Poor Man's Guide To Machine Learning

pehcy(CheeYung) @

References:

- An Introduction to Statistical Learning (with applications in R)
- Data Science from Scratch, *O' Reilly*
- Deep Learning with PyTorch, *Manning*
- Exam PA Study Manual, Fall 2021.
- Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow (2019), *O' Reilly*
- Regression Modeling with Actuarial and Financial Applications
- Statistics in A Nutshell, *O' Reilly*

Contents

1

Simple Linear Regression

2

Multiple Linear Regression

pehcy (Chee Yung) @ https://github.com/pehcy

In the case when we have multiple explanatory variables, we then have to fit this equation

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i \tag{2.1}$$

where n is the number of observations, or the size of data set. Each row of the data matrix is called an *observation* or *record*. Each observation consists of k explanatory variables. We can extend the simple linear regression concepts to fit these β's, which known as multiple linear regression.

To facilitate the calculation, we organize each observations into matrices. In general, they can be formulate into an $n \times (k+1)$ matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

Notice the intercept and dependent variable are now become a column vector,

$$\overrightarrow{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

2.1 The Least Squares Problem

Given that $A_{m,n}$ is an $m \times n$ matrix, and vector $\mathbf{b} \in \mathbb{R}^m$. $m \geq n \geq 1$ The least squares problem is the minimization problem as follow:

$$\hat{\mathbf{x}} = \arg\min_{x \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2 \tag{2.2}$$

This system has a solution if $\mathbf{b} \in \text{Span}(A)$, the column space of A, but normally this is not the case and we can only find an approximate solution. Our goal is to find a suitable solution $\hat{\mathbf{x}}$ that minimize the loss, $\hat{\mathbf{x}}$ is known as the least squares solution of $A\mathbf{x} = \mathbf{b}$. Soon, we will use geometric approach to figure out a general solution for this problem.

2.1.1 Geometric Interpretation

First we take a look on a simple example, we consider the following system of equations:

$$\begin{cases} 2x + y = 7 \\ x + 3y = 6 \\ 3x - y = 3 \end{cases} \tag{2.3}$$

It can be checked that this system has no solution; the first two equations only admit $x = 3, y = 1$, but this violates the third. As before, we write

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} x + \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} y = \begin{bmatrix} 7 \\ 6 \\ 3 \end{bmatrix} \tag{2.4}$$

Now we let $\mathbf{a}_1 = \langle 2, 1, 3 \rangle$, and $\mathbf{a}_2 = \langle 1, 3, -1 \rangle$. The column vectors $\mathbf{a}_1$ and $\mathbf{a_2}$ span the projection $\mathbf{b}$ such that

$$A\mathbf{x} = x\mathbf{a_1} + y\mathbf{a_2} = \mathbf{b} \tag{2.5}$$

Now it is evident that the linear combination $A\mathbf{x} = x\mathbf{a_1} + y\mathbf{a_2} = \mathbf{b}$ fill out a plane in 3D space. However, our target vector $\mathbf{b}$ lies outside this plane! Thus, there is no way of combining $\mathbf{a}_1$ and $\mathbf{a}_2$ to get $\mathbf{b}$.

In order to see what's happening, let's look at the span of the column vectors $\mathbf{a}_1$ and $\mathbf{a}_2$. we saw that

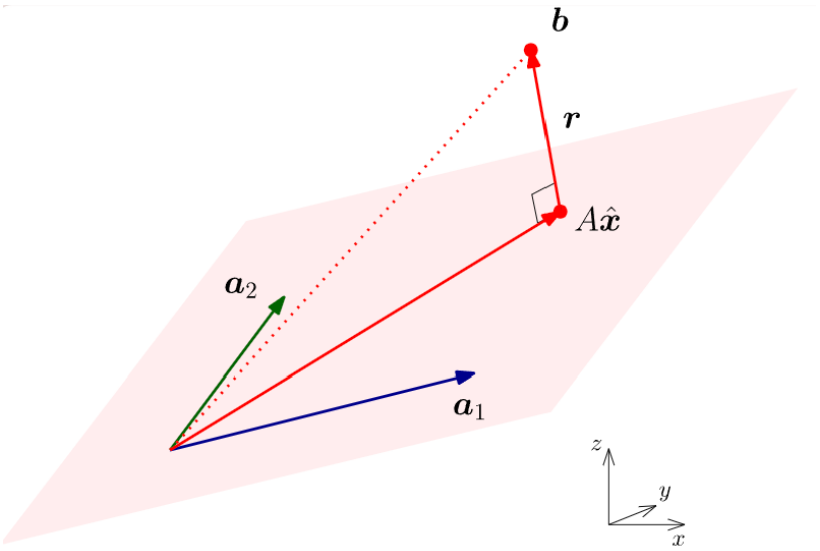


Figure 2.1: The pink plane, which is understood to extend outwards in all directions, denotes the span of the blue and green vectors. Note that the target point b b in red does not lie on this plane. It's 'shadow' on the plane, $A\hat{\mathbf{x}}$, has been indicated for perspective.

our target $\mathbf{b}$ lies outside our span, the set of vectors $A\mathbf{x}$. Thus, we cannot extract an exact solution; instead, we look for the best possible solution $\hat{\mathbf{x}}$. This is chosen in such a way that the corresponding point on the plane, $A\hat{\mathbf{x}}$, lies as close as possible to $\mathbf{b}$. Note that this corresponds to minimizing the square of the length $\|A\hat{\mathbf{x}} - \mathbf{b}\|_2^2$, which is exactly the sum of squares of deviations component-wise. Thus, we break down the vector $\mathbf{b}$ into

$$\mathbf{b} = A\hat{\mathbf{x}} + \mathbf{r} \tag{2.6}$$

In order to minimize the distance of $\mathbf{b}$ from the plane, i.e. the length of the component $\mathbf{r}$, it is intuitively

clear that the closest point $A\hat{\mathbf{x}}$ must be the perpendicular projection of $\mathbf{b}$onto the plane. Thus, the difference $\mathbf{r}$ must be perpendicular to the plane.

In particular, $\mathbf{r}$ is perpendicular to both $\mathbf{a}_1$ and $\mathbf{a}_2$. We write this in terms of the dot product: $\mathbf{a}_1 \cdot \mathbf{r} = 0$ and $\mathbf{a}_2 \cdot \mathbf{r} = 0$. The dot product can be rewritten in the form

$$\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{n} v_i w_i = \mathbf{v}^\intercal \mathbf{w} \tag{2.7}$$

Thus, we express our condition as $A \perp \mathbf{r}$, this means that

$$A^\intercal \mathbf{b} = (A^\intercal A)^{-1} \hat{\mathbf{x}} + A^\intercal \mathbf{r} = A^\intercal A \hat{\mathbf{x}} \tag{2.8}$$

Notice that $A^\intercal A$ is a square matrix. Whenever $A^\intercal A$ is invertible. i.e. there is a unique solution $\hat{\mathbf{x}}$, we must have

$$\hat{\mathbf{x}} = (A^\intercal A)^{-1} A^\intercal \mathbf{b} \tag{2.9}$$

And thus, this is the general solution for least squares problem.

Definition 2.1 Least Squares Method for Multiple Linear Regression
The coefficients (or slopes) of multiple linear equation is

$$\hat{\beta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} \tag{2.10}$$

where $\mathbf{X}_{n,p}$ is explanatory data, and $\mathbf{y}_{p,1}$ is the response data.

The computational complexity of inverting such a matrix is typically about $O(n^{2.4})$ to $O(n^3)$. This is why victories your data help to reduce CPU / GPU calculation time.

2.2 Analysis of Variance (ANOVA)

The question then is how can we choose the best values of β_i? First of all, we should define what we mean by the best. The most ideal strategy is that we choose the values of β that will create close predictions of Y on new future data.

In an observational study, both response and predictor data can be obtained via observation. For ANOVA, the test statistic is the F ratio, which can be used to determine whether significant differences exist between two or more population. An ANOVA test will provides an F-ratio comparing the population means. We would test the null hypothesis on a predetermined significant level such as $p < 0.05$ or $p < 0.01$, depends on the level of significance α.

2.2.1 Sum of Squares

One useful aspect of linear regression is that it can decomposed the variation into two parts: the variation of the predicted values, and the variation of prediction errors.

If all such deviations are squared, the squared deviations $(y_i - \overline{y})^2$ will provide the basis for measuring the spread of the data.

Squaring again both sides of the preceding equation and sum over all $i = 1, 2, ..., n$, we can obtain the following

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \tag{2.11}$$

After squaring the right-hand side of the equation, we need some algebraic calculation to evaluate the cross-product term $(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$.

$$\begin{aligned}
\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\,(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}) \\
&= \sum [y_i - \overline{y} - \hat{\beta}_1 (x_i - \overline{x})]\,[\hat{\beta}_1 (x_i - \overline{x})] \\
&= \hat{\beta}_1 \, \mathrm{SS}_{xy} - \hat{\beta}_1^2 \, \mathrm{SS}_x \\
&= 0
\end{aligned}$$

this implies that the sum of the cross-products over all response variables is zero [1]. And hence, the finalized equation will be

$$\underbrace{\sum (y_i - \overline{y})^2}_{SS_T} = \underbrace{\sum (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum (\hat{y}_i - \overline{y})^2}_{SS_R} \tag{2.12}$$

Let us interpret each SS one by one:

- **Total Sum of Squares** (SS_T), also called the sum of squares treatment. It represents the sum of squares in all of the responses. It measures the amount of variability inherent in the response prior to performing regression.

$$SS_T = \sum (y_i - \overline{y})^2 = \mathbf{y}^\intercal \mathbf{y} - \frac{\mathbf{y}^\intercal \mathbf{J} \mathbf{y}}{n} = \mathbf{y}^\intercal \mathbf{y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \tag{2.13}$$

  where $\mathbf{J}$ is a $n \times n$ square matrix with all elements 1, and n is the length of vector $\mathbf{y}$.

- **Sum of Squares Error** (SS_E), also called the L2-norm, is the variation of all the response values about the fitted regression line. It describes how well the model fits the data.

$$\begin{aligned}
SS_E = \mathbf{e}^\intercal \, \mathbf{e} &= \sum (y_i - \hat{y}_i)^2 \\
&= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal \, (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{y}^\intercal \, \mathbf{y} - \hat{\boldsymbol{\beta}}^\intercal \, \mathbf{X}^\intercal \mathbf{y}
\end{aligned}$$

- **Regression Sum Of Squares** (SS_R), the difference between SS_T and SS_R, or the total variation explained through knowledge of x. SS_R measure how effective the SLR model is in explaining the variation of y.

$$\begin{aligned}
SS_R = SS_T - SS_E &= \sum (y_i - \overline{y}_i)^2 - \sum (y_i - \hat{y}_i)^2 \\
&= \mathbf{y}^\intercal \mathbf{y} - \frac{\mathbf{y}^\intercal \mathbf{J} \mathbf{y}}{n} - (\mathbf{y}^\intercal \, \mathbf{y} - \hat{\boldsymbol{\beta}}^\intercal \, \mathbf{X}^\intercal \mathbf{y}) \\
&= \hat{\boldsymbol{\beta}}^\intercal \, \mathbf{X}^\intercal \mathbf{y} - \left(\frac{1}{n}\right) \mathbf{y}^\intercal \, \mathbf{J} \, \mathbf{y}
\end{aligned}$$

[1] In linear algebra, this implies the residuals are orthogonal to the regressor.

or you may use the computational formula to calculate SS_R

$$SS_R = \hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{2.14}$$

2.2.2 Estimation of σ^2

The mean squares error, $\hat{\sigma}^2 = MS_E = \dfrac{SS_E}{n-k-1}$ estimates the true variance of errors σ^2.

2.2.3 ANOVA table

One of the outputs of multiple regression is the ANOVA table. The following shows the general template of an ANOVA table.

Source	Degree of Freedom (df)	Sum of Squares(SS)	Mean Sum of Squares (MS)
Regression	k	SS_R	MS_R
Error	$n-p=n-(k-1)$	SS_E	MS_E
Treatment	$n-1$	SS_T	

With this table, we can compute the test statistic for F-test.

$$F_0 = \frac{MS_R}{MS_E} \tag{2.15}$$

$p = k+1$ is the number of betas (or coefficients) in the model.

where all mean squares are the sum of squares divided by its degree of freedom. Notice the $n-k-1$ instead of simple linear regression's $n-2$, which is the main change. The graphical interpretation of when $k=2$ is shown as the diagrams below.

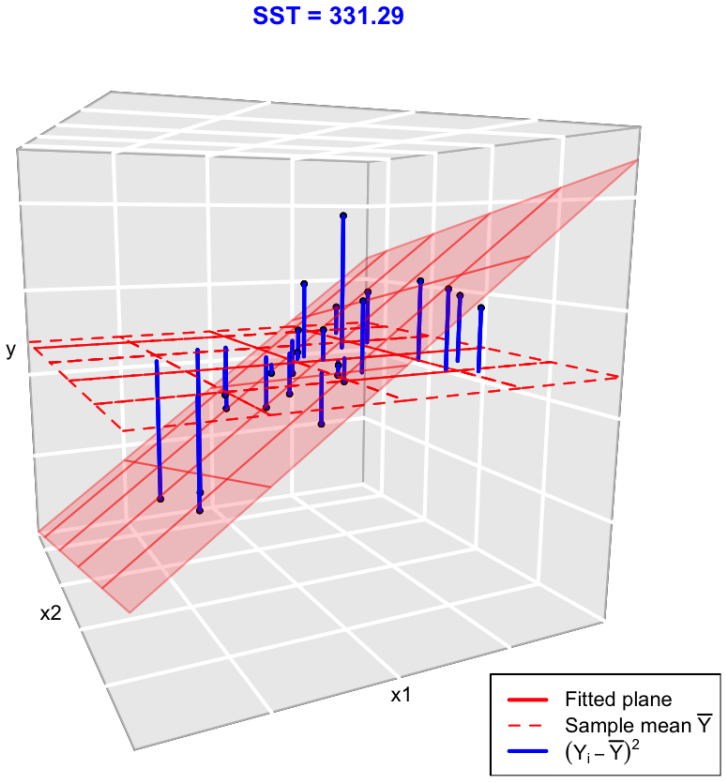


Figure 2.2: Visualization of the ANOVA decomposition when $k = 2$. SS_T measures the variation of $Y_1, Y_2, \ldots, Y_n$ with respect to $\overline{Y}$. SS_T measures the variation with respect to the conditional means, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$.

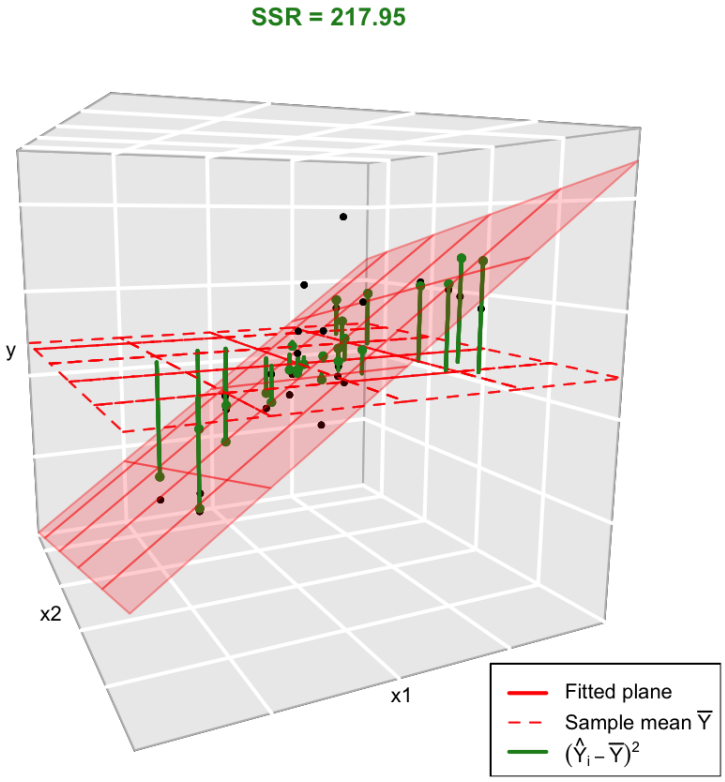


Figure 2.3: Visualization of SS_E. SS_E collects the variation of the residuals.

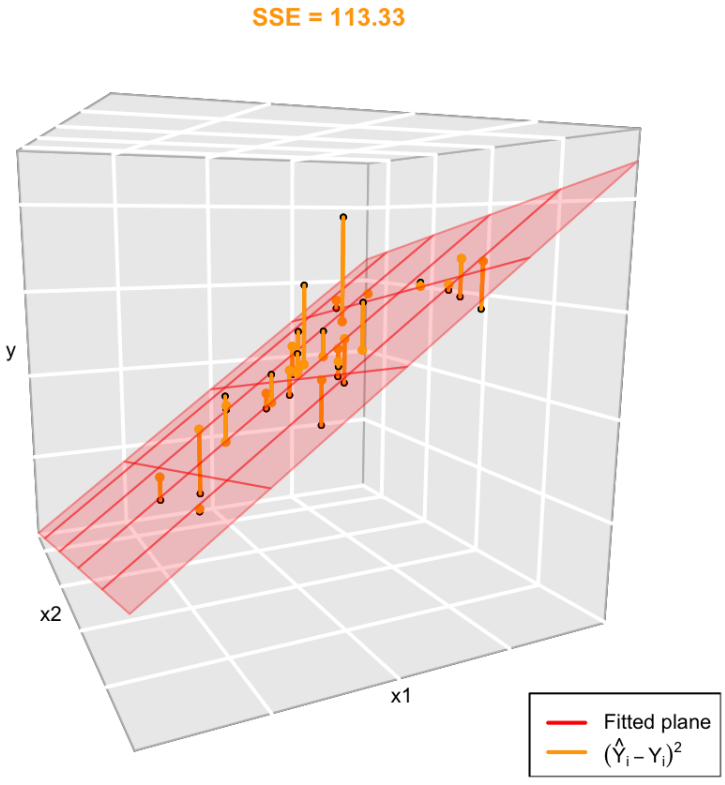


Figure 2.4: Visualization of SS_R.

Example 2.2.1. We conducted a small study of the relation between the yearly income of full-stack developer (per RM1,000) and their age and their number of years of work experience.

(https://www.kaggle.com/datasets/hussainnasirkhan/multiple-linear-regression-dataset)

Answer the following questions, you may use Python or R for numerical calculation.

1. Find $\hat{\boldsymbol{\beta}}$, how would you interpret the values of $\hat{\beta}_1$ and $\hat{\beta}_2$?
2. Find the estimated yearly income for a developer who is 28 years old and owned 5 years of work experience.
3. Estimate σ^2.
4. Find $Var(\hat{\beta}_2)$, and $Cov(2\beta_1, 3\beta_2)$.

2.3 Test for Significance of Regression model

Within a multiple regression model, we may want to know whether a particular x-variable is making a useful contribution to the model. That is, given the presence of the other x-variables in the model, does a particular x-variable help us predict or explain the y-variable? For instance, suppose that we have three x-variables in the model. The general structure of the model could be

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{2.16}$$

As an example, to determine whether variable is a useful predictor variable in this model, we could test

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0 \tag{2.17}$$

If the null hypothesis above were the case, then a change in the value of x1 would not change y, so y and X_1 are not linearly related (taking into account X_2 and X_3). Also, we would still be left with variables X_2 and X_3 being present in the model. When we cannot reject the null hypothesis above, we should say that we do not need variable X_1 in the model given that variables X_2 and X_3 will remain in the model. In general, the interpretation of a slope in multiple regression can be tricky. Correlations among the predictors can change the slope values dramatically from what they would be in separate simple regressions.

To carry out the test, statistical software will report p-values for all coefficients in the model. Each p-value will be based on a t-statistic calculated as

$$t_0 = \frac{(\text{sample coefficient} - \text{hypothesized value})}{\text{standard error of coefficient}} \tag{2.18}$$

Definition 2.2 Test statistics of one slope (Marginal test)
The null and alternative hypothesis is

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \tag{2.19}$$

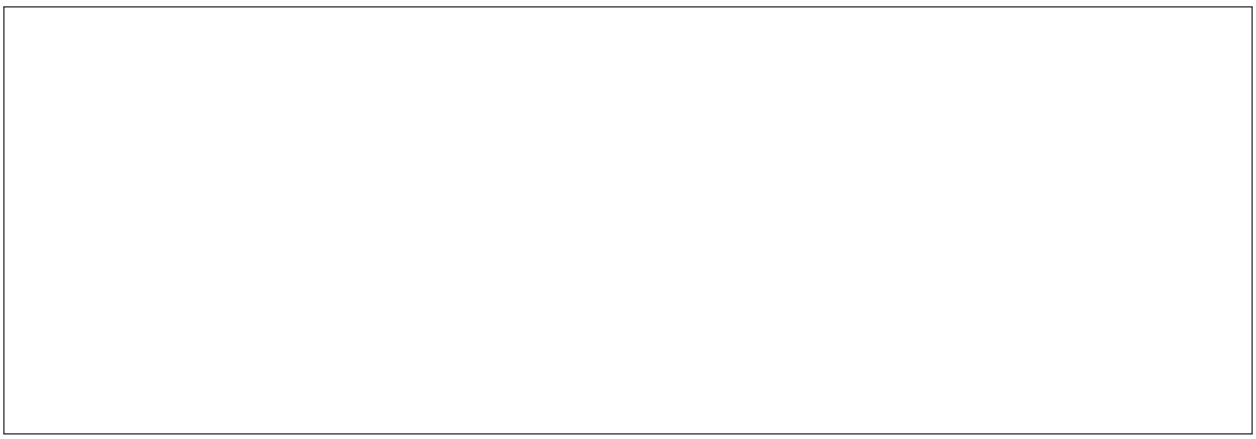
The test statistic for t-test is

$$t_0 = \frac{\hat{\beta}_j}{\hat{\sigma}^2 c_{jj}} \sim t_{\alpha/2;n-k-1} \tag{2.20}$$

where c_{jj} is the $(j+1)th$ diagonal element of the covariance matrix. $\mathbf{C} = (\mathbf{X}^\intercal \mathbf{X})^{-1}$. Reject H_0 if $t_0 > t_{\alpha/2;n-k-1}$.
Reject H_0 denotes that β_j does contribute significantly to the model. Otherwise, if H_0 is not rejected, that means x_j is unnecessary and can be removed from the model.

Notice that the value we test is not necessary equal to zero. In fact, we can even use t-test to check whether β_j equal to certain constant (e.g. $H_0 : \beta_j = 2; \quad H_1 : \beta_j \neq 2$).

Example 2.3.1. Refer to Example 2.1.1, Find the test statistics testing $\beta_1 = 0$ at the significance level $\alpha = 0.01$.



2.3.1 Test for significance of Parameters

Earlier in the example, we test a regression model to see if y and x are linearly related with the testing

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0 \tag{2.21}$$

with the t-test (or the equivalent F-test). In multiple linear regression, there are several partial slopes and the t-test and F-test are no longer equivalent. Our question changes: Is the regression equation that uses information provided by the predictor variables $X_1, X_2, \ldots, X_k$ better than the simple predictor $\hat{y}$ (the sample mean of response data), which does not rely on any of these independent variables?

Definition 2.3 Test statistics of Parameters
The null and alternative hypothesis is

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0 \\ H_1 : \text{At least one } \beta_j \neq 0 \end{cases} \tag{2.22}$$

The test statistic is

$$F_0 = \frac{MS_R}{MS_E} \sim F_{k,n-k-1} \tag{2.23}$$

Reject H_0 if $F_0 > F_{\alpha;k,n-k-1}$.
Reject H_0 implies that the regression model conatining at least one predictor useful in predicting y.

2.4 General F-Test

Generally, the F-test involves three basic steps:

1. Define a larger full model. Larger means we include more than one parameters.
2. Then, we define a smaller reduced model. By smaller, we mean one model with parameters lesser than the full model.
3. Use an F-statistic to decide whether or not to reject the smaller reduced model in favor of the larger full model.

Recall back the fundamental rule to make decision

2.4.1 Full model

The ”full model”, which is also sometimes referred to as the ”unrestricted model,” is the model thought to be most appropriate for the data. For simple linear regression, the full model is:

2.4.2 The Reduced model

The ”reduced model,” which is sometimes also referred to as the ”restricted model,” is the model described by the null hypothesis H_0. For simple linear regression, a common null hypothesis is $H_0 : \beta_1 \neq 0$. In this case, the reduced model is obtained by ”zeroing-out” the slope β_1 that appears in the full model. That is, the reduced model is:

Definition 2.4 Partial F-test on Reduced model
The null and alternative hypothesis is

$$\begin{cases} H_0 : y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\ H_1 : y = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \end{cases} \tag{2.24}$$

or

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases} \tag{2.25}$$

The test statistic is

$$F_0 = \frac{SS_R(\boldsymbol{\beta_2} \,|\, \boldsymbol{\beta_1})/(k-r+1)}{MS_E} \sim F_{k-r+1,\, n-k-1} \tag{2.26}$$

the $SS_R(\boldsymbol{\beta_2} \,|\, \boldsymbol{\beta_1})$ is regression sum of squares of β_2 given that β_1 exists in the model, the formula is

$$SS_R(\boldsymbol{\beta_2} \,|\, \boldsymbol{\beta_1}) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta_1}) \tag{2.27}$$

Reject H_0 if $F_0 > F_{\alpha;k-r+1,n-k-1}$.
This is called “extra sum of squares” because it measures the increase in the regression sum of squares that results from adding the parameters $X_r, \ldots, X_k$ to a model that already contains $X_1, X_2, \ldots, X_{r-1}$.

Example 2.4.1. The following regressions models have been fitted to 40 observations:

- Model A: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
- Model B: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- Model C: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Given that:

(i) $SS_E = 220$ for Model A.

(ii) The F-ratio for testing that $\beta_2 = 0$ in going from Model A to Model B is 30.83.

(iii) The F-ratio for testing that $\beta_3 = 0$ in going from Model B to Model C is 12.

Find the test statistic for testing that $\beta_2 = \beta_3 = 0$.

Example 2.4.2. Refer to Example 2.1.1, use reduced model $\tilde{Y} = \beta_0 + \beta_1 X_1 + \varepsilon$ to find the test statistic for testing $\beta_2 = 0$. (You may use R or Python for calculation).

2.5 Statistical Inference about Regression Coefficients

2.5.1 Prediction vs Estimation

What is the main reason we fitting a model to data? It is often to accomplish one of two goals. We can either use a model to **estimate** the relationship between response and the predictors, or to **predict** the response based on the predictors. Often, a good model can do both, but here we'll discuss both goals separately since the process of finding models for explaining and predicting are slightly different. In most cases, prediction is less precise than estimation with a bigger standard error.

Normally, The confidence interval for the mean $\mathbb{E}[\hat{Y}_h]$ is narrower than the prediction interval for y_h because of the additional uncertainty attributable to the random error ε when predicting some future value of y_h.

2.5.2 Confidence Interval

Does the explanatory variable have a real inference on the response? The fact is that constructing a confidence interval for the slope β_i is more useful than using t-score or the p-value. The confidence interval with a certain degree of confidence estimated the range of values for the slope β_i. Notes that using a higher confidence level will generate a wider confidence interval but produced less precise estimation.

Starting with the definition of confidence interval,

$$\Pr\left(-t_{n-2,\alpha/2} < \frac{\hat{\beta}_i - \beta_i}{Se(\hat{\beta}_i)} < t_{n-2,\alpha/2}\right) = 1-\alpha \tag{2.28}$$

A $100(1-\alpha)\%$ confidence interval for β_i can be calculated with the formula

$$\text{LSE} \pm t\text{-quantile} \times \text{Standard Error} = \hat{\beta}_i \pm t_{n-2,\,\alpha/2} \times Se(\hat{\beta}_i) \tag{2.29}$$

For the t-quantile, you need to make sure that you use $\alpha/2$ as the probability level since this is a two-tails test.

Definition 2.5 Confidence Interval for slopes (β_i)
At significance level $0 < \alpha < 1$, the $(1-\alpha)100\%$ confidence interval for β_i is

$$\hat{\beta}_i - t_{\alpha/2;n-k-1}\, Se(\hat{\beta}_i) \;\le\; \beta_i \;\le\; \hat{\beta}_i + t_{\alpha/2;n-k-1}\, Se(\hat{\beta}_i) \tag{2.30}$$

where

$$Se(\hat{\beta}_i) = \sqrt{MSE * c_{ii}} \tag{2.31}$$

c_{ii} is the $(i+1)th$ diagonal element of the covariance matrix. $\mathbf{C} = (\mathbf{X}^\intercal \mathbf{X})^{-1}$.

Example 2.5.1. Refer to Example 2.1.1, construct a 95% confidence interval for β_1 .

2.5.3 Prediction Interval

The variance

$$\begin{aligned}
\hat{\text{Var}}[\hat{f}(x) + \varepsilon] &= \text{Var}[\hat{f}(x)] + \text{Var}[\varepsilon] \\
&= \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\text{SS}_x}\right) + \sigma^2 \\
&= \hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\text{SS}_x}\right]
\end{aligned}$$

and

$$\hat{y}(x) + \varepsilon \sim \mathcal{N}\left(\beta_0 + \beta_1 x,\, \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{SS_x}\right)\right) \tag{2.32}$$

so the standard error will be

$$Se[\hat{y}(x)+\varepsilon] = \sqrt{s^2\left(1+\frac{1}{n}+\frac{(x-\overline{x})}{\mathrm{SS}_x}\right)} = s\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})}{\mathrm{SS}_x}} \tag{2.33}$$

where s is the sample standard variance.

For multiple linear regression, the prediction interval is similar,

Definition 2.6 Confidence Interval on mean response
At significance level $0 < \alpha < 1$, the $(1-\alpha)100\%$ confidence interval on $\mathbb{E}[Y_h]$, the mean response at the point x_h is

$$\mathbb{E}[\hat{Y}_h] - t_{\alpha/2;n-k-1}\, Se(\mathbb{E}[\hat{Y}_h]) \le \mathbb{E}[\hat{Y}_h] \le \mathbb{E}[\hat{Y}_h] + t_{\alpha/2;n-k-1}\, Se(\mathbb{E}[\hat{Y}_h]) \tag{2.34}$$

where $\mathbb{E}[\hat{Y}_h] = \mathbf{x}_h^\intercal \hat{\boldsymbol{\beta}}$, and

$$Se(\hat{y_h}) = \sqrt{MSE * \mathbf{x}_h^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{x}_h} \tag{2.35}$$

Definition 2.7 Prediction Interval on y_h
At significance level $0 < \alpha < 1$, the $(1-\alpha)100\%$ prediction interval on the prediction y_h at a specified value of x_h is

$$\hat{y}_h - t_{\alpha/2;n-k-1}\, Se(\hat{y}_h) \le y_h \le \hat{y}_h + t_{\alpha/2;n-k-1}\, Se(\hat{y}_h) \tag{2.36}$$

where

$$Se(\hat{y_h}) = \sqrt{MSE(1+\mathbf{x}_h^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{x}_h)} \tag{2.37}$$

There are two sources of uncertainty akin to prediction problem:

1. *Estimation of the true regression line at* x_*, The prediction will become less and less accurate if the data points started to moves away from the sample mean $\overline{x}$.
2. *Variability of the random error*. The future response y cannot be predicted perfectly even if we know the exact values of β_0 and β_1, this is due to the inherent random error ε with variance σ^2.

Example 2.5.2. Refer to Example 2.1.1, construct a 95% prediction interval for the mean response Y_h . Given that $\mathbf{x}_h = (27, 4)$.

Example 2.5.3. A sample of 25 observations has been represented by a model of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where is a random error term with mean 0 and variance σ^2. Given that:

$$\mathbf{C} = (\mathbf{X}^\intercal \mathbf{X})^{-1} = \begin{bmatrix} 188.9832 & 0.8578 & -28.0275 \\ 0.8578 & 0.2500 & -0.6 \\ -28.0275 & -0.6 & 5.0625 \end{bmatrix}, \quad MS_E = 0.0361, \quad \hat{\beta} = \begin{bmatrix} -4.04 \\ 0.14 \\ 0.45 \end{bmatrix}$$

(i) Determine the 95% confidence interval of β_1.

(ii) Given $x_1 = 1.5$, $x_2 = 3.2$. Find the 95% prediction interval for prediction y_h .

(iii) Find the test statistics for testing $H_0 : \beta_2 = 2$.

(iv) Construct a 95% confidence interval for $3\beta_0 + 5\beta_1 + 2\beta_2$.

2.6 Coefficient of Determination

R-squared (R^2) measures how well an estimated regression fits the data. It is also known as the coefficient of determination and can be formulated as:

$$R^2 = 1 - \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})} \tag{2.38}$$

R^2 measures the **proportion of variation** of the response variable Y that is **explained** by the predictors $X_1, X_2, \ldots, X_k$ through the regression. Intuitively, R^2 measures the **tightness of the data cloud around the regression plane**.

Definition 2.8 Adjusted R-squared

$$R^2_{Adj} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2) \tag{2.39}$$

Example 2.6.1. (CFA Level-2) Which of the following is most appropriate for adjusted R^2?

(A) It is always positive.

(B) It may or may not increase when one adds an independent variable.

(C) It is non-decreasing in the number of independent variables.

[Ans: B]

Example 2.6.2. An engineer is investigating how the amount of a product Y depends on temperature X_1 (in $^\circ C$) and production time X_2 (in hours). He has collected a sample of size $n = 32$ with $m = 6$ different groups of values for X_1 and X_2. He proposed and obtained an interaction model as:

$$\hat{y} = 95 + 15x_1 + 55x_2 - x_1x_2$$

Some findings are given below.

$$\sum\sum(y_{ij} - \overline{y})^2 = 688.2; \quad \sum\sum(\hat{y}_{ij} - \overline{y})^2 = 213.6$$

1. Predict the production amount when the temperature is $25^\circ C$ and 90 minutes production time. State also the corresponding $\mathbf{x}_h$.
2. Construct a 99% prediction interval on y_h when the temperature is $25^\circ C$ and 90 minutes production time. Given that $\mathbf{x}_h^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_h = 0.8$.
3. Construct the ANOVA table and test the significance of regression at the significance level $\alpha = 0.05$.
4. Compute R^2_{Adj} and interpret what it means.
5. Estimate the change in the production amount at $38^\circ C$ for each additional hour of production time.
6. Test the hypothesis that the production–time slope decreases as the temperature increase. Use the significance level $\alpha = 0.01$. Assume that the standard error of the coefficient estimate is 0.35.

2.6.1 Relationship between F-test and t-test

We performed both F-test and t-test for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Are these hypothesis tests always provide the same conclusion? Table below summarizes the differences between the F-test and t-test for testing $H_0 : \beta_1 = 0$.

In fact, the t-statistic and the F-statistic enjoy a one-to-one relationship given by

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(n-k-1)R^2}{(1-R^2)\,k} \tag{2.40}$$

F-test	t-test
$F = \frac{SS_R/1}{SS_E/(n-2)}$	$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{\sqrt{s^2/SS_x}}$
More convenient for testing whether $\beta_1 = 0$	Equally convenient for testing whether β_1 equals to any hypothesized value including zero value, e.g. $\beta_1 = 3.25$.
The alternative hypothesis is always two-sided, e.g. $H_1 : \beta_1 \neq 1$	The alternative hypothesis can be either two-sided or one-sided.

Table 2.1: The differences between t-test and F-test a