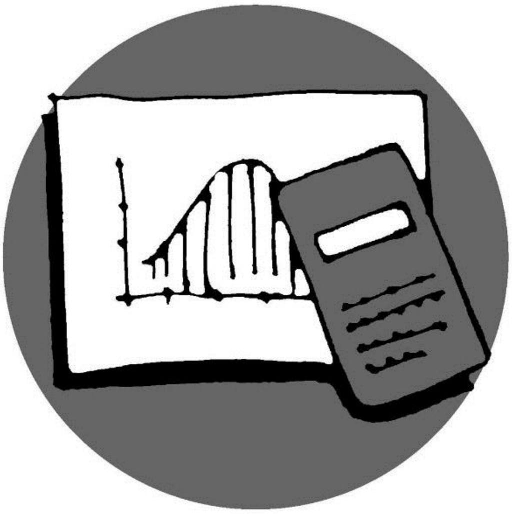
Library, Teaching and Learning

Regression Analysis and Confidence Intervals

QMET201



Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH • NEW ZEALAND
New Zealand's specialist land-based university

Regression Analysis and Confidence Intervals Summary

After calculating the regression equation, the next process is to analyse the variation. For Simple Linear Regression, there are three sources of variation:

- Total Variation *(i.e. variation between the observed Y_i values)*
- Variation due to the Regression
- Residual variation

Recall that in statistics 'variance' is the average of the squared deviations. The sum of the squared deviations (or differences) is $\Sigma\left(x_i - \bar{x}\right)^2$, which is abbreviated to sum of squares (SS).

Recall also SS$_{XY}$ $= \Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \Sigma XY - \dfrac{\Sigma X \times \Sigma Y}{n}$

SS$_X$ $= \Sigma\left(X - \bar{X}\right)^2 = \Sigma X^2 - \dfrac{\left(\Sigma X\right)^2}{n}$

SS$_Y$ $= \Sigma\left(Y - \bar{Y}\right)^2 = \Sigma Y^2 - \dfrac{\left(\Sigma Y\right)^2}{n}$

To calculate each of the above variations (Total, Regression and Residual) we need to calculate 'sums of squares' as follows:

- Total Variation requires $Total\ SS = SS_{\mathbf{total}} = \Sigma\left(Y_i - \bar{Y}\right)^2$
- Variation due to Regression requires $SS\ due\ to\ regression = SS_{\mathbf{reg}} = \Sigma\left(\hat{Y} - \bar{Y}\right)^2$
- Residual Variation requires $SS\ due\ to\ error = SS_{\mathbf{error}}$ **or** $\mathbf{SS}_{\mathbf{residual}} = \Sigma\left(Y_i - \hat{Y}_i\right)^2$

Calculation of these sums of squares can be managed as follows:

- $\boxed{SS_{total} = \Sigma Y^2 - \dfrac{\left(\Sigma Y\right)^2}{n} = SS_Y}$
- $\boxed{SS_{regression} = \dfrac{\left\{\Sigma XY - \dfrac{\Sigma X \times \Sigma Y}{n}\right\}^2}{\left\{\Sigma X^2 - \dfrac{\left(\Sigma X\right)^2}{n}\right\}} = \dfrac{\left(SS_{XY}\right)^2}{SS_X}}$
- $\boxed{SS_{error} = SS_{total} - SS_{regression}}$ or $\Sigma Y^2 - b_0\left(\Sigma Y\right) - b_1\left(\Sigma XY\right)$

A table is now used to summarise the **AN**alysis **O**f **Va**riation

ANOVA Table

Source of variation	Degrees of Freedom	Sum of Squares	…
Regression (Explained)	1	$SS_{reg} = \dfrac{\left\{\Sigma XY - \dfrac{\Sigma X \times \Sigma Y}{n}\right\}^2}{\left\{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}\right\}}$	…
Error or residual (Unexplained)	$n-2$	$SS_{residual} = SS_{total} - SS_{regression}$	…
Total	$n-1$	$SS_{total} = \Sigma Y^2 - \dfrac{(\Sigma Y)^2}{n}$	…

which is then completed:

Source of variance	df	Sum of Squares	Mean Square	F ratio	*p**
Regression (Explained)	1	$SS_{regression} = \dfrac{(SS_{XY})^2}{SS_X}$	$MR_{regression} = \dfrac{SS_{regresison}}{df}$ $= \dfrac{SS_{regression}}{1}$	$\dfrac{MS_{regression}}{MS_{residual}}$	
Residual or Error (Unexplained)	$n-2$	$SS_{error} = SS_{total} - SS_{regression}$	$MS_{residual} = \dfrac{SS_{residual}}{df}$ $= \dfrac{SS_{residual}}{n-2}$		
Total	$n-1$	$SS_{total} = \sum Y^2 - \dfrac{(\Sigma Y)^2}{n}$			

*The significance of the F test is determined by comparing the F ratio from the table above (F*calc*) with the F*table* value for a chosen value of α (usually 0.01 or 0.05). As with the χ^2 and t test, if the test value is greater than the table value, the null hypothesis is rejected. Two values are needed as degrees of freedom for the F test:

DF for the *numerator* =1 for simple linear regression (always) and
DF for the *denominator* = n-2 (= DF for the Residual line in the ANOVA)

Note on Residual Analysis

Residual = observed *Y* - predicted *Y*. Standardised residual = $\dfrac{residual}{\sqrt{MS_{(error)}}}$

If the model holds, about 95% of standardised residuals should have a value between -2 and 2.

Using the worked example from the previous booklet, recall the required totals were
$\Sigma X = 425, \quad \Sigma Y = 25, \quad \Sigma XY = 2550, \quad \Sigma X^2 = 43975, \quad \Sigma Y^2 = 151$
and $b_1 = 0.054, b_O = 0.398$. That is, $\hat{y} = 0.398 + 0.054x$.

$SS_{xy} = 2550 - \frac{425 \times 25}{5} = 425$, $SS_x = 43975 - \frac{425^2}{5} = 7850$, $SS_y = 151 - \frac{25^2}{5} = 26$

Hence:

Source	df	Sum of Squares	Mean Square	F ratio	*p*
Regression	1	$SS_{reg} = \frac{\left(2550 - \frac{425 \times 25}{5}\right)^2}{\left(43975 - \frac{425^2}{5}\right)}$ $= 23$	$MR_{reg} = \frac{23}{1}$ $= 23$	$\frac{23}{1} = 23$	
Residual	3	$SS_{error} = 26 - 23 = 3$	$MS_{error} = \frac{3}{3} = 1$		
Total	4	$SS_{total} = 151 - \frac{25^2}{5} = 26$			

Comment: a perfect fit occurs if $SS_{regression} = SS_Y$;a perfect fit occurs if $\text{SS}_{\text{residual}} = 0$.

From here, any of the following may be calculated:

- **Coefficient of Determination:** $R^2 = \frac{SS_{regression}}{SS_{total}}$

This represents the ***proportion*** of the total variation in Y that is ***explained*** by the fitted simple linear regression model. It always lies between 0 and 1.

Note: R^2 ranges from 0 to 1 inclusive.
$R^2 = 1$ if a perfect linear relationship exists.
$R^2 = 0$ if *no* perfect linear relationship exists.

In the above example, $R^2 = \frac{23}{26} = 0.88$

This indicates that 88% of the variation can be explained by the model.

- **Correlation Coefficient:** $r = \pm\sqrt{R^2}$ **.**

  This measures the ***strength*** of the **linear** relationship between *X* and *Y.*

  Points to note:

  - r ranges from -1 to +1 (perfect negative correlation to perfect positive correlation).
  - The closer r is to 1, the stronger the **linear** relationship between *X* and *Y.*
  - r = 0 implies no apparent linear relationship between *X* and *Y*, and *X* is not useful for predicting *Y*).
  - If *r* = 1, all points lie on a line with a positive slope.
  - If r = -1, all points lie on a line with a negative slope.
  - Note: It is possible to have a perfect relationship, which is not linear.

In the above example, $r = \sqrt{0.88} = 0.94$, which indicates a strong positive relationship. Note the "sign" of r is same as for the slope, b_1.

Alternative calculation: $r = \dfrac{Covariance}{s_x \times s_y}$

where $Cov(x, y) = \dfrac{SS_{xy}}{n-1} = \dfrac{\sum xy - \left(\sum x\right)\left(\sum y\right)/n}{n-1} = \dfrac{425}{4} = 106.25$ and

$s_X^2 = \dfrac{SS_x}{n-1} = \dfrac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1} \Rightarrow s_X = \sqrt{\dfrac{7850}{4}} = 44.3$ and

$s_Y = \sqrt{\dfrac{SS_Y}{n-1}} = \sqrt{\dfrac{26}{4}} = 2.55$

that is, $r = \dfrac{Covariance}{s_x \times s_y} = \dfrac{106.25}{44.3 \times 2.55} = 0.94$ as above.

- **Confidence Intervals**
  - For β1 (slope): $C.I.(slope) = b_1 \pm t_{(n-2)}\sqrt{\dfrac{MS_{error}}{SS_X}}$
  - For $\mu_{Y/X}$ (the mean of the population of Y values corresponding to X_i):

$$C.I.(mean\ prediction) = \hat{Y}_i \pm t_{(n-2)}\sqrt{MS_{error}} \times \sqrt{\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_X}}$$

  - For Y_i (an individual predicted value):

$$C.I.(individual\ \ \mathbf{p.v.}) = \hat{Y}_i \pm t_{(n-2)}\sqrt{MS_{error}} \times \sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_X}}$$

Return to the worked example again, with

$$SS_x = 43975 - \frac{425^2}{5} = 7850 \text{ and } MS_{error} = 1$$

Confidence Interval for Prediction of Slope

95% confidence interval for β_1 would be: 0.054 ± 3.182 × $\sqrt{\left(\dfrac{1}{7850}\right)}$

= [0.018, 0.090]

⇒ We can be 95% confident that for each increase of 1 ml in alcohol the increase in time taken is between 0.018 and 0.090 mins.

Interpretation - if the confidence interval does not include 0, there is good evidence that X and Y are related. If X and Y are not related, β_1 will be 0. So the confidence interval checks whether the model is useful for prediction.

Confidence Interval for Prediction of Mean Response

The main use of regression is to predict the value of Y corresponding to a particular x-value.

Use the given x-value in the equation to calculate an estimate for $\hat{y}$ and note, or calculate, $\bar{x}$. Use these values in the formula.

Note: the given x-value = x_i in the formula for the confidence interval.

Suppose we wish to estimate with 95% confidence, the true mean time taken for an intake of 100 mls of alcohol. Using the regression equation, $\hat{y} = 0.398 + 0.054x$ with $x = 100$, the point estimate of $\mu_{Y/X}$ in our example is 5.798 mins.

To form the 95% confidence-interval estimate for the true mean response we have

$$x_i = 100, \ \bar{x} = \frac{425}{5} = 85:$$

$$\Rightarrow C.I.(mean) = 5.798 \pm 3.182\sqrt{1} \times \sqrt{\frac{1}{5} + \frac{(100-85)^2}{7850}} = (4.276, \ 7.312)$$

That is, we can be 95% confident that the true mean time taken is between 4.3 and 7.3 mins.

Confidence Interval for the Individual Response

The previous confidence interval is for an **average**. Sometimes we want an interval estimate for an individual response $\hat{Y}$ corresponding to a given value X_i (rather than an estimate for the mean response). The best estimate of an individual response is still $\hat{y}$, but the confidence interval is much wider because individual values vary much more than the mean. i.e. it is harder to predict an *individual* value than an *average*.

eg for $x = 100$, calculate estimate for $\hat{y}$ as 5.798 as before.

Then 95% confidence-interval estimate for an individual response is:

$$C.I.(i.p.v) = 5.798 \pm 3.182\sqrt{1} \times \sqrt{1 + \frac{1}{5} + \frac{(100-85)^2}{7850}} = (2.27, \ 9.33)$$

That is, we can be 95% confident that the true time taken for an individual is between 2.27 and 9.33 mins.

Note the considerable increase in width of the interval. By increasing the sample size, this could be reduced. A sample size of 5 is inappropriate for testing, but is used here merely to demonstrate the process.

Other versions of formulae:

$SS_{total} = (n-1)\times s_x$ $\qquad SS_{reg} = Cov(X,Y)\times SS_{total}$

For testing coefficient of determination: $t = r\times\sqrt{\dfrac{n-2}{1-r^2}}$

For testing $H_0 : \beta_1 = 0\, vs\, H_A : \beta_1 \neq 0$: $t = \dfrac{b_1 - 0}{se(b_1)} = \dfrac{b_1}{\sqrt{\left(\frac{MS_{error}}{SS_X}\right)}}$

Practice Questions

The following data describes the flowering score (Y) for plants of spearmint (Mentha spicata) sown during various weeks (X).

Week sown (X)	Flowering score (Y)
2	5
3	20
4	24
5	21
6	13

1. For the flowering score data the sum of ($X\times Y$) values $\Sigma XY = 349$.
   What is S_{xy}?
   A. 69.8 B. 17 C. -262.2 D. -1311 E. 241

2. Calculate the Sums of squares for X, i.e., SS_X
   A. 2.500 B. 233.2 C. 5.342 D. 10.00 E. 1.581

The relationship between male mortality rate per 100,000 (in years 1958-64) and water hardness was studied by Hills et al.. (Open University). 61 cities were used in the study. The following partial regression analysis shows some of the results.

```
MTB > Regress ’Mortalit’ 1 ’Ca(ppm)’

The regression equation is Mortalit = 1676 - 3.23 Ca(ppm)
Predictor     Coef         StDev        T            P
Constant      1676.36      29.30        57.22        0.000
Ca(ppm)        -3.2261     *.***        -6.66        0.000

Analysis of Variance
Source        DF    SS           MS           F            P
Regression     1    906185       906185       44.30        0.000
Error         59    1206988      20457
Total         60    2113174
```

The sums of squares for Calcium (ppm) is 87069.0.

3. What is the standard error of the regression coefficient (-3.2261) ?
   A. 0.2350 B. 0.4847 C. 143.0 D. 29.30 E. -6.66

4. What is the CORRELATION COEFFICIENT?
A. -0.655 B. -3.2261 C. 1676.36 D. +0.429 E. -0.429

5. What would be the estimated mortality for a city with a calcium level of 100 ppm?
A. 1999 B. 1354 C. 167631 D. 1576 E. 1644

The dry weights (in mg) of successive leaves of a wheat plants were recorded as

	L1	L2	L3	L4	L5	L6	L7
At emergence	1.4	1.5	2.0	2.7	5.1	7.3	12.4
At maturity	12	18	36	62	76	89	109

6. From the following data calculate the "sums of products", SS_{xy}.
n = 7, $\Sigma x = 32.4, \Sigma x^2 = 248.56 \quad \Sigma y = 402.0, \Sigma y^2 = 31186 \quad \Sigma xy = 2672$
A. -10352.7 B. 1860.7 C. 2672.1 D. 13024.8 E. 811.414

7. In the previous example, what would be the degrees of freedom for the Regression SS, Error SS, and Total SS, respectively ?
A. 2, 5, 7 B. 1, 6, 7 C. 2, 4, 6 D. 1, 5, 6 E. 1, 5, 5

Answers:

1	B	2	D	3	B		
4	A	5	B	6	E	7	D

Exam Question ***[total 26 marks]***

A number of Weddell seals were captured in the Antarctic in 1998 and blood samples taken. Several measures were made of the blood, but here we consider cortisol levels (µM). Cortisol increases in animals under stress, and part of the stress is induced by the capture. In order to determine this, the animals were re-sampled over a period. Here is the data for a seal named “Pam”.

	Mean Cortisol µM	Time post capture minutes	$X \times Y$
	2.3	218	501.4
	2.4	265	636.0
	2.7	296	799.2
	2.8	326	912.8
	3.0	350	1050.0
	3.1	380	1178.0
	2.5	410	1025.0
	3.2	414	1324.8
	3.2	446	1427.2
Sum	25.2	3105	8854.4
S.D.	0.346410	75.6042	$\Sigma Y^2 = 71.52 \quad \Sigma X^2 = 1116953$

(a) Calculate the regression of Mean Cortisol on Time post capture.
You can check the data entry on your calculator by checking that *r* = 0.76555

(b) Calculate the **Total** sums of squares for the regression analysis of variance. Use this template in your answer.

Source	DF	SS	MS	F	P
Regression					
Error(Residual)					
Total					

(c) Given *r* (above) calculate the Regression sum of squares.

(Hint: what is R^2?). **Or** use some other way of calculating the Regression SS.

(d) Plot the data on the graph.

(e) Calculate $\hat{Y}$ for $X = 220$ and for $X = 440$.

(f) Use the values of Ŷ to draw the fitted line on the graph.

(g) Comment briefly on the fit of the line to the data. (Just a few lines).

Answers:

a) *Using the calculations:* $b_1 = \dfrac{8854.4 - 25.2\times 3105/9}{1116953 - 3105^2/9} = \dfrac{160.4}{45728} = 0.003508$

Hence: $b_0 = \dfrac{25.2}{9} - 0.003508\times\dfrac{3105}{9} = 1.590$

You can check your data entry on your calculator by checking that *r* =0.76556

b) $SS_{total} = 71.52 - \dfrac{25.2^2}{9} = 0.96$ *OR* $SS_{total} = (n-1)\times s.d._{x} = 8\times 0.34641^2 = 0.96$

c) $SS_{reg} = \dfrac{160.4^2}{45728} = 0.5626$ *OR*

$SS_{reg} = Cov(X,Y)\times SS_{total} = 0.76556\times 0.96 = 0.5627$

d) $SS_{error} = SS_{total} - SS_{regression} = 0.96 - 0.5626 = 0.3974$

and $MS_{error} = \dfrac{0.3974}{7} = 0.0568$

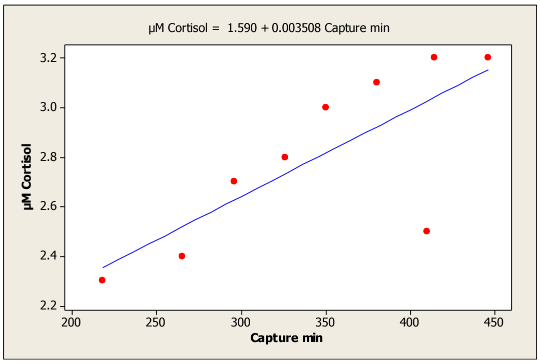
Compare with Minitab output:

```
Analysis of Variance

Source           DF    SS         MS         F      P
Regression       1     0.56263    0.56263    9.91   0.016
Residual Error   7     0.39737    0.05677
Total            8     0.96000
```

You could put p<0.05 for P.

e)



f)

$$\hat{Y} = 1.59 + 0.003508 \times 220 = 2.36$$

$$\hat{Y} = 1.59 + 0.003508 \times 440 = 3.13$$

g) *The seventh value seems seriously in error, especially considering the other points. The straight line does not seem to give a good indication of the response which seems to be more like a sigmoid or s-shaped response than a straight line. There does not seem to be a simple transformation (like logs or square root) that straightens the response out. The R^2 value of 0.766^2=0.59 implies that there is still 40% of the variation in Y not accounted for. One would like an R^2 closer to .8 or more.*

Extra question:

A real estate agent in Templeton has found that section prices in a new subdivision change with the size of the section. The following output represents a linear regression analysis of the section prices (in thousands of dollars) against the section size (in m^2).

The regression equation is: Cost = -27.1 + 0.125 Sect-size.

Predictor	Coef	Stdev	t-ratio	p
Constant	-27.11	13.12	-2.07	0.055
Sect-size	0.12514	0.01747	7.16	0.000

s = 3.352 R-sq = 76.2% R-sq(adj) = 74.7%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	576.78	576.78	51.32	0.000
Error	16	179.83	11.24		
Total	17	756.61			

a) If the mean section size is 749.8 m^2, and the mean cost is $66720, SHOW how to determine the *y* intercept b_0.

b) SHOW how to calculate the coefficient of determination $\left(r^2\right)$.

c) What is the estimated cost of a section size of 774 m^2?

d) In the section size data used to generate the regression equation, it was measured that a section with a size of 774 m^2 has a cost of $77000. Determine the standardised residual for this point.

e) Given $SS_X = 36831.5$, calculate a 95% confidence interval for the slope.

f) Interpret the confidence interval obtained in (e).

g) Construct a 95% confidence interval for an individual section with a size of 735 m^2.

Solutions:

a) $b_0 = \overline{Y} - b_1 X = 66.720 - 0.125 \times 749.8 = -27.005$

b) $r^2 = \dfrac{SS_{reg}}{SS_{total}} = \dfrac{576.78}{756.61} = 0.762\,(\approx 76\%)$

c) $Cost = -27.1 + 0.125 \times 774 = 69.65(thousand\ dollars)$ ie $69,650

d)

$$residual = 77.000 - 69.650 = 7.350$$

$$\Rightarrow \textbf{standardised residual} = \frac{residual}{\sqrt{MS_{error}}} = \frac{7.350}{\sqrt{11.24}} = 2.19$$

e) $C.I. = b_1 \pm t_{(\alpha,n-2)} \times \sqrt{\dfrac{MS_{error}}{SS_x}} = 0.125 \pm 2.1199 \times \sqrt{\dfrac{11.24}{36831.5}} = (0.088, 0.162)$

ie. Between $88 and $162

f) We can be 95% confident that the rate of change of price of section is between $88 and $162 per m^2 increase in section size.

g)

$$C.I.(Y_i) = \hat{Y} \pm t_{(n-2,\alpha)} \times \sqrt{MS_{error}} \times \sqrt{1 + \frac{1}{n} + \frac{(x_i - \overline{x})}{SS_x}}$$

$$= 64.775 \pm 2.1199 \times \sqrt{11.24} \times \sqrt{1 + \frac{1}{18} + \frac{(749.8 - 735)^2}{36831.5}}$$

$$= (57.453, 72.097)\ \ ie\ \ \$57,453\ \ and\ \ \$72,097$$

Summary of Formulae

Regression line:

From $\Sigma x^2, \Sigma x, n, \Sigma y^2, \Sigma y, \Sigma xy$,

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{\Sigma xy - \frac{\Sigma x \times \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}, \qquad b_0 = \bar{y} - b_1 \bar{x}$$

$$\Rightarrow \hat{y} = b_0 + b_1 x$$

Note also $SS_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) \qquad SS_x = \Sigma(x - \bar{x})^2 \qquad SS_y = \Sigma(y - \bar{y})^2$

Analysis of Variance:

$$SS_{reg} = \Sigma\left(\hat{Y} - \bar{Y}\right)^2 = \frac{(SS_{xy})^2}{SS_x} = \frac{\left(\Sigma xy - \frac{\Sigma x \times \Sigma y}{n}\right)^2}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$SS_{total} = \Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$\Rightarrow SS_{error} = \Sigma\left(Y_i - \hat{Y}\right)^2 = SS_{total} - SS_{regression}$$

$$MS_{reg} = \frac{SS_{reg}}{df_{reg}}, \; MS_{error} = \frac{SS_{error}}{df_{error}} \qquad \Rightarrow F = \frac{MS_{reg}}{MS_{error}}$$

For F table comparison, use DF for Regression, DF for Error

Coefficient of Determination: $r^2 = \frac{SS_{regression}}{SS_{total}}$

Correlation Coefficient: $r = \sqrt{r^2}$ (remember you need to add the sign, +/-)

Confidence Intervals:

For β_1(slope):

$$C.I.(slope) = b_1 \pm t_{(n-2)} \times \sqrt{\frac{MS_{error}}{SS_x}}$$

For $\mu_{Y/X}$ (the population mean):

$$C.I.(mean) = \hat{Y} \pm t_{(n-2)} \sqrt{MS_{error}} \times \sqrt{\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_X}}$$

For Y_i (an individual predicted value):

$$C.I.(indivdual\ \mathbf{p.v.}) = \hat{Y} \pm t_{(n-2)} \sqrt{MS_{error}} \times \sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_X}}$$

Note $s.e.\left(slope\right) = \sqrt{\frac{MS_{error}}{SS_x}}$

$$s.e.\left(pop.\,mean\right) = \sqrt{MS_{error}} \times \sqrt{\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_x}}$$

$$s.e.\left(i.p.v\right) = \sqrt{MS_{error}} \times \sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_x}}$$