Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

Inference for mean response

$\hat{y} = b_0 + b_1 x_i$

Recall our model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

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

$$E(y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + \underbrace{E(\epsilon_i)}_{=0}$$

true mean response

$$\Rightarrow \mu_{Y|x} = \beta_0 + \beta_1 x_i$$

Now, if some new covariate value x is within the range of the x_i's (we don't extrapolate), we can estimate the true mean response at this new x. i.e

estimate of the mean response

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

But how good is the estimate?

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

Inference for mean response

$\mu_{Y|X} = E(y)$

Under the model, $\hat{\mu}_{Y|x}$ is Normally distributed with

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

and Note: $Var(y) = \sigma^2$

$$\text{Var}\,\hat{\mu}_{Y|x} = \sigma^2\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right)$$

Where x is the individual value of x that we care about estimating $\mu_{Y|x}$ at, and x_i are all x_i's in our data.

So we can construct a $N(0, 1)$ random variable by standardizing.

$$Z = \frac{\hat{\mu}_{Y|x} - \mu_{Y|x}}{\sigma\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}} \sim N(0, 1)$$

$SE(\hat{\mu}_{Y|x})$

Inference for mean response

And when σ is unknown (i.e. basically always), we replace σ with $S_{LF} = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$ where we can get from JMP as **root mean square error (MSE)**. Then

$$T = \frac{\hat{\mu}_{Y|x} - \mu_{Y|x}}{s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}} \sim t_{(n-2)}$$

$\sqrt{MSE}$ $=$ s_{LF}

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

true mean response

estimate of mean response

$$K = \frac{\hat{\mu}_{Y|x} - \#}{s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}}$$

$SE(\hat{\mu}_{y|x}) =$ $s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)} \to \sqrt{var(\hat{\mu}_{y|x})}$

which has a $\boxed{t_{n-2}}$ distribution if 1) H_0 is true and 2) the model is correct.

Simple Linear Regression

Variance Estimation

MSE

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

Inference for Parameters

Inference for mean response

$\hat{\mu}_{Y|X}$

Inference for mean response

SE (estimate)

A 2-sided $(1-\alpha)100\%$ CI for $\mu_{y|x}$ is

* estimate

$$\hat{\mu}_{Y|x} \pm t_{(n-2,1-\alpha/2)} \; * \; s_{LF}\sqrt{(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2})}$$

dist. quantile

not given in JMP

and the one-sided the CI are analogous.

Note:

> in the above formula, $\sum(x_i-\overline{x})^2$ is not given by default in JMP.

JMP Shortcut Notice

Inference for mean response

MSE in JMP

individual value we care about estimating at. $\hat{y}|x$

Using JMP we can get

$$s_{LF}\sqrt{\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}=\sqrt{\left(\frac{s_{LF}^2}{n}+(x-\overline{x})^2\frac{s_{LF}^2}{\sum(x_i-\overline{x})^2}\right)}$$

easy to calculate

Note that: * not given in JMP

> We can get $\hat{Var}(b_1)$ from JMP as $(SE(b_1))^2$

$\hat{Var}(b_1)$

Example:[Ceramic powder pressing]

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

Return to the ceramic density problem. We will make a 2-sided 95% confidence interval for the true mean density of ceramics at 4000 psi and interpret it. (Note: $\overline{x} = 6000$)

x=

solution: first find an estimate for true mean density of ceramics at x = 4000.

$$\hat{\mu}_{Y|x=4000} = \hat{y} = b_0 + b_1 x$$

$$= 2.375 + 4.8667 \times 10^{-5} \times (4000) = 2.569668$$

and next, find the SE of $\hat{\mu}_{Y|x=4000}$:

$$s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$$

JMP MSE

$$= \sqrt{\left(\frac{s_{LF}^2}{n} + (x-\overline{x})^2\frac{s_{LF}^2}{\sum(x_i-\overline{x})^2}\right)}$$

JMP:

4000

6000

$var(b_1) = [SE(b_1)]^2$

Example:[Ceramic powder pressing]

$$= \sqrt{\frac{0.000396}{15} + (\underbrace{4000}_{x} - \underbrace{6000}_{\bar{x}})^2 \underbrace{(1.817 \times 10^{-6})}_{SE(b_1)}^2}$$

$$= \sqrt{0.000039606}$$

$$SE(\hat{\mu}_{Y|x=4000}) = 0.0062933 \checkmark$$

Therefore, a two-sided 95% confidence interval for the true mean density at 4000 psi is

$$\hat{\mu}_{Y|x=4000} \pm t_{(n-2,1-\alpha/2)} \times \overbrace{s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}}^{SE(\hat{\mu}_{Y|x=4000})}$$

$$= 2.569648 \pm \boxed{t_{(15-2,0.975)}} \times (0.0062933)$$

$$= 2.569648 \pm 2.160 \times (0.0062933) = \underbrace{(2.5561\ ,\ 2.5833)}_{\mu_{Y|x=4000}}$$

We are 95% cofident that the true mean density of the ceramics at 4000 psi is between 2.5561 and 2.5833.

Example:[Ceramic powder pressing]

Now calculate and interpret a 2-sided 95% confidence interval for the true mean density at 5000 psi. = x

① Find an estimate for true mean density at x=5000.

$$\hat{\mu}_{Y|x=5000} = \hat{y} = b_0 + b_1 x$$

$$= 2.375 + 4.8667 \times 10^{-5} \times (5000) = 2.618335$$

and ② Find SE($\hat{\mu}_{Y|x=5000}$)

$$s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$$

$$= \sqrt{\left(\frac{s_{LF}^2}{n} + (x-\overline{x})^2 \frac{s_{LF}^2}{\sum(x_i-\overline{x})^2}\right)}$$

→ JMP: $[SE(b_1)]^2$ *

x: 5000, $\overline{x}$: 6000

$$= \sqrt{\frac{0.00395}{15} + (5000-6000)^2(1.817 \times 10^{-6})^2}$$

$$= \sqrt{0.00002970} = 0.005449$$ $= SE(\hat{\mu}_{Y|x=5000})$

Example:[Ceramic powder pressing]

Therefore, a two-sided 95% confidence interval for the true mean density at 4000 psi is

$$\hat{\mu}_{Y|x=4000} \pm t_{(n-2,1-\alpha/2)} \times \underbrace{s_{LF}\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}}_{SE(\hat{\mu}_{Y|x=5000})}$$

$$= \boxed{2.618335} \pm \boxed{t_{(15-2,0.975)}} \times (0.005449)$$

$$= 2.618335 \pm 2.160 \times (0.005449)$$

$$= (2.60656\ ,\ 2.63011)\ *$$

We are 95% cofident that the true mean density of the ceramics at 4000 psi is between 2.60656 and 2.63011

Multiple Linear Regression

Multiple linear regression

Recall the summarization the effects of several different quantitative variables $x_1, \ldots, x_{p-1}$ on a response y.

$$y_i \approx \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i}$$

Where we estimate $\beta_0, \ldots, \beta_{p-1}$ using the *least squares principle* by minimizing the function

$$S(b_0, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1,i} - \cdots$$

$$- \beta_{p-1} x_{p-1,i})^2$$

to find the estimates $b_0, \ldots, b_{p-1}$.

We can formalize this now as

$$\rightarrow Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i$$

where we assume $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Variance Estimation in MLR

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Variance estimation

Based on our multiple regression model, the residuals are of the form

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1\,i} + \cdots + b_{p-1} x_{p-1\,i}$$

And we can estimate the variance similarly to the SLR case.

Definition:

For a set of n data vectors $(x_{11}, x_{21}, \ldots, x_{p-11}, y), \ldots, (x_{1n}, x_{2n}, \ldots, x_{p-1n}, y)$ where least squares fitting is used to fit a surface,

previously: s^2_{LF}

$$s^2_{SF} = \frac{1}{n-p} \sum (y - \hat{y})^2 = \frac{1}{n-p} \sum e_i^2$$

is the **surface-fitting sample variance** (also called mean square error, MSE). Associated with it are $\nu = n - p$ degrees of freedom and an estimated standard deviation of response $s_{SF} = \sqrt{s^2_{SF}}$.

Variance estimation

Note: the SLR fitting sample variance s_{LF}^2 is the special case of s_{SF}^2 for $\boxed{p = 2.}$

Example:[Stack loss]

Consider a chemical plant that makes nitric acid from ammonia. We want to predict stack loss (y, 10 times the % of ammonia lost) using

> x_1: air flow into the plant
>
> x_2: inlet temperature of the cooling water
>
> x_3: modified acid concentration (% circulating acid -50%) $\times$ 10

$$\overset{1}{y = \beta_0} + \overset{2}{\beta_1 x_1} + \overset{3}{\beta_2 x_2} + \overset{4}{\beta_3} x_3 + \varepsilon$$

$p = ?$

$p - 1 = 3 \Rightarrow \boxed{p = 4}$

$t_{(n-p)}$

Example:[Stack loss]

Summary of Fit

RSquare	0.975006
RSquare Adj	0.969238
Root Mean Square Error	1.252714
Mean of Response	14.47059
Observations (or Sum Wgts)	17

$\sqrt{MSE} = \sqrt{s_{SF}^2} = s_{SF}$

n =

Analysis of Variance

ANOVA table

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

$MSE = s_{SF}^2$

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

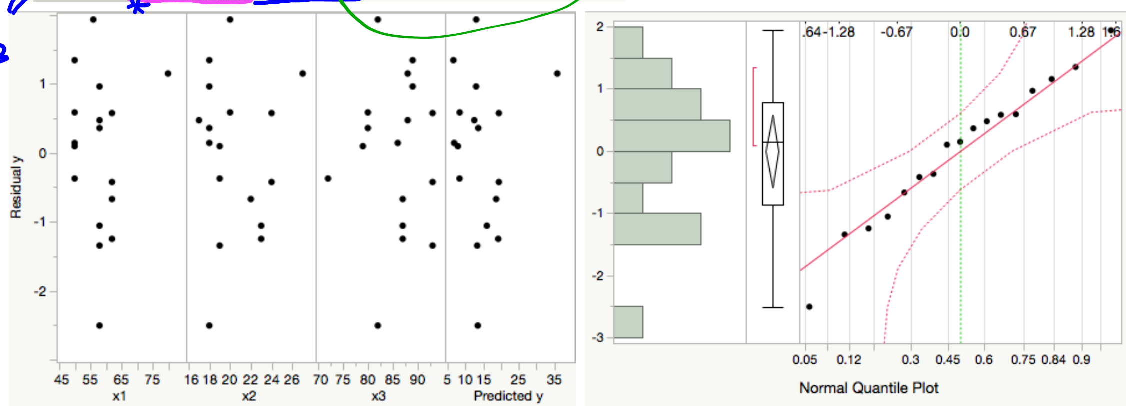
$\hat\beta_0 = b_0 =$

$b_1 =$

b_2

b_3

$SE(b_i)$; $i = 0,1,2,3$



Example:[Stack loss]

Then we have the fitted model as

$$\hat{y} = -37.65246 + 0.7977x_1 + 0.5773x_2 - 0.0971x_3$$

> The residual plots VS. x_1 , x_2 x_3 and $\hat{y}$ look like random scatter around zero.
>
> The QQ-plot of the residuals looks linear, indicating that the residuals are Normally distributed.
>
> This model is valid.

Inference for Parameters in MLR

Inference for parameters

We are often interested in answering questions (doing formal inference) for $\beta_0, \ldots, \beta_{p-1}$ individually. For example, we may want to know if there is a significant relationship between y and x_2 (holding all else constant).

~~\vspace{.2in}~~

previously, in SLR:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$$

d_1

Under our model assumptions,

$$b_i \sim N(\beta_i, d_i\sigma^2) \quad var(b_i) \rightarrow$$

for some positive constant $d_i, i = 0, 1, \ldots, p-1$. That are hard to compute analytically, but JMP can help)

That means

$$\sqrt{\widehat{var}(b_i)} \rightarrow \quad \frac{b_i - \beta_i}{s_{SF}\sqrt{d_i}} = \frac{b_i - \beta_i}{SE(b_i)} \sim t_{(n-p)}$$

JMP

Inference for parameters

So, a test statistic for $\mathrm{H}_0 : \beta_i = \#$ is

$$K = \frac{b_i - \#}{s_{LF}\sqrt{d_i}} = \frac{b_i - \#}{SE(b_i)} \sim t_{(n-p)}$$

if 1) H_0 is true and 2) the model is valid, and a 2-sided $(1-\alpha)100\%$ CI for β_i is

$$b_i \pm t_{(n-p,1-\alpha/2)} \times \underbrace{s_{LF}\sqrt{d_i}}_{SE(b_i)}$$

or

$$b_i \pm t_{(n-p,1-\alpha/2)} \times SE(b_i)$$

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

Using the model fit on slide 35, answer the following questions:

1.Is the average change in stack loss (y) for a one unit change in air flow into the plant (x_1) less than 1 (holding all else constant)? Use a significance testing framework with $\alpha = .1$.

solution:

1- $H_0: \boxed{\beta_1 = 1}\ vs. \quad H_1: \ \beta_1 < 1$

2- $\alpha = 0.1$

3- I will use the test statistics $K = \frac{b_1 - 1}{SE(b_1)}$ which has a $t_{n-p} = t_{17-4}$ distribution assuming that

1 • H_0 is true and

2 • The regression model $y_i = \beta_0 + \beta_1 x_{i\,1} + \beta_2 x_{i\,2} + \beta_3 x_{i\,3} + \epsilon_i$ is valid

Example:[Stack loss, cont'd]

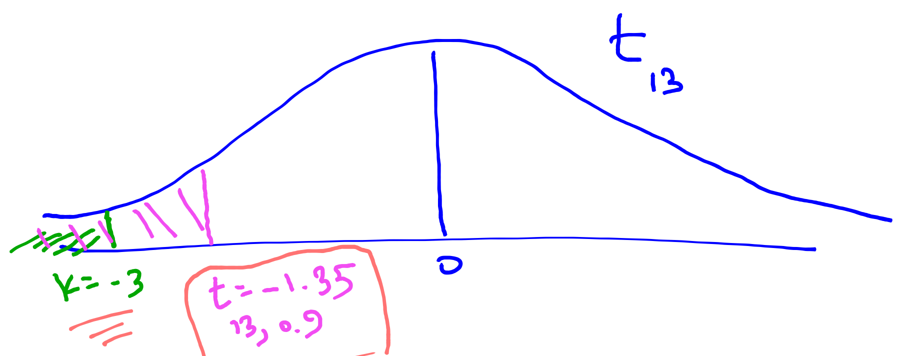
4- $K = \frac{\boxed{0.7977}-1}{0.06744} = \boxed{-3}$ and $t_{(13,.9)} = 1.35$. So,

p-value
$= P(T < K) < P(T < -3) < 0.1 = \alpha$

5- Since $K = -3 < -1.35 = -t_{(13,.9)}$, we **reject** H_0.

6- There is enough evidence to conclude that the slope on airflow is less than one unit stackloss/unit airflow. With each unit increase in airflow and all other covariates held constant, we expect stack loss to increase by less than one unit.

$\beta_1 < 1$



is the slope for $x_3 = 0$ or $\neq 0$

Example:[Stack loss, cont'd]

2. Is the there a significant relationship between stack loss (y) and modified acid concentation (x_3) (holding all else constant)? Use a significance testing framework with $\alpha = .05$.

solution:

1- $H_0 : \boxed{\beta_3 = 0}$ *vs.* $H_1 : \boxed{\beta_3 \neq 0}$

2- $\boxed{\alpha = 0.05}$

3- I will use the test statistics $K = \frac{b_3 - 1}{SE(b_3)}$ which has a $t_{n-p} = t_{17-4}$ distribution assuming that

- H_0 is true and
- The regression model $y_i = \beta_0 + \beta_1 x_{i\,1} + \beta_2 x_{i\,2} + \beta_3 x_{i\,3} + \epsilon_i$ is valid

Example:[Stack loss, cont'd]

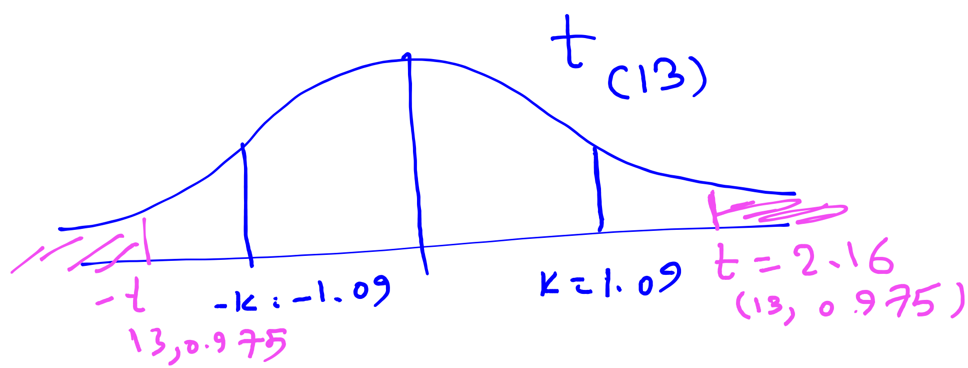
4- $K = \frac{-0.06706-0}{0.0616} = -1.09$ and
$t_{(13,.975)} = 2.16$. So,

p-value $= P(|T| > |K|) =$

$P(|T| > 1.09) > P(|T| > t_{(13,.975)}) = 0.05 = \alpha$

5- Since p-value $> \alpha$, we **fail to reject H_0**.

6- There is not enough evidence to conclude that, with all other covarates held constant, there is a significant linear relatinoship between stack loss and acid concentration.



Example:[Stack loss, cont'd]

3.Construct and interpret a 99% two-sided confidence interval for β_3.

solution:

$$t_{(n-p,1-\alpha/2)} = t_{(13,.995)} = 3.012$$

then

JMP JMP

$$b_3 \pm t_{(n-p,1-\alpha/2)} SE(b_3) = -.0.06706 \pm 3.62(0.0616)$$
$$= (-0.2525 \;\; 0.1185)$$

We are 99% confident that for every unit increase in acid concentration, **with all other covariates held constant**, we expect stack loss to increase anywehre from -0.2525 units to 0.1185 units.

Example:[Stack loss, cont'd]

4.Construct and interpret a two-sided 90% confidence interval for β_2

solution:

For a 90% two-sided CI for β_2,

$$\alpha = 0.1 \ , \ t_{(n-p,1-\alpha/2)} = t_{(13,0.95)} = 1.77$$

Then

JMP JMP

$$b_2 \pm t_{(n-p,1-\alpha/2)} \times SE(b_2) = 0.5773 \pm 1.77(0.166)$$
$$= (0.2834 \ \ 0.87.127)$$

We are 90% confident that for every one degree increase in temprature **with all other covariates held constant**, stack loss is expected to increase by anywhere from 0.2834 units to 0.8713 units.

Inference for Mean Response

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Previously: $\hat{\mu}_{y|x}$ → only one x.

Inference for mean response

$\mu_{y|\underline{x}}$ → multiple x_i's.

We can also 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, $\hat{\mu_{y|\boldsymbol{x}}}$, at $\boldsymbol{x} = (x_1, x_2, \ldots, x_{p-1})$ is **Normally distributed** with:

$$\mathbb{E}(\hat{\mu_{y|\boldsymbol{x}}}) = \mu_{y|\boldsymbol{x}} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \cdots \beta_{p-1} \boldsymbol{x}_{p-1}$$

and

$$Var(\hat{\mu_{y|\boldsymbol{x}}}) = \sigma^2 A^2$$

for some constant A, that is hard to compute by hand.

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Inference for mean response

$\mu_{y|\underline{x}}$ ← vector.

Then, under the model assumptions

$$\boxed{Z} = \frac{\boxed{\hat{\mu_{y|\boldsymbol{x}}}} - \boxed{\mu_{y|\boldsymbol{x}}}}{\sigma A} \sim N(0,1)$$

and

$$T = \frac{\hat{\mu_{y|\boldsymbol{x}}} - \mu_{y|\boldsymbol{x}}}{s_{LF} A}$$

S_{SF}

And a test statistic for testing $\mathrm{H}_0 : \mu_{y|\boldsymbol{x}} = \#$ is

$$K = \frac{\hat{\mu_{y|\boldsymbol{x}}} - \#}{s_{LF} A}$$

which has a $t_{(n-p)}$ distribution under H_0 if the model holds true.

Inference for mean response

A 2-sided $(1-\alpha)100\%$ CI for $\mu_{y|\boldsymbol{x}}$ is

$$\hat{\mu_{y|\boldsymbol{x}}} \pm t_{(n-p,1-\alpha/2)} \times s_{LF}A$$

Note that the one-sided CI will be analogous.

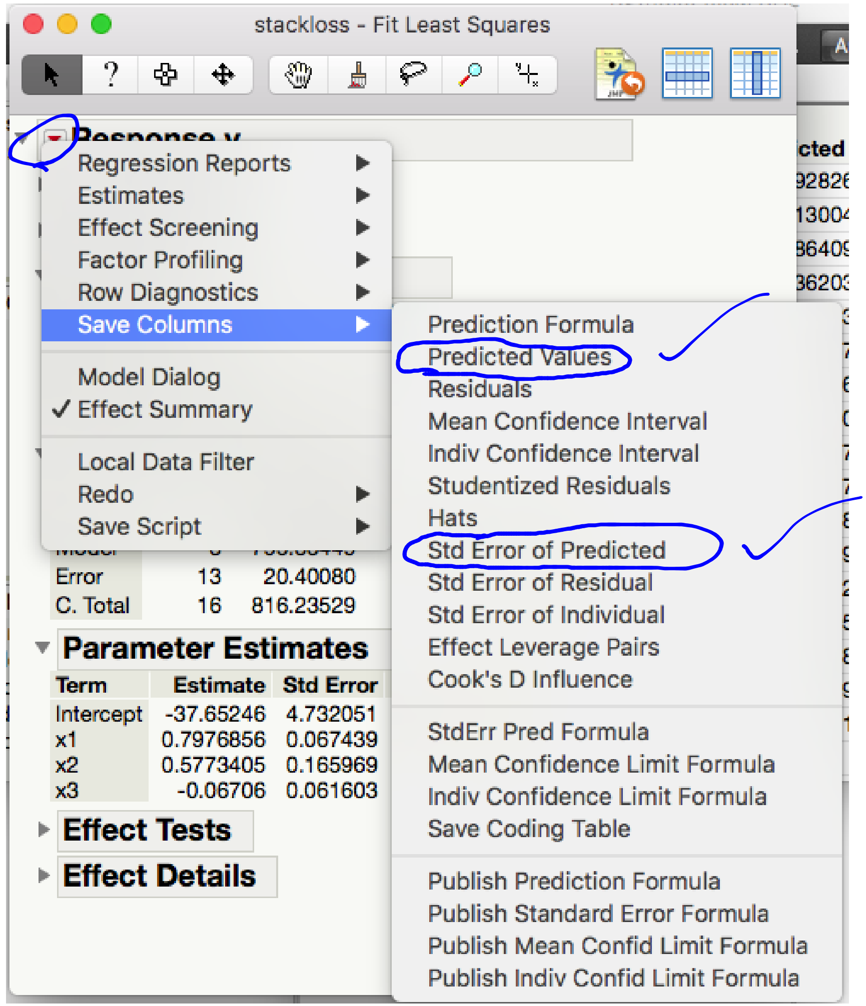
> Note: $S_{LF}A = SE(\hat{\mu_{y|\boldsymbol{x}}})$, and we can use JMP to get this.

Example:[Stack loss, cont'd]

We can use JMP to compute a 2-sided 95\% CI around the mean response at point 3:

$$x_1 = 62, x_2 = 23, x_3 = 87, y = 18$$

Example:[Stack loss, cont'd]



How to get predicted values and standard errors

Example:[Stack loss, cont'd]

$y = 18$

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

$SE(\hat{\mu}_{y|x})$

stackloss

stackloss

Source

Columns (6/0)

x1

x2

x3

y

Predicted y

StdErr Pred y

Rows

All rows	17
Selected	1
Excluded	0
Hidden	0
Labelled	0

Point 3:

	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

Predicted values and standard errors.

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

With $t_{(n-p,1-\alpha/2)} = t_{(13,.975)} = 2.16$, the 95% condidence interval is

$\mu_{y|x}$

$$\hat{\mu_{y|\boldsymbol{x}}} \pm t_{(n-p,1-\alpha/2)} \boxed{SE(\hat{\mu_{y|\boldsymbol{x}}})}$$

$$= 19.2486 \pm 2.16 \times (0.41785)$$

$$= (18.343\ ,\ 20.151)$$

> We are 95% confident that when air flow is 62 units, temperature is 23 degrees and the adjusted percentage of circulating acid is 87 units, the true mean stack loss is between 18.343 and 20.151 units.