STAT 3008
Applied Regression Analysis

Tutorial 2: Simple Linear Regression

LAI Chun Hei*

Department of Statistics, The Chinese University of Hong Kong

1 Model Assumption

To quantify the relationship between two factors, say X and Y, we have to at least assume the type of relationship they have, could it be linear, logarithm, quadratic, etc.

In the experiment, the factor Y is likely to be affected by the factor X. In usual experiment context, X and Y are known as independent and dependent factors respectively. But in linear regression, X and Y are known as predictor and response respectively. Now, we state our model assumption:

Simple Linear Regression Model
The model relating the two factors are assumed to be:

$$y_i = Y_{|X=x_i} = \beta_0 + \beta_1 x_i + e_i$$

where $\mathbb{E}(e_i) = 0$, $\mathrm{Var}(e_i) = \sigma^2$ and e_i's are i.i.d. Therefore, we have

$$\mathbb{E}(Y|X = x_i) = \beta_0 + \beta_1 x_i$$
$$\mathrm{Var}(Y|X = x_i) = \sigma^2$$

Remark 1.1. The values x_i here are known constants, instead of some realised observations from a random variable X. For regression of random predictors, you may refer to [1]. ■

Remark 1.2. It should be noticed that the response y_i is a random variable. The data set $\{x_i, y_i\}_{i=1}^n$ consists of realised values from the random variables. ■

2 Least Square Estimator

Essentially, we want to fit a straight line to the set of points on the Cartesian plane. However, there are many ways to define "good" in a fit. The simplest way is to consider the total vertical distance between the points and the line. The best fit line is therefore the line which minimises the distance, i.e. which minimises

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_i^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Minimising the total distance is equivalent to minimising the total squared distance. Hence, we have the Least Square Method. By elementary multivariate calculus and statistical concepts, the derivation performed in the lesson yields the following results.

*s1155002282@sta.cuhk.edu.hk

Least Squares Estimator

If we define the following notations,

$$\bar{x} = \frac{1}{n}\sum_i x_i \qquad \mathrm{SXX} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\,\bar{x}^2$$

$$\bar{y} = \frac{1}{n}\sum_i y_i \qquad \mathrm{SYY} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\,\bar{y}^2$$

$$\mathrm{SXY} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\,\bar{x}\,\bar{y}$$

then the estimators are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\,\bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{SXY}}{\mathrm{SXX}}$$

$$\hat{\sigma^2} = \frac{\sum_i \hat{e_i}^2}{n-2} = \frac{\mathrm{SYY} - \mathrm{SXY}^2 / \mathrm{SXX}}{n-2}.$$

Under our estimated model, we therefore have the fitted values given x_i and the residual, i.e. the difference between the fitted value and the realised value.

$$\hat{y}_i = \hat{\mathbb{E}}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1\,x_i$$

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1\,x_i$$

Remark 2.1. For the detailed derivation, please refer to the lecture notes. You should be familiar with their derivation as they may be tested in midterm and final exam. ■

Remark 2.2. $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as linear combinations of y_i:

$$\hat{\beta}_1 = \sum_i \left(\frac{x_i - \bar{x}}{\mathrm{SXX}}\right) y_i \qquad \text{and} \qquad \hat{\beta}_0 = \sum_i \left[\frac{1}{n} - \bar{x}\left(\frac{x_i - \bar{x}}{\mathrm{SXX}}\right)\right] y_i.$$

This is useful when deriving the distribution and consistency of the estimators in Exercise 4.1. ■

Remark 2.3. By the derivative condition of RSS with respect to β_0, we have

$$\sum_i \hat{e}_i = 0 \qquad \text{and} \qquad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\,\bar{x}\,. \qquad ■$$

Exercise 2.1. (2012 Fall Midterm #3) Use the simple linear regression model to fit a straight line on two data points: $(-2, 4), (-1, 3)$. What are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$?

Exercise 2.2. Show that

$$\sum_i x_i \, \hat{e_i} = 0 \qquad \text{and, therefore} \qquad \sum_i \hat{y_i} \, \hat{e_i} = 0.$$

Exercise 2.3. Show that $\hat{\sigma^2}$ is an unbiased estimator of σ^2, i.e.

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

3 Analysis of Variance(ANOVA)

It is natural that we are interested in whether our model assumption is correct. The basic question will be if Y is really related to X. Mathematically, this question is equivalent to asking whether β_1 is zero in our model. The most common way is the Analysis of Variance, which compares two models of different mean functions.

We want to test the following hypothesis:

$$H_0 : \mathbb{E}(Y|X=x) = \beta_0 \qquad \text{vs} \qquad H_1 : \mathbb{E}(Y|X=x) = \beta_0 + \beta_1 x.$$

Therefore, we need some test statistics related to these two hypothesis and, most importantly, of known distributions. Consider the residual sum of square for the two models. For H_0, we have

$$\hat{\mathbb{E}}(Y|X=x) = \hat{\beta_0} = \bar{y} \qquad \Rightarrow \qquad \mathrm{RSS}_{H_0} = \sum_i (y_i - \bar{y})^2 = \mathrm{SYY}\,.$$

On the other hand, for H_1, we have

$$\mathrm{RSS}_{H_1} = \sum_i \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1\, x_i \right) \right]^2 = \mathrm{SYY} - \frac{\mathrm{SXY}^2}{\mathrm{SXX}}.$$

Since we use more variables to fit the points in H_1, it must be true that $\mathrm{RSS}_{H_0} \geq \mathrm{RSS}_{H_1}$. Therefore, H_1 is valid only when $\mathrm{RSS}_{H0} \ggg \mathrm{RSS}_{H_1}$. For easy comparison, we define the Sum of Square due to Regression as

$$\mathrm{SSreg} = \mathrm{RSS}_{H_0} - \mathrm{RSS}_{H_1} = \frac{\mathrm{SXY}^2}{\mathrm{SXX}}.$$

Equivalently, H_1 is valid only when $\mathrm{SSreg} \ggg 0$.

3.1 Distributions of Estimators

To define how large is large, we need the distributions as well so that we can define "large" in a probabilistic sense. Under H_0, the distribution of the sum of squares are given below,

By simple algebra, we rewrite

$$\mathrm{SSreg} = \frac{\mathrm{SXY}^2}{\mathrm{SXX}} = \left[\sum_i \left(\frac{x_i - \bar{x}}{\sqrt{\mathrm{SXX}}} \right) y_i \right]^2 .$$

By Central Limit Theorem, we have

$$\sum_i \left(\frac{x_i - \bar{x}}{\sqrt{\mathrm{SXX}}} \right) y_i \rightsquigarrow \mathcal{N}(0, \sigma^2).$$

Therefore, we know $\mathrm{SSreg}\,/\sigma^2 \sim \chi^2(1)$. Also, it is known that $(n-2)\,\hat{\sigma^2}\,/\sigma^2 \sim \chi^2(n-2)$. The test statistic is thus given by

$$F = \frac{\frac{\mathrm{SSreg}\,/\sigma^2}{1}}{\frac{(n-2)\,\hat{\sigma^2}_{H_1}\,/\sigma^2}{n-2}} = \frac{\mathrm{SSreg}}{\hat{\sigma^2}_{H_1}} \sim F(1, n-2).$$

For significance level α, we reject H_0 if the p-value is smaller than $F_{1-\alpha}(1, n-2)$.

Remark 3.1. That $(n-2)\,\hat{\sigma^2}\,/\sigma^2 \sim \chi^2(n-2)$ is due to a fact that the degree of freedom is the number of values in the statistics that are free to vary. The detailed proof of the distribution of $(n-2)\,\hat{\sigma^2}\,/\sigma^2$ includes the use of quadratic forms, which is beyond the scope. Interested students may refer to [2]. ■

Remark 3.2. It should be noticed that the $\hat{\sigma^2}$ here is under the model in H_1. The reason of using this instead of $\hat{\sigma^2}$ under H_0 is examined in Exercise 3.1. ■

Exercise 3.1. What is the estimator of σ^2 under H_0? Explain why the use of it in the denominator makes no sense.

Exercise 3.2. Show that

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

3.2 ANOVA Table

Thanks to the result of Exercise 3.2, we have a neat and tidy representation of ANOVA, which is called the ANOVA table.

Source	df	SS	MS	F	p-value
Regression	1	SSreg	SSreg /1	SSreg / $\hat{\sigma^2}$	$P(F_{1,n-1} > F)$
Residual	$n-1$	RSS_{H_1}	$\hat{\sigma^2} = \text{RSS}/(n-2)$		
Total	$n-2$	SYY			

You should be extremely familiar with the above table because it appears in every midterm. We will practice this in Exercise 4.3 and 4.4.

Remark 3.3. Therefore, we can define the "Coefficient of Determination" to be

$$R^2 = \frac{\text{SSreg}}{\text{SYY}} \in [0, 1].$$

The realised value summarises the strength of relationship between the sampled response and predictors. ■

4 Intervals, Tests and Band

Besides testing the mean functions by ANOVA, we will also want to perform test on individual parameters. Therefore, we need the distributions of the estimator. We begin this section with an exercise.

Exercise 4.1. Prove that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and find their variance and asymptotic distribution.

4.1 Confidence Intervals and Tests for Intercept and Slope

With the distributions of the estimator and some facts in statistics, we can construct the test statistics from the distribution derived in Exercise 4.1.

Confidence Interval and Test for Intercept

If we want to test whether the intercept is a certain value β_0^*, i.e.

$$H_0 : \beta_0 = \beta_0^* \qquad \text{vs} \qquad H_1 : \beta_0 \neq \beta_0^*,$$

then the test statistic is

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0)} \sim t(n-2) \qquad \text{where} \qquad \text{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}}.$$

Therefore, for significance level α, we reject H_0 when $|t| > t_{1-\frac{\alpha}{2}}(n-2)$. Also, the $(1-\alpha)\times 100\%$ confidence interval of β_0 is given by

$$\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\,\text{se}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\,\text{se}(\hat{\beta}_0).$$

Confidence Interval and Test for Slope

Similarly, for the test of slope, i.e.

$$H_0 : \beta_1 = \beta_1^* \qquad \text{vs} \qquad H_1 : \beta_1 \neq \beta_1^*,$$

the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\text{se}(\hat{\beta}_1)} \sim t(n-2) \qquad \text{where} \qquad \text{se}(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{\text{SXX}}}.$$

Therefore, for significance level α, we reject H_0 when $|t| > t_{1-\frac{\alpha}{2}}(n-2)$. Also, the $(1-\alpha)\times 100\%$ confidence interval of β_1 is given by

$$\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\,\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\,\text{se}(\hat{\beta}_1).$$

Remark 4.1. Obviously, a test of zero slope, i.e.

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_1 : \beta_1 \neq 0,$$

is equivalent to testing

$$H_0 : \mathbb{E}(Y|X = x) = \beta_0 \qquad \text{vs} \qquad H_1 : \mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x,$$

which is our ANOVA F-test in Section 3. Therefore, they should give the same result. Mathematically, if we look at the t-statistics,

$$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{\text{SXX}}}$$

$$t^2 = \frac{\hat{\beta}_1{}^2}{\hat{\sigma^2}\,/\,\text{SXX}} = \frac{\hat{\beta}_1{}^2\,\text{SXX}}{\hat{\sigma^2}} = \frac{\text{SXY}^2}{\hat{\sigma^2}\,\text{SXX}} = \frac{\text{SSreg}}{\hat{\sigma^2}} = F.$$

In general, we have

$$F(1, m) = \frac{\chi^2(1)}{\chi^2(m)/m} = \frac{Z^2}{\chi^2(m)/m} = \left(\frac{Z}{\sqrt{\chi^2(m)/m}}\right)^2 = t(m)^2 \qquad \blacksquare$$

Exercise 4.2. Construct a 95% confidence interval for the slope from the data set $\{(1, 1), (4, 9), (10, 10)\}$, given $t_{0.975}(1) = 12.7062$. Bosco argues that the confidence interval you construct has a 95% probability of including the true slope. Explain whether he is correct.

Exercise 4.3. (2013 Fall Midterm #1) Fill in the missing values in the following tables of regression output from a data set of size 100.

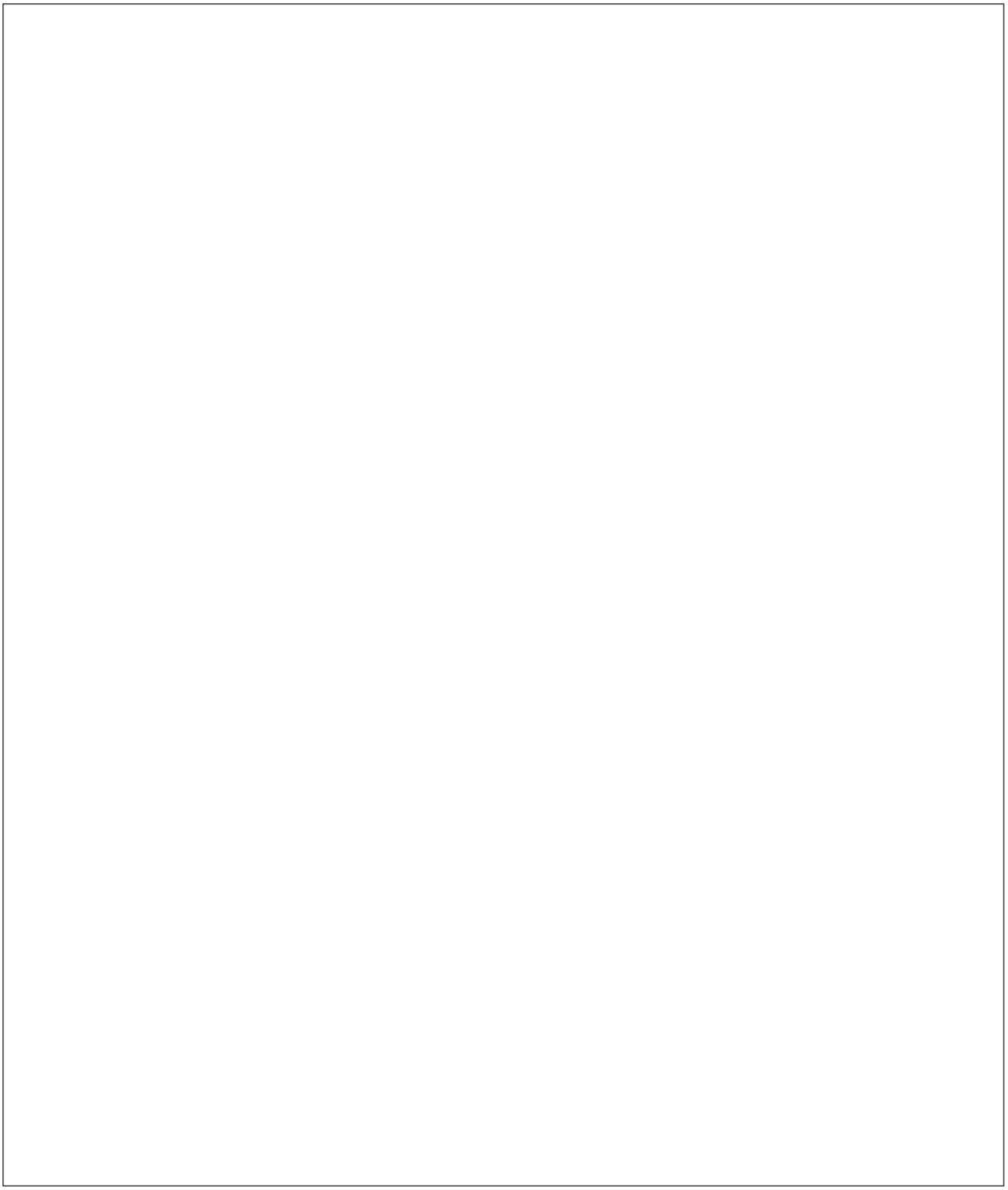
ANOVA Table				
Source	df	SS	MS	F
Regression				
Residual				
Total				

Coefficient Table				
Variable	Coefficient	s.e.	t-statistics	p-value
Constant	0.5854			0.2188
X		0.4927		
$n =$	$\hat{\sigma} = 4.714$	$R^2 = 0.03294$		

Exercise 4.4. (2012 Spring Midterm #1) Fill in the missing values in the following tables of regression output. In R, it is found that $qf(1-9.5e^{-9}, 1, 6) = 1917.3$. Also, $\bar{x} = 5.125, \bar{y} = -9.1974, \text{SXX} = 54.875$.

ANOVA Table					
Source	df	SS	MS	F	p-value
Regression	_____	_____	_____	_____	9.5e-09
Residual	_____	_____	_____		
Total	_____	_____			

Coefficient Table				
Variable	Coefficient	s.e.	t-statistics	p-value
Constant	_____	_____	_____	0.00322
X	-2.04245	_____	_____	
$n =$_____	$\hat{\sigma} =$_____	$R^2 =$_____		



4.2 Confidence and Prediction Intervals

Besides the true intercept and slope, we are often interested in the mean of response given a predictor x_*, i.e. $\mathbb{E}(Y|X = x_*)$. This can be constructed from the fitted value $\hat{y}_*$ since it is an unbiased estimator of the mean. For example, we consider the mean IQ of a student who scores 98 in the midterm.

On the other hand, instead of the mean, we may also be interested in the response itself , i.e. $y_{|X=x_*}$. In this case, we want to make a prediction on what the outcome will be, given x_*. Here, for example, we are looking for the IQ of a student who scores 98 in the midterm.

Therefore, we want to construct intervals for the mean and the prediction. The results are listed below.

Confidence Interval for Mean

Given a predictor value x_*, the true value and estimation of the mean are respectively

$$\mathbb{E}(Y|X = x_*) = \beta_0 + \beta_1 x_* \qquad \text{and} \qquad \hat{y_*} = \hat{\mathbb{E}}(Y|X = x_*) = \hat{\beta}_0 + \hat{\beta}_1\, x_*.$$

The estimation uncertainty of the mean is

$$\operatorname{Var}\left(\hat{y_*} - \mathbb{E}(Y|X = x_*)\right) = \operatorname{Var}(\hat{\beta}_0 + \hat{\beta}_1\, x_*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}}\right).$$

Define the standard error of fit as

$$\operatorname{sefit}(\hat{y_*}|x_*) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}}} \quad \Rightarrow \quad \frac{(n-2)\,\hat{\sigma^2}}{\sigma^2} = \frac{(n-2)\operatorname{sefit}(\hat{y_*}|x_*)^2}{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}}} \sim \chi^2(n-2).$$

Therefore, the $(1-\alpha)\times 100\%$ confidence interval for the mean is given by

$$\hat{y_*} - t_{1-\frac{\alpha}{2}}\operatorname{sefit}(\hat{y_*}|x_*) \le \mathbb{E}(Y|X = x_*) \le \hat{y_*} + t_{1-\frac{\alpha}{2}}\operatorname{sefit}(\hat{y_*}|x_*).$$

Prediction Interval for Response

Given a predictor value x_*, the response and its estimation are respectively

$$y_{|X=x_*} = \beta_0 + \beta_1 x_* + e \quad \text{and} \quad \hat{y_*} = \hat{\beta}_0 + \hat{\beta}_1\, x_*.$$

The estimation uncertainty of the prediction is

$$\operatorname{Var}\left(\hat{y_*} - y_{|X=x_*}\right) = \operatorname{Var}(\hat{\beta}_0 + \hat{\beta}_1\, x_* + e) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}}\right).$$

Define the standard error of prediction as

$$\operatorname{sepred}(\hat{y_*}|x_*) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\text{SXX}}} \quad \Rightarrow \quad \frac{(n-2)\,\hat{\sigma^2}}{\sigma^2} = \frac{(n-2)\operatorname{sefit}(\hat{y_*}|x_*)^2}{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}}} \sim \chi^2(n-2).$$

Therefore, the $(1-\alpha)\times 100\%$ confidence interval for the mean is given by

$$\hat{y_*} - t_{1-\frac{\alpha}{2}}\operatorname{sepred}(\hat{y_*}|x_*) \le y_{|X=x_*} \le \hat{y_*} + t_{1-\frac{\alpha}{2}}\operatorname{sepred}(\hat{y_*}|x_*).$$

Remark 4.2. The estimation uncertainty of prediction and mean differs only by σ^2. The extra uncertainty comes from the error term in the new observation that we wants to predict. Compare

$$y_{|X=x} = \beta_0 + \beta_1 x + e \qquad \text{and} \qquad \mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x.$$

Due to the extra uncertainty, prediction interval includes and is larger than the confidence interval for mean. ■

4.3 Confidence Band

In the previous subsection, we construct confidence interval of mean for a certain point x. It is tempting to connect all the upper limits and lower limits of confidence intervals, i.e.

$$(\hat{\beta}_0 + \hat{\beta}_1\, x) \pm t_{1-\frac{\alpha}{2}}\, \text{sefit}(\hat{y}|x), \quad \forall x$$

and say that this random band has a $(1-\alpha)\times 100\%$ probability of including the true mean line $\mathbb{E}(Y|X=x) = \beta_0 + \beta_1 x$. However, this is wrong (see Remark 4.3 and Exercise 4.5). The correct band is given below.

Confidence Band for Mean Function

The $(1-\alpha)\times 100\%$ confidence band of the mean function is given by

$$\mathcal{C}(x) = (\hat{\beta}_0 + \hat{\beta}_1\, x) \pm \sqrt{2F_{1-\alpha}(2, n-2)}\, \text{sefit}(\hat{y}|x), \quad \forall x.$$

Therefore, it is true that

$$\Pr(\text{The mean line lies in the confidence band}) = \Pr\Big(\forall x,\, \mathbb{E}(Y|X=x) \in \mathcal{C}(x)\Big) = 1-\alpha.$$

Remark 4.3. For confidence interval $C(x) = (\hat{\beta}_0 + \hat{\beta}_1\, x) \pm t_{1-\frac{\alpha}{2}}\, \text{sefit}(\hat{y}|x_*)$, we have by definition

$$\forall x,\, \Pr\Big(\mathbb{E}(Y|X=x) \in C(x)\Big) = 1-\alpha.$$

This relationship holds for each point, i.e. pointwise. While for the confidence band, we have

$$\Pr\Big(\forall x,\, \mathbb{E}(Y|X=x) \in \mathcal{C}(x)\Big) = 1-\alpha.$$

Here, the inclusion is for the entire line. The two cases are different. ■

Exercise 4.5. Explain, why it is wrong to say the band,

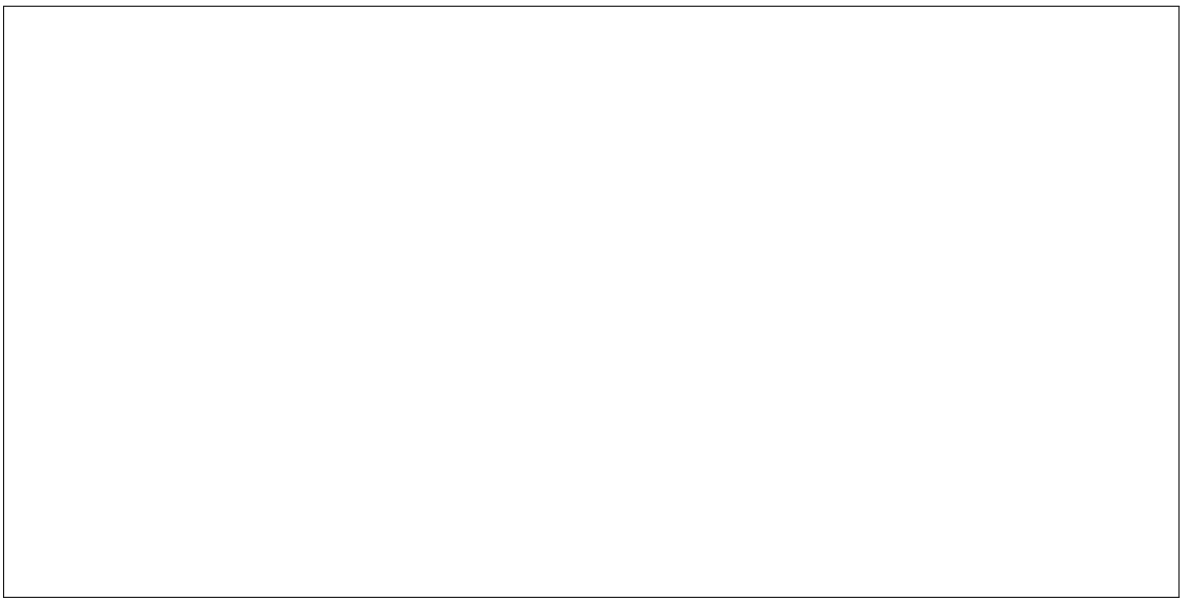
$$(\hat{\beta}_0 + \hat{\beta}_1\, x) \pm t_{1-\frac{\alpha}{2}}\, \text{sefit}(\hat{y}|x), \quad \forall x$$

has a $(1-\alpha)\times 100\%$ probability of including the mean line $\mathbb{E}(Y|X=x) = \beta_0 + \beta_1 x$.

Exercise 4.6. For the data set $\{(1,1),(4,9),(10,10)\}$, construct

1. a 95% confidence interval and a 95% interval for the point $x = 3$, and
2. a 95% confidence band.
3. What is the value of the band when $x = 3$?

You are given $t_{0.975}(1) = 12.7062$ and $F_{0.95}(2,1) = 199.5$.



5 Residuals

To check whether our model assumption is valid, a good way is to look at the residual plot. Recall that the residuals

$$\hat{e_i} = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i$$

so this gives

$$\mathbb{E}(\hat{e_i}) = \mathbb{E}(y_i) - \beta_0 - \beta_1 x_i = 0 \qquad \text{and} \qquad \mathrm{Var}(\hat{e_i}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\mathrm{SXX}}\right).$$

The data set gives a set of realised $\hat{e_i}$. According to our observation above, these realised residuals should

- have mean close to zero, and
- have constant variance for all value x_i.

A plot that satisfies the above criteria is a null plot, which indicates that the model assumption is valid and the regression is a good fit.

6 Appendix

For more reference, you may refer to the following text books.

References

[1] Douglas C. MONTGOMERY, Elizabeth A. PECK and G. Geoffrey VINING (2006). *Introduction to Linear Regression Analysis*, Wiley.

[2] Robert V. HOGG and Allen T. CRAIG *Introduction to Mathematical Statistics*, Pearson.