Math 4501 - Probability and Statistics II

7.4 - Sample size

Overview: confidence intervals for a proportion

We will see how to select the sample size in order to estimate an unknown proportion to a given degree of accuracy.

↑
maximum error E.

Confidence intervals for a proportion

Problem: Denote by $\hat{p} = \frac{y}{n}$ the proportion of observed successes in n trials and choose the sample size n so that the $100(1-\alpha)\%$ confidence interval for p

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is not longer than $\bar{x} \pm \varepsilon$, for some fixed $\varepsilon > 0$.

Partial solution: Given the maximum error ε, we would like to solve for n the inequality

$$\frac{z_{\alpha/2}\sqrt{\hat{p}\,(1-\hat{p})}}{\sqrt{n}} \leq \varepsilon$$

Difficulty: we do not know what $\hat{p} = \frac{y}{n}$ actually is until we perform the experiment, and we first need to decide what the sample size n needs to be.

First solution: If, based on past experience, we know of a reasonable approximate p^* for the unknown proportion p, then we can set $\hat{p} = p^*$ in the maximum error of the point estimate to get

$$\frac{z_{\alpha/2}\sqrt{p^*\left(1-p^*\right)}}{\sqrt{n}} \leq \varepsilon \ .$$

Solving for n we get

$$n \geq \frac{z_{\alpha/2}^2 p^*\left(1-p^*\right)}{\varepsilon^2} \ .$$

and then take n at least as large as the smallest integer greater than the quantity on the right hand side of the inequality above.

Second solution: Solve

$$\frac{z_{\alpha/2}\sqrt{\hat{p}\,(1-\hat{p})}}{\sqrt{n}} \leq \varepsilon\,.$$

for n to get

$$n \geq \frac{z_{\alpha/2}^2\,\hat{p}\,(1-\hat{p})}{\varepsilon^2}\,.$$

Note that

$$\hat{p}\,(1-\hat{p}) \leq \frac{1}{4}$$

and use the the maximum value of $1/4$ to get

$$n \geq \frac{z_{\alpha/2}^2}{4\varepsilon^2}\,.$$

$\hat{p}(1-\hat{p}) = \hat{p} - \hat{p}^2$

Maximum occurs when $\hat{p} = 1/2$ and is equal to $\frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}$

worst possible case

choose n such that $n \geq \frac{z_{\alpha/2}^2}{4\varepsilon^2}$

Take n at least as large as the smallest integer greater than the quantity on the right hand side of the inequality above.

Note: this second approach is easier to implement, but yields a larger sample size than the first due to the lack of information regarding $\hat{p}$.

Example

Suppose it is known that the unemployment rate on a certain country has been about 8%

An updated estimate is needed. How large should the sample be so that a 99% confidence interval provides an estimate for the actual unemployment rate to within 0.001?

Let y be the number of unemployed people in a sample of size n.

Using that $z_{0.005} = 2.576$, we find that the interval is given by

$$\frac{y}{n} \pm 2.576\sqrt{\frac{(y/n)(1-y/n)}{n}}\,.$$

The margin of error is

$$2.576\sqrt{\frac{(y/n)(1-y/n)}{n}} \approx 2.576\sqrt{\frac{(0.08)(0.92)}{n}}\,,$$

where the right hand side is obtained using the knowledge that the unemployment rate is close to 8%.

Solving

$$2.576\sqrt{\frac{(0.08)(0.92)}{n}} \leq 0.001$$

for n, we get

$$n \geq 488,394\ ,$$

which is a very large sample size...

To bring the sample to a more manageable size, we can:

- decrease the confidence level (from 99% to, for instance, 98%)
- increase the allowed error (from 0.001 to, for instance, 0.01)

Such choices would yield (note that $z_{0.01} = 2.236$):

$$2.236\sqrt{\frac{(0.08)(0.92)}{n}} \leq 0.01$$

and

$$n \geq 3,982\ ,$$

a much more reasonable sample size.

Example

Without access to any additional information, how large should a sample be so that a 95% confidence interval for a proportion provides an estimate of the true proportion to within 0.03?

$\alpha = 0.05$

→ max error $\varepsilon = 0.03$

Since we have no information about the approximate value of the true proportion, we require that

$$n \geq \frac{z_{\alpha/2}^2}{4\varepsilon^2}.$$

endpoints for confidence interval

$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\Downarrow$

$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \varepsilon$

$\Rightarrow n \geq \frac{z_{\alpha/2}^2\,\hat{p}(1-\hat{p})}{\varepsilon^2}$

Since $\hat{p}$ is unknown → $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$

Take $n \geq \frac{z_{\alpha/2}^2}{4\varepsilon^2}$

Setting $\alpha = 0.05$ (and so $z_{0.025} = 1.96$) and $\varepsilon = 0.03$, we get

$$n \geq \frac{(1.96)^2}{4(0.03)^2} \approx 1067.11$$

and choose to take a sample of size at least 1068.

Math 4501 - Probability and Statistics II

7.6 - Confidence intervals for regression parameters

(related with Sec. 7.1)

based mostly on finding confidence intervals for the mean.

Simple linear regression: review

Problem: Given the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estimate the parameters α and β of the linear model

$$E[Y|x] = \alpha + \beta(x - \bar{x})$$

Standing assumption: for each particular value of x, the value of Y differs from its mean by a random amount $\varepsilon \sim N(0, \sigma^2)$.

Consequence: For each $i = 1, 2, \ldots, n$, we have

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i,$$

where $\alpha + \beta(x_i - \bar{x})$ are nonrandom and $\varepsilon_i \sim N(0, \sigma^2)$ are independent (with unknown variance σ^2).

The random variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent normal variables

$$Y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2), \quad i = 1, 2, \ldots, n$$

Proposition (Maximum likelihood estimators of α, β, and σ^2)

[From Chp. 6]

Under the conditions described above, the maximum likelihood estimators of α, β and σ^2 are given by:

$$\widehat{\alpha} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

→ linear combination of normally distr. r.v.

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \widehat{\alpha} - \widehat{\beta}\left(x_i - \bar{x}\right)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{Y}_i\right]^2$$

where $\hat{Y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$

Proposition (Distributions of $\widehat{\alpha}$ and $\widehat{\beta}$) [From Chp. 6]

Under the conditions described earlier, we have that:

1) $\widehat{\alpha}$ *is normally distributed with mean* α *and variance* $\frac{\sigma^2}{n}$. i.e. $\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{n}\right)$

2) $\widehat{\beta}$ *is normally distributed with mean* β *and variance*

$$\mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

i.e. $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

Proof Chp. 6

Sec. 7.1

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\mu \in \bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$

Goal: to find confidence intervals for α and β !!!

BUT α and β are just the unknown means of $\hat{\alpha}$ and $\hat{\beta}$

Distribution of $\widehat{\sigma^2}$

Proposition [Proof is on the notes in Blackboard : very similar to the proof of $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$]

Under the conditions described earlier, we have that:

1) $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$ are mutually independent; → analogue to $\bar{x}$ and s^2 independent!

2) $\frac{n\widehat{\sigma^2}}{\sigma^2}$ has a $\chi^2(n-2)$ distribution.

compare with

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \overbrace{(\hat{\alpha}+\hat{\beta}(x_i-\bar{x}))}^{\hat{y}_i}\right]^2$$

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

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n}\left[\frac{y_i-(\hat{\alpha}+\hat{\beta}(x_i-\bar{x}))}{\sigma}\right]^2 \sim \chi^2(n-2)$$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

-1 due to estimate $\bar{x}$ for μ

$\hat{\alpha}$ and $\hat{\beta}$ are two estimated parameters

Chp 5 $\Rightarrow x_i \sim N(\mu,\sigma^2) \Rightarrow \frac{x_i-\mu}{\sigma} \sim N(0,1)$

$\Rightarrow \left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2(1) \Rightarrow \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$

Confidence interval for α ⇝ parameter of linear regression

We have seen that

1) $\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{n}\right)$ with σ^2 unknown (α is the unknown mean from a normal distribution with unknown variance)

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

3) $\hat{\alpha}$ and $\widehat{\sigma^2}$ are mutually independent

Then

$$Z = \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{\sigma^2}{n}}} \quad \text{is} \quad N(0,1)$$

and

$$U = \frac{n\widehat{\sigma^2}}{\sigma^2} \quad \text{is} \quad \chi^2(n-2)$$

and so

$$T_\alpha = \frac{Z}{\sqrt{U/(n-2)}} \quad \text{is} \quad t(n-2)\,.$$

Analogue to Sec. 7.1

$n-2$ is the # of d.o.f. of $U \sim \chi^2(n-2)$

Note that

$$T_\alpha = \frac{Z}{\sqrt{U/(n-2)}} = \frac{\frac{\widehat{\alpha}-\alpha}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2(n-2)}}}$$

unknown σ^2 cancels out

can be rewritten as

$$T_\alpha = \frac{\widehat{\alpha}-\alpha}{\sqrt{\frac{\widehat{\sigma^2}}{n-2}}} \, .$$

Set

$$S_{\hat{\alpha}} = \sqrt{\frac{\widehat{\sigma^2}}{n-2}} \, .$$

then $\boxed{T_\alpha = \dfrac{\hat{\alpha}-\alpha}{S_{\hat{\alpha}}} \sim t(n-2)}$

Pick $t_0 = t_{\gamma/2}(n-2)$ and note that

$$P\left(-t_0 \leq \frac{\widehat{\alpha}-\alpha}{S_{\hat{\alpha}}} \leq t_0\right) = 1-\gamma$$

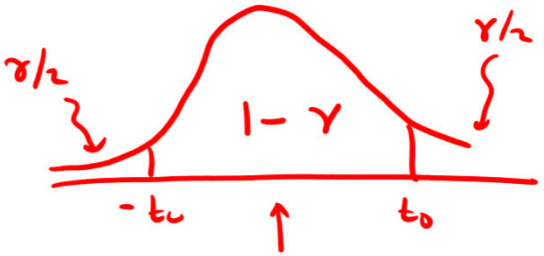
can be rewritten as

$$P\left(\widehat{\alpha} - t_0 S_{\hat{\alpha}} \leq \alpha \leq \widehat{\alpha} + t_0 S_{\hat{\alpha}}\right)$$

We conclude that

$$[\widehat{\alpha} - t_0 S_{\hat{\alpha}}, \widehat{\alpha} + t_0 S_{\hat{\alpha}}]$$

is a $100(1-\gamma)\%$ confidence interval for α.



confidence level is $1-\gamma$

where $S_{\hat{\alpha}} = \sqrt{\frac{\hat{\sigma}^2}{n-2}}$

Confidence interval for β ← the other parameter of the linear regression

We have seen that

1) $\hat{\beta} \sim N\left(\beta, \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$ with σ^2 unknown

[we want a confidence interval for the mean β of a normal distribution with unknown variance]

↑ suggest the use of t distr.

2) $\dfrac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$

3) $\widehat{\beta}$ and $\widehat{\sigma^2}$ are mutually independent

standardization

Then

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \quad \text{is} \quad N(0,1)$$

and

$$U = \frac{n\widehat{\sigma^2}}{\sigma^2} \quad \text{is} \quad \chi^2(n-2)$$

and so

$$T_\beta = \frac{Z}{\sqrt{U/(n-2)}} \quad \text{is} \quad t(n-2)\,.$$

# of dof of U

Note that

$$T_\beta = \frac{Z}{\sqrt{U/(n-2)}} = \frac{\frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}}}{\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2(n-2)}}}$$

can be rewritten as

$$T_\beta = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^n (x_i-\bar{x})^2}}}\,.$$

Set

$$S_{\hat{\beta}} = \sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sum_{i=1}^n (x_i-\bar{x})^2}}$$

← depends only on known quantities

so that $T_\beta = \dfrac{\hat{\beta}-\beta}{S_{\hat{\beta}}} \sim t(n-2)$

Pick $t_0 = t_{\gamma/2}(n-2)$ and note that

$$P\left(-t_0 \leq \frac{\widehat{\beta}-\beta}{S_{\hat{\beta}}} \leq t_0\right) = 1-\gamma$$

confidence level $1-\gamma$

solve for β

can be rewritten as

$$P\left(\widehat{\beta} - t_0 S_{\hat{\beta}} \leq \beta \leq \widehat{\beta} + t_0 S_{\hat{\beta}}\right)$$

We conclude that

$$\left[\widehat{\beta} - t_0 S_{\hat{\beta}}, \widehat{\beta} + t_0 S_{\hat{\beta}}\right]$$

is a $100(1-\gamma)\%$ confidence interval for β.

when $S_{\hat{\beta}} = \sqrt{\dfrac{n\,\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

Confidence interval for $\mu(x) = \alpha + \beta(x - \bar{x})$

Since

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}(x - \bar{x})$$

$\mu(x) = E[Y|x]$

think of $\hat{y}$ as the point estimator for $\mu(x) = E[Y|x]$ just like $\bar{x}$ is for μ or $\hat{p}$ is for p

is a linear combination of the independent and normally distributed random variables $\widehat{\alpha}$ and $\widehat{\beta}$, $\widehat{Y}$ is also normally distributed.

The mean of $\widehat{Y}$ is

out of $\hat{y}$ — $E[\hat{\alpha}] + (x - \bar{x})E[\hat{\beta}] =$

$$E[\widehat{Y}] = E[\widehat{\alpha} + \widehat{\beta}(x - \bar{x})] = \alpha + \beta(x - \bar{x})$$

$\mu(x)$

and its variance is

$\text{Var}(\hat{y})$ — independence of $\hat{\alpha}$ and $\hat{\beta}$

$$\begin{aligned}\text{Var}(\widehat{Y}) &= \text{Var}[\widehat{\alpha} + \widehat{\beta}(x - \bar{x})] = \text{Var}(\widehat{\alpha}) + (x - \bar{x})^2 \text{Var}(\widehat{\beta}) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(x - \bar{x})^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\end{aligned}$$

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

σ^2 is unknown!!

Since

$$Z = \frac{\widehat{\alpha} + \widehat{\beta}(x - \bar{x}) - [\alpha + \beta(x - \bar{x})]}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim N(0, 1)$$

and

$$U = \frac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n - 2) \,,$$

independence $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$, allow us to conclude that

$$T_\mu = \frac{Z}{\sqrt{U/(n-2)}} = \frac{\frac{\widehat{\alpha}+\widehat{\beta}(x-\bar{x})-[\alpha+\beta(x-\bar{x})]}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}}{\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sigma^2}}}$$

$$= \frac{\widehat{\alpha} + \widehat{\beta}(x - \bar{x}) - [\alpha + \beta(x - \bar{x})]}{\sqrt{\frac{n\widehat{\sigma^2}}{n-2}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

has a $t(n - 2)$ distribution.

Set $t_0 = t_{\gamma/2}(n-2)$ and

$$S_\mu = \sqrt{\frac{n\widehat{\sigma^2}}{n-2}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

so that $T_\mu = \dfrac{\hat{y} - (\alpha + \beta(x-\bar{x}))}{S_\mu}$ (unknown mean)

and note that

$$P(-t_0 \le T_\mu \le t_0) = 1-\gamma$$

can be rewritten as

$$P\left(\widehat{\alpha} + \widehat{\beta}(x-\bar{x}) - t_0 S_\mu \le \alpha + \beta(x-\bar{x}) \le \widehat{\alpha} + \widehat{\beta}(x-\bar{x}) + t_0 S_\mu\right)$$

($\widehat{\alpha} + \widehat{\beta}(x-\bar{x}) = \hat{y}$, $\widehat{\alpha} + \widehat{\beta}(x-\bar{x}) = \hat{y}$)

We conclude that

$$\left[\widehat{\alpha} + \widehat{\beta}(x-\bar{x}) - t_0 S_\mu, \widehat{\alpha} + \widehat{\beta}(x-\bar{x}) + t_0 S_\mu\right] \rightarrow \hat{y} \pm t_0 S_\mu$$

is a $100(1-\gamma)\%$ confidence interval for $\mu(x) = \alpha + \beta(x-\bar{x})$.

Note: Since S_μ depends on x, so does the width of the confidence interval constructed above.

Prediction interval

Problem: Having used the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to estimate α, β, and σ^2, we now want to estimate the value of Y corresponding to a given a value of $x = x_{n+1}$.

$\mu(x) = E[Y|x] = \alpha + \beta(x - \bar{x})$ → estimate is $\hat{y} = \hat{\alpha} + \hat{\beta}(x - \bar{x})$

A point estimate of the corresponding value of Y is given by

$$\widehat{y}_{n+1} = \widehat{\alpha} + \widehat{\beta}\,(x_{n+1} - \bar{x}) \ ,$$

} estimate of Y at a distinct value of $x = x_{n+1}$

but this is just one possible value of the random variable

$$Y_{n+1} = \alpha + \beta\,(x_{n+1} - \bar{x}) + \varepsilon_{n+1} \ .$$

prediction for the value of Y at a distinct value for x

The construction of a *prediction interval* for Y_{n+1} when $x = x_{n+1}$ is similar to that of the confidence interval for the mean of Y when $x = x_{n+1}$.

previous case.

Since $Y_{n+1} = \alpha + \beta\,(x_{n+1} - \bar{x}) + \varepsilon_{n+1}$ with $\varepsilon_{n+1} \sim N\left(0, \sigma^2\right)$,

is normally distributed with mean $\alpha + \beta\,(x_{n+1} - \bar{x})$ and variance σ^2, then

$$W = Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x}) \ ,$$

being a linear combination of independent and normally distributed random variables, is also normally distributed.

The mean of W is

$$\begin{aligned} E(W) &= E\left[Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x})\right] \\ &= E\left[Y_{n+1}\right] - E\left[\widehat{\alpha}\right] - E\left[\widehat{\beta}\,(x_{n+1} - \bar{x})\right] \\ &= \alpha + \beta\,(x_{n+1} - \bar{x}) - \alpha - \beta\,(x_{n+1} - \bar{x}) = 0 \ . \end{aligned}$$

Since Y_{n+1}, $\widehat{\alpha}$, and $\widehat{\beta}$ are independent, the variance of W is

$$\begin{aligned}
\mathrm{Var}(W) &= \mathrm{Var}\left(Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x})\right) \\
&= \mathrm{Var}\,(Y_{n+1}) + \mathrm{Var}\,(\widehat{\alpha}) + \mathrm{Var}\left(\widehat{\beta}\,(x_{n+1} - \bar{x})\right) \\
&= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(x_{n+1} - \bar{x})^2 \\
&= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].
\end{aligned}$$

As a consequence, we get that

$$Z = \frac{W - 0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} = \frac{Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1} - \bar{x})}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim N(0,1)$$

Recalling that

$$U = \frac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)\ ,$$

independence of Y_{n+1}, $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$, allows us to conclude that

$$T_Y = \frac{Z}{\sqrt{U/(n-2)}} = \frac{\frac{Y_{n+1}-\widehat{\alpha}-\widehat{\beta}(x_{n+1}-\bar{x})}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}}}{\sqrt{\frac{n\widehat{\sigma^2}}{(n-2)\sigma^2}}}$$

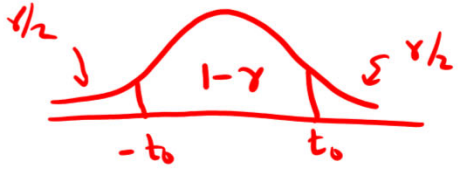
$$= \frac{Y_{n+1} - \widehat{\alpha} - \widehat{\beta}\,(x_{n+1}-\bar{x})}{\sqrt{\frac{n\widehat{\sigma^2}}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}}$$

has a $t(n-2)$ distribution.

Set $t_0 = t_{\gamma/2}(n-2)$ and

$$S_Y = \sqrt{\frac{n\widehat{\sigma^2}}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

so that $T_Y = \dfrac{Y_{n+1} - \hat{\alpha} - \hat{\beta}(x_{n+1}-\bar{x})}{S_Y}$



and note that

$$P\left(-t_0 \leq T_Y \leq t_0\right) = 1-\gamma$$

can be rewritten as

$$P\left(\widehat{\alpha}+\widehat{\beta}(x_{n+1}-\bar{x}) - t_0 S_Y \leq Y_{n+1} \leq \widehat{\alpha}+\widehat{\beta}(x_{n+1}-\bar{x}) + t_0 S_Y\right)$$

We conclude that

$$\left[\widehat{\alpha}+\widehat{\beta}(x_{n+1}-\bar{x}) - t_0 S_Y, \widehat{\alpha}+\widehat{\beta}(x_{n+1}-\bar{x}) + t_0 S_Y\right]$$

$\hat{\alpha} + \hat{\beta}(x_{n+1}-\bar{x}) \pm t_0 S_Y$

pt estimate for Y_{n+1}

is a $100(1-\gamma)\%$ confidence interval for Y_{n+1}.