Week -3 Lecture

08 February 2024 10:53

Inference about the regression parameters
(Statistical Modelling I)

Dr Lubna Shaheen

Week 3, Lecture 2

Inference about the regression parameters

Outline

1. Standardised Residuals & Q-Q Plot
   - Exams Style Question
2. Inference
   - Confidence Interval
   - Test of Significance for Parameters
   - Prediction Intervals

ahw811@qmul.ac.uk.

office Hours → Thursday: 1:00 - 2:00pm
Learning Cafe

Standardised Residuals

Three useful plots

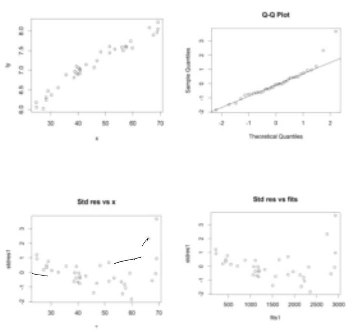
d_i against x_i	• Check whether a linear model is appropriate • Check the Normal assumptions
d_i against $\hat{y}_i$	• Check for constant variance • Called homoscedasticity
QQ plot in R	• Good first indication of Normal residuals • Looking for a straight line

Standardised Residuals

Exams Style Question (2022)

The thickness (x) and hardness (y) of 36 woods are plotted in the table below. We are interested in establishing the relationship between the y and x values. For these data, using R, we obtained the following output.

$y \propto x$.




```
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-417.10 -142.03  -13.83  103.70  814.42
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1298.282    139.496  -9.307  7.1e-11 ***
x              61.127      2.927  20.882  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 235.2 on 34 degrees of freedom
Multiple R-squared:  0.9277,Adjusted R-squared:  0.9255
F-statistic:   436 on 1 and 34 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0 = -1298.282$

$\hat{\beta}_1 = 61.127$

$\hat{y}_i = -1298.282 + 61.127 x_i$

$R^2 = 92.77\%$ Explain much of the variation.

$\log(\hat{y}_i) = \hat{c}_1 + \hat{c}_2 x_i$

Transfor (y) — $\log(y)$.

Shapiro-Wilk: H_0: Data has been generated from the Normal Distribution

$p \leq 0.05$

Reject Null Hypothesis

Standardised Residuals

MTH5120 (2022)

(a) Looking at the value of R^2 above, is this linear model a reasonable fit? ✓ **[3]**

(b) Viewing the residual plot, is there a possible problem with the constancy of variance? **[5]**

(c) Using the Q-Q plot and the Shapiro-Wilk test, check if there is a possible problem with the assumption of normality? **[4]**

(d) Looking at the plots above, is there any other transformation that you would like to consider? Give reasons for your answer. **[3]**

Normal assumption is composoed as $P < 0.05$.

Normal assumption is composed as $p < 0.05$.

$y_i \sim \log(y_i)$

Inference

Inference: A conclusion we reached on the basis of evidence and reasoning

Conclusions we would like to make:

1. Confidence Intervals for Parameters or the mean response
   - CI for β_1
   - CI for β_0
   - CI for $\hat{\mu}_0$
   - CI for $\hat{y}_0$
2. Tests of significance for parameters
   - Hypothesis testing using t-Distribution for β_0 and β_1
3. Prediction intervals for a new observation
   - Prediction interval for a new value

Standardisation of β_1

In the linear regression model the sampling distribution of the $\hat{\beta}_1$ of β_1 is normal with $E(\hat{\beta}_1) = \beta_1$ and var$(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, that is $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

We can standardized $\hat{\beta}_1$ by doing standardization, i.e

$$\hat{\beta}_1 - \beta_1 \sim N(0, \frac{\sigma^2}{S_{xx}})$$

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma^2}{S_{xx}}} \sim N(0, 1)$$

The variance usually is not known and it is replaced by its estimate then then Normal distribution changes to a **student t distribution**.

Applying Student t distribution

From Probability and Statistics II we have the followings

if $Z \sim N(0,1)$ and $U \sim \mathcal{X}_\nu^2$ and we have Z and U independent, then

$$\frac{Z}{\sqrt{\frac{U}{\nu}}} \sim t_\nu$$

where $\nu =$ degree of freedom.

We will see later that $U = \frac{(n-2)S^2}{\sigma^2} \sim \mathcal{X}_{n-2}^2$ and S^2 and $\hat{\beta}_1$ are independent. The student t distribution applies here and we have

$$T = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}\frac{1}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}$$

is distributed with (n-2) degrees of freedom.

Developing a confidence interval for β_1

This forms the basis for testing hypotheses and constructing confidence intervals for β_1.

We need to compute the CI for β_1 and to find a CI for unknown parameter θ means to find boundaries a and b such that

$|x| < a$

$-a < x < a.$

$$P(a < \theta < b) = 1 - \alpha$$

for some small values of α. If $\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}$ and we define $t_{\alpha/2}$ to be the quantity such that

$$P(|t_\nu| < t_{\frac{\alpha}{2}}) = 1 - \alpha$$

This gives

$$P\Big(-t_{\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} < t_{\alpha/2}\Big) = 1 - \alpha$$

Developing a confidence interval for β_1

$$P\left(-t_{\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} < t_{\alpha/2}\right) = 1-\alpha.$$

$$P\left(-t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}} < \hat{\beta}_1 - \beta_1 < t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right) = 1-\alpha.$$

$$P\left(-t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}} - \hat{\beta}_1 < -\beta_1 < -\hat{\beta}_1 + t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right) = 1-\alpha.$$

$$\|$$

$$P\left(\hat{\beta}_1 - t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right) = 1-\alpha.$$

Developing a confidence interval for $\hat{\beta}_1$

$$P(\hat{\beta}_1 - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}) = 1 - \alpha$$

Suppose we want to generate a 95% confidence interval estimate for $\hat{\beta}_1$. This means that there is a 95% probability that the confidence interval will contain the true value of β_1. Thus,

$$P(\text{[mean of estimate]} - \text{margin of error} < \beta_1 < \text{[mean estimate]} + \text{margin of error}) = 0.95 = 1-\alpha.$$

$$P(\hat{\beta}_1 - \underline{t_{0.025}}\frac{S}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{0.025}\frac{S}{\sqrt{S_{xx}}}) = 0.95$$

where we define $t_{0.025}$ to be the quantity such that

$$P(|t_\nu| < t_{0.025}) = 0.95$$

$$t_{\alpha/2} = t_{\alpha/2,\, n-2}$$

Confidence interval for $\hat{\beta}_1$

Confidence interval = $\Big\{$mean of estimate $\pm$ margin of error (the variation in that estimate)$\Big\}$

For a particular data set with $\hat{\beta}_1$ and S^2 calculated for that data

$$[a, b] = (\hat{\beta}_1 - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{xx}}})$$

Confidence interval for β_1

Comments: The confidence interval for β_1 based on $t_{\frac{\alpha}{2}}$, $\hat{\beta}_1$ and S^2. $t_{\alpha/2, n-2}$

1. $t_{\alpha/2}$: This also known as the critical value of t
2. $\hat{\beta}_1$: which in general is a random variable and
3. S^2: which depends on our observed data.

This means that it only makes sense to calculate the confidence interval given a particular set of observed data.

Remark: If the confidence interval (CI) does not contain null hypothesis value, then the results of β_1 are statistically significant.

Estimated Standard error of β_1

The estimate of the **standard error** is the square root of the estimated variance

$$\widehat{se}(\hat{\beta}_1) = \sqrt{\frac{S^2}{S_{xx}}}$$

$S/\sqrt{S_{xx}} = \sqrt{\frac{S^2}{S_{xx}}} = \widehat{se}(\hat{\beta}_1)$

We can then re-frame the confidence interval and the test statistic for β_1 in terms of this estimated standard error

[a,b]= $[\hat{\beta}_1 - t_{\frac{\alpha}{2}}\widehat{se}(\hat{\beta}_1), \hat{\beta}_1 + t_{\frac{\alpha}{2}}\widehat{se}(\hat{\beta}_1)]$ and $T = \frac{\hat{\beta}_1}{\widehat{se}(\hat{\beta}_1)} \sim t_{n-2}$

- 68% Confidence Interval: $\beta_1 \pm 1 \times \widehat{se}(\hat{\beta}_1)$
- 95% Confidence Interval: $\beta_1 \pm 2 \times \widehat{se}(\hat{\beta}_1)$
- 99% Confidence Interval: $\beta_1 \pm 3 \times \widehat{se}(\hat{\beta}_1)$

The confidence interval provides you with a set of plausible values for the parameters

Confidence interval for β_1

Example:

Using the R, we obtained the following output.

```
> mody <- lm(y ~ x)
> summary(mody)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-67.022 -31.346  -0.631  33.654  54.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  429.048     26.519  16.179 1.69e-08 ***
x             18.244      5.643   3.233  0.00898 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.2 on 10 degrees of freedom
Multiple R-squared:  0.511,     Adjusted R-squared:  0.4621
```

1st 2nd

$\hat{\beta}_0$ $\hat{\beta}_1$

$S/\sqrt{S_{xx}} = \widehat{se}(\beta_1) = 5.643$

$\hat{\beta}_0 = 429.048$

$\hat{\beta}_1 = 18.244$

$t_{\alpha/2, n-2}$

$\widehat{se}(\beta_1) = 5.643$ $\widehat{se}(\beta_0) = 26.519$

Confidence interval for β_1

$\alpha =$ two tail

$\alpha/2 =$ one-tail.

$t_{\alpha/2, n-2}$

```
F-statistic: 10.45 on 1 and 10 DF,  p-value: 0.008979

> anova(mody)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1  16059 16058.9  10.451 0.008979 **
Residuals 10  15366  1536.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$n = 12$

95% CI for β_1 $\quad \hat{\beta}_1 = 18.244$.

$$\left(\hat{\beta}_1 - t_{\alpha/2}\, S/\sqrt{S_{xx}},\ \hat{\beta}_1 + t_{\alpha/2} \frac{S}{\sqrt{S_{xx}}}\right)$$

$$\left(\hat{\beta}_1 - t_{\alpha/2}\, \hat{se}(\beta_1),\ \hat{\beta}_1 + t_{\alpha/2}\, \hat{se}(\beta_1)\right)$$

Use the R output above to answer the questions below.

(a) By looking at the summary output, write down the fitted model. [2]

(b) Write down the formula to compute the 95% confidence interval for β_1? [3]

(c) Compute the 95% confidence interval for β_1. [4]

$$18.244 \pm (2.2284)(5.643)$$

$$= [5.6706, 30.82]$$

(d) Fill in the blanks in the following table. [5]

Source of Variation	D F	Sum of Squares	Mean Square	F Value
Regression	1	SSR = 16059	MSR = 16058.9	?
Residual	12 - ? = ?	SSE = ?	MSE = $\frac{15366}{?}$ = ?	

$t_{\alpha/2, n-2} = t_{0.025, 10} = 2.2284$ $\qquad \alpha = 0.05$.

Confidence interval of β_1

Developing the test statistics

Last week we used the ANOVA table and F statistics to test the null hypothesis $H_0 : \beta_1 = 0$.

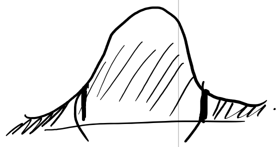
Now that we have a confidence interval for β_1 there is another way to test this same hypothesis.

We have already seen $T = \dfrac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$

Developing the test statistics

Now under $H_0 = \beta_1 = 0$ this test statistics becomes

T_{cal}



$$T = \frac{\hat{\beta}_1}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

which we can calculate for any particular data set. We then reject H_0 if

$$|T| > t_{n-2,\frac{\alpha}{2}}$$

This is mathematically equivalent to the F statistics test

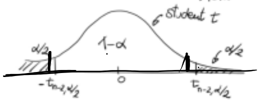
P-values:

1. A p-value is a statistical measurement used to validate a hypothesis against observed data.
2. Small p-values are evidence against the null hypothesis.
3. A p-value of 0.05 or lower is generally considered statistically significant.

Confidence Interval and Student t Distribution

CI for β_1

$$P\left(-t_{n-2,\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} < t_{n-2,\alpha/2}\right) = 1-\alpha$$



Remarks: Looking at the confidence interval. If the hypothesized value is outside the confidence interval you reject the null hypothesis.

Notice that this is equivalent to the t-test. An absolute value for t greater than 2 implies that the proposed value is outside the confidence interval therefore reject. In fact, a 95% confidence interval contains all the values for a parameter that are not rejected by hypothesis test with a false positive rate of 5%

$\hat{\beta}_1 + 2se(\hat{\beta}_1)$ — 95% values.

Confidence interval for β_0

Because we are usually employing statistical modelling to better understand the relationship between Y and X, we are generally more interested in β_1 than β_0 However, we can also develop confidence intervals and test hypotheses for β_0 Last week we found the sampling distribution for β_0

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}))$$

We can now use the same methodology with β_0 as earlier for β_1

Confidence interval for β_0

The $100(1-\alpha)\%$ confidence interval for β_0 is

$$[a,b] = [\hat{\beta}_0 - t_{\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}, \hat{\beta}_0 + t_{\frac{\alpha}{2}}\widehat{se(\hat{\beta}_0)}]$$

where

$$\widehat{se(\hat{\beta}_0)} = \sqrt{S^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})}$$

Test statistics for β_0

The test statistic to test the null hypothesis $H_0 : \beta_0 = B$ for some value B (which may or may not be zero) is

$$T = \frac{\hat{\beta}_0 - B}{\widehat{se(\hat{\beta}_0)}} \sim t_{n-2}$$

$H_0: \beta_0 = 0$

Confidence interval for β_0

$\left(\hat{\beta}_0 - 2.131 \times 8.321815,\ \hat{\beta}_0 + 2.131 \times 8.3281\right)$

$T_{cal} = 17.98$

Exercise

The following are the R output of the data givings the one-way airfare (in US dollars) and distance (in miles) from city A to 17 other cities in the US.

- Write down the formula to compute 95% confidence interval for β_0?
- Compute the 95% confidence interval for β_0.

Regression Output from R

The least squares estimates for the production data were calculated using R, giving the following results:

```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 149.74770   8.32815   17.98  6.00e-13 ***
RunSize       0.25924   0.03714    6.98  1.61e-06 ***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-Squared: 0.7302,      Adjusted R-squared: 0.7152
F-statistic: 48.72 on 1 and 18 DF, p-value: 1.615e-06
```

$\hat{\beta}_1$

$n = 17.$

$1 - \alpha = 0.95$

$\alpha = 0.05$

$t_{n-2, \alpha/2} =$

$t_{15, \frac{0.05}{2}} = 2.131$

Confidence interval for β_0

Confidence interval for the mean response μ_i

We can also develop confidence intervals and test hypotheses for the mean response, that is for $E[Y_i|X_i = x_i]$ which is often written μ_i.

Under the simple linear regression model,

$$\mu_i = E[Y_i|X_i = x_i] = \beta_0 + \beta_1 x_i$$

and μ_i is estimated by least squares at a particular value of x_i as

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sampling distribution for μ_i

Under the simple linear regression model, the sampling distribution of μ_i is also normal

$$\hat{\mu}_i \sim N(\mu_i, \sigma^2(\frac{1}{n} + \frac{x_i - \overline{x}^2}{S_{xx}}))$$

which leads to a $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_i$ of

$$[a, b] = \left[\hat{\mu}_i - t_{\frac{\alpha}{2}} \widehat{se(\hat{\mu}_i)}, \hat{\mu}_i + t_{\frac{\alpha}{2}} \widehat{se(\hat{\mu}_i)}\right]$$

$$\widehat{se}(\hat{\mu}_i) = s\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

Test statistics for the mean response μ_i

where $\widehat{se\hat{\mu}_i} = \sqrt{s^2(\frac{1}{n} + \frac{x_i - \bar{x}^2}{s_{xx}})}$
we can test the null hypothesis, $H_0 : \mu_i = M$ for some value M (which is not necessary zero), with the test statistics

$$T = \frac{\hat{\mu}_i - M}{\widehat{se(\hat{\mu}_i)}} \sim t_{n-2}$$

A note of caution

1. For the estimation of the mean response to be valid, The value of x_i used should be within the range of observed values for X
2. The model has said nothing about the applicability of linear regression outside of this range for x_i
3. **We should not use inference about μ_i as a method of extrapolation**

Extrapolation: The action of estimating or concluding something by assuming that existing trends will continue or a current method will remain applicable.

However we can now turn to using the model to predict the response value for some new value of x_i for which y_i has not yet been observed.

Prediction Interval for a new observation

Motivation:

- Simple linear regression models can be used to predict the response at any given value of the predictor
- **Beware predicting far beyond the range of the data**
- Point predictions should be accompanied by corresponding prediction intervals, providing a range of plausible values
- Suppose that we want to predict $y_0 = y(x_0)$ when the predictor x takes the value x_0
- Note that predicting a response is about estimating the value of a random variable, say y_0, and not the value of a parameter, say μ_0

Prediction Interval for a new observation

For a simple linear regression:

- The standard deviation of the sampling distribution of $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is

$$\sigma_{\hat{\mu}_0} = \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

- The standard error estimate for $\hat{\mu}_0$ is

$$se(\hat{\mu}_0) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}}{S_{xx}}}$$

- It can be shown that the sampling distribution of $\hat{\mu}_0$ is defined by

$$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_{n-2}$$

- we can use a linear regression model to predict the response value for some new value of x_i for which y_i has not yet been observed
- This is called a **Prediction Interval** sometimes just PI for a new observation

Prediction interval for a new observation

The prediction was

- $$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon_0$$
- But $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is also the natural estimate of $E(y_0) = \mu_0$ and $E(\epsilon_0) = 0$.
- Hence $\hat{y}_0$, the predicted value of y_0 is the same as $E(\hat{y}_0) = \hat{\mu}_0$, the estimated mean response of $E(y_0)$.
- However, difference arise if we want to construct corresponding confidence intervals.
- We seek $y_0 = \mu_0 + \epsilon_0$. The " point prediction" would be $\hat{y}_0 = \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- We know that
$$\hat{\mu}_0 \sim N(\mu_0, \sigma^2(\frac{1}{n} + \frac{x_0 - \overline{x}^2}{s_{xx}}))$$
- Therefore the distribution of $\hat{\mu}_0 - \mu_0$ is $\hat{\mu}_0 - \mu_0 \sim N(0, \sigma^2(\frac{1}{n} + \frac{x_0 - \overline{x}^2}{S_{xx}}))$

From μ_0 to y_0

- But rather than $\hat{\mu}_0 - \mu_0$ we would prefer the distribution of $\hat{y}_0 - y_0$
- If we add and subtract ϵ_0 to the distribution equation for $\hat{\mu}_0 - \mu_0$ we have $\hat{\mu}_0 - \mu_0 = \hat{\mu}_0 - (\mu_0 + \epsilon_0) + \epsilon_0$
$$= \hat{y}_0 - y_0 + \epsilon_0 \sim N(0, \sigma^2(\frac{1}{n} + \frac{x_0 - \overline{x}^2}{S_{xx}}))$$
- But we know that $\epsilon_0 \sim N(0, \sigma^2)$ from the original model definition, so
$$\mathbf{\hat{y}_0 - y_0 \sim N(0, \sigma^2(1 + \frac{1}{n} + \frac{x_0 - \overline{x}^2}{s_{xx}}))}$$

From distribution to PI

To get the prediction interval we need to:

1. standardise the normal distribution
2. replace the unknown variance σ^2 with its estimator S^2

1. leads to $\dfrac{\hat{y}_0 - y_0}{\sqrt{\sigma^2(1 + \frac{1}{n} + \frac{x_0 - \overline{x}^2}{s_{xx}})}} \sim N(0,1)$

2. gives us $\dfrac{\hat{y}_0 - y_0}{\sqrt{s^2(1 + \frac{1}{n} + \frac{x_0 - \overline{x}^2}{s_{xx}})}} \sim t_{n-2}$

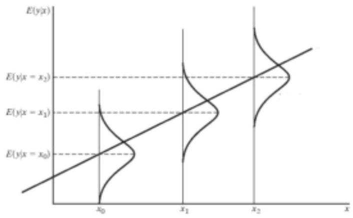
Prediction interval for y_0

The $100(1-\alpha)\%$ prediction interval for y_0 is then

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}} \sqrt{S^2(1 + \frac{1}{n} + \frac{x_0 - \overline{x}^2}{S_{xx}})}$$

Note the prediction interval for y_0 is usually much wider than the confidence interval for μ_0 because the random variability term ϵ_0 impacts the PI.

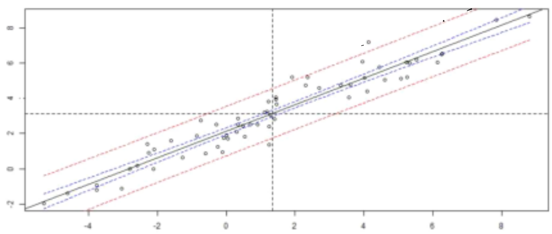
Confidence Interval and Prediction interval plot



1. μ_0 is a fixed mean parameter of normal curve when $x = x_0$ and Y_0 is a random variable around the normal curve with mean μ_0, i,e $y_0 \sim N(\mu_0, \sigma^2)$ with variability due to σ^2.
2. Different samples give different regression lines with slightly different $\hat{\mu}_0$
3. $\hat{\mu}_0$ has variability due to sampling distribution. $\hat{y}_0$ has two sources of variability from the model distribution and the sampling distribution of $\hat{\mu}_0$
4. **this explain why the PI for $\hat{y}_0$ is wider than the CI for the mean $\hat{\mu}_0$**

Confidence Interval and Prediction interval plot

Confidence interval for mean response and prediction.



Note that the CI for mean response is narrower in the middle and is much narrower than the PI for predicted response.

Confidence Interval Versus Prediction Interval

Example: For the given small data set calculate the 90% confidence interval and prediction intervals for the response when $x^* = 1$ given the simple linear regression line have equation

$$\hat{y} = 1.5 + 1.5x, \quad s^2 = 0.25 \quad \text{and} \quad \sum (x_i - \bar{x})^2 = 2 = S_{xx}.$$

$n=4$

x	y
1	3
2	4
2	5
3	6

$\hat{y} = 1.5 + 1.5(1) = 3.$ Point estimate.

$\hat{y} \pm t^* s_{\hat{y}}$

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$= 0.5\sqrt{\frac{1}{4} + \frac{(1-2)^2}{2}}$$

$$= 0.433$$

$t_{0.1,2} = 2.920$

$t_{\alpha/2, n-2}$

CI

90% CI $\quad 3 \pm 2.9290(0.433) = (1.74, 4.26)$.

90% PI $\quad \hat{y} \pm t^* \sqrt{s^2 + s_{\hat{y}}^2} = 3 \pm 2.9290\sqrt{0.687}$

PI.

$= (0.5825, 5.42)$

PI.

Critical values of *t* for two-tailed tests

Significance level (α)

Degrees of freedom (*df*)	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850