Inference in Regression Analysis

Paweł Polak

September 18, 2017

Linear Regression Models - Lecture 3

Content: ALRM Book Chapter 2 (Sec. 2.7-2.11)

- ANOVA
- General linear test approach.
- Descriptive measures of linear association between X and Y.
- Normal correlations models.

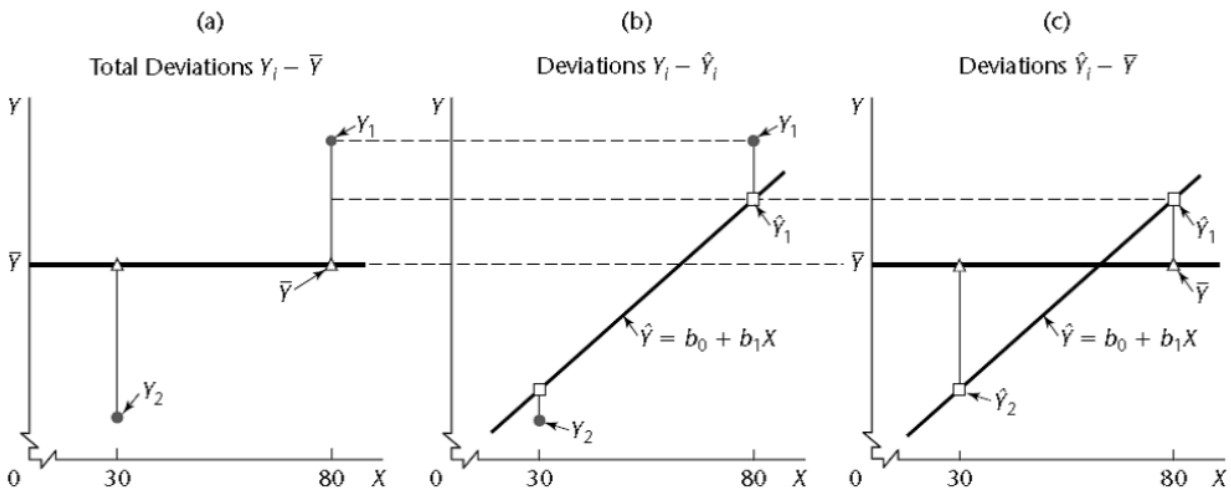
ANOVA

- ANOVA is an inference in linear regression models from the perspective of analysis of variance.
- In Simple Linear Regression it does not add anything new.
- It will come to its own in the multiple regreession models and other types of linear statistical models.

Partitioning of Total Deviations

Decomposition of Total Deviations:

$$\underbrace{Y_i - \bar{Y}}_{\text{total deviation}} = \underbrace{Y_i - \widehat{Y}_i}_{\text{deviation from the fitted regreesion line}} + \underbrace{\widehat{Y}_i - \bar{Y}}_{\text{deviation of fitted regression line from } \bar{Y}}$$



Partitioning of Total Deviations

- The measure of total variation is denoted by

$$SSTO = \sum_{i=1}^{N} (Y_i - \bar{Y})^2.$$

- $SSTO$ stands for total sum of squares.
- If all $Y_i^{'}s$ are the same, $SSTO = 0$.
- The greater the variation of the $Y_i^{'}s$ the greater $SSTO$.

Variation after predictor effect

- The measure of variation of the $Y_i^{'}s$ that is still present when the predictor variable X is taken into account is the sum of the squared deviations

$$SSE = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

- SSE denotes error sum of squares

Regression Sum of Squares

- The difference between $SSTO$ and SSE is SSR

$$SSR = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$$

- SSR stands for regression sum of squares

Remarkable Property

- Recall:

$$Y_i - \bar{Y} = \left(\widehat{Y}_i - \bar{Y}\right) + \left(Y_i - \widehat{Y}_i\right).$$

- But the sums of the same deviations squared has the same property!

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

or

$$SSTO = SSR + SSE$$

Proof: SSTO=SSR+SSE

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2$$

$$= \sum_{i=1}^{N}[(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)]$$

$$= \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

but

$$\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum_{i=1}^{N}\hat{Y}_i(Y_i - \hat{Y}_i) - \sum_{i=1}^{N}\bar{Y}(Y_i - \hat{Y}_i) = 0$$

by properties previously demonstrated, namely

$$\sum_{i=1}^{N}\hat{Y}_i e_i = 0 \text{ and } \sum_{i=1}^{N} e_i = 0.$$

Breakdown of Degrees of Freedom

- $SSTO = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$
  - 1 linear constraint due to the calculation and inclusion of the mean
    - N-1 degrees of freedom
- $SSE = \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$
  - 2 linear constraints arising from the estimation of β_1 and β_0
    - N-2 degrees of freedom
- $SSR = \sum_{i=1}^{N}(\widehat{Y}_i - \bar{Y})^2$
  - Two degrees of freedom in the regression parameters, one is lost due to linear constraint
    - 1 degree of freedom

Mean Squares

A sum of squares divided by its associated degrees of freedom is called a mean square

The regression mean square is

$$MSR = \frac{SSR}{1} = SSR$$

The mean square error is

$$MSE = \frac{SSE}{N-2}$$

ANOVA table for simple lin. regression

The ANOVA approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y.

Source of Variation	Sum of Squares (SS)	df	Mean Squares (MS)	Expected MS $\mathbb{E}$(MS)
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	MSR = SSR/1	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$N-2$	$MSE = SSE/(N-2)$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$N-1$		

- In order to make inference based on the analysis of variance approach, we need to know the expected value of each of the mean squares (as it is given in the last column in the table above).

$\mathbb{E}\{MSE\} = \sigma^2$

- Last time we said that

Theorem

For the normal error regression model, $\frac{SSE}{\sigma^2}$ *is distributed as* χ^2 *with* $N-2$ *degrees of freedom*

$$SSE/\sigma^2 \sim \chi^2(N-2)$$

and is independent of both b_0 *and* b_1.

- That means that $\mathbb{E}\{SSE/\sigma^2\} = N - 2$
- And thus that $\mathbb{E}\{SSE/(N-2)\} = \mathbb{E}\{MSE\} = \sigma^2$, i.e., MSE is an unbiased estimator of σ^2.

$\mathbb{E}\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{N}(X_i - \bar{X})^2$

- To begin, we take an alternative but equivalent form for SSR (see below)

$$SSR = b_1^2 \sum_{i=1}^{N}(X_i - \bar{X}^2)$$

Proof:

$$\begin{aligned} SSR = \sum_{i=1}^{N}\left(\widehat{Y}_i - \bar{Y}\right)^2 &= \sum_{i=1}^{N}\left(b_0 + b_1 X_i - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{N}\left[(\bar{Y} - b_1\bar{X}) + b_1 X_i - \bar{Y}\right]^2 \\ &= \sum_{i=1}^{N}\left[b_1 X_i - b_1\bar{X}\right]^2 \\ &= b_1^2 \sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 . \end{aligned}$$

- And note that, by definition of variance we can write

$$\sigma^2\{b_1\} = E\{b_1^2\} - (E\{b_1\})^2.$$

$\mathbb{E}\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{N}(X_i - \bar{X})^2$

- But we know that b_1 is an unbiased estimator of β_1 so $\mathbb{E}\{b_1\} = \beta_1$.
- We also know (from previous lectures) that

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

- So we can rearrange terms and plug in

$$\sigma^2\{b_1\} = \mathbb{E}\{b_1^2\} - (\mathbb{E}\{b_1\})^2.$$

$$\mathbb{E}\{b_1^2\} = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} + \beta_1^2.$$

$\mathbb{E}\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^N (X_i - \bar{X})^2$

- From the previous slide

$$\mathbb{E}\{b_1^2\} = \frac{\sigma^2}{\sum_{i=1}^N (X_i - \bar{X})^2} + \beta_1^2$$

- Combining we get the result

$$\mathbb{E}\{MSR\} = \mathbb{E}\{SSR/1\} = \mathbb{E}\{b_1^2\} \sum_{i=1}^N (X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum_{i=1}^N (X_i - \bar{X})^2$$

Comments and Intuition

- The mean of the sampling distribution of MSE is σ^2 regardless of whether X and Y are linearly related (i.e. whether $\beta_1 = 0$)
- The mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$
  - When $\beta_1 = 0$ the sampling distributions of MSR and MSE tend to be the same.
  - Hence, we can use the ratio of MSR and MSE as a test statistic.

F Test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

ANOVA provides a battery of useful tests. For example, ANOVA provides an easy test for

Two-sided test

$$H_0 : \beta_1 = 0 \quad \text{v.s.} \quad H_a : \beta_1 \neq 0$$

Test statistic from before

$$t^* = \frac{b_1 - 0}{s\{b_1\}}$$

ANOVA test statistic

$$F^* = \frac{MSR}{MSE}$$

Sampling distribution of F^*

- The sampling distribution of F^* when $H_0 : \beta_1 = 0$ holds can be derived from Cochran's theorem
- Cochran's theorem:

Theorem

If all N observations Y_i come from the same normal distribution with mean μ and variance σ^2 , and SSTO is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if

$$\sum_{r=1}^{k} df_r = N - 1$$

The F Test

We have decomposed SSTO into two sums of squares SSR and SSE and their degrees of freedom are additive, hence, by Cochran's theorem:

If $\beta_1 = 0$ so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables.

F^* Test Statistic

- F^* can be written as follows

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2}{1}}{\frac{SSE/\sigma^2}{N-2}}$$

- But by Cochran' s theorem, we have when H_0 holds

$$F^* \sim \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(N-2)}{N-2}}$$

F Distribution

Definition

The F distribution is the ratio of two independent χ^2 variables normalized by their corresponding degrees of freedom.

Hence, the test statistic F^* follows the distribution

$$F^* \sim F(1, N-2)$$

Hypothesis Test Decision Rule

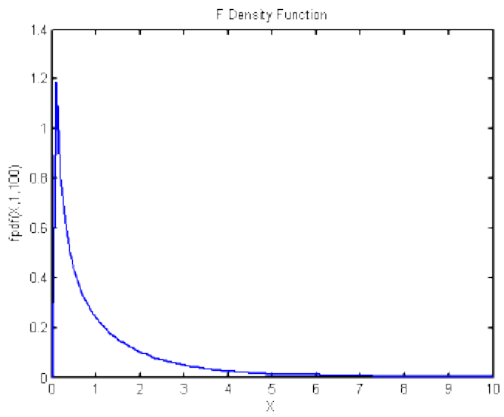
Since F^* is distributed as $F(1, N-2)$ when H_0 holds, the decision rule to follow when the risk of a Type I error is to be controlled at α is:

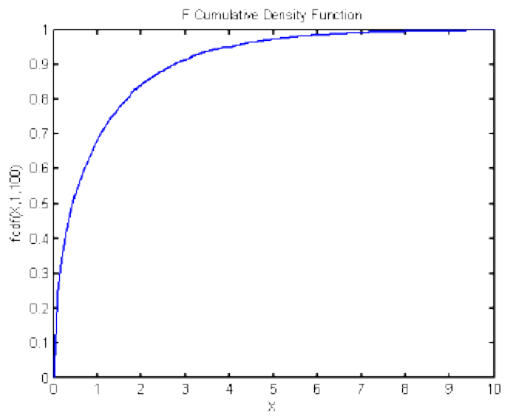
$$\text{If } F^* \leq F(1-\alpha; 1; N-2), \text{ then conclude } H_0$$

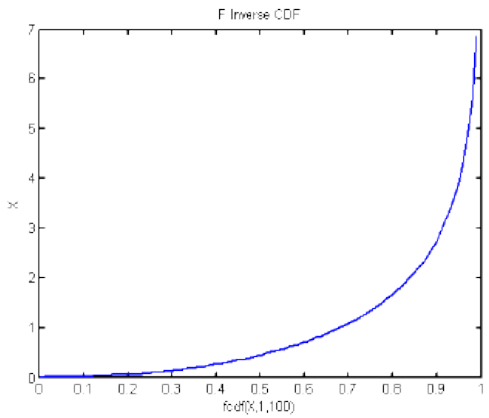
$$\text{If } F^* > F(1-\alpha; 1; N-2), \text{ then conclude } H_1$$

F distribution

- PDF, CDF, Inverse CDF of F distribution



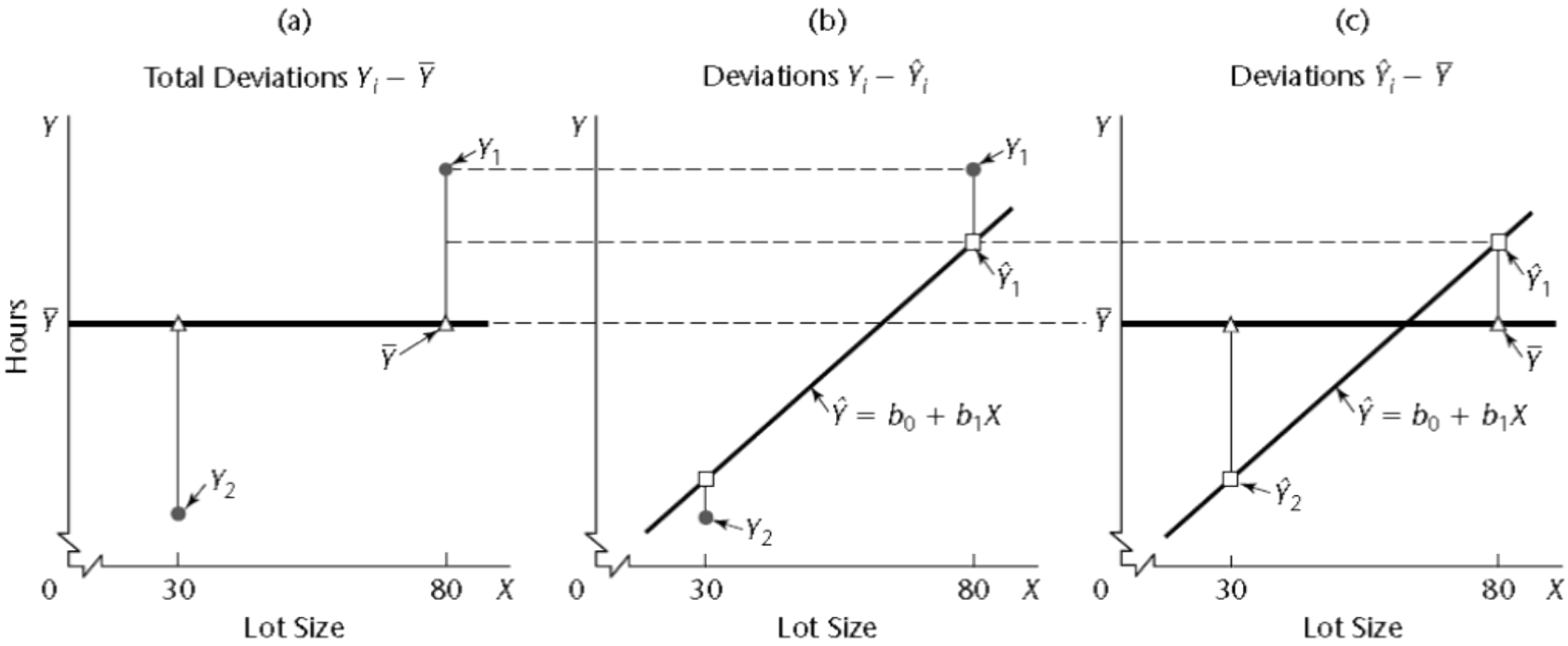




- Note, MSR/MSE must be big in order to reject hypothesis.

Partitioning of Total Deviations

Does this make sense? When is MSR/MSE big?



Equivalence of F test and two-sided t test

$$
\begin{aligned}
F^* &= \frac{MSR}{MSE} \\
&= \frac{b_1^2 \sum_{i=1}^{N} (X_i - \bar{X})^2}{MSE} \\
&= \frac{b_1^2}{s^2\{b_1\}} \\
&= \left(\frac{b_1}{s\{b_1\}} \right)^2 \\
&= (t^*)^2
\end{aligned}
$$

In addition: $F(1-\alpha; 1; N-2) = t(1-\alpha/2; N-2)^2$.

General Linear Test

- The test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is a simple example of a general linear test.
- The general linear test has three parts
  (1) Full Model
  (2) Reduced Model
  (3) Test Statistic

Full Model Fit

- A full linear model is first fit to the data

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

- Using this model the error sum of squares is obtained, here for example the simple linear model with non-zero slope is the "full" model

$$SSE(F) = \sum_{i=1}^{N} [Y_i - (b_0 + b_1 X_1)] = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = SSE$$

Fit Reduced Model

- One can test the hypothesis that a simpler model is a "better" model via a general linear test (which is really a likelihood ratio test in disguise). For instance, consider a "reduced" model in which the slope is zero (i.e. no relationship between input and output).

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

- The model when H_0 holds is called the reduced or restricted model.

$$Y_i = \beta_0 + \varepsilon_i.$$

- The SSE for the reduced model is obtained

$$SSE(R) = \sum_{i=1}^{N} (Y_i - b_0)^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 = SSTO.$$

Test Statistic

- The idea is to compare the two error sums of squares SSE(F) and SSE(R).
- Because the full model F has more parameters than the reduced model, $SSE(F) \leq SSE(R)$ always
- In the general linear test, the test statistic is

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

  which follows the F distribution when H_0 holds.
- df_R and df_F are those associated with the reduced and full model error sums of squares respectively

General Linear Test: Again $\beta_1 = 0$ vs. $\beta_1 \neq 0$

For testing whether or not $\beta_1 = 0$, we therefore have:

$$SSE(R) = SSTO \qquad SSE(F) = SSE$$
$$df_R = N - 1 \qquad df_F = N - 2$$

So that we obtain:

$$F^* = \frac{SSTO - SSE}{(N-1) - (N-2)} \div \frac{SSE}{N-2} = \frac{SSR}{1} \div \frac{SSE}{N-2} = \frac{MSR}{MSE}$$

which is identical to the analysis of variance test statistic.

R^2 (Coefficient of Determination)

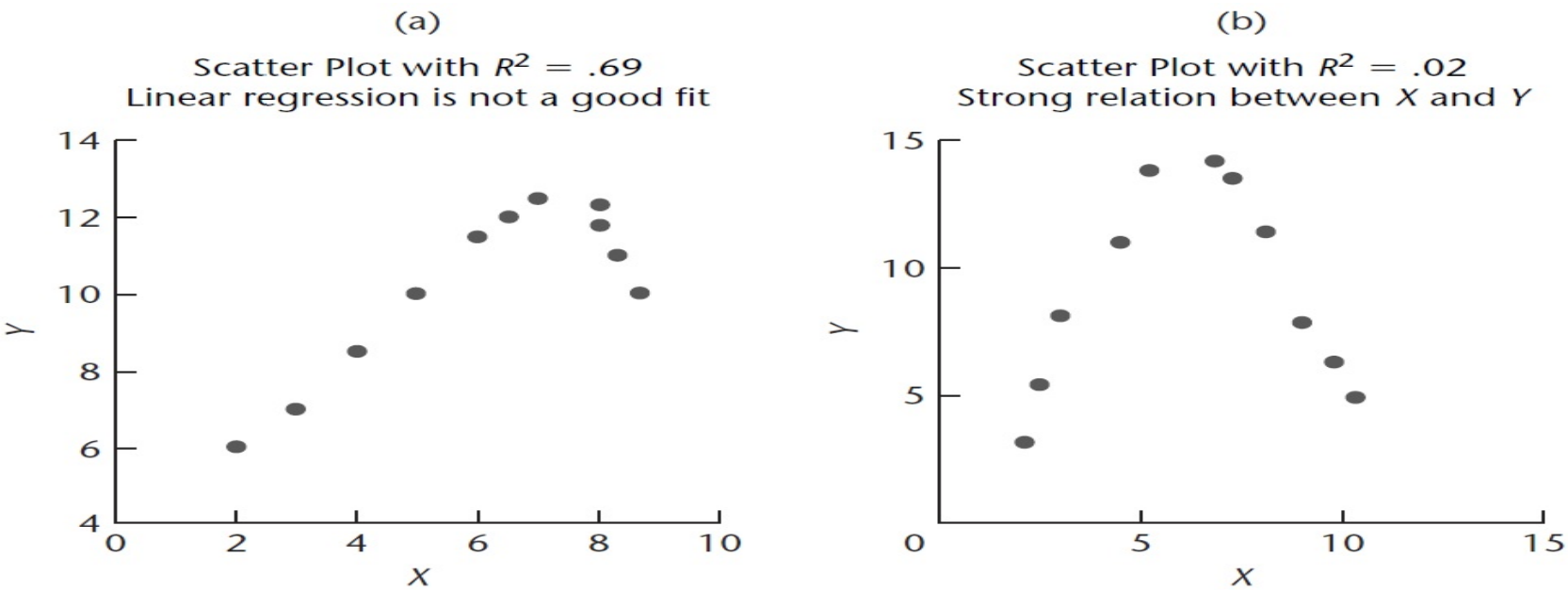
- $SSTO$ measures the variation in the observations Y_i when X is not considered
- SSE measures the variation in the Y_i after a predictor variable X is employed
- A natural measure of the effect of X in reducing variation in Y is to express the reduction in variation ($SSTO - SSE = SSR$) as a proportion of the total variation

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- Note that since $0 \leq SSE \leq SSTO$ then $0 \leq R^2 \leq 1$

Limitations of and misunderstandings about R^2

1. Claim: high R^2 indicates that useful predictions can be made. The prediction interval for a particular input of interest may still be wide even if R^2 is high.
2. Claim: high R^2 means that there is a good linear fit between predictor and output. It can be the case that an approximate (bad) linear fit to a truly curvilinear relationship might result in a high R^2.
3. Claim: low R^2 means that there is no relationship between input and output. Also not true since there can be clear and strong relationships between input and output that are not well explained by a linear functional relationship.



Source: Figure 2.9 in ALRM book.

Coefficient of Correlation

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

where $\text{sign}\,(x) = 1$ if $x > 0$, $\text{sign}\,(x) = -1$ if $x < 0$, and $\text{sign}\,(x) = 0$ if $x = 0$

Range:

$$-1 \leq r \leq 1$$

It is a measure of linear association between Y and X when both X and Y are random.

Some Considerations in Applying Regression Analysis

(1) When one uses the regression analysis to make inferences for the future, one has to be careful if the linear relation (or more generally any causal conditions) between Y and X will be similar to those in existence during the period upon which the regression analysis is based.

(2) In predicting new values of Y, the predictor variable X itself often has to be predicted. One has to see these predictions of Y as conditional on the specific realization of X.

(3) Another caution deals with inferences pertaining to levels of the predictor variable X that fall outside the range of observations.

(4) A statistical test that leads to the conclusion that $\beta_1 \neq 0$ does not establish a cause-and-effect relation between the predictor and response variables. With nonexperimental data both X and Y variables may be simultaneously influenced by other variables not in the regression model. On the other hand, the existence of a regression relation in controlled experiment is often a good evidence of a cause-and-effect relation.

Normal Correlation Models

When we cannot control the values of X, then correlation models are more natural than regression models, e.g.,

- relationship between sales of gasoline and sales of auxiliary products;
- relationship between blood pressure and age;
- relationship between two stock returns.

The difference between Regression models and Correlation models:

- Regression models: X values are known constants.
- Correlation models: Both X and Y are random.

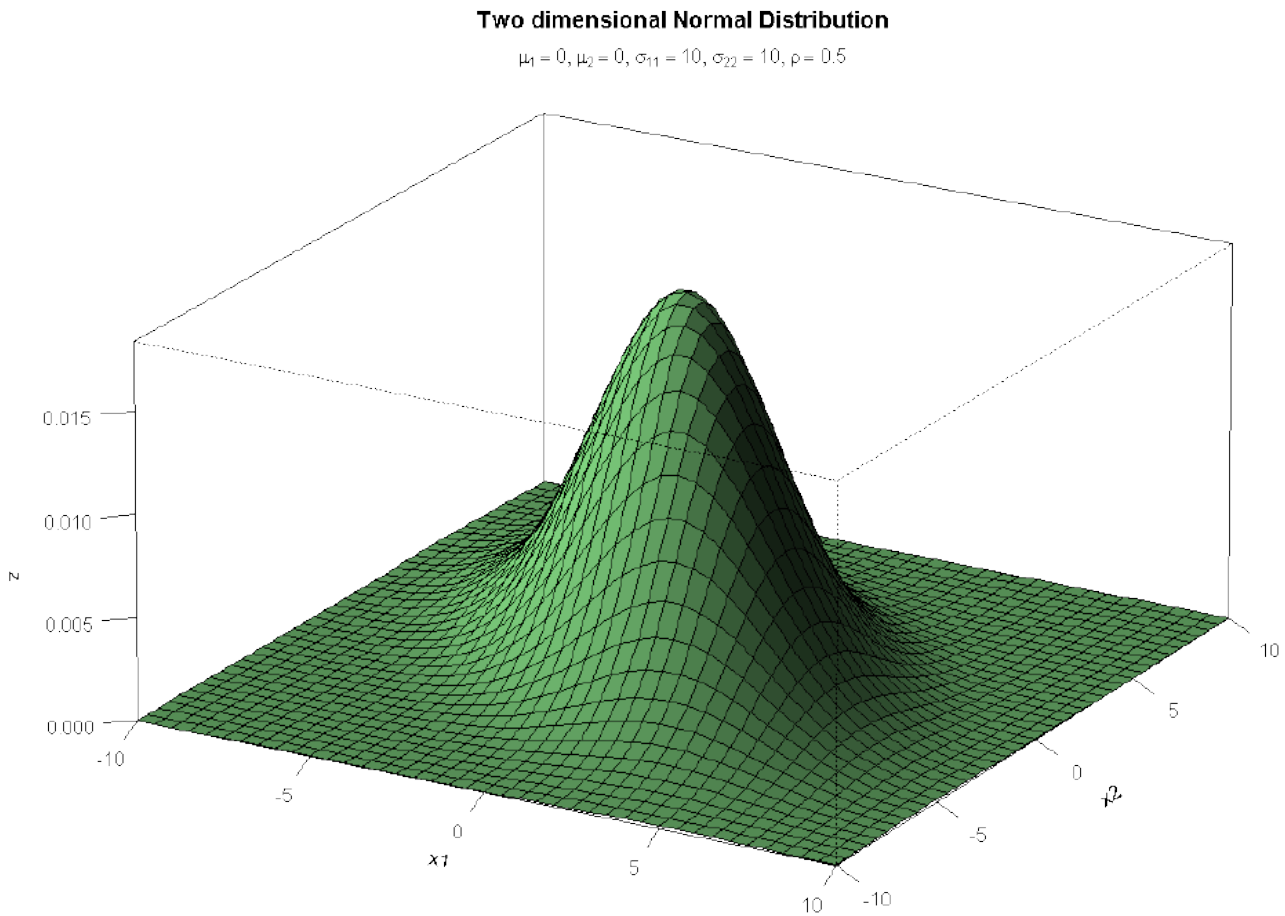
Bivariate Normal Density

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} exp\{-\frac{1}{2(1-\rho_{12}^2)}[(\frac{Y_1-\mu_1}{\sigma_1})^2$$
$$-2\rho_{12}(\frac{Y_1-\mu_1}{\sigma_1})(\frac{Y_2-\mu_2}{\sigma_2}) + (\frac{Y_2-\mu_2}{\sigma_2})^2]\}$$

Parameters: μ_1, μ_2, σ_1, σ_2, ρ_{12}

Here, ρ_{12} is the coefficient of correlation between variable Y_1 and Y_2.

$$\rho_{12} = \rho\{Y_1, Y_2\} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$
$$\sigma_{12} = \sigma\{Y_1, Y_2\} = E\{(Y_1-\mu_1)(Y_2-\mu_2)\}$$



$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\sigma_{11}\,\sigma_{22}\,(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_{11}} - 2\rho\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}}\right]\right\}$$

Marginal Density

Marginal distribution of Y_1 is normal with mean μ_1 and standard deviation σ_1:

$$f_1(Y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} exp[-\frac{1}{2}(\frac{Y_1 - \mu_1}{\sigma_1})^2]$$

How to get it from the bivariate density function?

Conditional Probability Distribution

Theorem

Let $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *be a multivariate normal vector with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*. Consider partitioning* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *into*

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2]' \quad \text{and} \quad \boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12};\ \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{22}]$$

$$\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]'$$

Then $\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2$*, the conditional distribution of the first partition given the second, is* $N\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}\right)$*, with mean*

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right),$$

and covariance matrix

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Conditional Inferences

- One principal use of a bivariate correlation model is to make conditional inferences regarding one variable, given the other variable, e.g.,
  - Suppose Y_1 represents a service station's auxiliary products sales, and Y_2 represents its gasoline sales.
  - To predict sales of auxiliary products Y_1 given that the gasoline sales are $Y_2 = 5,500$ USD we need to use conditional probability distribution from the previous theorem:

$$f(Y_1|Y_2) = \frac{f(Y_1, Y_2)}{f_2(Y_2)} = \frac{1}{\sqrt{2\pi}\sigma_{1|2}} exp[-\frac{1}{2}(\frac{Y_1 - \alpha_{1|2} - \beta_{12} Y_2}{\sigma_{1|2}})^2]$$

where

$\alpha_{1|2} = \mu_1 - \mu_2 \rho_{12} \frac{\sigma_1}{\sigma_2}$ - is an intercept of the line of regression of Y_1 on Y_2,

$\beta_{12} = \rho_{12} \frac{\sigma_1}{\sigma_2}$ - is the slope of this line, and

$\sigma^2_{1|2} = \sigma^2_1(1 - \rho^2_{12})$ - is the conditional variance of the error term.

Equivalence to Normal Error Regression Model

If we want to make conditional inferences about Y_1, given Y_2. Based on the above results it is clear that the normal error regression model is applicable because

- The Y_1 observations are independent.
- The Y_1 observations when Y_2 is considered given or fixed are normally distributed with mean $\mathbb{E}[Y_1 \mid Y_2] = \alpha_{1|2} + \beta_{12} Y_2$, and constant variance $\sigma^2_{1|2}$.

Inferences on Correlation Coefficients

- ρ_{12} provides information about the degree of the linear relationship between the two variables. It is worth to have a good estimator of ρ_{12}.
- Maximum Likelihood Estimator of ρ_{12} (Pearson Product-moment Correlation Coefficient):

$$r_{12} = \frac{\sum_{i=1}^{N}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{[\sum_{i=1}^{N}(Y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^{N}(Y_{i2} - \bar{Y}_2)^2]^{1/2}}$$

Usually biased, but bias is small when N is large.
Test whether $\rho_{12} = 0$

$$H_0 : \rho_{12} = 0$$

$$H_1 : \rho_{12} \neq 0$$

Test Statistics:

$$t^* = \frac{r_{12}\sqrt{N-2}}{\sqrt{1 - r_{12}^2}}$$

If H_0 holds, t^* follows the $t(N-2)$ distribution.

Interval Estimation of ρ_{12}

If the true $\rho_{12} \neq 0$ the sampling distribution of r_{12} is skewed and complicated, therefore, one makes the Fisher z transformation:

$$z' = \frac{1}{2} log_e(\frac{1 + r_{12}}{1 - r_{12}})$$

When N is large ($N > 25$), z' is approximately normally distributed with

$$E\{z'\} = \zeta = \frac{1}{2} log_e(\frac{1 + \rho_{12}}{1 - \rho_{12}})$$

$$\sigma^2\{z'\} = \frac{1}{N - 3}$$

Then we can make interval estimate

$$\frac{z' - \zeta}{\sigma\{z'\}}$$

is approximately standard normal. Then $1 - \alpha$ confidence limits for ζ are

$$z' \pm z(1 - \alpha/2)\sigma\{z'\}$$

Spearman Rank Correlation Coefficient

- Denote the rank of Y_{i1} by R_{i1} and the rank of Y_{i2} by R_{i2}. The ranks are from 1 to N.
- The Spearman Rank Correlation Coefficient r_s is then defined as

$$r_s = \frac{\sum_{i=1}^{N}(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{[\sum_{i=1}^{N}(R_{i1} - \bar{R}_1)^2 \sum_{i=1}^{N}(R_{i2} - \bar{R}_2)^2]^{1/2}}$$

- as before $-1 \leq r_s \leq 1$.
- More robust compared with the Pearson correlation coefficient.