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

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

$$y = \beta_0 + \beta_1 x + \epsilon \qquad \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

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.

Note, in SLR $y = \beta_0 + \beta_1 x + \epsilon$, if $\beta_1 = 0$, $y = \beta_0 + \epsilon$. i.e there's no relationship between x, y. → So it's important to test if $\beta_1 = 0$.

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

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

is the line-fitting sample variance.

Variance estimation

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

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

Inference for Parameter θ in θ_1

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

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Both of these require

1. An estimate for β_1 and
2. a **standard error** for β_1

$SE(\hat{\beta}_1) = SE(b_1)$

Inference for β_1:

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

$$K = \frac{b_1 - \#}{\frac{s_{LF}}{\sqrt{\sum(x_i - \bar{x})^2}}} \sim t_{(n-2)} \text{ under } H_0$$

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

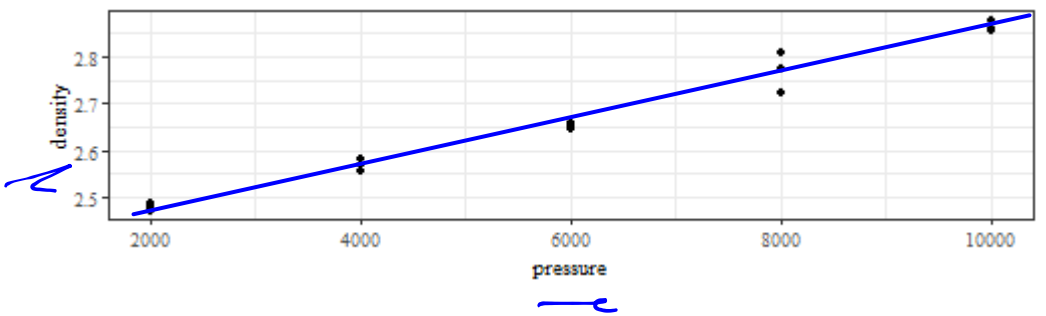
$$\begin{cases} x = & \text{the pressure setting used (psi)} \\ y = & \text{the density obtained (g/cc)} \end{cases}$$

in the dry pressing of a ceramic compound into cylinders.

Example:[Ceramic powder pressing]

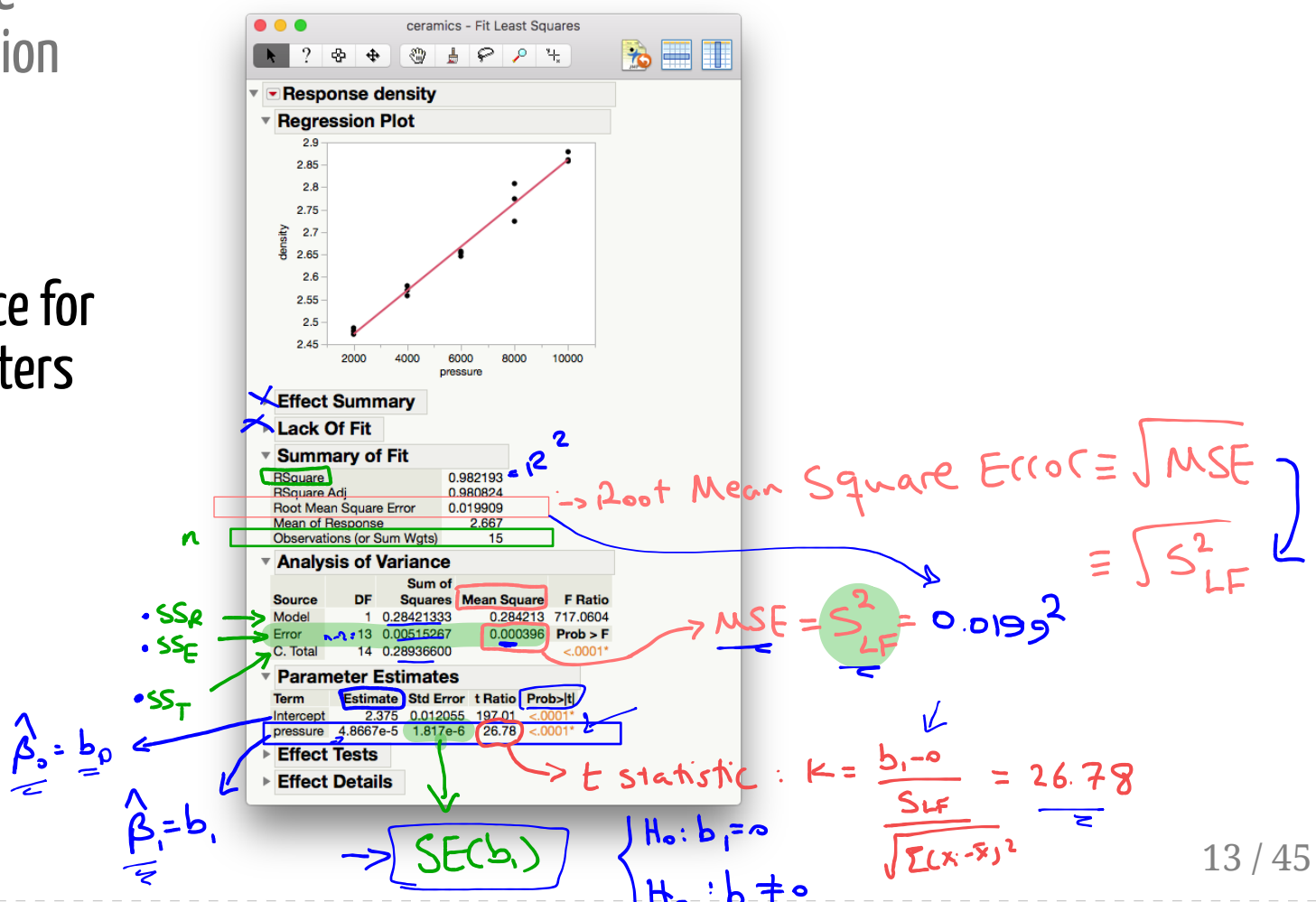
pressure	density	pressure	density
2000	2.486	6000	2.653
2000	2.479	8000	2.724
2000	2.472	8000	2.774
4000	2.558	8000	2.808
4000	2.570	10000	2.861
4000	2.580	10000	2.879
6000	2.646	10000	2.858
6000	2.657		

n=15



Example:[Ceramic powder pressing]

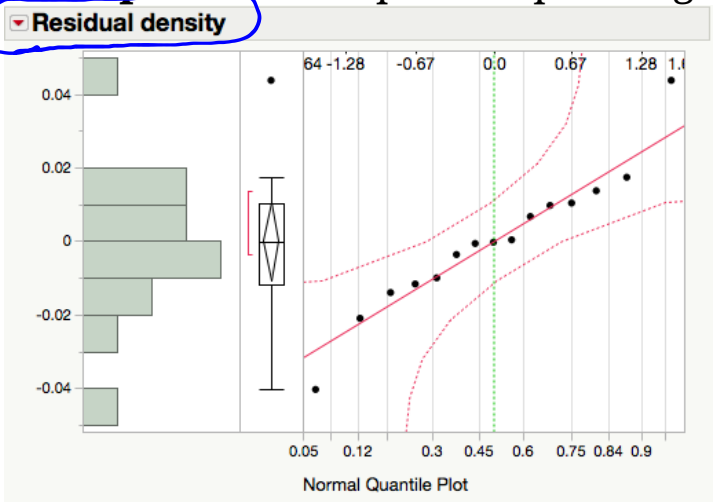
A line has been fit in JMP using the method of least squares.

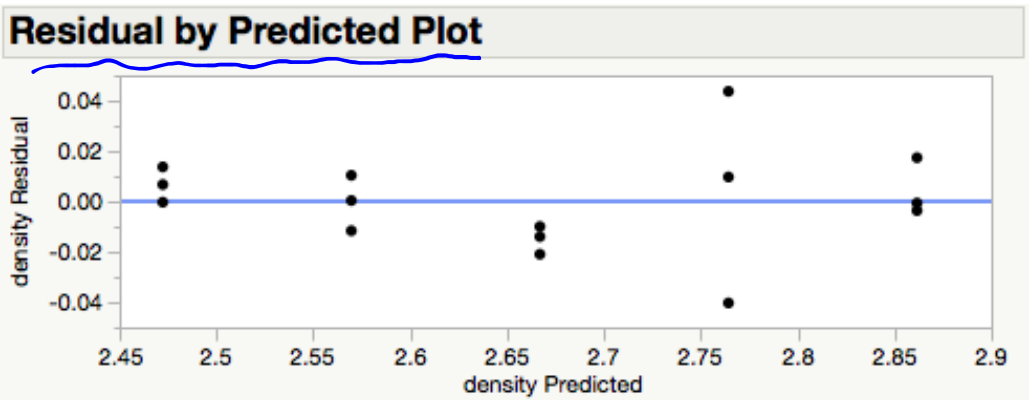


$y = \beta_0 + \beta_1 x + \epsilon$ (density ↓, pressure ↪)

p-value < 0.0001 ↓ we'll reject H_0. i.e $\beta_1 \neq 0$ and there's a significant relationship between pressure & density

Example:[Ceramic powder pressing]





Least squares regression of density on pressure of ceramic cylinders

Example:[Ceramic powder pressing]

1.Write out the model with the appropriate estimates.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = b_0 + b_1 x$

$= 2.375 + 4.8667 \times 10^{-5} x$

2.Are the assumptions for the model met?

Yes, the residual plot shows random scatter around zero & normal QQ-plot looks relatively linear indicating that residuals are normal

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

$R^2 = 98.21\%$

Example:[Ceramic powder pressing]

4.What is the correlation between x and y?

In SLR: $r = \sqrt{R^2}$

$= \sqrt{0.9821} = .9911$

5.Estimate σ^2.

$\hat{\sigma}^2 = s^2_{LF} = MSE = 0.000396$

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

$$\widehat{Var}(b_1) = \frac{s^2_{LF}}{\sum(x_i - \bar{x})^2} = (SE(b_1))^2$$

$= (1.817 \times 10^{-6})^2$

$= 3.3015 \times 10^{-12}$

JMP doesn't give us this.

Example:[Ceramic powder pressing]

7.Calculate and interpret the 95% CI for β_1

$b_1 \pm t_{(n-2,1-\alpha/2)} \frac{S_{LF}}{\sqrt{\sum(x_i-\bar{x})^2}}$: $4.8667 \times 10^{-5} \pm t_{(15-2,0.975)} (3.3015 \times 10^{-6})$

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

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

2- $\alpha = 0.05$

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

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

- H_0 is true and
- The regression model is valid

Example:[Ceramic powder pressing]

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

So,

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**