PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE19CS203 – STATISTICS FOR DATA SCIENCE

Unit-5 - Power of Test and Simple Linear Regression

QUESTION BANK

Predictions using regression models - Uncertainties in Regression Coefficients.

Exercises for section 7.3: [Text Book Exercise 7.3– Pg. No. [554 – 559]]

1. A chemical reaction is run 12 times, and the temperature x_i (in ◦C) and the yield y_i (in percent of a theoretical maximum) is recorded each time. The following summary statistics are recorded:

$$\bar{x} = 65.0 \qquad \bar{y} = 29.05 \qquad \sum_{i=1}^{12}(x_i - \overline{x})^2 = 6032.0$$

$$\sum_{i=1}^{12}(y_i - \bar{y})^2 = 835.42 \qquad \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 1988.4$$

Let β_0 represent the hypothetical yield at a temperature of 0◦C, and let $\beta_1$represent the increase in yield caused by an increase in temperature of 1◦C. Assume that assumptions for errors in linear model holds.

   a. Compute the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
   b. Compute the error variance estimate s^2.
   c. Find 95% confidence intervals for β_0 and β_1
   d. A chemical engineer claims that the yield increases by more than 0.5 for each 1◦C increase in temperature. Do the data provide sufficient evidence for you to conclude that this claim is false?
   e. Find a 95% confidence interval for the mean yield at a temperature of 40◦C.

f. Find a 95% prediction interval for the yield of a particular reaction at a temperature of 40◦C.

2. The following output (from MINITAB) describes the fit of a linear model $y = \beta_o + \beta_1 x + \varepsilon$ that expresses the length of a spring in cm (y) in terms of the load upon it in kg (x). There are $n = 15$ observations.

Predictor	Coef	StDev	T	P
Constant	6.6361	1.1455	5.79	0.000
Load	2.9349	0.086738	33.8	0.000

a. How many degrees of freedom are there for the Student's t statistics?
b. Find a 98% confidence interval for β_1
c. Find a 98% confidence interval for β_o
d. Someone claims that if the load is increased by 1 kg, that the length will increase by exactly 0.35 cm. Use the given output to perform a hypothesis test to determine whether this claim is plausible.
e. Someone claims that the unloaded (load = 0) length of the spring is more than 1.5 cm. Use the given output to perform a hypothesis test to determine whether this claim is plausible.

3. Ozone (O_3) is a major component of air pollution in many cities. Atmospheric ozone levels are influenced by many factors, including weather. In one study, the mean percent relative humidity (x) and the mean ozone levels (y) were measured for 120 days in a western city. Mean ozone levels were measured in ppb. The following output (from MINITAB) describes the fit of a linear model to these data. Assume that assumptions for errors in Lineal models hold.

```
The regression equation is
Ozone = 88.8 – 0.752 Humidity

Predictor          Coef       SE Coef            T           P
Constant         88.761         7.288        12.18       0.000
Humidity        −0.7524       0.13024        −5.78       0.000

S = 11.43          R-Sq = 22.0%       R-Sq(adj) = 21.4%

Predicted Values for New Observations

New Obs      Fit    SE Fit       95.0% CI           95.0% PI
1          43.62      1.20    ( 41.23 46.00)    ( 20.86, 66.37)

Values of Predictors for New Observations

New Obs    Humidity
1              60.0
```

a. What are the slope and intercept of the least-squares line?
b. Is the linear model useful for predicting ozone levels from relative humidity? Explain.
c. Predict the ozone level for a day when the relative humidity is 50%.
d. What is the correlation between relative humidity and ozone level?
e. The output provides a 95% confidence interval for the mean ozone level for days where the relative humidity is 60%. There are n = 120 observations in this data set. Using the value "SE Fit," find a 90% confidence interval.
f. Upon learning that the relative humidity on a certain day is 60%, someone predicts that the ozone level that day will be 80 ppb. Is this a reasonable prediction? If so, explain why. If not, give a reasonable range of predicted value

4. In an study similar to the one in Exercise 3, the relative humidity and ozone levels were measured for 120 days in another city. The MINITAB output follows. Assume that assumptions for errors in Linear Models .

```
The regression equation is
Ozone = 29.7 – 0.135 Humidity

Predictor          Coef     SE Coef          T          P
Constant         29.703       2.066      14.38      0.000
Humidity      −0.13468     0.03798      −3.55      0.001

S = 6.26          R−Sq = 9.6%       R−Sq(adj) = 8.9%
```

a. What is the slope of the least-squares line?
b. Find a 95% confidence interval for the slope.
c. Perform a test of the null hypothesis that the slope is greater than or equal to −0.1. What is the P-value?

5. Refer to Exercises 3 and 4. An atmospheric scientist notices that the slope of the least-squares line in the study described in Exercise 4 differs from the one in the study described in Exercise 3. He wishes to test the hypothesis that the effect of humidity on ozone level differs between the two cities. Let β_A denote the change in ozone level associated with an increase of 1 percent relative humidity for the city in Exercise 3, and β_B denote the corresponding increase for the city in Exercise 4.
    a. Express the null hypothesis to be tested in terms of β_A and β_B.
    b. Let β_A and β_B denote the slopes of the least-squares lines. Assume these slopes are independent. There are 120 observations in each data set. Test the null hypothesis in part (a). Can you conclude that the effect of humidity differs between the two cities?

6. The article "Withdrawal Strength of Threaded Nails" (D. Rammer, S.Winistorfer, and D. Bender, Journal of Structural Engineering, 2001:442–449) describes an experiment to investigate the relationship between the diameter of a nail (x) and its ultimate withdrawal strength (y). Annularly threaded nails were driven into Douglas fir lumber, and then their withdrawal strengths were measured in N/mm. The following results for 10 different diameters (in mm) were obtained.

x	2.52	2.87	3.05	3.43	3.68	3.76	3.76	4.50	4.50	5.26
y	54.74	59.01	72.92	50.85	54.99	60.56	69.08	77.03	69.97	90.70

    a. Compute the least-squares line for predicting strength from diameter.
    b. Compute the error standard deviation estimate s.
    c. Compute a 95% confidence interval for the slope.
    d. Find a 95% confidence interval for the mean withdrawal strength of nails 4 mm in diameter.

e. Can you conclude that the mean withdrawal strength of nails 4 mm in diameter is greater than 60 N/mm? Perform a hypothesis test and report the P-value.
f. Find a 95% prediction interval for the withdrawal strength of a particular nail whose diameter is 4 mm.

7. The coefficient of absorption (COA) for a clay brick is the ratio of the amount of cold water to the amount of boiling water that the brick will absorb. The article "Effects of Waste Glass Additions on the Properties and Durability of Fired Clay Brick" (S. Chidiac and L. Federico, Can J Civ Eng, 2007:1458–1466) presents measurements of the (COA) and the pore volume (in cm3 /g) for seven bricks. The results are presented in the following table.

Pore volume	COA
1.750	0.80
1.632	0.78
1.594	0.77
1.623	0.75
1.495	0.71
1.465	0.66
1.272	0.63

a. Compute the least-squares line for predicting COA from pore volume.
b. Compute the error standard deviation estimate s.
c. Compute a 95% confidence interval for the slope.
d. Find a 95% confidence interval for the mean COA for bricks with pore volume 1.5 cm3 /g.
e. Can you conclude that the mean COA for bricks with pore volume 1.5 cm3 /g is less than 0.75? Perform a hypothesis test and report the P-value.
f. Find a 95% prediction interval for the COA of a particular brick whose pore volume is 1.5 cm3 /g.

8. The article "Application of Radial Basis Function Neural Networks in Optimization of Hard Turning of AISI D2 ColdWorked Tool Steel With a Ceramic Tool" (S. Basak, U. Dixit, and J. Davim, Journal of Engineering Manufacture, 2007:987–998) presents the results of an experiment in which the surface roughness (in μm) was measured for 27 D2 steel specimens and compared with the roughness predicted by a neural network model. The results are presented in the following table.

True Value (x)	Predicted Value (y)	True Value (x)	Predicted Value (y)	True Value (x)	Predicted Value (y)
0.45	0.42	0.52	0.51	0.57	0.55
0.82	0.70	1.02	0.91	1.14	1.01
0.54	0.52	0.60	0.71	0.74	0.81
0.41	0.39	0.58	0.50	0.62	0.66
0.77	0.74	0.87	0.91	1.15	1.06
0.79	0.78	1.06	1.04	1.27	1.31
0.25	0.27	0.45	0.52	1.31	1.40
0.62	0.60	1.09	0.97	1.33	1.41
0.91	0.87	1.35	1.29	1.46	1.46

To check the accuracy of the prediction method, the linear model $y = \beta 0 + \beta 1\, x + \varepsilon$ is fit. If the prediction method is accurate, the value of β_0 will be 0 and the value of β_1 will be 1.

a. Compute the least-squares estimates β_0 and β_1 .
b. Can you reject the null hypothesis $H_0 : \beta_0 = 0$?
c. Can you reject the null hypothesis $H_O : \beta_1 = 1$?
d. Do the data provide sufficient evidence to conclude that the prediction method is not accurate?
e. Compute a 95% confidence interval for the mean prediction when the true roughness is 0.8μm. f. Someone claims that when the true roughness is 0.8 μm, the mean prediction is only 0.75μm. Do these data provide sufficient evidence for you to conclude that this claim is false? Explain.

9. In a study to determine the relationship between ambient outdoor temperature and the rate of evaporation of water from soil, measurements of average daytime temperature in $\circ\, C$ and evaporation in mm/day were taken for 40 days. The results are shown in the following table.

Temp.	Evap.	Temp.	Evap.	Temp.	Evap.	Temp.	Evap.
11.8	2.4	11.8	3.8	18.6	3.5	14.0	1.1
21.5	4.4	24.2	5.0	25.4	5.5	13.6	3.5
16.5	5.0	15.8	2.6	22.1	4.8	25.4	5.1
23.6	4.1	26.8	8.0	25.4	4.8	17.7	2.0
19.1	6.0	24.8	5.4	22.6	3.2	24.7	5.7
21.6	5.9	26.2	4.2	24.4	5.1	24.3	4.7
31.0	4.8	14.2	4.4	15.8	3.3	25.8	5.8
18.9	3.0	14.1	2.2	22.3	4.9	28.3	5.8
24.2	7.1	30.3	5.7	23.2	7.4	29.8	7.8
19.1	1.6	15.2	1.2	19.7	3.3	26.5	5.1

a. Compute the least-squares line for predicting evaporation (y) from temperature (x).
b. Compute 95% confidence intervals for: $\beta_0$and β_1 .
c. Predict the evaporation rate when the temperature is $20°C$.
d. Find a 95% confidence interval for the mean evaporation rate for all days with a temperature of $20°C$.
e. Find a 95% prediction interval for the evaporation rate on a given day with a temperature of $20°C$.

10. Three engineers are independently estimating the spring constant of a spring, using the linear model specified by Hooke's law. Engineer A measures the length of the spring under loads of 0, 1, 3, 4, and 6 lb, for a total of five measurements. Engineer B uses the same loads, but repeats the experiment twice, for a total of 10 independent measurements. Engineer C uses loads of 0, 2, 6, 8, and 12 lb, measuring once for each load. The engineers all use the same measurement apparatus and procedure. Each engineer computes a 95% confidence interval for the spring constant.

a. If the width of the interval of engineer A is divided by the width of the interval of engineer B, the quotient will be approximately --------
b. If the width of the interval of engineer A is divided by the width of the interval of engineer C, the quotient will be approximately -----
c. Each engineer computes a 95% confidence interval for the length of the spring under a load of 2.5 lb. Which interval is most likely to be the shortest? Which interval is most likely to be the longest?

11. In the weld data(given below), imagine that 95% confidence intervals are computed for the mean strength of welds with oxygen contents of 1.3, 1.5, and 1.8 parts per thousand. Which of the confidence intervals would be the shortest? Which would be the longest?

Oxygen Content	Strength	Oxygen Content	Strength	Oxygen Content	Strength
1.08	63.00	1.16	68.00	1.17	73.00
1.19	76.00	1.32	79.67	1.40	81.00
1.57	66.33	1.61	71.00	1.69	75.00
1.72	79.67	1.70	81.00	1.71	75.33
1.80	72.50	1.69	68.65	1.63	73.70
1.65	78.40	1.78	84.40	1.70	91.20
1.50	72.00	1.50	75.05	1.60	79.55
1.60	83.20	1.70	84.45	1.60	73.95
1.20	71.85	1.30	70.25	1.30	66.05
1.80	87.15	1.40	68.05		

12. A chemical reaction is run 12 times, and the temperature x_i (in ◦C) and the yield y_i (in percent of a theoretical maximum) is recorded each time. The following summary statistics are recorded:

$$\bar{x} = 65.0 \qquad \bar{y} = 29.05 \qquad \sum_{i=1}^{12}(x_i - \overline{x})^2 = 6032.0$$

$$\sum_{i=1}^{12}(y_i - \bar{y})^2 = 835.42 \qquad \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 1988.4$$

If 95% confidence intervals are constructed for the yield of the reaction at temperatures of 45◦C, 60◦C, and 75◦C, which confidence interval would be the shortest? Which would be the longest?

13. In a study of copper bars, the relationship between shear stress in ksi (x) and shear strain in $\%\,(y)$ was summarized by the least-squares line $y = -20.00 + 2.56x$. There were a total of $n = 17$ observations, and the coefficient of determination was r^2= 0.9111. If the total sum of squares was$\sum(y_i - \bar{y})2 = 234.19$, compute the estimated error variance s^2 .

14. In the manufacture of synthetic fiber, the fiber is often "set" by subjecting it to high temperatures. The object is to improve the shrinkage properties of the fiber. In a test of 25 yarn specimens, the relationship between temperature in ◦C (x) and shrinkage in % (y) was summarized by the least-squares line $y = -12.789 + 0.133x$. The total sum of squares was $\sum(y_i - \bar{y})2 = 57.313$, and the estimated error variance was s^2 = 0.0670. Compute the coefficient of determination r^2.

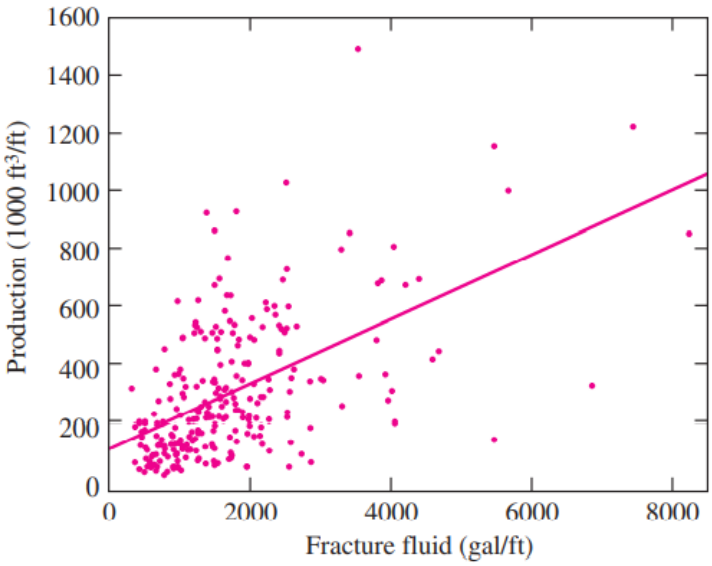
15. In the following MINITAB output, some of the numbers have been accidentally erased. Recompute them, using the numbers still available. There are $n = 25$ points in the data set.

```
The regression equation is
Y = 1.71 + 4.27 X

Predictor          Coef     SE Coef         T       P
Constant        1.71348     6.69327       (a)     (b)
X               4.27473         (c)     3.768     (d)

S = 0.05749     R-Sq = 38.2%
```

16. In the following MINITAB output, some of the numbers have been accidentally erased. Recompute them, using the numbers still available. There are $n = 20$ points in the data set.

```
Predictor          Coef      SE Coef         T       P
Constant            (a)      0.43309     0.688     (b)
X               0.18917     0.065729       (c)     (d)

S = 0.67580     R-Sq = 31.0%
```

17. In order to increase the production of gas wells, a procedure known as "fracture treatment" is often used. Fracture fluid, which consists of fluid mixed with sand, is pumped into the well. The following figure presents a scatterplot of the monthly production versus the volume of fracture fluid pumped for 255 gas wells. Both production and fluid are expressed in units of volume per foot of depth of the well. The least-squares line is superimposed. The equation of the least-squares line is $y = 106.11 + 0.1119x$.



a. From the least-squares line, estimate the production for a well into which 4000 gal/ft are pumped.
b. From the least-squares line, estimate the production for a well into which 500 gal/ft are pumped.
c. A new well is dug, and 500 gal/ft of fracture fluid are pumped in. Based on the scatterplot, is it more likely that the production of this well will fall above or below the least-squares estimate?
d. What feature of the scatterplot indicates that assumption that the errors have the same variance is violated?