a biased estimator of the error term variance.

Inference on β_0 and β_1

- We knew how to **point estimate** of β_0, β_1 in simple linear regression from data.
 - Using Ordinary Least Squares (OLS)
 - maximum likelihood estimators (MLE).
- Now we take up the inferences concerning the regression parameters β_0, β_1
 - How accurate are their estimates?
 - How to obtain an interval estimate for each of them?
 - How to test a specific parameter value of interest?

Sampling distribution of b_1

- From earlier, we know (no need to show)
 - $b_1 = \frac{S_{XY}}{S_{XX}} = \sum_{i=1}^n k_i Y_i$, so b_1 is normally distributed.
 - b_1 is unbiased $E(b_1) = \beta_1$, $\sigma^2(b_1) = V(b_1) = \frac{\sigma^2}{S_{XX}}$
 - $\hat{\sigma}^2 = \text{MSE} = \text{SSE}/(n-2)$, so $s^2(b_1) = \text{MSE}/S_{XX}$, $\sqrt{s^2(b_1)} = s(b_1)$
- (Common) Sampling distribution of b_1 following normal error model assumption when σ^2 is unknown:

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$$

- (Not common) Sampling distribution of b_1 following normal error model assumption when σ^2 is known: $\frac{b_1 - \beta_1}{\sigma(b_1)} \sim N(0, 1)$

Inferences concerning β_1

- $1 - \alpha$ confidence limits for β_1 are

$$b_1 \pm t_{1-\alpha/2, n-2} s(b_1) = b_1 \pm t_{1-\alpha/2, n-2} s(b_1)$$

- Testing concerning β_1

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

Note that when $\beta_1 = 0$, there is no linear association between Y and X .

- Test statistics

$$t^* = \frac{b_1 - 0}{s(b_1)}|_{H_0} \sim t_{n-2}$$

- If $|t^*| \leq t_{1-\alpha/2, n-2}$, fail to reject H_0 .
- If $|t^*| > t_{1-\alpha/2, n-2}$, reject H_0 .
- Or find relevant p-value, and we reject H_0 if p-value is less than α .
- If we change $H_a : \beta > 0$ then we reject H_0 if $t^* > t_{1-\alpha, n-2}$ otherwise we fail to reject null hypothesis.

Sampling distribution of b_0

- From earlier, we know (no need to show)
 - $b_0 = \bar{Y} - b_1 \bar{X} = \sum_{i=1}^n w_i Y_i$, so the b_0 is normally distributed.
 - b_0 is unbiased. $E(b_0) = \beta_0$, $\sigma^2(b_0) = V(b_0) = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right\}$
 - $s^2(b_0) = MSE \left\{ \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right\}$
- (Common) Sampling distribution of b_0 Normal error model assumption when σ^2 is unknown:

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$$

- (Not Common) Sampling distribution of b_0 Normal error model assumption when σ^2 is known:

$$\frac{b_0 - \beta_0}{\sigma(b_0)} \sim N(0, 1)$$

Inferences concerning β_0

There are only infrequent occasions when we wish to make inferences concerning β_0 . We do inferences concerning β_0 only when the scope of the model includes $X = 0$.

- $1 - \alpha$ confidence limits for β_0 are

$$b_0 \pm t_{1-\alpha/2, n-2} s(b_0) = b_0 \pm t_{1-\alpha/2, n-2} s(b_0)$$

- Testing concerning β_0 is less interest.

Summarizing and interpreting a linear regression model

Key information from a simple linear regression model

- Test whether the slope is zero (establishing the presence linear relationship)
 - significance is the presence or absence of a linear relationship
 - interpretation is the amount of per unit increase/decrease in Y due to a per unit increase in X
- Calculate the coefficient of determination (assess the strength of the linear relationship)
 - interpretation is the amount of variability in Y that is explained by X
 - for a simple linear regression the pearson's correlation coefficient is exactly the same as the coefficient of determination
- The intercept provides a baseline for discussion.
 - the interpretation is the baseline value of Y when $X = 0$ (this might not make sense in certain applications when range of X does not contain 0).

Example 2. Annual Salary (Y) vs years of service (X)

```
X=c(3,6,10,8,13)  # assign predictor observations to object X
Y=c(34,34,38,37,47) # assign response observations to object Y
lmfit = lm(Y~X)     # fitting data with a simple linear regression
summary(lmfit)      # summary of the fitted model
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      1      2      3      4      5
## 2.293 -1.483 -2.517 -1.000  2.707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.9310     3.1002   9.01 0.00289 **
## X            1.2586     0.3566   3.53 0.03864 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.715 on 3 degrees of freedom
## Multiple R-squared:  0.806, Adjusted R-squared:  0.7413
## F-statistic: 12.46 on 1 and 3 DF, p-value: 0.03864
```

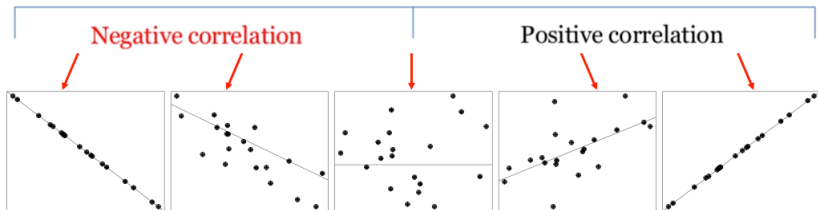
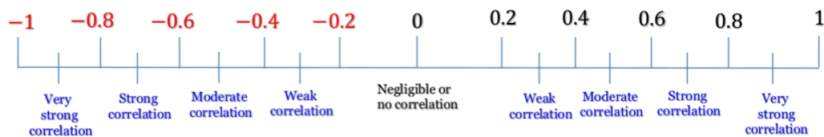
(Pearson's) Coefficient of Correlation

- r is the basic quantity in correlation analysis

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum(X_i - \bar{X})^2][\sum(Y_i - \bar{Y})^2]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

- r takes values in $[-1,1]$, the \pm signs are used for positive and negative linear correlations, respectively.
- r measures strength & direction of linear relationship
 - Value of -1 or +1 indicates a perfect positive or negative correlation, all data points all lie exactly on a straight line
 - Value of 0.0 indicates no linear correlation
 - Positive values indicate a direct relationship, X and Y move in the same directions.
 - Negative values indicate an inverse relationship, X and Y move in opposite directions.
- Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

Interpretation of Pearson's correlation coefficient



- $0.2 \leq |r| \leq 0.4$: weak correlation.
- $0.4 < |r| \leq 0.6$: moderate correlation.
- $0.6 < |r| \leq 0.8$: strong correlation.
- $|r| > 0.8$: very strong correlation.

Coefficient of Determination

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- The measure R^2 is called the **coefficient of determination**.
- R^2 is a proportion. Since $0 \leq SSE \leq SSTO$, it follows that $0 \leq R^2 \leq 1$
- R^2 measures strength of linear relationship.
 - $R^2 = 1$, all data points fall perfectly on the regression line and X accounts for all of the variation in Y.
 - $R^2 = 0$, the estimated regression line is perfectly horizontal. X accounts none of the variation in Y.
- Interpretation of R^2 :
 - " $R^2 \times 100$ percent of the variation in Y is reduced by taking into account predictor X"
 - " $R^2 \times 100$ percent of the variation in Y is 'explained by' the variation in predictor X"

Exercise: show $R^2 = r^2$ under a simple linear regression model.

Hints:

$$R^2 = \frac{\sum_{i=1}^n (b_0 + b_1 X_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$
$$r_{X,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

where

$$S_{YY} = SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$$

From R^2 to get r

$$r = \pm\sqrt{R^2} = \text{sign}(b_1)\sqrt{R^2}$$

- If b_1 is negative, then r takes a negative sign.
- If b_1 is positive, then r takes a positive sign.

$$r = b_1 \sqrt{\frac{S_{XX}}{S_{YY}}}$$

- The estimated slope b_1 of the regression line and the correlation coefficient r always share the same sign.
- If the estimated slope b_1 of the regression line is 0, then the correlation coefficient r must also be 0.

Limitations of R^2

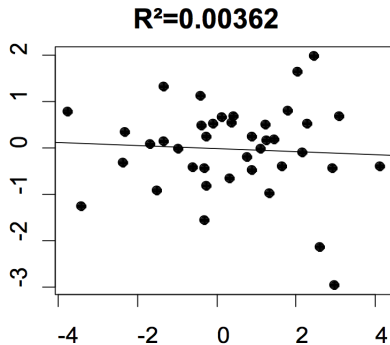
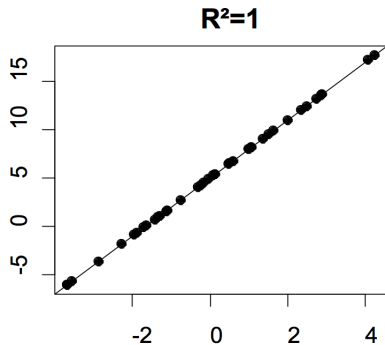
- R^2 is frequently used for assessing and comparing model fits, but
 - A high R^2 does not necessarily imply that you can make useful predictions.
 - A high R^2 does not necessarily imply that the estimated line is a good fit.
 - A low R^2 does not necessarily imply that X and Y are not related or independent.
- R^2 measures degree of *linear* association but does not measure the evidence of a linear relationship between X and Y.
- R^2 usually can be made larger by including a larger number of predictors. (More predictors, MSE goes down, SSTO remains unchanged, so R^2 goes up)
- Adjusted R^2 , R_a^2 (p=number of predictors in the model+1, +1 for β_0)

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{n-1}{n-p} \frac{SSE}{SSTO} = 1 - (n-1) \frac{MSE}{SSTO}$$

$$R_a^2 = 1 - \frac{(1 - R^2)(n-1)}{n-p}$$

Interpretation of R^2

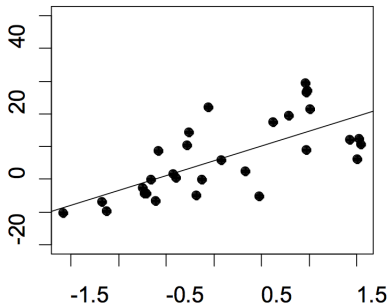
- R^2 measures degree of *linear* association between X and Y.



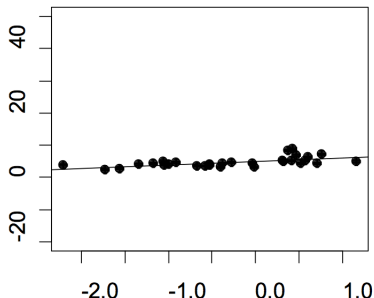
Interpretation of R^2

- R^2 measures only a relative reduction from SSTO.
- R^2 might be large but MSE may still be too large for inference to be useful in prediction.
- R^2 might be small but MSE may still be small which is useful in prediction

$R^2=0.4808$, $MSE=8.531$



$R^2=0.3711$, $MSE=1.219$



Interpretation of R^2

- R^2 measures degree of *linear* association.
- R^2 might be large or small if the true regression association between X and Y is curvilinear.

