STAC67: Regression Analysis

Lecture 2

Sohee Kang

Jan. 14, 2021

Review

- Covariance and Correlation Coefficient

Suppose we have observations on n subjects consisting of a **dependent** or **response variable** Y and an **independent** or **explanatory variable** X.

- Measure both direction and strength of the relationship between Y and X.

Obs	Y	X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Covariance and Correlation

$\Gamma \int_R \int_R (x-\mu_x)(y-\mu_y)\, f(x,y)\, dx\, dy$

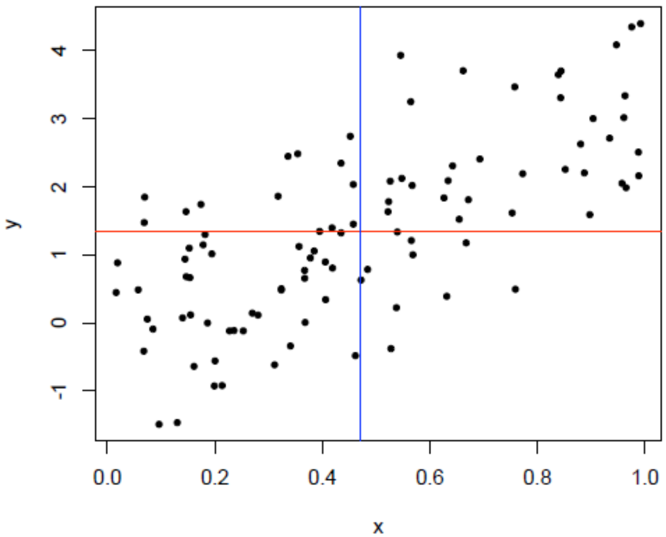
Def. $Cov(X, Y) = E((X - \mu_x)(Y - \mu_y))$, where $\mu_x = E(X)$, $\mu_y = E(Y)$

$Z_x = \frac{X-\mu_x}{\sqrt{Var(X)}}, \quad Z_y = \frac{Y-\mu_y}{\sqrt{Var(Y)}}$

$$Cov(Z_x, Z_y) = \rho_{xy} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

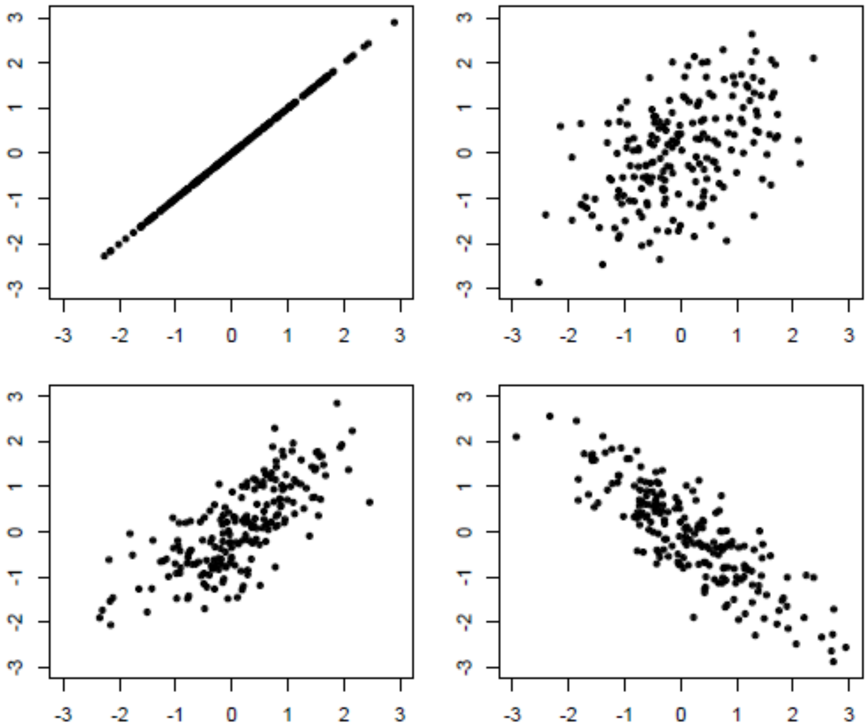
- $-1 \leq \rho_{xy} \leq 1$
- When the relationship is perfectly linear then $|\rho| = 1$.
- if two variables are independent then $\rho = 0$. (Note: the inverse does not hold)

Sample Covariance and Correlation



$$Cor(Y, X) = Cov(Z_y, Z_x) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s_y}\right) \left(\frac{x_i - \bar{x}}{s_x}\right)$$

Correlation



Question: what are main differences between correlation and regression model?

Test for Population correlation

When $\rho = 0$, and the joint distribution of (X, Y) is bivariate normal, and it can be shown that:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a student's t distiruibution with $n-2$ degress of freedom

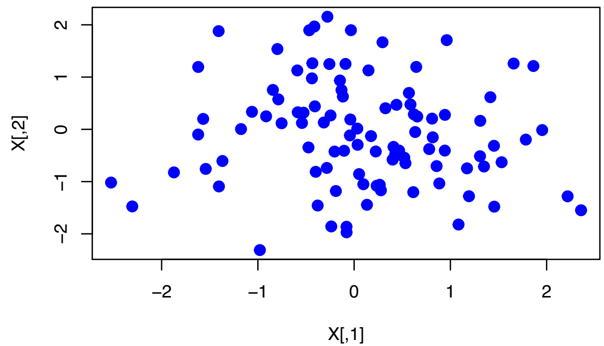
$$H_0 : \rho = 0 \qquad vs \qquad H_1 : \rho \neq 0$$

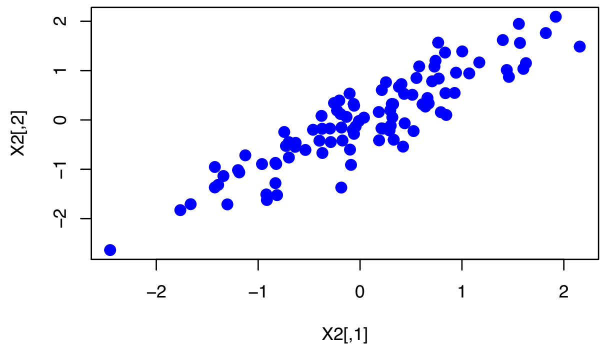
- Testing Procedure

- Calculating the observed value of t (call this t_{obs})
- Compute the p-value for the test

Simulation

```
par(mfrow=c(2,2))
library(mvtnorm)

sigma.1 = matrix(c(1, 0, 0, 1), ncol=2)
sigma.2 = matrix(c(1, 0.9, 0.9, 1), ncol=2)

X = rmvnorm(100, mean=c(0,0), sigma.1)
plot(X, pch=20, cex=2, col="blue")

X2 = rmvnorm(100, mean=c(0,0), sigma.2)

plot(X2, pch=20, cex=2, col="blue")
```





Simulation

```r
x = X2[,1]
y = X2[, 2]
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 20.733, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8580898 0.9333868
## sample estimates:
##       cor
## 0.9024115
```

```r
r= cor(x, y)
t = r*sqrt(98)/sqrt(1-r^2)
t
```

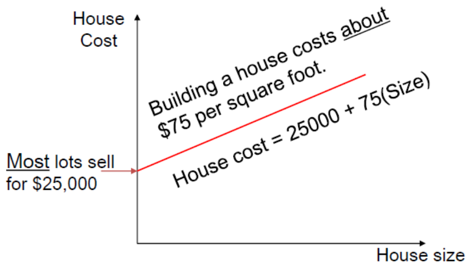
```
## [1] 20.73319
```

Relationship between variables

What factor or variable affects the price of house?

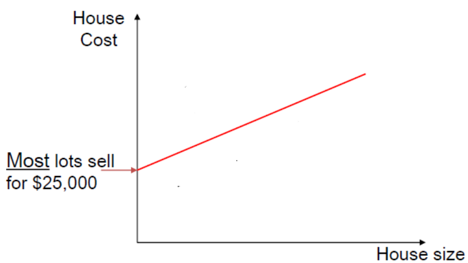
- Relation of the form

1. Mathematical Relation: $Y = f(X)$, where X, Y are variables and f is a function



2. Statistical Relation:

$$Y = f(X) + \epsilon$$



Data Collection for regression analysis

- **Observational study**
  - Investigator has no control over the explanatory variables (X)
  - Limitation: not adequate for cause-and-effect. A strong association does not necessarily means a cause-and-effect relationship
- **Experiment**
  - Investigator exercises control over the explanatory variables (X) through random assignment
  - Random assignment balances out effect of other variables that might affect Y
  - Gold standard for cause-and-effect conclusions

The Regression Process

1. The researcher must clearly define the question(s) of interest in the study
2. The response variable Y must be decided on, based on the question of interest.
3. A set of potentially relevant covariates, which can be measured, needs to be defined.
4. Data is collected.
5. Model Specification.
6. Decide on a method for fitting the specified model
7. Fit the model - typically using software such as R
8. Examine the fitted model for violations of assumptions.
9. Conduct hypothesis testing for questions of interest.
10. Report the results from statistical inference.

Background Review: distributions

RVs

- using capital letters
- observed values denoted using Lower letters
- each RV has a distribution

Distributions

- has density func $f(x)$

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
3. $P(X \in A) = \int_A f(x)\,dx$ $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$

Background Review: distributions

Random Vectors

- two RVs have joint distribution, $f(x,y)$
- Marginal PDF : $\int_{-\infty}^{\infty} f(x,y)\,dy = f(x)$
- Conditional PdF : $\dfrac{f(x,y)}{f(y)} = f(x \mid y)$
- Independence : $f(x,y) = f(x) \cdot f(y)$

Background Review: distributions

Statistical Inference

- With random sample $\{x_1, x_2 \ldots x_n\}$ that's i.i.d.
  - they have Likelihood fnc. $f(x;\theta)$
  - θ is parameter
- Inference of θ
  1. Estimation of θ
  2. Confidence interval of θ
  3. Hypothesis testing about θ taking a certain value

Simple Linear Regression

Suppose we have n observed pairs $(X_i, Y_i), i = 1, \ldots, n$.

Assumptions

1. A linear relationship

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

   - Y_i is the value of the reponse variable in the *ith* trial.
   - X_i is a known constant, namely, the value of the preditor variable in *ith* trial.
   - β_0, β_1: regression model cofficients (population parameters)
     β_0: intercept
     β_1: slope

2. ϵ_i are random errors that zero mean, $E(\epsilon_i) = 0$, with common variance, $Var(\epsilon_i) = \sigma^2$, and pairwise independent.

$$Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

Important Features

- Simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$E(y) = B_0 + B_1 X$

Regression Equation

1. The response variable Y is a sum of two terms:

   ① Constant term : $B_0 + B_1 X_i$

   ② Random term: ε_i

2. $E(Y_i) = \beta_0 + \beta_1 X_i$, where $E(Y_i)$ is a shortcut for $E(Y_i|X_i)$, the mean of Y when $X = X_i$

   $E(Y_i) = E(B_0 + B_1 X_i + \varepsilon_i) = B_0 + B_1 X_i$

3. $Var(Y_i) = \sigma^2$, where $Var(Y_i)$ is a shortcut for $Var(Y_i|X_i)$, the variance of Y when $X = X_i$

   $Var(Y_i) = Var(B_0 + B_1 X_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2$

4. The outcomes Y_i are pairwise independent because ϵ_i are pairwise independent.

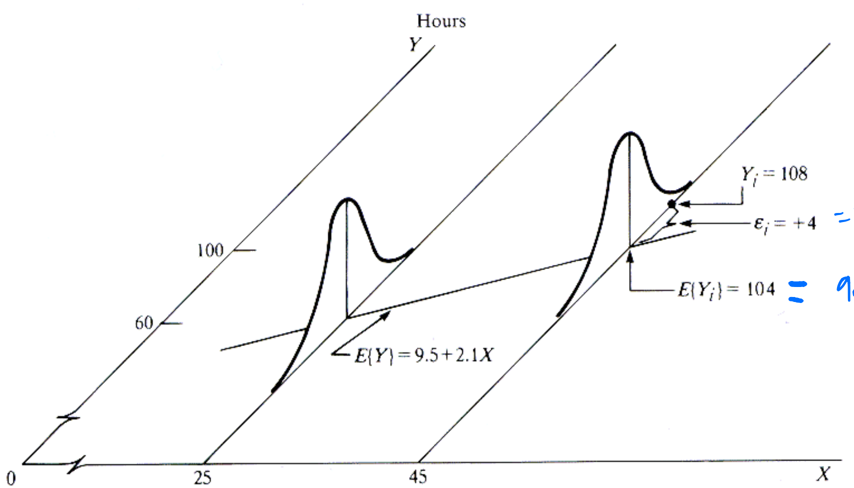
   $Cov(Y_i, Y_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$

Important Features

Example) Y = the time required to prepare for the bid, X =the number of bids requested

- Regression function: $E(Y) = 9.5 + 2.1X$
- $\beta_1 = 2.1$ indicates: The preparation of the additional bid leads to an increase in the mean of Y of 2.1 hours
- $\beta_0 = 9.5$ indicates: $E(Y) = 9.5$, when $X = 0$

FIGURE 1.6 **Illustration of Simple Linear Regression Model (1.1).**



$= Y_i - E(Y_i)$

$= 9.5 + 2.1 \cdot 45$

Exercise

- The regression model applies with $\beta_0 = 100, \quad \beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for $X = 5$.

$Y_i = 100 + 20 X_i + \varepsilon_i = 200 + \varepsilon_i$

a. Can you state the exact probability that Y will fall between 195 and 205? Explain. No, since we didn't define the distribution of error.

b. If the normal error regression model is applicable, can you now state the exact probability that Y will fall betwen 195 and 205? If so, state it.

$\varepsilon_i \sim N(0, 25)$

$Y|X=5 \sim N(200, 25)$

$P(195 < Y < 205) = P(-1 < Z < 1)$

$\simeq 0.68$

Simple linear model with normal errors

- The random errors are sometimes assumed to be normally distributed.
- Simple linear model with normal errors:

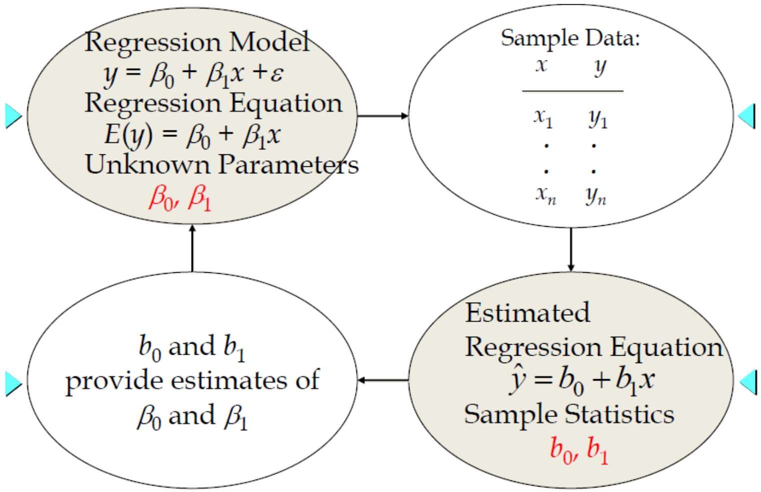
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

,where ϵ_i are independent and identically distributed (i.i.d) with normal distribution with mean 0 and variance σ^2.

- In terms of Y, this means that the conditional distribution

$$Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

Parameter Estimation



- Maximum Likelihood Estimation

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

- Likelihood Function:

$$L(\beta_0, \beta_1, \sigma^2) = \Pi_{i=1}^{n} f(y_i; \beta_0, \beta_1, \sigma^2)$$

Parameter Estimation

Log-likelihood Function:

$$\ln L(\beta_0, \beta_1, \sigma^2) = K - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Maximizing Likelihood function w.r.t β_0, β_1 is equivalent to minimizing,

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

.

Find m.l.e for β_0, β_1, and σ^2.

Parameter Estimation

- Method of Least Squres

Simple linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- **Goal**: Find the best estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ given the data.

- What does it mean, "best"?
- Least squares: best by criterion

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Find the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the criterion Q.
    1. Write the normal equations (derivatives of Q set to 0)
    2. Find the solution of the normal equations

Least Squre Estimators

Least Square Estimators

- Other Criteria

Why square the residuals?

we could use least absolute deviations estimates, minimizing

$$Q_1(\beta_0, \beta_1) = \sum_{i=1}^{n} |(y_i - \beta_0 - \beta_1 x_i)|$$

- **Convenience**
- **Optimality**

Gauss-Markov Theorem

- Theorem 1

Consider the simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

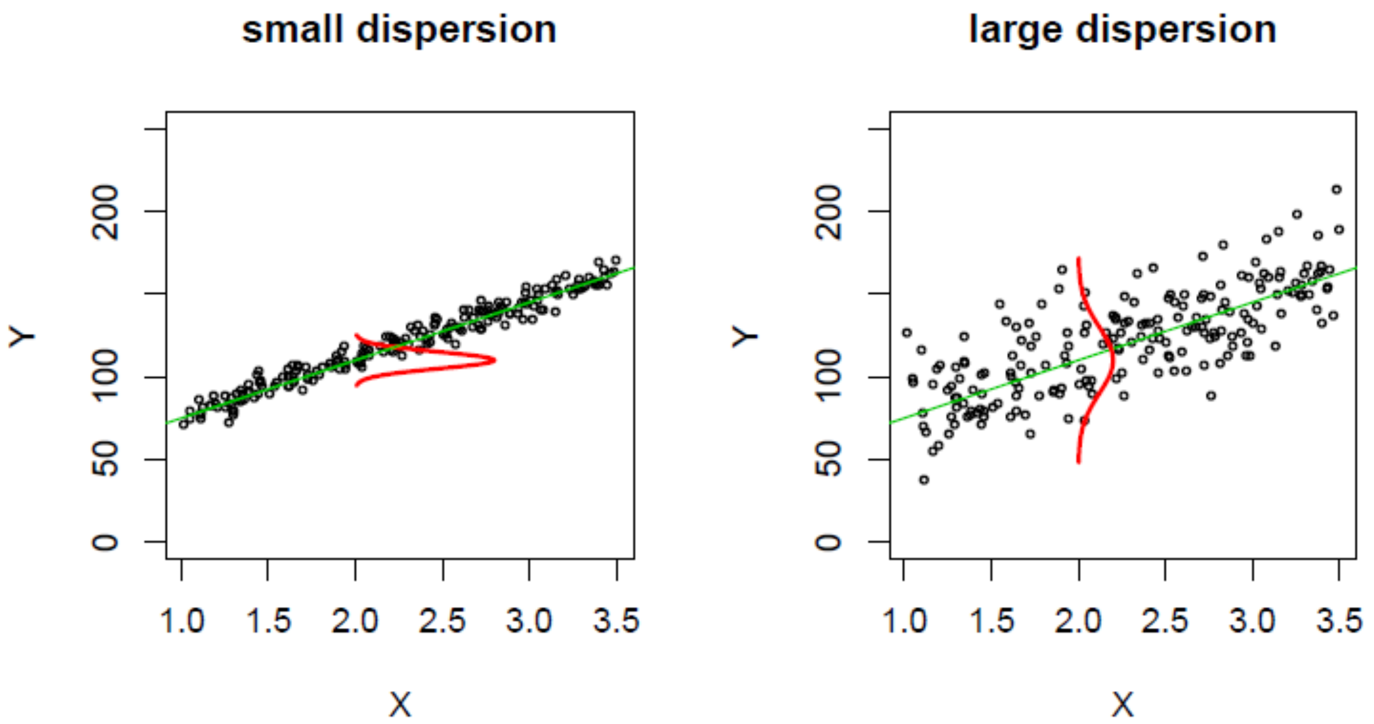
Suppose that the following assumptions (called Gauss-Markov assumptions) concerning the random errors are satisfied:

- Mean zero: $E(\epsilon_i) = 0$
- Constant variance: $Var(\epsilon_i) = \sigma^2$
- Uncorrelated: $Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$

Then the least squre estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators (BLUE).

The interpretation of σ^2

- The variance, σ^2 controls the disperson of Y around $\beta_0 + \beta_1 X$



Fitted values and Residuals

- **Regression equation** or **fitted regression line**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

, where $\hat{Y}$ is the estimated mean of the response variable at level X of the explanatory.

- For each observation, we can compute the fitted value:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n$$

- The vertical distance from the observed y_i to the fitted value is called: residual

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \ldots, n$$

The residuals can be thought of as predicted (observed) valus of the unknown error, $\epsilon_1, \ldots, \epsilon_n$.