Inference for Multiple Regression

Will Landau

Iowa State University

Apr 18, 2013

Outline

Multiple Regression: a Review

Estimating σ^2

Standardized Residuals

Inference for $\beta_0, \beta_1, \dots, \beta_{p-1}$

Inference for Mean Responses
- Individual mean responses
- Multiple mean responses

Multiple Regression

- **Multiple Regression**: regression on multiple variables:

$$y_i \approx b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3} + \cdots + b_{p-1} x_{i,p-1}$$

  - The p coefficients $b_0, b_1, \ldots, b_{p-1}$ are estimated by minimizing the loss function below using the least squares principle:

$$S(b_0, b_1, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i - b_0 + b_1 x_{i,1} + \cdots + b_{p-1} x_{i,p-1})^2$$

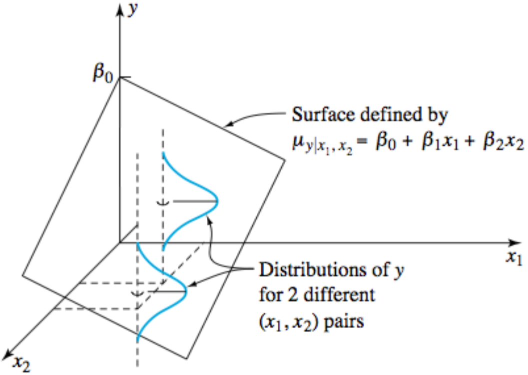
  - In practice, we make a computer find the coefficients for us. This class uses JMP 10, a statistical software tool.

Formalizing the multiple regression model

- Now, we'll work with a formal multiple regression model:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$$

- Assume $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim$ iid $N(0, \sigma^2)$.



Outline

Estimating σ^2

- Now, the residuals are of the form:

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - (b_0 + b_1 x_{1,i} + \cdots + b_{p-1} x_{p-1,i}) \end{aligned}$$

- We estimate the variance with the **surface-fitting sample variance**, also called **mean squared error (MSE)**:

$$s_{SF}^2 = \frac{1}{n-p} \sum e_i^2$$

- The estimated standard deviation is $s_{SF} = \sqrt{s_{SF}^2}$.
- Note: the line fitting sample variance s_{LF}^2 is the special case of s_{SF}^2 for $p = 2$.

Example: stack loss

1. Consider a chemical plant that makes nitric acid from ammonia.
2. We want to predict stack loss (y, 10 times the % ammonia that escapes from the absorption column) using:
   - x_1: air flow, the rate of operation of the plant
   - x_2, inlet temperature of the cooling water
   - x_3: (% circulating acid - 50%)$\times$10

Example: stack loss

i, Observation Number	x_{1i}, Air Flow	x_{2i}, Cooling Water Inlet Temperature	x_{3i}, Acid Concentration	y_i, Stack Loss
1	80	27	88	37
2	62	22	87	18
3	62	23	87	18
4	62	24	93	19
5	62	24	93	20
6	58	23	87	15
7	58	18	80	14
8	58	18	89	14
9	58	17	88	13
10	58	18	82	11
11	58	19	93	12
12	50	18	89	8
13	50	18	86	7
14	50	19	72	8
15	50	19	79	8
16	50	20	80	9
17	56	20	82	15

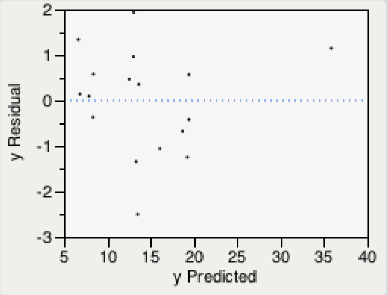
Fitted surface:
$$\widehat{y}_i = -37.65 + 0.797x_{1,i} + 0.577x_{2,i} - 0.067x_{3,i}$$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-37.65246	4.732051	-7.96	<.0001*
x1	0.7976856	0.067439	11.83	<.0001*
x2	0.5773405	0.165969	3.48	0.0041*
x3	-0.06706	0.061603	-1.09	0.2961

Effect Tests

Residual by Predicted Plot



Example: stack loss

Summary of Fit

RSquare	0.975006
RSquare Adj	0.969238
Root Mean Square Error	$\boxed{1.252714}$
Mean of Response	14.47059
Observations (or Sum Wgts)	17

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	795.83449	265.278	169.0432
Error	13	20.40080	$\boxed{1.569}$	**Prob > F**
C. Total	16	816.23529		<.0001*

- $s_{SF}^2 = 1.569$ ("Mean Square Error", blue)
- $s_{SF} = \sqrt{1.569} = 1.25$, also under "Root Mean Square Error" (red).

Outline

Standardized residuals

- As with simple linear regression, $\text{Var}(e_i)$ is not constant even though $\text{Var}(\varepsilon_i) = \sigma^2$.
- There are some constants $a_1, a_2, \ldots, a_n$ such that:

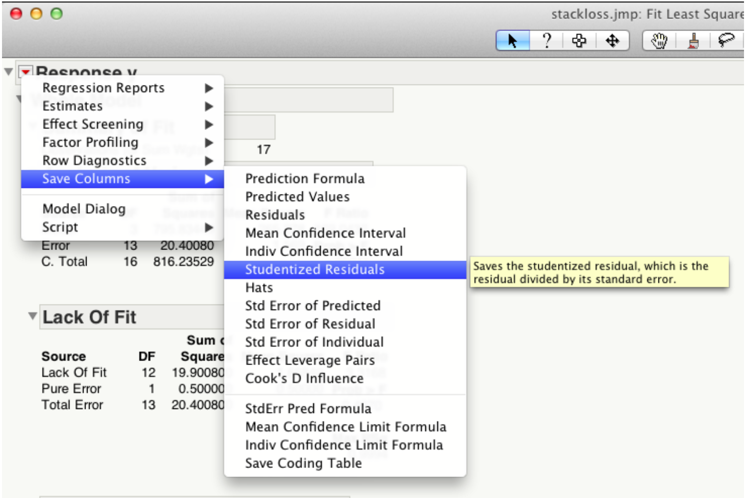
$$Var(e_i) = a_i \sigma^2$$

- Hence, we compute the standardized residuals as:

$$e_i^* = \frac{e_i}{s_{SF}\sqrt{a_i}}$$

- In practice, $a_1, \ldots, a_n$ are hard to compute. We'll make JMP do all the hard work.

Example: stack loss



Example: stack loss

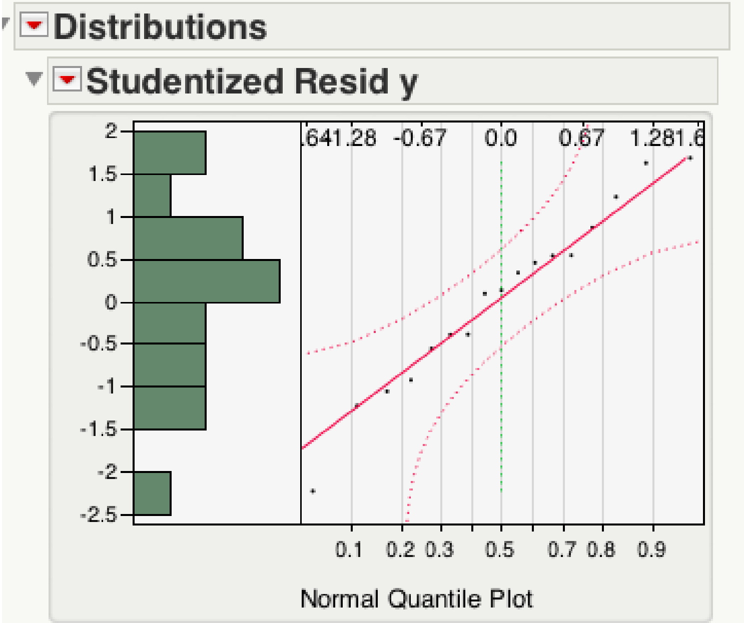
stackloss.jmp

stackloss.jmp

	x1	x2	x3	y	Studentized Resid y
1	80	27	88	37	1.6699301413
2	62	22	87	18	-0.559158154
3	62	23	87	18	-1.057298349
4	62	24	93	19	-0.391146767
5	62	24	93	20	0.5321959564
6	58	23	87	15	-0.928385357
7	58	18	80	14	0.3282961878
8	58	18	89	14	0.8586448878
9	58	17	88	13	0.4481300771
10	58	18	82	11	-2.228855416
11	58	19	93	12	-1.227567579
12	50	18	89	8	1.2152152611
13	50	18	86	7	0.1241939003
14	50	19	72	8	-0.394983744
15	50	19	79	8	0.0849968417
16	50	20	80	9	0.5271747049
17	56	20	82	15	1.6133138635

Columns (5/0)
- x1
- x2
- x3
- y
- Studentized Resid

Example: stack loss



Outline

Inference for $\beta_0, \beta_1, \ldots, \beta_{p-1}$

- Our formal model is:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$$

- Our estimated model is:

$$\widehat{y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_{p-1} x_{p-1,i}$$

- How close are the estimates to their true values?

Inference for $\beta_0, \beta_1, \ldots, \beta_p$

- Under our model assumptions:

$$b_l \sim N(\beta_l, d_l \sigma^2)$$

for some positive constant d_l, $l = 0, 1, 2, \ldots, p-1$.

- That means:

$$\frac{b_l - \beta_l}{s_{SF}\sqrt{d_l}} = \frac{b_l - \beta_l}{\widehat{SD}(b_l)} \sim t_{n-p}$$

- A test statistic for testing $H_0 : b_l = \#$ is:

$$K = \frac{b_l - \#}{s_{SF}\sqrt{d_l}} = \frac{b_l - \#}{\widehat{SD}(b_l)} \sim t_{n-p}$$

- A 2-sided $1 - \alpha$ confidence interval for β_l is:

$$b_l \pm t_{n-p,\ 1-\alpha/2} \cdot s_{SF}\sqrt{d_l}$$

i.e.,

$$b_l \pm t_{n-p,\ 1-\alpha/2} \cdot \widehat{SD}(d_l)$$

Your turn

- $n = 17$
- x_1: air flow, the rate of operation of the plant
- x_2, inlet temperature of the cooling water
- x_3: (% circulating acid - 50%)$\times$10

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-37.65246	4.732051	-7.96	<.0001*
x1	0.7976856	0.067439	11.83	<.0001*
x2	0.5773405	0.165969	3.48	0.0041*
x3	-0.06706	0.061603	-1.09	0.2961

1. Test $H_0 : \beta_1 = 1$ vs. $H_a : \beta_1 < 1$ using $\alpha = 0.1$.
2. Test $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$ by hand ($\alpha = 0.05$), and compare your t statistic to the one in the output table.
3. Construct and interpret a 2-sided 99% confidence interval for β_3.
4. Construct and interpret a 1-sided lower 90% confidence interval for β_2

Answers: 1

1. $H_0 : \beta_1 = 1$, $H_a : \beta_1 < 1$
2. $\alpha = 0.1$
3. I use the test statistic:

$$K = \frac{b_1 - 1}{\widehat{SD}(b_1)}$$

- I assume:
  - H_0 is true.
  - The model $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_1 x_{2,i} + \beta_1 x_{3,i} + \varepsilon_i$, with $\varepsilon_1, \ldots, \varepsilon_{17} \sim N(0, \sigma^2)$ is correct.
- Under the assumptions, $K \sim t_{n-p} = t_{17-4} = t_{13}$.
- I will reject H_0 if $K < -t_{13,1-\alpha} = t_{13,0.9} = 1.35$.

Answers: 1

4. The moment of truth:

$$K = \frac{0.7977 - 1}{0.06744} = -3.00$$

5. With $K = -3 < -1.35 = -t_{13,0.9}$, we reject H_0 and conclude H_a.
6. There is enough evidence to conclude that the true slope on airflow is less than 1 unit stack loss / unit airflow. With each unit increase in airflow and all the other covariates held constant, we expect stack loss to increase by less than one unit.

Answers: 2

1. $H_0 : \beta_3 = 0, \; H_a : \beta_3 \neq 0$
2. $\alpha = 0.05$
3. I use the test statistic:

$$K = \frac{b_3 - 0}{\widehat{SD}(b_3)}$$

- I assume:
  - H_0 is true.
  - The model $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_1 x_{2,i} + \beta_1 x_{3,i} + \varepsilon_i$, with $\varepsilon_1, \ldots, \varepsilon_{17} \sim N(0, \sigma^2)$ is correct.
- Under the assumptions, $K \sim t_{n-p} = t_{17-4} = t_{13}$.
- I will reject H_0 if $|K| > |t_{13,1-\alpha/2}| = t_{13,0.975} = 2.16$.

Answers: 2

4. The moment of truth:

$$K = \frac{-0.06706 - 0}{0.0616} = -1.089 \quad \text{(agrees with the ``t Ratio'')}$$

5. With $|K| = 1.089 < 2.16$, we fail to reject H_0.
6. There is not enough evidence to conclude that the true slope on circulating acid (shifted and scaled) is nonzero. With each unit increase acid and all the other covariates held constant, there is no evidence that the stack loss should change.

Answers: 3

- For a confidence level of 99%, $\alpha = 0.01$ and so $t_{n-p,1-\alpha/2} = t_{13,0.995} = 3.012$.

$$\begin{aligned}&\left(b_3 - t_{n-p,1-\alpha/2} \cdot \widehat{SD}(b_3),\ b_3 + t_{n-p,1-\alpha/2} \cdot \widehat{SD}(b_3)\right)\\ &= (-0.06706 - 3.012 \cdot 0.0616,\ -0.06706 + 3.012 \cdot 0.0616)\\ &= (-0.2525,\ 0.1185)\end{aligned}$$

- We're 99% confident that, for every unit increase in acid with all other covariates held constant, stack loss increases by anywhere from -0.2525 units to 0.1185 units.

Answers: 4

- For a confidence level of 90%, $\alpha = 0.1$ and so $t_{n-p,1-\alpha/2} = t_{13,0.95} = 1.77$.

$$\begin{aligned}
&\left(b_2 - t_{n-p,1-\alpha/2} \cdot \widehat{SD}(b_2),\ b_2 + t_{n-p,1-\alpha/2} \cdot \widehat{SD}(b_2)\right) \\
&= (0.5573 - 1.77 \cdot 0.166,\ 0.5573 + 1.77 \cdot 0.166) \\
&= (0.26348,\ 0.8511)
\end{aligned}$$

- We're 90% confident that, for every 1-degree increase in temperature with all other covariates held constant, stack loss increases by anywhere from 0.26348 units to 0.8511 units.

Outline

Individual mean responses

- We want to estimate the mean response at the set of covariate values, $(x_1, x_2, \ldots, x_{p-1})$
- Under the model assumptions, the estimated mean response, $\widehat{\mu}_{y|\mathbf{x}}$, at $\mathbf{x} = (x_1, x_2, \ldots, x_{p-1})$ is normally distributed with:

$$E(\widehat{\mu}_{y|\mathbf{x}}) = \mu_{y|\mathbf{x}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$
$$Var(\widehat{\mu}_{y|\mathbf{x}}) = \sigma^2 A^2$$

  for some constant A.
- Under the model assumptions:

$$Z = \frac{\widehat{\mu}_{y|\mathbf{x}} - \mu_{y|\mathbf{x}}}{\sigma A} \sim N(0,1) \quad T = \frac{\widehat{\mu}_{y|\mathbf{x}} - \mu_{y|\mathbf{x}}}{s_{SF} A} \sim t_{n-p}$$

- A test statistic for testing $H_0 : \mu_{y|\mathbf{x}} = \#$ is:

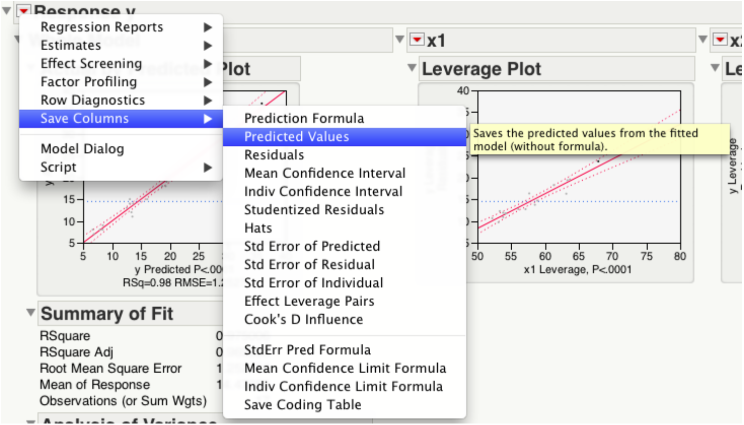
$$\frac{\widehat{\mu}_{y|\mathbf{x}} - \#}{s_{SF} A}$$

  which has a t_{n-p} distribution under H_0.
- A 2-sided, $1 - \alpha$ confidence interval for $\mu_{y|\mathbf{x}}$ in compact form is $\widehat{\mu}_{y|\mathbf{x}} \pm t_{n-p,1-\alpha/2} \cdot s_{SF} \cdot A$.
- Note: $s_{SF} A = \widehat{SD}(\widehat{\mu}_{y|\mathbf{x}})$, which you can get directly from JMP output.

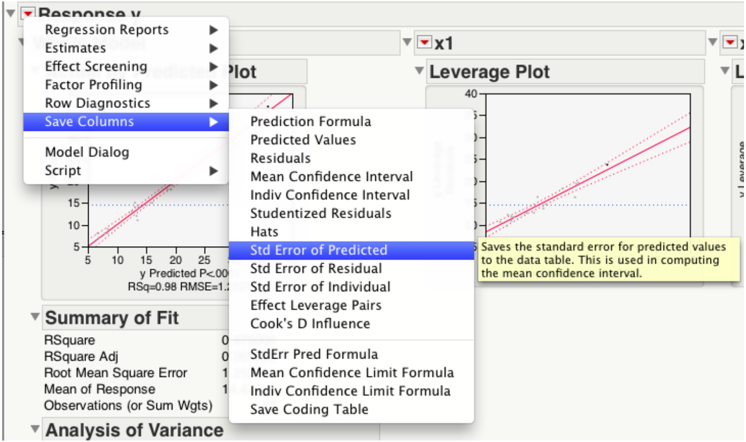
Example: stack loss

- I will use JMP to compute a 2-sided 95% confidence interval around the mean response at point 3: $x_1 = 62, x_2 = 23, x_3 = 87, y = 18.$

Example: stack loss



Example: stack loss



Example: stack loss

	x1	x2	x3	y	Predicted y	StdErr Pred y
1	80	27	88	37	35.849282687	1.0461642094
2	62	22	87	18	18.671300496	0.35771273
3	62	23	87	18	19.248640953	0.417845385
4	62	24	93	19	19.423620349	0.6295687471
5	62	24	93	20	19.423620349	0.6295687471
6	58	23	87	15	16.057898713	0.5204068064
7	58	18	80	14	13.640617664	0.6090546656
8	58	18	89	14	13.037076072	0.5582571612
9	58	17	88	13	12.526795792	0.6739851764
10	58	18	82	11	13.50649731	0.5519432283
11	58	19	93	12	13.346175822	0.6055705716
12	50	18	89	8	6.6555915917	0.5876767248
13	50	18	86	7	6.8567721223	0.4891659484
14	50	19	72	8	8.3729550563	0.8232400377
15	50	19	79	8	7.903533818	0.5302896274
16	50	20	80	9	8.4138140985	0.5769617708
17	56	20	82	15	13.065807105	0.3632418427

Example: stack loss

- With $t_{n-p,1-\alpha/2} = t_{13,0.975} = 2.16$ the confidence interval is:

$$
\begin{aligned}
&(\widehat{\mu}_{y|\mathbf{x}} - 2.16 \cdot \widehat{SD}(\widehat{\mu}_{y|\mathbf{x}}),\ \widehat{\mu}_{y|\mathbf{x}} + 2.16 \cdot \widehat{SD}(\widehat{\mu}_{y|\mathbf{x}})) \\
&= (19.249 - 2.16 \cdot 0.419,\ 19.249 + 2.16 \cdot 0.419) \\
&= (10.199,\ 28.299)
\end{aligned}
$$

- We're 95% confident that when the air flow is 62, the temperature is 23 degrees, and the adjusted percentage of circulating acid is 87, the true mean stack loss is anywhere between 10.199 and 28.299 units.

Multiple mean responses

- The multiple $1-\alpha$ confidence interval formula for $\mu_{y|x_1,\ldots,x_{p-1}}$ is:

$$\widehat{\mu}_{y|\mathbf{x}} \pm \sqrt{p \cdot F_{p,\ n-p,\ 1-\alpha}} \cdot s_{SF} \cdot A$$

- Since there's no simple formula for A, we'll make JMP do all the work for us.
- First, we'll need to write $\widehat{SD}(\widehat{\mu}_{y|\mathbf{x}}) = s_{SF} \cdot A$ and write the interval as:

$$\widehat{\mu}_{y|\mathbf{x}} \pm \sqrt{p \cdot F_{p,\ n-p,\ 1-\alpha}} \cdot \widehat{SD}(\widehat{\mu}_{y|\mathbf{x}})$$

Example: stack loss

- With p parameters and 95% confidence intervals, $F_{p,\ n-p,\ 1-\alpha} = F_{4,13,0.95} = 3.18$.
- The multiple confidence interval becomes:

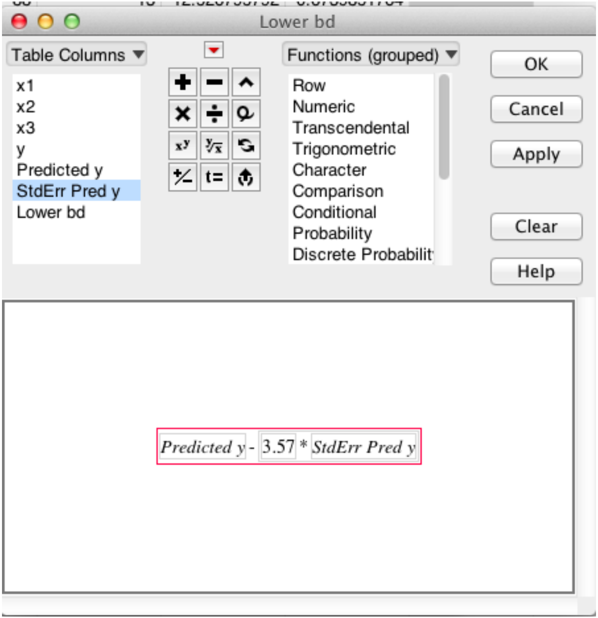
$$\widehat{\mu}_{y|\mathbf{x}} \pm \sqrt{4 \cdot 3.18} \cdot \widehat{SD}(\widehat{\mu}_{y|\mathbf{x}})$$

i.e.,

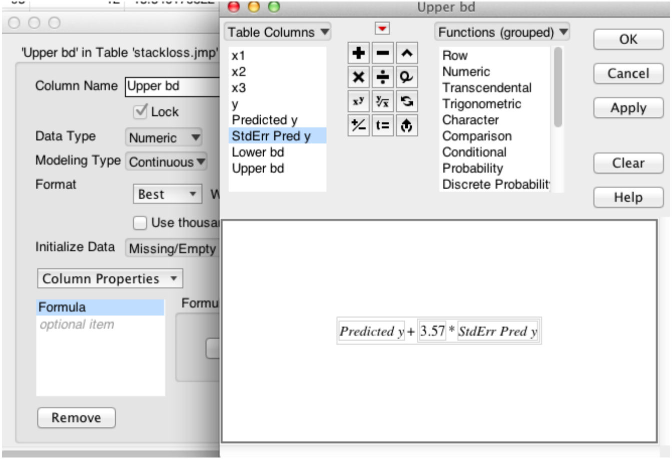
$$\widehat{\mu}_{y|\mathbf{x}} \pm 3.57 \cdot \widehat{SD}(\widehat{\mu}_{y|\mathbf{x}})$$

- $\widehat{\mu}_{y|\mathbf{x}}$ and $\widehat{SD}(\widehat{\mu}_{y|\mathbf{x}})$ vary from point to point.

Example: stack loss



Example: stack loss



Example: stack loss

- The columns, "Lower bd" and "Upper bd", give the endpoints for the confidence intervals.

x1	x2	x3	y	Predicted y	StdErr Pred y	Lower bd	Upper bd
80	27	88	37	35.849282687	1.0461642094	32.114476459	39.584088915
62	22	87	18	18.671300496	0.35771273	17.39426605	19.948334942
62	23	87	18	19.248640953	0.417845385	17.756932929	20.740348978
62	24	93	19	19.423620349	0.6295687471	17.176059922	21.671180776
62	24	93	20	19.423620349	0.6295687471	17.176059922	21.671180776
58	23	87	15	16.057898713	0.5204068064	14.200046414	17.915751012
58	18	80	14	13.640617664	0.6090546656	11.466292508	15.814942821
58	18	89	14	13.037076072	0.5582571612	11.044098007	15.030054138
58	17	88	13	12.526795792	0.6739851764	10.120668712	14.932922872
58	18	82	11	13.50649731	0.5519432283	11.536059986	15.476934635
58	19	93	12	13.346175822	0.6055705716	11.184288881	15.508062763
50	18	89	8	6.6555915917	0.5876767248	4.557585684	8.7535974993
50	18	86	7	6.8567721223	0.4891659484	5.1104496867	8.603094558
50	19	72	8	8.3729550563	0.8232400377	5.4339881219	11.311921991
50	19	79	8	7.903533818	0.5302896274	6.0103998483	9.7966677878
50	20	80	9	8.4138140985	0.5769617708	6.3540605768	10.47356762
56	20	82	15	13.065807105	0.3632418427	11.769033727	14.362580484