

STAT 305: Chapter 9

Inference for curve and surface fitting

Amin Shirazi

ashirazist.github.io/stat305.github.io

Chapter 9:

Inference for curve and surface fitting

Previously, we have discussed how to describe relationships between variables (Ch. 4). We now move into formal inference for these relationships starting with relationships between two variables and moving on to more.

Simple linear regression

Recall, in Ch. 4, we wanted an equation to describe how a dependent (response) variable, y , changes in response to a change in one or more independent (experimental) variable(s), x .

We used the notation

$$y = \beta_0 + \beta_1 x + \epsilon$$

Simple Linear Regression

where β_0 is the intercept.

| It is the expected value for y when $x = 0$.

β_1 is the slope.

| It is the expected increase (decrease) in y for every **one** unit change in x

ϵ is some error. In fact,

$$\epsilon \sim \text{iid } N(0, \sigma^2)$$

Recall:

| Cheking if residuals are normally distributed is one of our model assessment technique.

Goal: We want to use inference to get interval estimates for our slope and predicted values and significance tests that the slope is not equal to zero.

Variance Estimation

Simple Linear Regression

Variance Estimation

Variance estimation

In the simple linear regression $y = \beta_0 + \beta_1 x + \epsilon$, the parameters are β_0 , β_1 and σ^2 .

We already know how to estimate β_0 and β_1 using least squares.

We need an estimate for σ^2 in a *regression*, or "*line-fitting*" context.

Definition:

For a set of data pairs $(x_1, y_1), \dots, (x_n, y_n)$ where least squares fitting of a line produces fitted values $\hat{y}_i = b_0 + b_1 x_i$ and residuals $e_i = y_i - \hat{y}_i$,

$$s_{LF}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

is the **line-fitting sample variance**.

Simple Linear Regression

Variance Estimation

MSE

Variance estimation

Associated with s_{LF}^2 are $\nu = n - 2$ degrees of freedom and an estimated standard deviation of response

$$s_{LF} = \sqrt{s_{LF}^2}.$$

This is also called **Mean Square Error (MSE)** and can be found in *JMP* output.

It has $\nu = n - 2$ degrees of freedom because we must estimate 2 quantities β_0 and β_1 to calculate it.

s_{LF}^2 estimates the level of basic background variation σ^2 , whenever the model is an adequate description of the data.

Inference for Parameter β_0 and β_1

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for parameters

Inference for β_1 :

We are often interested in testing if $\beta_1 = 0$. This tests whether or not there is a *significant linear relationship* between x and y . We can do this using

1. $100^* (1 - \alpha) \%$ confidence interval

2. Formal hypothesis tests

Both of these require

1. An estimate for β_1 and

2. a **standard error** for β_1

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for β_1 :

It can be shown that since $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, then

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

Note that we never know σ^2 , so we must estimate it using $\sqrt{\text{MSE}} = S_{LF}$.

So, a $(1 - \alpha)100\%$ CI for β_1 is

$$b_1 \pm t_{(n-2, 1-\alpha/2)} \frac{s_{LF}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

and the test statistic for $H_0 : \beta_1 = \#$ is

$$K = \frac{b_1 - \#}{\sqrt{\frac{s_{LF}}{\sum(x_i - \bar{x})^2}}}$$

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Example:[Ceramic powder pressing]

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. Consider a pressure/density study of $n = 15$ data pairs representing

$$x = \text{the pressure setting used (psi)}$$
$$y = \text{the density obtained (g/cc)}$$

in the dry pressing of a ceramic compound into cylinders.

Simple Linear Regression

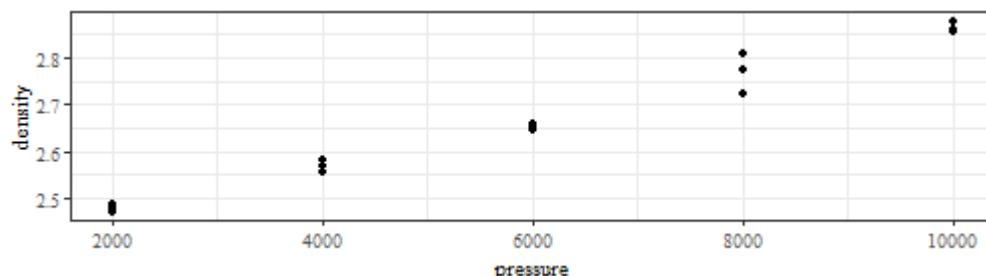
Variance Estimation

MSE

Inference for Parameters

Example:[Ceramic powder pressing]

pressure	density	pressure	density
2000	2.486	6000	2.653
2000	2.479	8000	2.724
2000	2.472	8000	2.774
4000	2.558	8000	2.808
4000	2.570	10000	2.861
4000	2.580	10000	2.879
6000	2.646	10000	2.858
6000	2.657		



Simple Linear Regression

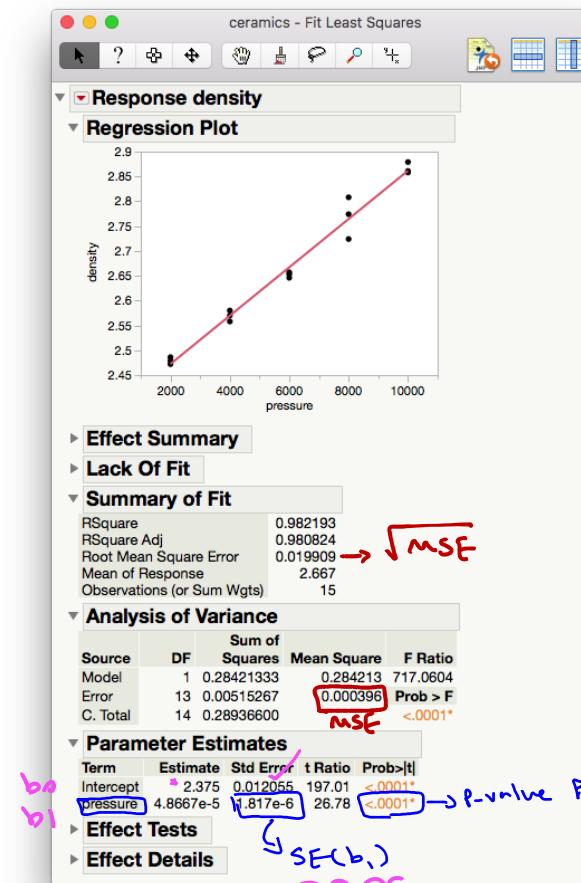
Variance Estimation

MSE

Inference for Parameters

Example:[Ceramic powder pressing]

A line has been fit in JMP using the method of least squares.



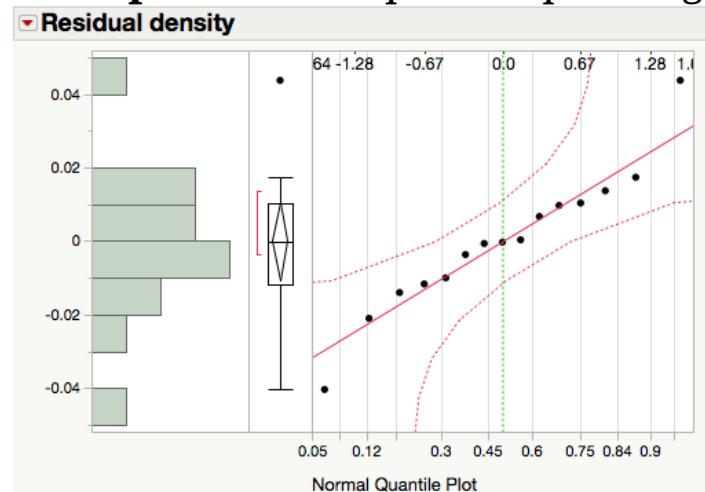
Simple Linear Regression

Variance Estimation

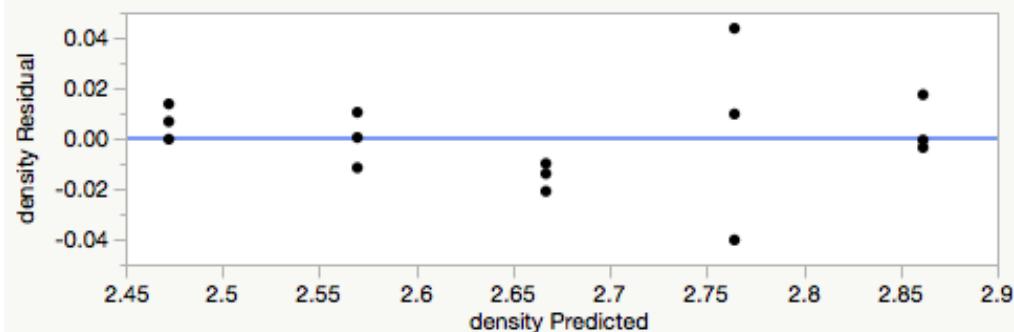
MSE

Inference for Parameters

Example:[Ceramic powder pressing]



Residual by Predicted Plot



Least squares regression of density on pressure of ceramic cylinders

Simple Linear Regression

Example:[Ceramic powder pressing]

1. Write out the model with the appropriate estimates.

Variance Estimation

2. Are the assumptions for the model met?

MSE

Inference for Parameters

3. What is the fraction of raw variation in y accounted for by the fitted equation?

Simple Linear Regression

Example:[Ceramic powder pressing]

4.What is the correlation between x and y ?

Variance Estimation

5.Estimate σ^2 .

MSE

Inference for Parameters

6.Estimate $\text{Var}(b_1)$.

Simple Linear Regression

Example:[Ceramic powder pressing]

7.Calculate and interpret the 95% CI for β_1

Variance Estimation

MSE

Inference for Parameters

8.Conduct a formal hypothesis test at the $\alpha = .05$ significance level to determine if the relationship between density and pressure is significant.

1- $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

2- $\alpha = 0.05$

3- I will use the test statistics $K = \frac{b_1 - \#}{\frac{s_{LF}}{\sum(x_i - \bar{x})^2}}$

which has a t_{n-2} distribution assuming that

- H_0 is true and
- The regression model is valid

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Example:[Ceramic powder pressing]

4-

$$K = \frac{4.8667 \exp -5}{1.817 \exp -6} = 26.7843 > t_{(13,.975)}=2.160.$$

So,

$$\text{p-value} = P(|T| > K) < 0.05 = \alpha$$

5- Since $K = 26.7843 > 2.160 = t_{(13,.975)}$, we **reject H_0** .

6- There is **enough evidence** to conclude that there is **a linear relationship between density and pressure**

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

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

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

$$E(Y) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \underbrace{\beta_0}_{\text{true mean response}} + \underbrace{\beta_1 x_i}_{\mu_{Y|x}} + \underbrace{E(\epsilon_i)}_{=0}$$

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

$$\text{estimate of } \mu_{Y|x} \leftarrow \hat{y} = b_0 + b_1 x$$

the mean response
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

$$\text{Note: } \text{Var}(Y) = \sigma^2$$

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

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

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

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

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

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

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

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

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

true mean response $\hat{\mu}_{Y|x}$ *estimate of mean response*

$SE(\hat{\mu}_{Y|x}) = \sqrt{\text{var}(\hat{\mu}_{Y|x})}$

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

Simple Linear Regression

Variance Estimation

$$MSE = \hat{Y}_T = b_0 + b_1 X$$

Inference for Parameters

Inference for mean response

$\hat{Y}_{|x}$ Inference for mean response

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

* estimate

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

dist. quantile

$t_{(n-2, 1-\alpha/2)}$

*

s_{LF}

$\sqrt{(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2})}$

SE (estimate)

and the one-sided the CI are analogous.

Note:

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

not given
in JMP

JMP Shortcut Notice

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

Inference for mean response

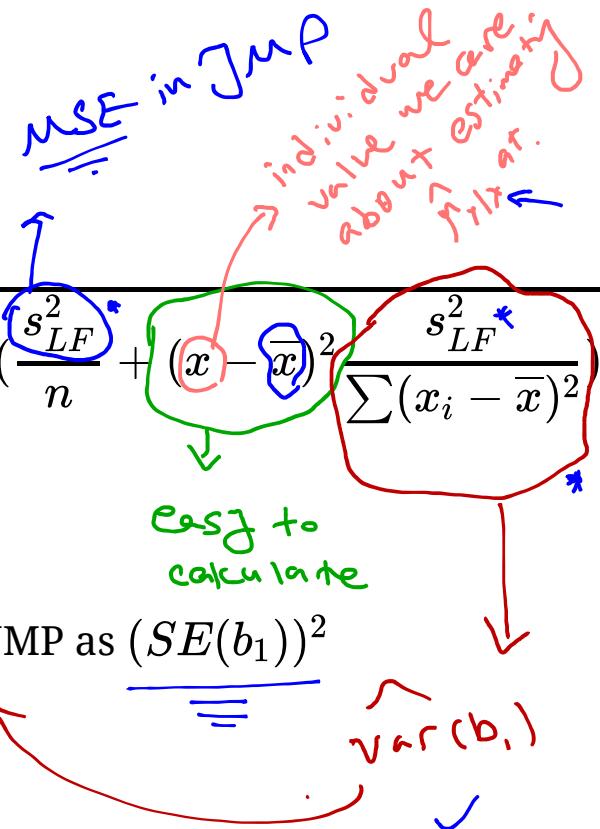
Using JMP we can get

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

Note that:

* not given
in JMP

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



Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

Example:[Ceramic powder pressing]

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: $\bar{x} = 6000$) ✓

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 $\hat{\mu}_{Y|x=4000}$ *

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

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

• 4000

6000

$$\text{var}(b_1) = [SE(b_1)]^2$$

Simple Linear Regression

Variance Estimation

MSE

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

Inference for Parameters

Inference for mean response

Example:[Ceramic powder pressing]

$$\begin{aligned} &= \sqrt{\frac{0.000396}{15} + (4000 - 6000)^2 (1.817 \times 10^{-6})^2} \\ &= \sqrt{0.000039606} \end{aligned}$$

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

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

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

$$= 2.569648 \pm 2.160 \times (0.0062933) = (2.5561, 2.5833)$$

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

Example:[Ceramic powder pressing]

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

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 $SE(\hat{y}|x=5000)$

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

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

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

$$= \sqrt{0.00002970} = 0.005449 = SE(\hat{y}|x=5000) \quad 27 / 53$$

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

Example: [Ceramic powder pressing]

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

$$\hat{\mu}_{Y|x=4000} \pm t_{(n-2, 1-\alpha/2)} \times s_{LF} \sqrt{\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$
$$= 2.618335 \pm t_{(15-2, 0.975)} \times (0.005449) \quad SE(\hat{\mu}_{Y|x=4000})$$
$$= 2.618335 \pm 2.160 \times (0.005449)$$
$$= (2.60656, 2.63011) *$$

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

Multiple Linear Regression

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Multiple linear regression

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

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

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

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

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

We can formalize this now as

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

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

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

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

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

Definition:

For a set of n data vectors

$(x_{11}, x_{21}, \dots, x_{p-11}, y), \dots, (x_{1n}, x_{2n}, \dots, x_{p-1n}, y)$
where least squares fitting is used to fit a surface,

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

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Variance estimation

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



Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

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%) $\times 10$

$$\rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$\begin{aligned} P &=? \\ P-1 &= 3 \Rightarrow P=4 \end{aligned}$$

$$t_{(n-P)}$$

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss]

Summary of Fit					
RSquare RSquare Adj Root Mean Square Error Mean of Response Observations (or Sum Wgts)					
0.975006 0.969238 1.252714 14.47059 17					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	3	795.83449	265.278	169.0432	
Error	13	20.40080	1.569		<.0001*
C. Total	16	816.23529			
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	



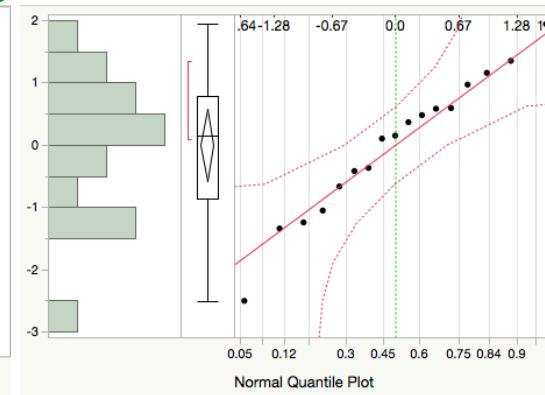
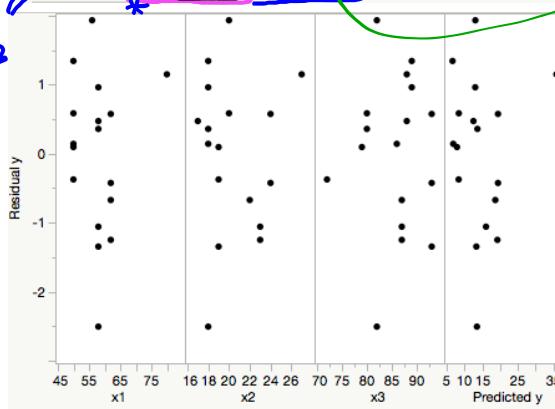
$$\sqrt{MSE} = \sqrt{S_{SF}^2} = S_{SF}$$

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	
C. Total	16	816.23529		

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



$$SE(\hat{b}_i); i=0, 1, 2, 3$$

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss]

Then we have the fitted model as

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

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Inference for parameters

We are often interested in answering questions (doing formal inference) for $\beta_0, \dots, \beta_{p-1}$ 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:

Under our model assumptions,

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

$b_i \sim N(\beta_i, d_i \sigma^2)$ $\text{var}(b_i) \rightarrow \sum (x_i - \bar{x})^2$

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

That means

$$\frac{b_i - \beta_i}{s_{\text{SE}} \sqrt{d_i}} = \frac{b_i - \beta_i}{\text{SE}(b_i)} \sim t_{(n-p)}$$

$\sqrt{\text{var}(b_i)} \rightarrow s_{\text{SE}}$ JMP

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Inference for parameters

So, a test statistic for $H_0 : \beta_i = \#$ is

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

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

$$\rightarrow b_i \pm t_{(n-p, 1-\alpha/2)} \times s_{\text{CF}} \sqrt{d_i}$$

\uparrow
 $s_{\text{Ec } b_i})$

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

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:

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

$$2 - \alpha = 0.1$$

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 $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$ is valid

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

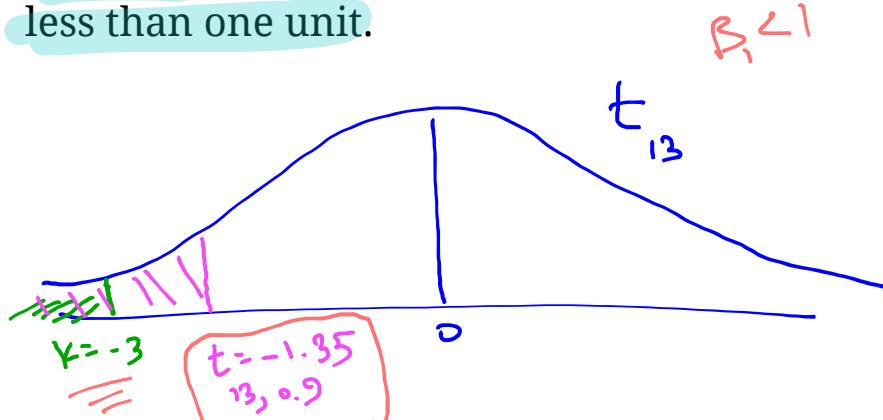
$$4- K = \frac{0.7977 - 1}{0.06744} = -3 \text{ and } t_{(13,9)} = 1.35. \text{ So,}$$

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.



Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

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

solution:

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

$$2 - \alpha = 0.05$$

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 $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$ is valid

is the slope for β_3

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example: [Stack loss, cont'd]

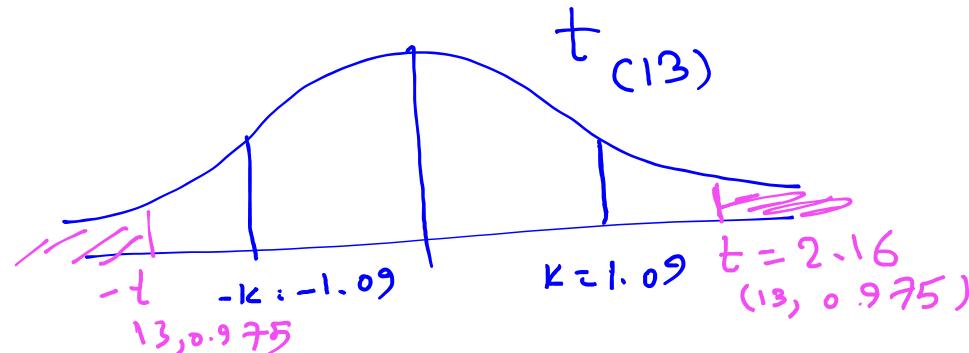
→ 4- $K = \frac{-0.06706 - 0}{0.0616} = -1.09$ and
 $t_{(13,.975)} = 2.16$. So,

$$\text{p-value} = P(|T| > |K|) =$$

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

5- Since p-value > α , we fail to reject H_0 .

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



Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

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

$$\hat{b}_3 \pm t_{(n-p, 1-\alpha/2)} \overbrace{SE(\hat{b}_3)}^{\text{JMP}} = -.06706 \pm 3.62(0.0616) \\ = (-0.2525 \ 0.1185)$$

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

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

$$b_2 \pm t_{(n-p, 1-\alpha/2)} \times \overbrace{SE(b_2)}^{\text{JMP}} = 0.5773 \pm 1.77(0.166) \\ = (0.2834 \ 0.87127)$$

We are 90% confident that for every one degree increase in temperature **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 we fit $y|x$ only one x .
Inference for mean response $\hat{y}|x \rightarrow$ multiple $x_i's$.

We can also estimate the mean response at the set of covariate values, $(x_1, x_2, \dots, x_{p-1})$. Under the model assumptions, the estimated mean response, $\hat{\mu}_{y|x}$, at $\mathbf{x} = (x_1, x_2, \dots, x_{p-1})$ is **Normally distributed** with:

$$\mathbb{E}(\hat{\mu}_{y|x}) = \mu_{y|x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

and

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

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

$\mu_{y|x}$ ← vector.

Then, under the model assumptions

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

and

$$S_{SF} \quad T = \frac{\hat{\mu}_{y|x} - \mu_{y|x}}{s_{SF} A}$$

And a test statistic for testing $H_0 : \mu_{y|x} = \#$ is

$$K = \frac{\hat{\mu}_{y|x} - \#}{s_{SF} A}$$

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Inference for mean response

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

$$\hat{\mu}_{y|x} \pm t_{(n-p, 1-\alpha/2)} \times s_{\text{CFA}}$$

Note that the one-sided CI will be analogous.

Note: $S_{\text{CFA}} = SE(\hat{\mu}_{y|x})$, and we can use JMP to get this.

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

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

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

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

The screenshot shows the JMP software interface for a "stackloss - Fit Least Squares" model. A context menu is open under the "Save Columns" option. The menu includes options like "Regression Reports", "Estimates", "Effect Screening", "Factor Profiling", "Row Diagnostics", "Save Columns", "Model Dialog", "Effect Summary", "Local Data Filter", "Redo", and "Save Script". Under "Save Columns", a sub-menu is displayed with the following items: "Prediction Formula", "Predicted Values" (which is circled in blue), "Residuals", "Mean Confidence Interval", "Indiv Confidence Interval", "Studentized Residuals", "Hats", "Std Error of Predicted" (which is also circled in blue), "Std Error of Residual", "Std Error of Individual", "Effect Leverage Pairs", "Cook's D Influence", "StdErr Pred Formula", "Mean Confidence Limit Formula", "Indiv Confidence Limit Formula", "Save Coding Table", "Publish Prediction Formula", "Publish Standard Error Formula", "Publish Mean Confid Limit Formula", and "Publish Indiv Confid Limit Formula".

How to get predicted values and standard errors

Simple Linear Regression

Variance Estimation

MSE

Inference for Parameters

Inference for mean response

MLR

Example:[Stack loss, cont'd]

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

$$\hat{\mu}_{y|x} \pm t_{(n-p,1-\alpha/2)} SE(\hat{\mu}_{y|x})$$
$$= \underline{19.2486} \pm \underline{2.16} \times \underline{(0.41785)}$$
$$= \underline{(18.343, 20.151)}$$

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.

