Oakland University

Department of Mathematics and Statistics

Linear Regression

Huong Tran

December 27, 2021

Contents

1 Preliminary: **2**
1.1 Basic Calculations: 2

2 Analysis of Variance (ANOVA): One way **3**
2.1 Model and Distribution Assumption: 4
2.2 The classic ANOVA Hypothesis: 4
2.3 Inferences Regarding Linear Combinations of Means: 5
2.4 The ANOVA F-test: 7

3 Simple Linear Regression: **8**
3.1 Introduction: 8
3.2 Data for Regression Analysis: 9
3.3 Least Squares Model: 10
3.4 A Statistical Solution: 10
3.5 Models and Distribution Assumptions: 11
3.6 Interpretation of β: 12
3.7 Estimation and Prediction: 15
3.8 Matrix Approach: 16
3.9 Multiple regression: 17
3.10 Least square model: 19

4 Appendix: **22**
4.1 t-distribution: 22
4.2 F-distribution: 22

1 Preliminary:

1.1 Basic Calculations:

Definition 1.1 (Joint pdf)

Definition 1.2 (Conditional pdf)

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

Definition 1.3 (Independence) *j X and Y are called independent variables if*

$$f(x,y) = f_X(x)f_Y(y)$$

In this event, we derive that:

$$f(y|x) = f_Y(y)$$

and moreover

$$E\Big(g(x)h(Y)\Big) = Eg(X)Eh(Y)$$

Theorem 1.1.1 *If X and Y are two random variables, then*

$$EX = E\big(E(X|Y)\big)$$

and

$$Var(X) = E\big(Var(X|Y)\big) + Var(E(X|Y))$$

Definition 1.4 (Covariance)

$$Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) = EXY - \mu_X\mu_Y$$

Note that: If X and Y are independent random variables, then $\mathrm{Cov}(X,Y) = 0$

Definition 1.5 (Correlation)

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Theorem 1.1.2 (Sum of normally distributed variables) *Let $X_1, \dots X_n$ be mutually independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$. Then*

$$Z = \sum_{i=1}^{n}\big(a_iX_i + b_i\big) \sim N\left(\sum_{i=1}^{n} a_i\mu_i + b_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

Theorem 1.1.3 (Independence criteria) *$X_1 \dots X_n$ are mutually independent variables if and only if there exists functions $g_i(x_i), i = 1, 2 \dots n$ such that*

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} g_i(x_i)$$

Theorem 1.1.4 *Let $X_1, \ldots, X_n$ be random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

$$E\bar{X} = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n} \qquad ES^2 = \sigma^2$$

Theorem 1.1.5 *If $X_1, \ldots X_n \sim N(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent and:*

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Theorem 1.1.6 (Facts about variables transformation) *The following are useful:*

1. *$Z \sim N(0,1)$ then $Z^2 \sim \chi^2_1$.*
2. *Let $X_1, \ldots X_n$ are random independent variables$X_i \sim \chi^2_{p_i}$, then $\sum_{i=1}^{n} X_i \sim \chi^2_{p_1 + \ldots p_n}$*
3. *If $U \sim N(0,1)$ and $V \sim \chi^2_p$ are independent , then $T = \dfrac{U}{\sqrt{\frac{V}{p}}} \sim t_p$.*
4. *If $U \sim \chi^2_p$ and $V \sim \chi^2_q$, then $\dfrac{U/p}{V/q} \sim F_{p-1,q-1}$*
5. *If $X \sim t_q$ then $X^2 \sim F_{1,q}$*
6. *If $X \sim F_{p,q}$ then $\dfrac{1}{X} \sim F_{q,p}$*
7. *A vector $U \sim N(\mu, \Sigma)$, then $AU \sim N(A\mu, A\Sigma A')$*

2 Analysis of Variance (ANOVA): One way

Basic idea of the ANOVA, that of partitioning variation, is a fundamental idea of experimental statistics. Note, it is not concerned with analyzing the variances, but rather with analyzing the **variation in the means**.

In its simplest form, ANOVA is a method of estimating the **means** of several populations (usually assumed to be normally distributed).

In the oneway ANOVA, we assumed that data Y_{ij} are observed according to the model

$$Y_{ij} = \theta_i + \epsilon_{ij}, i = 1, \ldots k, j = 1 \ldots n$$

where θ_i are unknown and ϵ_{ij} are error random variables.

Example 2.1 (Oneway ANOVA) *The data from oneway ANOVA table of k treatments (k levels) and n observation will look like this:*

Treatments				
1	*2*	*3*	...	*k*
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	
y_{1n}	y_{2n}	y_{3n}	$\cdots$	y_{kn}

Note that the number of observation in each treatment groups are not necessarily equal.

Without loss of generality, we can assume that $E\epsilon_{ij} = 0$, since if not, we can rescale the ϵ_{ij} an absorb the leftover mean into δ_{ij}. Thus it follows that

$$EY_{ij} = \theta_i$$

so θ_i is the **mean** of Y_{ij}, i.e, it is the mean of level i^{th} of the treatment.

2.1 Model and Distribution Assumption:

For the estimation to be done, we assumed that $E\epsilon_{ij} = 0$ and $\text{Var}\epsilon_{ij} < \infty$ for all i, j. For the confidence interval estimation, we need a distributional assumption.

Then, oneway ANOVA assumptions are:

- $E\epsilon_{ij} = 0, \text{Var}\epsilon_{ij} = \sigma_i^2 < \infty$ for all i, j. And $\text{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$, for all i, i', j, j'.
- The ϵ_{ij} are independent and normally distributed (**normal error**).
- $\sigma^2 = \sigma, \forall i$ (Equality of variance, also called **homoscedasticity**)

For short: *Error random variable in each level of the treatment are identically independent distributed, with mean of 0 and equal variance.*

Note that, the assumption of normal distribution is not a must, but it is more convenient for the interval estimation. The assumption of equal variance is quite important. In fact, the robustness of the ANOVA to the assumption of normality depends on how equal the variances are.

2.2 The classic ANOVA Hypothesis:

The classic ANOVA Hypothesis is a test of the null hypothesis:

$$\begin{cases} H_0: & \theta_1 = \theta_2 \cdots = \theta_k \\ H_a: & \theta_i \neq \theta_j, \text{ for some } i, j \end{cases}$$

However, it is not very interesting, when the experimenter is sure that there must be a difference between his treatments. Instead, we usually want to figure out which treatments are better. If H_0 is rejected, we can conclude only that there is *some* difference in the mean of treatments, but we can make no inference as to where this difference might be.

Theorem 2.2.1 *Let* $\theta = \left(\theta_1, \theta_2, \dots \theta_k\right)$ *be arbitrary parameters. Then*

$$\theta_1 = \theta_2 = \cdots = \theta_k \Leftrightarrow \sum_{i=1}^{k} a_i\theta_i = 0$$

as long as $\sum_{i=1}^{k} a_i = 0$*, in this situation,* $\sum_{i=1}^{k} a_i\theta_i$ *is called* ***contrasts****.*

By this theorem, the ANOVA null can be expressed as

$$\begin{cases} H_0: & \sum_{i=1}^{k} a_i\theta_i = 0 \text{ for all } (a_1, \dots, a_k) \text{ such that } \sum_{i=1}^{k} a_i = 0 \\ H_1: & \sum_{i=1}^{k} a_i\theta_i \neq 0 \text{ for some } (a_1, \dots a_k) \text{ such that } \sum_{i=1}^{k} a_i = 0 \end{cases}$$

We derived thinking in a univariate manner.

2.3 Inferences Regarding Linear Combinations of Means:

Using the assumption of ANOVA, we have

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

then

$$Y_{ij} = \theta_i + \epsilon_{ij} \sim N(\theta_i, \sigma^2), \forall i = 1, 2 \dots k, j = 1, 2 \dots n_i$$

Therefore

$$\bar{Y_{i\cdot}} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim N\left(\theta_i, \frac{\sigma^2}{n_i}\right), \forall i = 1, 2 \dots k$$

Then, $\sum_{i=1}^{k} a_i\bar{Y_{i\cdot}}$ is normal, with

$$E\left(\sum_{i=1}^{k} a_i\bar{Y_{i\cdot}}\right) = \sum_{i=1}^{k} a_i\theta_i \text{ and Var}\left(\sum_{i=1}^{k} a_i\bar{Y_{i\cdot}}\right) = \sigma^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}$$

Therefore

$$\frac{\sum_{i=1}^{k} a_i\bar{Y_{i\cdot}} - \sum_{i=1}^{k} a_i\theta_i}{\sqrt{\sigma^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}}} \sim N(0, 1)$$

In fact, we don't know about σ^2, and using an estimator of σ^2 would be easier Using the fact that

$$Y_{ij} \sim (\theta_i, \sigma^2)$$

we derive

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot} \right), \forall i = 1, 2 \dots k$$

is an estimator of σ^2 and

$$\frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}$$

And

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} (n_i - 1)S_i^2 = \sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi_{N-k}, \text{ where } N = \sum_{i=1}^{k} n_i$$

Let

$$S_p^2 = \frac{1}{N - k} \sum_{i=1}^{k} \left(n_i - 1 \right) S_i^2$$

which is a better estimator of σ^2 and also $\frac{(N-k)S_p^2}{\sigma^2} \sim \chi^2_{N-k}$ Thus

$$\frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}} = \frac{\frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i}{\sqrt{\sigma^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}}}{\sqrt{\frac{1}{N-k} \frac{(N-k)S_p^2}{\sigma^2}}} \sim t_{N-k}$$

Now, back to our test

$$\begin{cases} H_0 : & \sum_{i=1}^{k} a_i \theta_i = 0 \\ H_1 : & \sum_{i=1}^{k} a_i \theta_i \neq 0 \end{cases}$$

at level α. Rejection region is

$$\left| \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot}}{\sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}} \right| > t_{N-k,\alpha/2}$$

Example 2.2 *Choosing $a = (1, -1, 0, \dots, 0)$, we are able to test*

$$\begin{cases} H_0 : & \theta_1 = \theta_2 \\ H_1 : & \theta_1 \neq \theta_2 \end{cases}$$

and the rejection region is

$$\left| \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \right| > t_{N-k,\alpha/2}$$

The difference between this test and two-sample t-test is that here we used the information of all treatments, including $3, \ldots, k$ *to estimate* σ^2*. By choosing* a_i *such that* $\sum_{i=1}^{k} a_i = 0$*, we obtain several tests to compare the mean of treatments.*

2.4 The ANOVA F-test:

Recall the ANOVA hypothesis is:

$$\begin{cases} H_0 : & \sum_{i=1}^{k} a_i \theta_i = 0 \text{ for all } (a_1, \ldots, a_k) \in \mathcal{A} \\ H_1 : & \sum_{i=1}^{k} a_i \theta_i \neq 0 \text{ for some } (a_1, \ldots, a_k) \in \mathcal{A} \end{cases}$$

where

$$\mathcal{A} = \left\{ (a_1, \ldots, a_k) : \sum_{i=1}^{k} a_i = 0 \right\}$$

To see this more clearly as a union-intersection test, define

$$\Theta_a = \left\{ \theta = (\theta_1, \ldots, \theta_k) : \sum_{i=1}^{k} \theta_i a_i = 0 \right\}$$

Then, 2.2.1 derive that

$$\theta_1 = \cdots = \theta_k \Leftrightarrow \theta \in \bigcap_{a \in \mathcal{A}} \Theta_a$$

Then, the ANOVA test can be written as

$$H_0 : \theta \in \bigcap_{a \in \mathcal{A}} \Theta_a$$

we would reject H_0 if we can reject

$$\begin{cases} H_{0a} : \theta \in \Theta_a \\ H_{1a} : \theta \notin \Theta_a \end{cases} \quad \text{for any } a \in \mathcal{A}$$

To test H_{0a}, we use the test statistic

$$T_a = \left| \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}} \right|$$

with rejection $T_a > k$, for some constant k. We would reject for any a, which means that, if we could reject for $\sup_a T_a > k$. Therefore, find the $\sup_a T_a$ would be helpful for this union-intersection test. And maximizing T_a is equivalent to maximizing T_a^2:

$$T_a^2 = \frac{\left(\sum_{i=1}^k a_i \left(\bar{Y}_{i\cdot} - \theta_i\right)\right)^2}{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}$$

And we know that $T_a \sim t_{N-k}$, then $T_a^2 \sim F_{1,N-k}$

Theorem 2.4.1

$$\sup_{a \in \mathcal{A}} T_a^2 = \frac{\sum_{i=1}^k n_i \left(\left(\bar{Y}_{i\cdot} - \bar{\bar{Y}}\right) - \left(\theta_i - \bar{\theta}\right) \right)^2}{S_p^2}$$

Furthermore, ANOVA assumption derive that

$$\sup_{a \in \mathcal{A}} T_a^2 \sim (k-1) F_{k-1,N-k}$$

Theorem 2.4.1 deduce the rejection region of ANOVA test is

$$F = \frac{\frac{\sum_{i=1}^k n_i \left(\left(\bar{Y}_{i\cdot} - \bar{\bar{Y}}\right) \right)^2}{k-1}}{S_p^2} > F_{k-1,N-k,\alpha}$$

3 Simple Linear Regression:

3.1 Introduction:

In simple linear regression we have a relationship of the form:

$$Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, \dots n \tag{1}$$

where Y_i is a random variable (response) and x_i is another observable variable (predictor). α is called *intercept* while β is the *slope* of the regression and ϵ_i is supposed to be random with $E\epsilon_i = 0$.We have

$$EY_i = \alpha + \beta x_i \tag{2}$$

To keep straight the fact that inference about the relationship between Y_i and x_i assume knowledge of x_i, we could write 2 as

$$E(Y_i|x_i) = \alpha + \beta x_i \tag{3}$$

Note that, the term *linear regression* refers to a specification that is *linear in the parameters*. Thus, the specifications $E(Y_i|x_i) = \alpha + \beta x_i^2$ is also *linear regression* while $E(Y_i|x_i) = \alpha + \beta^2 x_i$ is not linear regression.

When we do a regression analysis, we are investigating the relationship between the response Y and the predictor X. There are 2 steps to do this analysis:

- Step 1: is data-oriented, in which we try to observe the data. Suppose we have n pairs of data points $(x_1, y_1), \ldots, (x_n, y_n)$

  1. The sample means are:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4}$$

  2. The sum of squares are:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad\qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5}$$

  3. The sum of cross-products is:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{6}$$

- Step 2: is statistical, which which we try to make inference about the relationship of response and predictor. To do this, we need some assumptions about the population. Different assumptions will lead to different models.

3.2 Data for Regression Analysis:

1. **Observational Data:** obtained from **nonexperimental studies**, which do not control the response or predictor variables. A major limitation of observational data is that they often do not provide adequate information about cause-and-effect relationships.

   For example, a positive relation between age of employee and number of days of illness in the company personnel example may not imply that the number of days of illness is the direct result of age.

2. **Experimental Data:** data was collected with the values of x's set in advance of observing y (i.e value of x are controlled).

   The procedure for selecting sample data with the x's set in advance is called the **design of the experiment**. The statistical procedure for comparing the population means is called an **analysis of variance**.

3. **Completely Randomized Design:**

3.3 Least Squares Model:

When we fit x_i into our model, $\alpha + \beta x_i$ is the predicted value for y_i for $i = 1, \dots n$, the *residual sum of squares (SSE)* is defined as follow

$$\text{SSE} = \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2 \tag{7}$$

In fact, we want RSS to be as small as possible. The *least squares estimates* of α and β is defined so that α and β minimizes RSS, i.e:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2$$

Solving the optimization problem, we obtain

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad\qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{8}$$

Please note that, at this step, the *least squares method* is considered as a method of "fitting data point" to a set of data, not a method of statistical inference. Because, until now, we have no basis for constructing confidence intervals or testing hypothesis.

3.4 A Statistical Solution:

Now, we will show that the estimate of α and β in 8 are optimal in the class of linear unbiased estimates under a fairly general statistical model. Recall the linear relationship we assumed between x and y is:

$$Y_i = \alpha + \beta x_i, + \epsilon_i \text{ with } E\epsilon_i = 0, i = 1, \dots n$$

now, we also assume that

$$\text{Var}(\epsilon_i) = \sigma^2, i = 1, \dots n$$

which means that all the ϵ_is have the same but unknown variance. But, we do not need specify a probability distribution for them.

Recall that an linear estimator of β must be as the form

$$\sum_{i=1}^{n} d_i Y_i, \text{ where } d_i \text{ are constant}$$

And an linear unbiased estimator of β must satisfy:

$$E(\sum_{i=1}^{n} d_i Y_i) = \beta$$

i.e

$$\beta = \alpha \left(\sum_{i=1}^{n} d_i \right) + \beta \left(\sum_{i=1}^{n} d_i x_i \right)$$

Which implies that constant d_i must satisfy:

$$\sum_{i=1}^{n} d_i = 0 \text{ and } \sum_{i=1}^{n} d_i x_i = 1$$

Under above assumption, we can show that

$$E\hat{\beta} = \beta \text{ and } E\hat{\alpha} = \alpha$$

Recall that

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

and

$$d_i = \frac{x_i - \bar{x}}{S_{xx}}$$

satisfy condition. Therefore $\hat{\beta}$ is a linear unbiased estimator of β. Moreover, it has smallest variance among all linear unbiased estimators. Similarly, we can show that $\hat{\alpha}$ is the best linear unbiased estimator of α.

3.5 Models and Distribution Assumptions:

Until now, we have passed through two steps:

1. Solving mathematical solution for α and β to minimize the residual sum of squares, with assumption of $E\epsilon_i = 0, i = 1, 2 \dots n$
2. Proved that $\hat{\alpha}$ and $\hat{\beta}$ found in the first step is the best linear unbiased estimator for α and β, under the assumption that $\text{Var}(\epsilon_i) = \sigma^2, i = 1, 2 \dots$

but we still can not derive any tests or confidence interval under this model, because the model does not specify enough about the probability distribution of the data. For further inference, we can assume that random error $\epsilon_i \sim N(0, \sigma^2), i = 1, \dots n$. Which then implies that

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), i = 1, \dots n$$

Define the *residuals from the regression* as follow:

$$\hat{\epsilon} = Y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right), i = 1, \dots n \tag{9}$$

Theorem 3.5.1 *An unbiased estimator for* σ^2 *is*

$$S = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

Theorem 3.5.2 *Under the above assumptions, we have*

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n}x_i^2\right) \text{ and } \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

with

$$Cov(\hat{\alpha}, \hat{\beta}) = \frac{-\sigma^2\bar{x}}{S_{xx}}$$

Furthermore, $\left(\hat{\alpha}, \hat{\beta}\right)$ *and* S^2 *are independent and*

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

In fact, the variance σ^2 is unknown, and using normal distribution to make inference about $\hat{\alpha}$ and $\hat{\beta}$ is impossible. Instead, we use S^2 to estimate σ^2 and inference regrading to α and β are based on the following t-distribution:

$$\frac{\hat{\alpha}-\alpha}{\sqrt{S\left(\sum_{i=1}^{n}x_i^2\right)/(nS_{xx})}} \sim t_{n-2} \text{ and } \frac{\hat{\beta}-\beta}{S/\sqrt{S_{xx}}} \sim t_{n-2}$$

This result is immediately from theorem 3.5.2.

3.6 Interpretation of β:

Value $x = 0$ may not be a reasonable value for predictor variable, therefore α may not be an interest. However, β is the **rate of change of** $E(Y|x)$ as a function of x. That is, β is the amount that $E(Y|x)$ changes if x is changed by one unit. Thus, this parameter contains the information about whatever linear relationship exists between Y and x.

The rejection region at level α for the following test

$$\begin{cases} H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \end{cases}$$

is

$$\left|\frac{\hat{\beta}}{S/\sqrt{S_{xx}}}\right| > t_{n-2,\alpha/2} \tag{10}$$

which is equivalent to

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} > F_{1,n-2,\alpha}$$

We have

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} = \frac{S_{xy}^2}{S_{xx}^2}\frac{S_{xx}}{\text{SSE}/(n-2)} = \frac{S_{xy}^2/S_{xx}}{\text{SSE}/(n-2)} = \frac{\sum_{i=1}^n(\hat{y}_i-\bar{y})^2}{\text{SEE}/(n-2)} = \frac{\text{Sum of square of model}}{\text{Sum of square of errors / df}}$$

The last equality can be derived by showing

$$\frac{S_{xy}^2}{S_{xx}} = \sum_{i=1}^n(\hat{y}_i-\bar{y})^2$$

In fact

$$\begin{aligned} S_{xx}\sum_{i=1}^n(\hat{y}_i-\bar{y})^2 &= S_{xx}\sum_{i=1}^n(\hat{\beta}x_i+\hat{\alpha}-\bar{y})^2 \\ &= S_{xx}\sum_{i=1}^2(\hat{\beta}x_i-\hat{\beta}\bar{x})^2 \\ &= S_{xx}\hat{\beta}^2\sum_{i=1}^n(x_i-\bar{x})^2 \\ &= S_{xx}\hat{\beta}^2 S_{xx} = S_{xy}^2 \end{aligned}$$

Therefore, F-statistic for our tests can be computed as

$$F = \frac{\hat{\beta}^2}{S^2/S_{xx}} = \frac{\text{Regression sum of square}}{\text{Residual sum of square/df}} = \frac{\text{MSM}}{\text{MSE}}$$

Moreover, using the fact that $\hat{y}_i = \hat{\alpha}+\hat{\beta}x_i, i = 1,\dots n$ and $\hat{\alpha} = \bar{y}-\hat{\beta}\bar{x}$, we have:

$$\begin{aligned} \sum_{i=1}^n(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \sum_{i=1}^n(y_i-\hat{\alpha}-\hat{\beta}x_i)(\hat{\alpha}+\hat{\beta}x_i-\bar{y}) \\ &= \sum_{i=1}^n\big(y_i-\bar{y}-\hat{\beta}(x_i-\bar{x})\big)(\hat{\beta}x_i-\hat{\beta}\bar{x}) \\ &= \hat{\beta}\sum_{i=1}^n(y_i-\bar{y})(x_i-\bar{x}) - \hat{\beta}^2\sum_{i=1}^n(x_i-\bar{x})^2 \\ &= \hat{\beta}S_{xy} - \hat{\beta}^2 S_{xx} \\ &= \frac{S_{xy}^2}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}} \\ &= 0 \end{aligned}$$

Which then implies the partition of the total sum of square:

$$\sum_{i=1}^n(y_i-\bar{y})^2 = \sum_{i=1}^2(y_i-\hat{y}_i)^2 + \sum_{i=1}^n(\hat{y}_i-\bar{y})^2 \tag{11}$$

$$\text{SST} = \text{SSE}+\text{SSM} \tag{12}$$

Sum of variation	df	Sum of squares	Mean square	F statistic
Regression (slope)	1	$\text{SSM} = \frac{S_{xy}^2}{S_{xx}}$	$\text{MSM} = \text{SSM}$	$\text{F} = \frac{\text{MSM}}{\text{MSE}}$
Residual	n - 2	$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$	
Total	n - 1	$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})$		

Table 1: ANOVA table for simple linear regression

Regression ANOVA table: The formulas of how we derive test statistic is summarized in table 1.

Remark 3.6.1

$$S^2 = MSE$$

Definition 3.1 (Coefficient of determination:)

$$r^2 = \frac{\textit{Regression sum of squares}}{\textit{Total sum of squares}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} S_{xy}} \tag{13}$$

r^2 is used to quantify how well the fitted line describes the data. It measures the proportion of the total variation in $y_1, \dots y_n$ (measured by S_{yy}, which is the total sum of squares) that is explained by the fitted line (measured by regression sum of squares). From 11, we know that $0 \leq r^2 \leq 1$, and:

1. If $y_1 = \hat{y}_1, \dots, y_n = \hat{y}_n$, $r^2 = 1$
2. If $y_1, \dots, y_n$ are not close to the line, which means that the residual sum of squares are large and $r^2 \approx 0$.

Remark 3.6.2 *r^2 does not take into account the size of the data. Also, it only describes how good the fit is, but it does not say anything about how good the prediction may be.*

Moreover, using 10, we can construct $100(1-\alpha)\%$ confidence interval for β:

$$\hat{\beta} - t_{n-1,\alpha/2} \frac{S}{\sqrt{S_{xx}}} < \beta < \hat{\beta} + t_{n-1,\alpha/2} \frac{S}{\sqrt{S_{xx}}}$$

Also, a level α test of $H_0 : \beta = \beta_0$ versus $\beta \neq \beta_0$ rejects H_0 if

$$\left| \frac{\hat{\beta} - \beta_0}{S/\sqrt{S_{xx}}} \right| > t_{n-2,\alpha/2}$$

3.7 Estimation and Prediction:

After observing the regression data $(x_1, y_1), \dots (x_n, y_n)$, and estimating α, β and σ, we will want to predict new observation $Y = y_0$ from new data $x = x_0$ or even estimate to mean of the population, from which Y_0 will be drawn. **Estimation:** Consider estimating the mean of the Y population associated with x_0, that is

$$E(Y|x_0) = \alpha + \beta x_0$$

Using our linear regression model, it is obvious that $\hat{\alpha} + \hat{\beta}x_0$ is an unbiased estimator for $E(Y|x_0)$:

$$E(\hat{\alpha} + \hat{\beta}x_0) = E(\hat{\alpha} + x_0 E(\hat{\beta})) = \alpha + \beta$$

Also, by theorem 3.5.2, we have

$$\begin{aligned}
\text{Var}(\hat{\alpha} + \hat{\beta}x_0) &= \text{Var}(\hat{\alpha}) + x_0^2 \text{Var}(\hat{\beta}) + 2x_0 \text{Cov}(\hat{\alpha}, \hat{\beta}) \\
&= \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2 + x_0^2 \frac{\sigma^2}{S_{xx}} - 2x_0 \frac{\sigma^2 \bar{x}}{S_{xx}} \\
&= \frac{\sigma^2}{S_{xx}} \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 + x_0^2 - 2x_0 \bar{x} \right) \\
&= \frac{\sigma^2}{S_{xx}} \left(\frac{1}{n} \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right] + (x_0 - \bar{x})^2 \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)
\end{aligned}$$

The last "=" is derived by

$$\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 = S_{xx}$$

Finally, $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of $Y_1, \dots, Y_n$, so is $\hat{\alpha} + \hat{\beta}x_0$. And by assumption, $Y_1, \dots, Y_n$ follow normal distribution, we then end up with the distribution of $\hat{\alpha} + \hat{\beta}x_0$ as follow:

$$\hat{\alpha} + \hat{\beta}x_0 \sim N \left(\alpha + \beta x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

Moreover, by theorem 3.5.2, $(\hat{\alpha}, \hat{\beta})$ are independent of S^2, therefore:

$$\frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

Prediction: we will give a prediction interval on a random variable, not a parameter. We assume the new observation Y_0 to be taken at $x = x_0$, independent of previous data,

thus Y_0 is independent of $\hat{\alpha}, \hat{\beta}, S$, by previous discussion, we know $Y_0 - \hat{\alpha} - \hat{\beta}x_0$ has normal distribution with mean 0, and

$$\text{Var}(Y_0 - \hat{\alpha} - \hat{\beta}x_0) = \text{Var}(Y_0) + \text{Var}(\hat{\alpha} + \hat{\beta}x_0) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

Using the independence of S^2 and $Y_0 - (\hat{\alpha} + \hat{\beta}x_0)$, we see that

$$T = \frac{Y_0 - (\hat{\alpha} + \hat{\beta}x_0)}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$$

Intuitively, random variable is more variable that a parameter, therefore, prediction gives larger variation than estimation.

3.8 Matrix Approach:

Simple linear regression model can be written as the form of matrices:

$$Y_{n\times 1} = X_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1}, \text{ with } \epsilon \sim N_n\left(0, \sigma^2 I_n\right)$$

The best linear unbiased estimate for β is

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{14}$$

and let

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = \left[X(X'X)^{-1}X'\right]Y = HY \tag{15}$$

where

$$H_{n\times n} = X(X'X)^{-1}X' \tag{16}$$

Theorem 3.8.1 (Properties of H) *We have following observations about matrix H:*

1. *H is symmetric. In fact,*

$$H' = \left(X(X'X)^{-1}X'\right)' = X''\left[(X'X)^{-1}\right]'X' = X(X'X)X' = H$$

2. *$H^2 = H$ (i.e H is an idempotent matrix):*

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}I_{2,2}X' = H$$

3. *$I - H$ is also symmetric:*

$$(I - H)' = I' - H' = I - H$$

4. *$(I - H)^2 = I - H$ (i.e $I - H$ is an idempotent matrix):*

$$(I - H)^2 = (I - H)(I - H) = I - H - H + H^2 = I - 2H + H = I - H$$

Now, consider the residual sum of squares:

$$\text{SSE} = (Y - \hat{Y})'(Y - \hat{Y}) = Y'(I - H)'(I - H)Y = Y'(I - H)Y \tag{17}$$

The total sum of squares is:

$$\text{SST} = Y' \left(I - \frac{1}{n} I_n I_n' \right) Y \tag{18}$$

and the regression sum of square is:

$$\text{SSM} = Y' \left(H - \frac{1}{n} II' \right) Y \tag{19}$$

And finally, we can derive that

$$\text{STT} = \text{SSE} + \text{SSM}$$

An unbiased estimate for σ^2 is:

$$S^2 = \text{MSE} = \frac{1}{n-2} Y'(I - H)Y \tag{20}$$

4 Multiple regression:

Suppose we want to model one response y as a function of several k independent variables $x_1, \dots x_k$. The collected data can be described as the table below:

Obs	y	x_1	x_2	$\dots$	x_k
1	y_1	x_{11}	x_{12}	$\dots$	x_{1k}
2	y_2	y_{21}	y_{22}	$\dots$	x_{2k}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
n	y_n	x_{n1}	x_{n2}	$\dots$	y_{nk}

Table 2: Data of Multiple regression

Thus, for the i^{th} observation, we have:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_k x_{ik} + \epsilon_i$$

Under the assumption that $\epsilon_i \sim N(0, \sigma^2), \forall i = 1, \dots n$ are independent, we obtain the model:

$$EY_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_k x_{ik}$$

In short, it can be written as

$$EY = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_k x_k$$

Remark 4.0.1 *This model can be interpreted in various context: For example, suppose there are 2 independent variables x_1 and x_2:*

1. *There is an interaction between x_1 and x_2, then the model is*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

2. *If the model is non-linear and we can approximate by a polynomial, then*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon$$

3. *The full model with interaction terms is:*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

4. *Model with transformed variable:*

$$Y = \beta_0 + \beta_1 \ln x_1 + \beta_2 e^{x_2} + \epsilon$$

   which can be written as

$$Y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^*, \text{ where } x_1^* = \ln x_1 \text{ and } x_2^* = e^{x_2}$$

5. *Or, if we want to compare the treatments effect in an experiment of 2 level A and B, we can add extra variable to indicate where the observation from:*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x + 2 + \beta_3 x_3 + \epsilon, \text{ where } x_3 = \begin{cases} 1, & \text{if level } A \\ 0, & \text{if level } B \end{cases}$$

   This single model can yield 2 models, one for each lab;

   - *For lab A:*
$$Y = (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 \epsilon$$
   - *For lab B:*
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

By matrix approach, we are able to obtain the model of Y depending on many predictors variables $x_1, \dots x_k$ as follow:

$$Y_{n\times 1} = X_{n\times(k+1)}\beta_{(k+1)\times 1} + \epsilon_{n\times 1} \text{ with } \epsilon \sim N_n\left(0, \sigma^2 I_n\right)$$

Denote $p = k + 1$, then:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \text{ with } \epsilon \sim N_n\left(0, \sigma^2 I_n\right)$$

Remark 4.0.2 *We have the following observations:*

1. *The 0^{th} column of X is constant, corresponding to the intersection and other i^{th} of X contains data on x_i predictors, where $i = 1, \dots k$.*
2. *$Y = X\beta + \epsilon$, thus*
$$Y \sim N_n(X\beta, \sigma^2 I_n)$$
which means that components $y_1, \dots y_n$ are all independent but not identically distributed.
3. *From now we assume that X is of full rank matrix: $R(X) = p$*

4.1 Least square model:

Using the same approach in Simple linear Regression, we want to minimized the residual sum of square, which is:

$$\text{SSE} = (Y - \hat{Y})'(Y - \hat{Y}) = (Y - X\beta)'(Y - X\beta)$$

i.e

$$\min_{\beta}(Y - X\beta)'(Y - X\beta) \tag{21}$$

We have

$$\begin{aligned}(Y - X\beta)'(Y - X\beta) &= Y'(Y - X\beta) - \beta' X'(Y - X\beta) \\ &= Y'Y - Y'X\beta - \beta' X'Y + \beta' X'X\beta \\ &= Y'Y - 2Y'X\beta + \beta' X'X\beta\end{aligned}$$

We have:

1. $Y'1 \times nX_{n\times p}\beta_{p\times 1}$ and $\beta'_{1\times p}X'_{p\times n}Y_{n\times 1}$, therefore they can be combined.
2. $X'X$ is a semi-positive definite and symmetric matrix, then $\beta'(X'X)\beta$ is as quadratic form.

Therefore

$$\begin{aligned}&\frac{\partial}{\partial\beta}\big((Y - X\beta)'(Y - X\beta)\big) = 0 \\ &\Leftrightarrow -2X'Y + 2X'X\beta = 0 \\ &\Leftrightarrow X'X\beta = X'Y\end{aligned}$$

This system has unique solution if and only if $X'X$ has it inverse. Under the assumption of $R(X) = p$, we know that $R(X'_{p\times n}X_{n\times p}) = p$, it means $X'X$ is a full rank matrix, which is invertible. Therefore

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{22}$$

$\hat{\beta}$ is a linear combination of vector $Y \sim N_n(X\beta, \sigma^2 I_n)$, using theorem 1.1.6 we obtain

$$\hat{\beta} \sim N\Big((X'X)^{-1}X'X\beta, (X'X)^{-1}X'\sigma^2 I_n((X'X)^{-1}X')'\Big)$$

i.e

$$\hat{\beta} \sim N\Big(\beta, \sigma^2(X'X)^{-1}\Big)$$

In conclusion, we have the following theorem

Theorem 4.1.1 *$\hat{\beta}$ is a unbiased estimator for β:*

1. $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$
2. $Var(\hat{\beta}_i) = \sigma^2\left[(X'X)^{-1}\right]_{ii}, \forall i = 1, \dots k$
3. *The estimators of β_i and β_j are correlated:* $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2\left[(X'X)^{-1}\right]_{ij}$

Now, the predicted value of Y, using least square model is

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY, \text{ where } H = X(X'X)^{-1}X'$$

matrix H is called *"hat-matrix"*.

Definition 4.1 (Residual vector:)

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

Theorem 4.1.2 (Properties of hat-matrix) *Recall that $H = X_{n\times p}(X'X)^{-1}_{p\times p}X'_{p\times n}$, then we have*

1. *Using the fact that $R(AB) \leq R(A)$, for ant matrix A, we have: $R(H) \leq R(X)$, which means that $n \leq p$. If $n > p$, then H is not a non-singular matrix, and its inverse does not exists.*
2. *H and $I - H$ are idempotent matrix.*
3. *$Rank(H) = p$ and $Rank(I - H) = n - p$.*

The last item can be derived as follow: Since H is idempotent, we have

$$\text{Rank}(H) = \text{Tr}(H)$$

Then

$$\begin{aligned}\text{Tr}(H) &= \text{Tr}\big(X(X'X)^{-1}X'\big)\\ &= \text{Tr}\big(X'X(X'X)^{-1}\big)(\text{ because } \text{Tr}(AB) = \text{Tr}(BA))\\ &= \text{Tr}(I_p) = p\end{aligned}$$

And finally,

$$\text{Tr}(I - H) = \text{Tr}(I) - \text{Tr}(H) = n - p$$

Theorem 4.1.3 (Estimate of σ^2)

$$\hat{\sigma}^2 = \frac{e'e}{n-p} = \frac{SSE}{n-p} = MSE$$

is an unbiased estimator of σ^2 *and*

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

From knowledge about distribution of $\hat{\beta}$ and $\hat{\sigma}^2$, we can derive the tests $\beta_i, i = 1, \dots k$:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

Under H_0, the test statistic

$$\frac{\hat{\beta}_i}{\sqrt{\text{MSE}\big((X'X)^{-1}\big)_{ii}}} \sim t_{n-p}$$

then rejection region is $|t| > t_{\alpha/2,n-p}$.

For the test of model

$$\begin{cases} H_0 : \beta_1 = \cdots = \beta_k = 0 \\ H_1 : \text{at least } \beta_i's \text{ not } 0 \end{cases}$$

we use the following test statistic

$$F = \frac{\text{MSM}}{\text{MSE}} \sim F_{p-1,n-p}$$

Regression ANOVA table: The formulas of how we derive test statistic is summarized in table 3.

Sum of variation	df	Sum of squares	Mean square	F statistic
Regression	p -1	SSM	$\text{MSM} = \frac{\text{SSM}}{p-1}$	$\text{F} = \frac{\text{MSM}}{\text{MSE}}$
Residual	n - p	SSE	$\text{MSE} = \frac{\text{SSE}}{n-p}$	
Total	n - 1	SST		

Table 3: ANOVA table for simple linear regression

Definition 4.2 ((Coefficient of determination:))

$$r^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

5 Appendix:

5.1 t-distribution:

Let $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$ be independent. Then a t-random variable can be defined as follow:

$$T = \frac{Z}{\sqrt{\frac{U}{\mu}}} \sim t_\mu$$

In fact, when we want to test the hypothesis:

$$\begin{cases} H_0 : & \nu = \nu_0 \\ H_1 : & \mu \neq \mu_0 \end{cases}$$

the test statistic under H_0 is:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

In many cases, σ^2 is unknown, therefore an estimate of σ^2 is used:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

which implies the test statistic

$$\frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

by theorem 1.1.6.

5.2 F-distribution:

Let $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$ be independent, then a F-random variable can be defined as follow:

$$F = \frac{U_1/\nu_1}{V_2/\nu_2} \sim F_{\nu_1,\nu_2}$$

In fact, if we have two normal distribution $N(\mu_1, \sigma_1^2)$ and $N(\mu_1, \sigma_2^2)$ corresponding to n and m data points, and want to test:

$$\begin{cases} H_0 : & \sigma_1^2 = \sigma_2^2 \\ H_1 : & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

The test statistic

$$\frac{S_1^2}{S_2^2} \sim F_{n-1,m-1} \text{ under } H_0$$

References