More Inference for Simple Linear Regression (Ch. 9.1)

Yifan Zhu

Iowa State University

Outline

SLR: Inference for the Mean Response at some x

Prediction interval for a new y at some x

Simultaneous Confidence Intervals for $\mu_{y|x}$

SLR: mean response at x

- Recall our model:

constant

$$Y_i = \boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i$$

$\varepsilon_1, \ldots, \varepsilon_n \sim$ iid $N(0, \sigma^2)$

- Under the model, the true mean response at some observed covariate value x_i is:

$$E(y_i) = \mu_{y|x_i} = \beta_0 + \beta_1 x_i$$

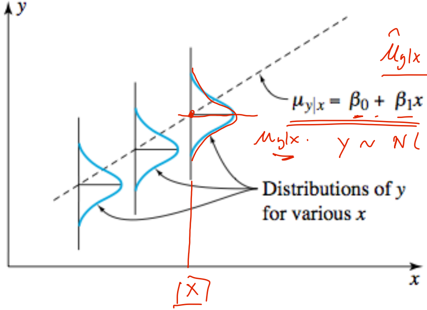
- Now, if some new covariate value x is within the range of the x_i's, we can estimate the true mean response at this new x:

$$\widehat{\mu}_{y|x} = b_0 + b_1 x$$

SLR: mean response at x

- But how good is the estimate?

$\hat{\mu}_{y|x} = b_0 + b_1 x$.



$y \sim N(\mu_{y|x}, \sigma^2_{y|x})$.

- That's why we do inference.

SLR: mean response at x

- Under the model, $\widehat{\mu}_{y|x}$ is normally distributed with:

$$E(\widehat{\mu}_{y|x}) = \mu_{y|x} = \beta_0 + \beta_1 x$$

→ unbiased estimator

$$Var(\widehat{\mu}_{y|x}) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)$$

- We can construct a $N(0,1)$ random variable by standardizing:

$$Z = \frac{\widehat{\mu}_{y|x} - \mu_{y|x}}{\sigma \sqrt{\frac{1}{n} \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}} \sim N(0,1)$$

- Replacing σ with $s_{LF} = \sqrt{\frac{1}{n-2} \sum_i (y_i - \widehat{y}_i)^2}$:

$$T = \boxed{\frac{\widehat{\mu}_{y|x} - \mu_{y|x}}{s_{LF} \sqrt{\frac{1}{n} \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}}} \sim t_{n-2}$$

→ test statistic

$se(\hat{\mu}_{y|x})$ →

SLR: mean response at x

- To test $H_0 : \mu_{y|x} = \#$, we can use the test statistic:

$$T = \frac{\widehat{\mu}_{y|x} - \#}{s_{LF}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}}$$

$\mu_{y|x}$

which has a t_{n-2} distribution if H_0 is true and the model is correct.

- A 2-sided $1-\alpha$ confidence interval for $\mu_{y|x}$ is:

$se(\hat{\mu}_{y|x})$ $=$ $\widehat{sd}(\hat{\mu}_{y|x})$

$$\left(\widehat{\mu}_{y|x} - t_{n-2,\ 1-\alpha/2} \cdot s_{LF}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}},\right.$$
$$\left.\widehat{\mu}_{y|x} + t_{n-2,\ 1-\alpha/2} \cdot s_{LF}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}\right)$$

and the one-sided intervals are analogous.

$(\hat{\mu}_{y|x} - t_{n-2,1-\alpha} \cdot se(\hat{\mu}_{y|x}), +\infty)$ $\leftarrow$ lower $1-\alpha$ C.I.

Pressing pressures and specimen densities for a ceramic compound

A mixture of Al_2O_3, polyvinyl alcohol, and water was prepared, dried overnight, crushed, and sieved to obtain 100 mesh size grains. These were pressed into cylinders at pressures from 2,000 psi to 10,000 psi, and cylinder densities were calculated.

predictor / covariate   response.

x (pressure in psi)	y (density in g/cc)
2000.00	2.49
2000.00	2.48
2000.00	2.47
4000.00	2.56
4000.00	2.57
4000.00	2.58
6000.00	2.65
6000.00	2.66
6000.00	2.65
8000.00	2.72
8000.00	2.77
8000.00	2.81
10000.00	2.86
10000.00	2.88
10000.00	2.86

Example: ceramics

- First, I'll make a 2-sided 95% confidence interval for the true mean density of the ceramics at 4000 psi. $\mu_{y|x=4000}$.

  b_0 b_1

  $$\widehat{\mu}_{y|x} = 2.375 + 0.0000487(4000) = 2.5697g/cc$$

  With $t_{n-2,\ 1-\alpha/2} = t_{13,0.975} = 2.160$, the margin of error in the confidence interval is:

  Looking up table B.4.

  $$t_{n-2,\ 1-\alpha/2}\, s_{LF}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

  $se(\hat{\mu}_{y|x=4000})$.

  $$= 2.160(0.0199)\sqrt{\frac{1}{15} + \frac{(4000-6000)^2}{1.2\times 10^8}} = 0.0136g/cc$$

  Hence, the 95% CI is:

  $$(2.5697 - 0.0136,\ 2.5697 + 0.0136) = (2.5561,\ 2.5833)$$

  $\hat{\mu}_{y|x} \pm$ margin of error.

- We're 95% confident that the true mean density of the ceramics at 4000 psi is between 2.5561 g/cc and 2.5833 g/cc.

Your turn: ceramics

- Calculate and interpret a 2-sided 95% confidence interval for the true mean density at 5000 psi, given:
  - $\widehat{\mu}_{y|x} = 2.375 + 0.0000487x$
  - The margin of error is $t_{n-2,1-\alpha/2} s_{LF} \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$
  - $\sum_i (x_i - \overline{x})^2 = 1.2 \times 10^8$
  - $n = 15$, $\overline{x} = 6000$.
  - $s_{LF} = 0.0199$
  - $t_{13,0.975} = 2.16$
- Test $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$ at significance level $\alpha = 0.05$ using the method of p-values.

Answers: ceramics

- Make a 2-sided 95% confidence interval for the true mean density of the ceramics at 5000 psi:

$$\widehat{\mu}_{y|x} = \underbrace{2.375}_{b_0} + \underbrace{0.0000487(5000)}_{b_1 \cdot x} = 2.6183g/cc$$

With $t_{n-2,\ 1-\alpha/2} = t_{13,0.975} = 2.160$, the margin of error in the confidence interval is:

$$t_{n-2,\ 1-\alpha/2} \cdot s_{LF}\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

$$= 2.160(0.0199)\sqrt{\frac{1}{15} + \frac{(5000 - 6000)^2}{1.2 \times 10^8}} = 0.0118g/cc$$

Hence, the 95% CI is:

$$(2.6183 - 0.0118,\ 2.6183 + 0.0118) = (2.6065,\ 2.6301)$$

$\hat{\mu}_{y|x}$ width is different from that in 4000 psi.

$\mu_{y|x}$

- We're 95% confident that the true mean density of the ceramics at 5000 psi is between 2.6065 g/cc and 2.6301 g/cc.

More Inference for Simple Linear Regression (Ch. 9.1)

Yifan Zhu

SLR: Inference for the Mean Response at some x

Prediction interval for a new y at some x

Simultaneous Confidence Intervals for $\mu_{y|x}$

Answers: ceramics

Now for the hypothesis test: $\mu_{y|x} = \beta_0 + \beta_1 x$.

1. $H_0 : \boxed{\beta_0 = 0,}\ H_a : \beta_0 \neq 0$ $\quad \mu_{y|x=0} = \beta_0 + \beta_1 \cdot 0 = \beta_0$.
2. $\alpha = 0.05$
3. β_0 is just $\mu_{y|x=0}$. The test statistic is: $\hat{\mu}_{y|x} - \mu_{y|x} = b_0 + b_1 \cdot 0 - \beta_0 = b_0 - 0$

$$T = \frac{b_0 - 0}{s_{LF}\sqrt{\frac{1}{n} + \frac{(0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}} = \boxed{\frac{b_0}{s_{LF}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}}}} \leftarrow \frac{b_0}{se(b_0)}$$

x

- $T \sim t_{n-2}$ assuming: $n = 15$, $\boxed{t_{13}}$ ← reference distribution.
  - H_0 is true.
  - The model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is correct, with $\varepsilon_1, \ldots \varepsilon_n \sim$ iid $N(0, \cancel{1})$. σ^2

Answers: ceramics

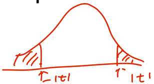
4. The observed test statistic:

$$b_0 = 2.375$$

$$t = \frac{2.375}{0.0199\sqrt{\frac{1}{15} + \frac{6000^2}{1.2\times 10^8}}} = 197.09$$

$$\text{p-value} = P(|t_{13}| > 197.09) \ll 0.0001$$

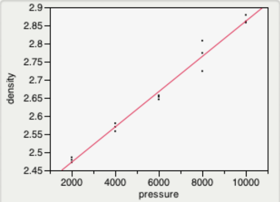
$|T|$ $|t|$



5. With a p-value $\ll 0.0001 < \alpha$, we reject H_0 and conclude H_a.
6. There is overwhelming evidence that the intercept of the "true" line is different from 0.

Ceramics: back to the JMP output

Bivariate Fit of density By pressure



Linear Fit

Linear Fit

density = 2.375 + 4.8667e-5*pressure

Summary of Fit

RSquare	0.982193
RSquare Adj	0.980824
Root Mean Square Error	0.019909
Mean of Response	2.667
Observations (or Sum Wgts)	15

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.28421333	0.284213	717.0604
Error	13	0.00515267	0.000396	**Prob > F**
C. Total	14	0.28936600		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.375	0.012055	197.01	<.0001*
pressure	4.8667e-5	1.817e-6	26.78	<.0001*

Ceramics: back to the JMP output

$H_0: \beta_0 = 0,\ H_a: \beta_0 \neq 0.$

b_0

$se(b_0).$ p-value.

Parameter Estimates $se(b_0)$

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	$b_0 \leftarrow$ 2.375	0.012055	197.01	<.0001*
pressure	4.8667e-5	1.817e-6	26.78	<.0001*

b_1 $se(b_1).$ $b_1/se(b_1).$ p-value.

$H_0: \beta_1 = 0,\ H_a: \beta_1 \neq 0$

- The observed test statistic t is under "t Ratio" for the intercept.
- "Prob$>|t|$" for the intercept is the p-value for the significance test you just did.
- "Estimate" for the intercept is b_0.
- "Std Error" for the intercept is:

$$\widehat{SD}(b_0) = s_{LF}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}}$$

Be careful with Inference on β_0

- In this case and many others, $\beta_0 = \mu_{y|x=0}$ is beyond the range of our data.
- Estimating beyond the range of our covariate values is called **extrapolation**, which is dangerous for linear regression.
- Only extrapolate when:
  - You know your process or system well, and can describe it with the right equations.
  - You estimate the parameters of the resulting model using *nonlinear* regression:
    - Example: special case of the Michaelis-Menten model for enzyme kinetics with reaction speed y and substrate concentration x:

$$Y_i = \boxed{\frac{\theta_1 x_i}{\theta_2 + x_i}} + \varepsilon_i$$

you know this is always true for all x.

  - See Nonlinear Regression Analysis and Its Applications by Bates and Watts for more information on nonlinear regression.

Outline

SLR: Inference for the Mean Response at some x

Prediction interval for a new y at some x

Simultaneous Confidence Intervals for $\mu_{y|x}$

Prediction interval for a new y at some x

In one-sample inference: a. P.I. is for a new / the next data point X_{n+1} (a random variable). P.I. is wider than C.I. due to more uncertainty.

- The prediction interval in SLR is trying to capture the next response at a given value of predictor variable.
  y is a random variable.
- A 2-sided $1 - \alpha$ prediction interval for a new response y at some x is:

one-sided: $t_{n-2, 1-\alpha}$.

$$\left(\widehat{\mu}_{y|x} - t_{n-2,\ 1-\alpha/2} \cdot s_{LF} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}, \right.$$

$$\left. \widehat{\mu}_{y|x} + t_{n-2,\ 1-\alpha/2} \cdot s_{LF} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}} \right)$$

wider than C.I.. also due to more uncertainty.

and the one-sided intervals are analogous.

$Y = \boxed{\mu_{y|x}} + \boxed{\varepsilon}$

extra error term

Example: ceramics

- We will make a 2-sided 95% prediction interval for the next density of the ceramics at 4000 psi.

$$\widehat{\mu}_{y|x} = 2.375 + 0.0000487(4000) = 2.5697 g/cc$$

With $t_{n-2,\ 1-\alpha/2} = t_{13,0.975} = 2.160$, the margin of error in the confidence interval is:

$$t_{n-2,\ 1-\alpha/2} \cdot s_{LF} \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

$$= 2.160(0.0199)\sqrt{1 + \frac{1}{15} + \frac{(4000-6000)^2}{1.2 \times 10^8}} = 0.0451 g/cc$$

Hence, the 95% CI is:

$$(2.5697 - 0.0451,\ 2.5697 + 0.0451) = (2.5246,\ 2.6148)$$

wider than c.i. for 4000 psi.

- We're 95% confident that the next collected density of the ceramics at 4000 psi is between 2.5246 g/cc and 2.6148 g/cc.

Outline

SLR: Inference for the Mean Response at some x

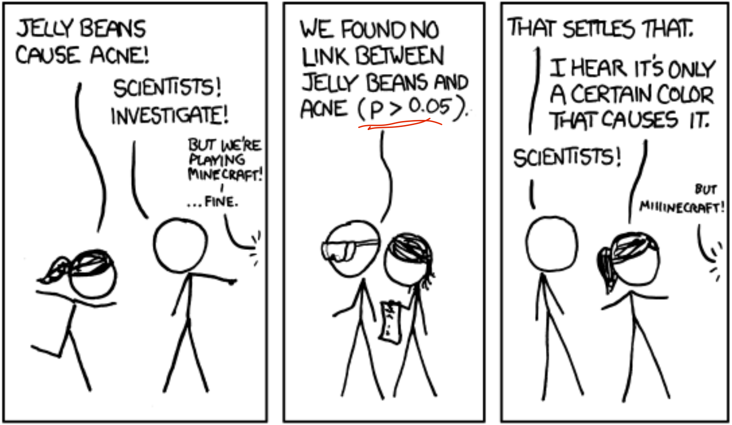
Prediction interval for a new y at some x

Simultaneous Confidence Intervals for $\mu_{y|x}$

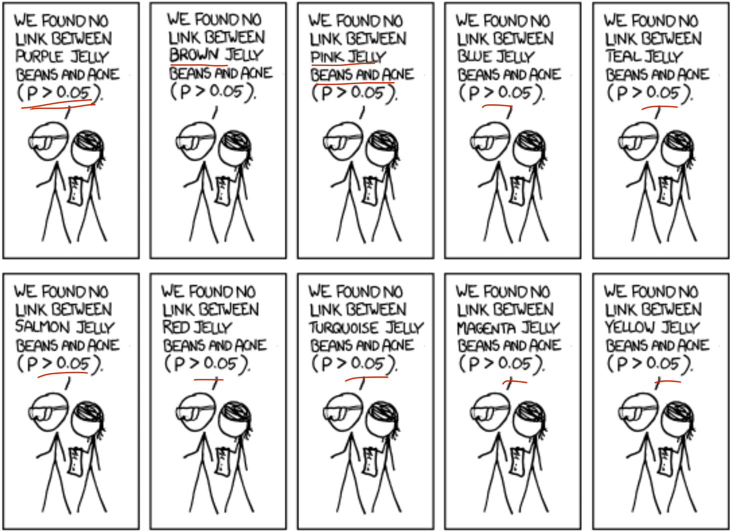
Simultaneous confidence intervals

- Situations will arise when you'll want to do inference on $\mu_{y|x=2000}, \mu_{y|x=4000}, \mu_{y|x=6000}, \ldots$, all at once.
- When you compute several confidence intervals at once or do multiple tests at once, you need to account for the simultaneity.
- On average, for every 20 tests you do independently at $\alpha = 0.05$, we expect 1 of those tests to conclude H_a by chance alone.
    - Remember: $\alpha = \text{P(reject } H_0 \text{ assuming } H_0 \text{ is true)}$.

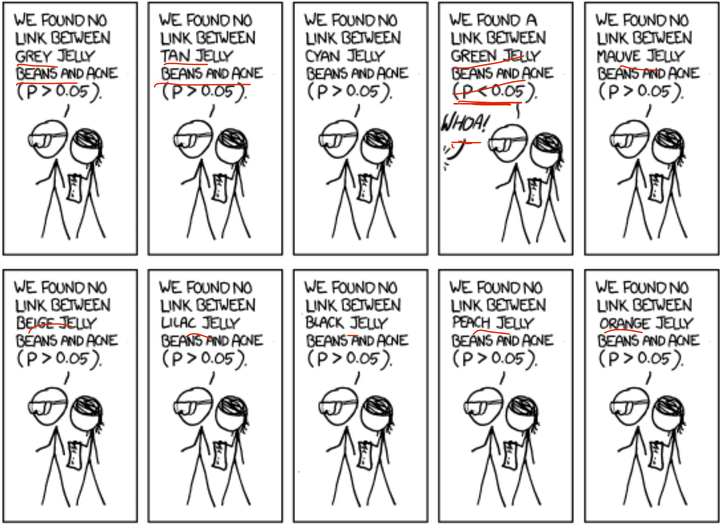
Example: http://xkcd.com/882/



Example: http://xkcd.com/882/

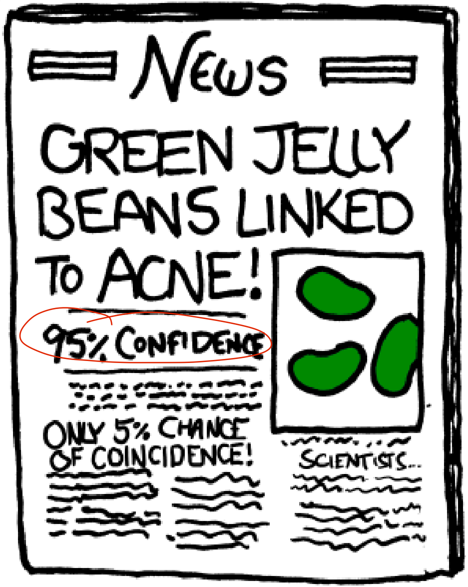


Example: http://xkcd.com/882/



20 tests

Example: http://xkcd.com/882/



Simultaneous confidence interval

- If we have k simutaneous tests, each with type I error α, then the type I error for the simutaneous tests is

$$P(\text{at least one rejection in these } k \text{ tests}) > \alpha$$

If these tests are independent, the actual type I error is

$$
\begin{aligned}
&P(\text{at least one rejection in these } k \text{ tests})\\
=&1 - P(\text{no rejection in these } k \text{ tests})\\
=&1 - \prod_{i=1}^{k} P(\text{fail to reject the } i\text{-th test}) = 1 - (1-\alpha)^k
\end{aligned}
$$

$0.95^{20} \approx 0.$

$\approx 1.$

- For k confidence intervals for $\mu_1, \mu_2, \dots, \mu_k$, denote the corresponding random intervals $I_1, I_2, \dots, I_k$. If the confidence level is $1-\alpha$, then

$$P(\mu_i \in I_i) = 1 - \alpha,\ i = 1, 2, \dots, k$$

And the simutaneous confidence level would be

extreme case:
when these events ind.
$P(\cdot\ ) = (1-\alpha)^k \approx 0.$

$$P(\mu_1 \in I_1 \text{ and } \mu_2 \in I_2 \text{ and } \cdots \mu_k \in I_k \text{ at the same time}) < 1 - \alpha$$

- To get the $1-\alpha$ simutaneous confidence level, the simultaneous confidence intervals should be **wider** then individual confidence interval.

Simultaneous confidence intervals for $\mu_{y|x}$

- Let I_x be the random intervals for the simultaneous $1 - \alpha$ confidence intervals for $\mu_{y|x}$. Then we want

$$P(\mu_{y|x} \in I_x \text{ at the same time for all } x) = 1 - \alpha$$

infinite events happening at the same time.

- The simultaneous confidence intervals for $\mu_{y|x}$ are given by:

before. not simultaneous case: $t_{n-1, 1-\alpha/2}$

$\hat{\mu}_{y|x}$

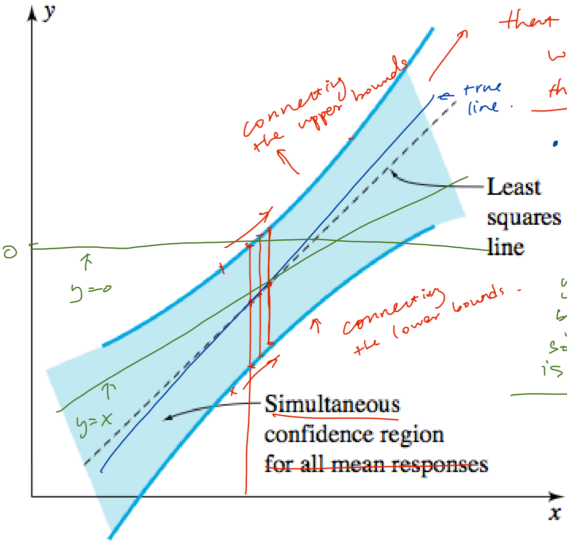
$$\underbrace{b_0 + b_1 x}_{\hat{\mu}_{y|x}} \pm \boxed{\sqrt{2F_{2,n-2,1-\alpha}}} \cdot \underbrace{s_{LF} \cdot \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}}_{se(\hat{\mu}_{y|x})}$$

- This formula accounts for the fact that we're computing k confidence intervals at the same time.

k can be arbitrary number.

Simultaneous confidence intervals for $\mu_{y|x}$

$1-\alpha = 0.95$



We are 95% confident that the true line will fall in this region.

- This can be used to test if ~~the~~ true line is $y = 0$.

$y=0$ is not covered by this region, so reject. the line is not $y = 0$.

$y = x$

fail to reject. plausible that the true line is $y=x$

Example: ceramics

- Given:
  - $n = 15$
  - $\overline{x} = 6000$
  - $\sum_i (x_i - \overline{x})^2 = 1.2 \times 10^8$
  - $\widehat{y} = 2.375 + 4.87 \times 10^{-5} x$, $s_{LF} = 0.0199$.
  - The simultaneous confidence interval formula is:

$$b_0 + b_1 x \pm \sqrt{2F_{2,k,1-\alpha/2}} \cdot s_{LF} \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}}$$

- I will calculate simultaneous 95% confidence intervals for the mean responses $\mu_{y|x}$ at $x = 2000$, 4000, 6000, 8000, and 10000.

Example: ceramics

Simultaneous C.I. is two-sided.

use $1-\alpha$ quantile.

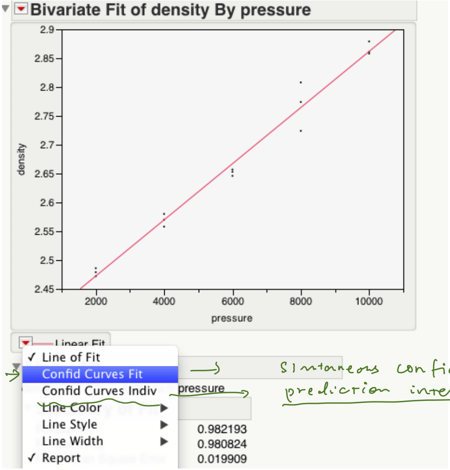
- Using $F_{2,n-2,1-\alpha} = F_{2,13,0.95} = 3.81$, the intervals are of the form:

$$2.375 + 4.87 \times 10^{-5} x \pm \sqrt{2 \cdot 3.81} \cdot 0.0199 \cdot \sqrt{\frac{1}{15} + \frac{(x - 6000)^2}{1.2 \times 10^8}}$$

$$= 2.375 + 4.87 \times 10^{-5} x \pm 0.0549\sqrt{0.066 + 8.33 \times 10^{-9}(x - 6000)^2}$$

x, pressure	CI, compact form	CI
2000	2.4723 ± 0.0246	(2.4477, 2.4969)
4000	2.5697 ± 0.0174	(2.5523, 2.5871)
6000	2.6670 ± 0.0142	(2.6528, 2.6812)
8000	2.7643 ± 0.0174	(2.7469, 2.7817)
10000	2.8617 ± 0.0246	(2.8371, 2.8863)

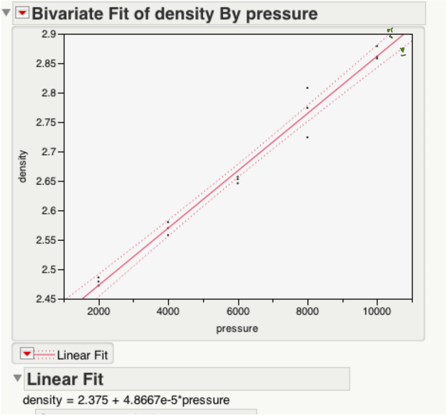
Ceramics; plotting simultaneous confidence regions



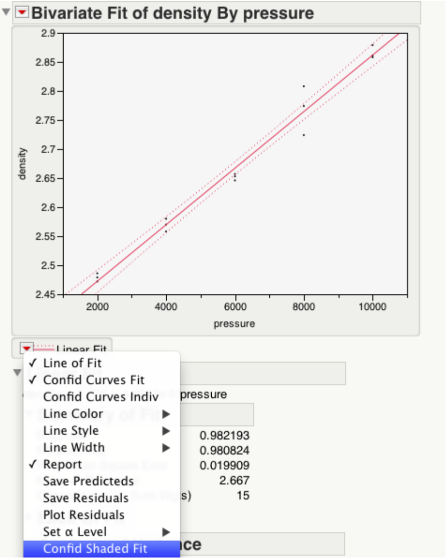
simultaneous confidence intervals / regions.

prediction intervals.

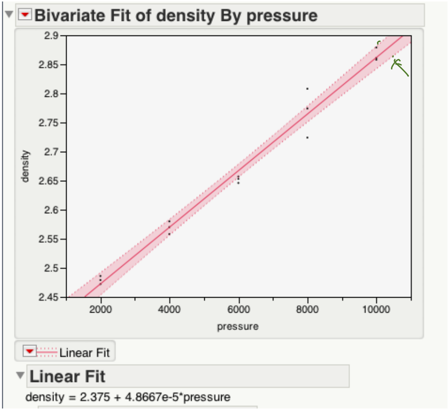
Ceramics; plotting simultaneous confidence regions



Ceramics; plotting simultaneous confidence regions

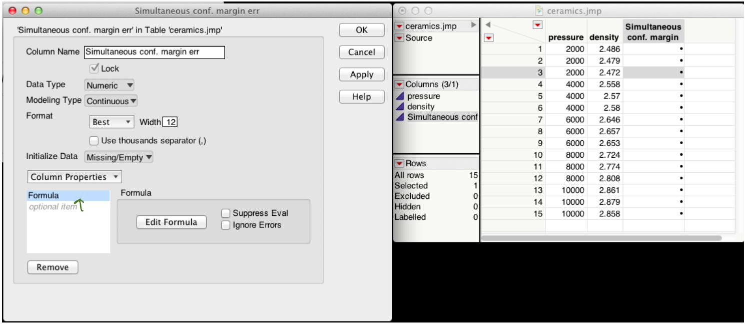


Ceramics; plotting simultaneous confidence regions

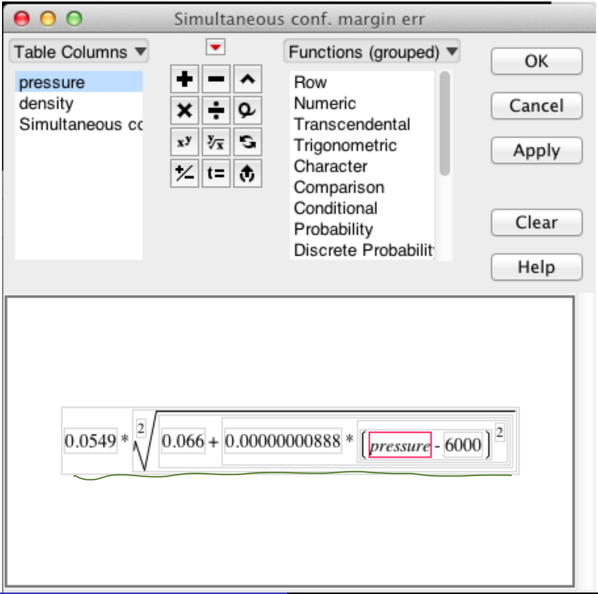


95% confidence region
by default.

Ceramics: calculating the margins of error in JMP



Ceramics: calculating the margins of error in JMP



Ceramics: calculating the margins of error in JMP

ceramics.jmp

ceramics.jmp
Source

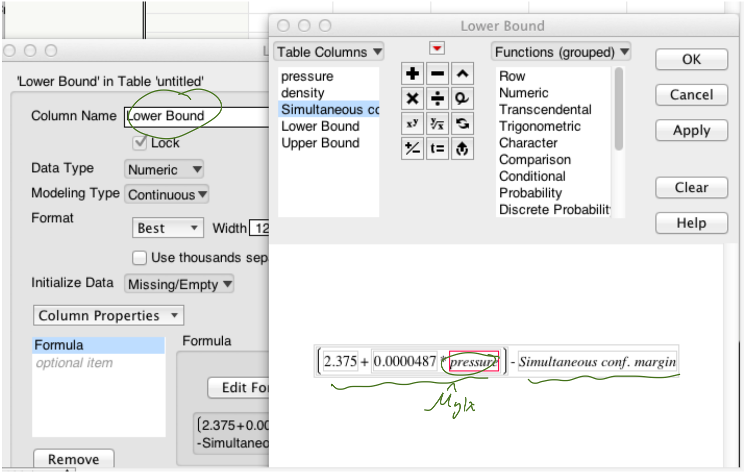
Columns (3/1)
pressure
density
Simultaneous conf

Rows

All rows	15
Selected	1
Excluded	0
Hidden	0
Labelled	0

	pressure	density	Simultaneous conf. margin
1	2000	2.486	0.0245077929
2	2000	2.479	0.0245077929
3	2000	2.472	0.0245077929
4	4000	2.558	0.0173017766
5	4000	2.57	0.0173017766
6	4000	2.58	0.0173017766
7	6000	2.646	0.0141040654
8	6000	2.657	0.0141040654
9	6000	2.653	0.0141040654
10	8000	2.724	0.0173017766
11	8000	2.774	0.0173017766
12	8000	2.808	0.0173017766
13	10000	2.861	0.0245077929
14	10000	2.879	0.0245077929
15	10000	2.858	0.0245077929

Ceramics: calculating the margins of error in JMP



Ceramics: calculating the margins of error in JMP

untitled

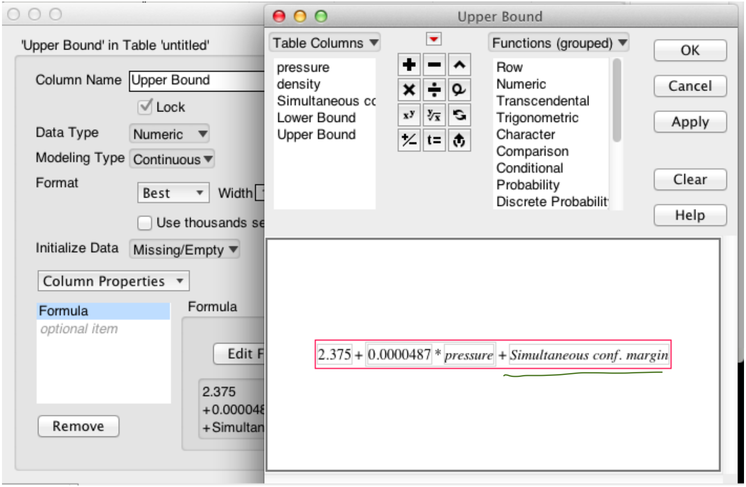
untitled

Columns (5/0)

- pressure
- density
- Simultaneous conf
- Lower Bound
- Upper Bound

	pressure	density	Simultaneous conf. margin	Lower Bound	Upper Bound
1	2000	2.486	0.0245077929	2.4478922071	•
2	2000	2.479	0.0245077929	2.4478922071	•
3	2000	2.472	0.0245077929	2.4478922071	•
4	4000	2.558	0.0173017766	2.5524982234	•
5	4000	2.57	0.0173017766	2.5524982234	•
6	4000	2.58	0.0173017766	2.5524982234	•
7	6000	2.646	0.0141040654	2.6530959346	•
8	6000	2.657	0.0141040654	2.6530959346	•
9	6000	2.653	0.0141040654	2.6530959346	•
10	8000	2.724	0.0173017766	2.7472982234	•
11	8000	2.774	0.0173017766	2.7472982234	•
12	8000	2.808	0.0173017766	2.7472982234	•
13	10000	2.861	0.0245077929	2.8374922071	•
14	10000	2.879	0.0245077929	2.8374922071	•
15	10000	2.858	0.0245077929	2.8374922071	•

Ceramics: calculating the margins of error in JMP



Ceramics: calculating the margins of error in JMP

untitled

Columns (5/1)

- pressure
- density
- Simultaneous conf
- Lower Bound +
- Upper Bound +

	pressure	density	Simultaneous conf. margin	Lower Bound	Upper Bound
1	2000	2.486	0.0245077929	2.4478922071	2.4969077929
2	2000	2.479	0.0245077929	2.4478922071	2.4969077929
3	2000	2.472	0.0245077929	2.4478922071	2.4969077929
4	4000	2.558	0.0173017766	2.5524982234	2.5871017766
5	4000	2.57	0.0173017766	2.5524982234	2.5871017766
6	4000	2.58	0.0173017766	2.5524982234	2.5871017766
7	6000	2.646	0.0141040654	2.6530959346	2.6813040654
8	6000	2.657	0.0141040654	2.6530959346	2.6813040654
9	6000	2.653	0.0141040654	2.6530959346	2.6813040654
10	8000	2.724	0.0173017766	2.7472982234	2.7819017766
11	8000	2.774	0.0173017766	2.7472982234	2.7819017766
12	8000	2.808	0.0173017766	2.7472982234	2.7819017766
13	10000	2.861	0.0245077929	2.8374922071	2.8865077929
14	10000	2.879	0.0245077929	2.8374922071	2.8865077929
15	10000	2.858	0.0245077929	2.8374922071	2.8865077929