Indian Institute of Technology Bombay

SI 422: Regression Analysis

Solution of Problem Set 1 for 'Puriry hydrocarbon' Data

February 3, 2020

Given data: A paired sample of size $n = 20$ are given for the purity of oxygen produced by a fractionation process and the percentage of hydrocarbons in the main condensor of the processing unit.

Defining variables: Let the response variable y be the purity of oxygen and the predictor variable x be the percentage of hydrocarbons. Observed data are denoted by $\{(x_i, y_i) : \ i = 1, 2, \ldots, n\}$.

Simple linear regression model: Simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ \forall i = 1, 2, \ldots, n$$

where β_0, β_1 are respectively unknown intercept and slope parameters of the model and ε_i is random error corresponding to the i-th observation satisfying $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Also we assume that x is non-stochastic.

(i) Scatter plot is attached.

(ii) **Fitting simple linear regression model:** The least squares estimator of β_0 and β_1 are

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ and $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$.

For the given data, they turn out to be $\widehat{\beta}_0 = 77.863, \widehat{\beta}_1 = 11.801$. This implies that the mean value of purity of oxygen is 77.863 unit when percentage of hydrocarbon is 0. Moreover, purity of oxygen increases by 11.801 unit when percentage of hydrocarbon increases by one unit.

Therefore the fitted simple linear regression model is

$$\widehat{y} = 77.863 + 11.801x. \tag{0.1}$$

(iii) **Fitted values and residuals:**

Fitted values $\{\widehat{y}_i : \ i = 1, 2, \ldots, n\}$ and residuals $\{e_i : \ i = 1, 2, \ldots, n\}$ are given by

$$\widehat{y}_i = 77.863 + 11.801x_i \quad \text{and} \quad e_i = y_i - \widehat{y}_i \quad \forall i = 1, 2, \ldots, n.$$

We have the following fitted values and residuals for the given data set.

Subjects	Fitted values	Residuals
1	89.9	-2.99
2	90.96	-1.11
3	94.74	-4.46
4	90.96	-4.62
5	89.78	2.798
6	90.96	-1.74
7	88.13	-4.67
8	88.13	3.73
9	94.74	0.87
10	89.9	-0.04
11	95.1	1.637
12	96.15	3.265
13	96.15	2.505
14	96.15	-0.085
15	94.38	-0.735
16	91.43	-4.124
17	89.78	5.218
18	89.55	7.304
19	89.07	-3.874
20	89.43	1.132

(iv) **Coefficient of determination:** Coefficient of determination, denoted by R^2, is defined by

$$R^2 = \frac{\text{SS}_R}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}.$$

where SS_R is the sum of squares due to regression fit, SSE is the residual sum of squares and TSS is the total sum of squares. For our data set, $R^2 = 0.3891$. This implies that 38.91% of the total sum of squares is explained by the fitted linear regression model (0.1).

(v) **Testing for intercept parameter:** Here we wish to test

$$H_0: \quad \beta_0 = 0 \quad \text{against} \quad H_1: \quad \beta_0 \neq 0. \tag{0.2}$$

To test (0.2), we use the test statistic

$$T_0 = \frac{\widehat{\beta}_0}{\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \overset{H_0}{\sim} t_{n-2}$$

and reject H_0 in (0.2) at 5% level of significance if $\mathbb{P}(|t_{n-2}| > |(T_0)_{\text{observed}}|) < 0.05$.

For our data set, $(T_0)_{\text{observed}} = 18.544$ and $\mathbb{P}(|t_{18}| > 18.544) = 3.54 \times 10^{-13} < 0.05$.

Therefore, in the light of the given data, we reject H_0 in (0.2) at 5% level of significance.

(vi) **Testing for slope parameter:** Here we wish to test

$$H_0: \quad \beta_1 = 0 \quad \text{against} \quad H_1: \quad \beta_1 \neq 0. \tag{0.3}$$

To test (0.3), we use the test statistic

$$T_1 = \frac{\widehat{\beta}_1 \sqrt{S_{xx}}}{\sqrt{\text{MSE}}} \overset{H_0}{\sim} t_{n-2}$$

and reject H_0 in (0.3) at 5% level of significance if $\mathbb{P}(|t_{n-2}| > |(T_1)_{\text{observed}}|) < 0.05$.

For our data set, $(T_1)_{\text{observed}} = 3.386$ and $\mathbb{P}(|t_{18}| > 3.386) = 0.00329 < 0.05$.

Therefore, in the light of the given data, we reject H_0 in (0.3) at 5% level of significance.

(vii) **ANOVA table:**

Source	SS	DF	MS	F stat	p-value
Regression	148.31	1	148.313	11.466	0.0033
Error	232.83	18	12.935		
Total	381.14	19			

Note that the above p-value is less than 0.05. Hence, in the light of the given data and at 5% level of significance, the regression model (0.1) is significant. This implies that even though there is significant linear effect of percentage of hydrocarbon on the purity of oxygen, better result could be obtained with the addition of higher order polynomial terms in the model.

(viii) **Confidence interval for the intercept parameter:**

A 95% confidence interval of β_0 is

$$\left(\widehat{\beta}_0 - (t_{n-2})_{0.025}\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \widehat{\beta}_0 + (t_{n-2})_{0.025}\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right).$$

For the given data, it turns out to be $(69.042, 86.68)$.

(ix) **Confidence interval for the slope parameter:**

A 95% confidence interval of β_1 is

$$\left(\widehat{\beta}_1 - (t_{n-2})_{0.025}\sqrt{\frac{MSE}{S_{xx}}}, \widehat{\beta}_1 + (t_{n-2})_{0.025}\sqrt{\frac{MSE}{S_{xx}}}\right).$$

For the given data, it turns out to be $(4.48, 19.123)$.

(x) **Confidence interval for the mean response:**

A 95% confidence interval of the mean purity of oxygen for 1% of hydrocarbon is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 - (t_{18})_{0.025}\sqrt{MSE\left(\frac{1}{n} + \frac{(1-\bar{x})^2}{S_{xx}}\right)}, \widehat{\beta}_0 + \widehat{\beta}_1 + (t_{18})_{0.025}\sqrt{MSE\left(\frac{1}{n} + \frac{(1-\bar{x})^2}{S_{xx}}\right)}\right).$$

For the given data, it turns out to be $(87.51, 91.82)$.

(xi) **Prediction interval for a new observation:**

A 95% prediction interval of purity of oxygen for 1% of hydrocarbon is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 - (t_{18})_{0.025}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(1-\bar{x})^2}{S_{xx}}\right)}, \widehat{\beta}_0 + \widehat{\beta}_1 + (t_{18})_{0.025}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(1-\bar{x})^2}{S_{xx}}\right)}\right).$$

For the given data, it turns out to be $(81.81, 97.52)$.

(xii) **Confidence band and Prediction band:**

Consider an equidistant partition of $(x_{\min}, x_{\max})$ of length k (say). Suppose it is $x_{new} = ((x_{new})_1, (x_{new})_2, \ldots, (x_{new})_k)$. Then a 95% confidence band is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 (x_{new})_j - (t_{18})_{0.025} \sqrt{MSE\left(\frac{1}{n} + \frac{((x_{new})_j - \bar{x})^2}{S_{xx}}\right)},\right.$$

$$\left.\widehat{\beta}_0 + \widehat{\beta}_1 (x_{new})_j + (t_{18})_{0.025} \sqrt{MSE\left(\frac{1}{n} + \frac{((x_{new})_j) - \bar{x})^2}{S_{xx}}\right)}\ \right), \quad j = 1, 2, \ldots, k.$$

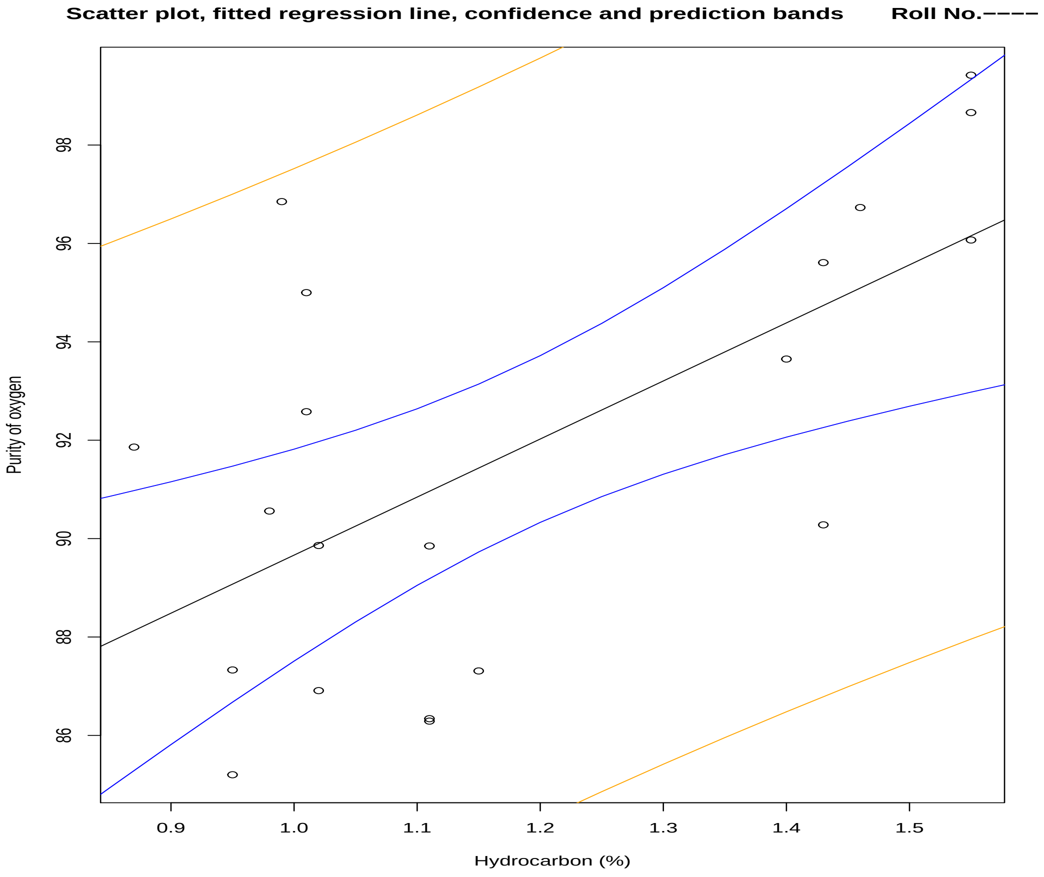
A 95% prediction band is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 (x_{new})_j - (t_{18})_{0.025} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{((x_{new})_j - \bar{x})^2}{S_{xx}}\right)},\right.$$

$$\left.\widehat{\beta}_0 + \widehat{\beta}_1 (x_{new})_j + (t_{18})_{0.025} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{((x_{new})_j) - \bar{x})^2}{S_{xx}}\right)}\ \right), \quad j = 1, 2, \ldots, k.$$

Plot of 95% confidence and prediction bands is attached.

(xiii) Prediction interval is wider than confidence interval as the former one encounters random error of the model.


Scatter plot, fitted regression line, confidence and prediction bands
Roll No.----
Purity of oxygen
Hydrocarbon (%)