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

Checking Assumptions and transforming data:

Exercises for section 7.4: [Text Book Exercise 7.4– Pg. No. [576 – 584]]

1. The following output (from MINITAB) is for the least-squares fit of the model $ln\, y = \beta_0 + \beta_1\, ln\, x + \varepsilon$, where y represents the monthly production of a gas well and x represents the volume of fracture fluid pumped in. (A scatterplot of these data is presented in Figure 7.22.)

```
Regression Analysis: LN PROD versus LN FLUID

The regression equation is
LN PROD = - 0.444 + 0.798 LN FLUID

Predictor          Coef      SE Coef          T          P
Constant        -0.4442       0.5853      -0.76      0.449
LN FLUID        0.79833      0.08010       9.97      0.000

S = 0.7459     R-Sq = 28.2%         R-Sq(adj) = 27.9%

Analysis of Variance

Source             DF         SS         MS         F         P
Regression          1     55.268     55.268     99.34     0.000
Residual Error    253    140.756      0.556
Total             254    196.024

Predicted Values for New Observations

New Obs        Fit   SE Fit        95.0% CI             95.0% PI
1           5.4457   0.0473   ( 5.3526, 5.5389)   ( 3.9738, 6.9176)

Values of Predictors for New Observations

New Obs   LN FLUID
1          7.3778
```

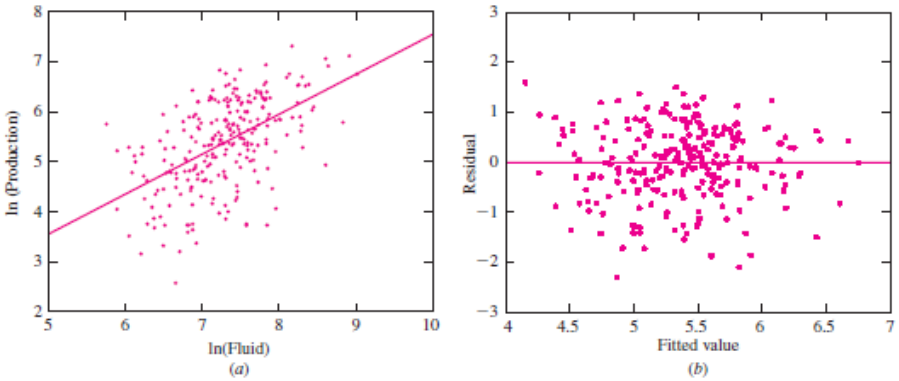


FIGURE 7.22 *(a)* Plot of the log of production versus the log of the volume of fracture fluid for 255 gas wells, with the least-squares line superimposed. *(b)* Plot of residuals versus fitted values. There is no substantial pattern to the residuals. The linear model looks good.

a. What is the equation of the least-squares line for predicting $ln\ y$ from $ln\ x$?
b. Predict the production of a well into which 2500 gal/ft of fluid have been pumped.
c. Predict the production of a well into which 1600 gal/ft of fluid have been pumped.
d. Find a 95% prediction interval for the production of a well into which 1600 gal/ft of fluid have been pumped. (Note: ln 1600 = 7.3778.)

2. The processing of raw coal involves "washing," in which coal ash (nonorganic, incombustible material) is removed. The article "Quantifying Sampling Precision for Coal Ash Using Gy's Discrete Model of the Fundamental Error" (Journal of Coal Quality, 1989:33–39) provides data relating the percentage of ash to the volume of a coal particle. The average percentage of ash for six volumes of coal particles was measured. The data are as follows:

Volume (cm^3)	0.01	0.06	0.58	2.24	15.55	276.02
Percent ash	3.32	4.05	5.69	7.06	8.17	9.36

a. Compute the least-squares line for predicting percent ash (y) from volume (x). Plot the residuals versus the fitted values. Does the linear model seem appropriate? Explain.

b. Compute the least-squares line for predicting percent ash from $ln\ volume$. Plot the residuals versus the fitted values. Does the linear model seem appropriate? Explain.
c. Compute the least-squares line for predicting percent ash from $\sqrt{volume}$. Plot the residuals versus the fitted values. Does the linear model seem appropriate? Explain.
d. Using the most appropriate model, predict the percent ash for particles with a volume of 50 m^3.
e. Using the most appropriate model, construct a 95% confidence interval for the mean percent ash for particles with a volume of 50 m^3.

3. To determine the effect of temperature on the yield of a certain chemical process, the process is run 24 times at various temperatures. The temperature (in ◦C) and the yield (expressed as a percentage of a theoretical maximum) for each run are given in the following table. The results are presented in the order in which they were run, from earliest to latest.

Order	Temp	Yield	Order	Temp	Yield	Order	Temp	Yield
1	30	49.2	9	25	59.3	17	34	65.9
2	32	55.3	10	38	64.5	18	43	75.2
3	35	53.4	11	39	68.2	19	34	69.5
4	39	59.9	12	30	53.0	20	41	80.8
5	31	51.4	13	30	58.3	21	36	78.6
6	27	52.1	14	39	64.3	22	37	77.2
7	33	60.2	15	40	71.6	23	42	80.3
8	34	60.5	16	44	73.0	24	28	69.5

a. Compute the least-squares line for predicting yield (y) from temperature (x).
b. Plot the residuals versus the fitted values. Does the linear model seem appropriate? Explain.
c. Plot the residuals versus the order in which the observations were made. Is there a trend in the residuals over time? Does the linear model seem appropriate? Explain.

4. In rock blasting, the peak particle velocity (PPV) depends both on the distance from the blast and on the amount of charge. The article "Prediction of Particle Velocity Caused by Blasting for an Infrastructure Excavation Covering Granite Bedrock" (A. Kahriman, Mineral Resources Engineering, 2001:205–218) suggests predicting PPV (y) from the scaled distance (x), which is equal to the distance divided by the square root of the charge. The results for 15 blasts are presented in the following table.

PPV (mm/s)	Scaled Distance ($m/kg^{0.5}$)
1.4	47.33
15.7	9.6
2.54	15.8
1.14	24.3
0.889	23.0
1.65	12.7
1.4	39.3
26.8	8.0
1.02	29.94
4.57	10.9
6.6	8.63
1.02	28.64
3.94	18.21
1.4	33.0
1.4	34.0

   a. Plot PPV versus scaled distance. Does the relationship appear to be linear?
   b. Compute the least-squares line for the model $ln\ PPV = \beta_0 + \beta_1$ ln scaled distance $+ \varepsilon$. Plot the residuals versus fitted values. Does this linear model seem appropriate?
   c. Use the least-squares line computed in part (b) to predict the PPV when the scaled distance is 20. Find a 95% prediction interval.

5. Good forecasting and control of preconstruction activities leads to more efficient use of time and resources in highway construction projects. Data on construction costs (in $1000s) and person-hours of labor required on several projects are presented in the following table and are taken from the article "Forecasting Engineering Manpower Requirements for Highway Preconstruction Activities" (K. Persad, J. O'Connor, and K. Varghese, Journal of Management Engineering, 1995:41–47). Each value represents an average of several projects, and two outliers have been deleted

Person-Hours (x)	Cost (y)	Person-Hours (x)	Cost (y)
939	251	1069	355
5796	4690	6945	5253
289	124	4159	1177
283	294	1266	802
138	138	1481	945
2698	1385	4716	2327
663	345		

a. Compute the least-squares line for predicting y from x.
b. Plot the residuals versus the fitted values. Does the model seem appropriate?
c. Compute the least-squares line for predicting $ln\ y$ from $ln\ x$.
d. Plot the residuals versus the fitted values. Does the model seem appropriate?
e. Using the more appropriate model, construct a 95% prediction interval for the cost of a project that requires 1000 person-hours of labor.

6. The article "Drift in Posturography Systems Equipped With a Piezoelectric Force Platform: Analysis and Numerical Compensation" (L. Quagliarella, N. Sasanelli, and V. Monaco, IEEE Transactions on Instrumentation and Navidi-1820036 book November 16, 2009 8:31 7.4 Checking Assumptions and Transforming Data 579 Measurement, 2008:997–1004), reported the results of an experiment to determine the effect of load on the drift in signals derived from a piezoelectric force plates. The correlation coefficient y between output and time was computed for various loads x in kN, as shown in the following table.

x	y
0.196	−0.9710
0.245	−0.9735
0.294	−0.9694
0.343	−0.9684
0.392	−0.9624
0.441	−0.9688
0.490	−0.9519
0.539	−0.9573
0.588	−0.9515

a. Compute the least-squares line for predicting y from x.
b. Plot the residuals versus the fitted values. Does the least-squares line seem appropriate?
c. Compute the least-squares line for predicting y from x^2
d. Plot the residuals versus the fitted values. Does the least-squares line seem appropriate?
e. For each model, find a 95% confidence interval for the mean value of y when $x = 0.32$. Are the confidence intervals similar?

7. The National Assessment for Educational Progress measured the percentage of eighth grade students who were proficient in reading and the percentage of students who graduated from high school in each state in the U.S. The results for the ten most populous states are as follows:

State	Reading Proficiency	Graduation Rate
California	60	75
Texas	73	74
New York	75	65
Florida	66	65
Illinois	75	79
Pennsylvania	79	83
Ohio	79	80
Michigan	73	73
Georgia	67	62
North Carolina	71	73

Reading data from 2005, graduation data from 2007

a. Construct a scatterplot of graduation rate (y) versus reading proficiency (x). Which state is an outlier?
b. Compute the least-squares line for predicting graduation rate from reading proficiency, using the data from all ten states.
c. Remove the outlier and compute the least-squares line, using the data from the other nine states.
d. Is the outlier an influential point? Explain.

8. The article "Oxidation State and Activities of Chromium Oxides in $CaO - SiO_2 - CrO_x$ Slag System" (Y. Xiao, L. Holappa, and M. Reuter, Metallurgical and Materials Transactions B, 2002:595–603) presents the amount x (in mole percent) and activity coefficient y of CrO1.5 for several specimens. The data, extracted from a larger table, are presented in the following table.

x	y	x	y	x	y
10.20	2.6	7.13	5.8	5.33	13.1
5.03	19.9	3.40	29.4	16.70	0.6
8.84	0.8	5.57	2.2	9.75	2.2
6.62	5.3	7.23	5.5	2.74	16.9
2.89	20.3	2.12	33.1	2.58	35.5
2.31	39.4	1.67	44.2	1.50	48.0

   a. Compute the least-squares line for predicting y from x.
   b. Plot the residuals versus the fitted values.
   c. Compute the least-squares line for predicting y from $1/x$.
   d. Plot the residuals versus the fitted values.
   e. Using the better fitting line, find a 95% confidence interval for the mean value of y when $x = 5.0$.

9. A windmill is used to generate direct current. Data are collected on 45 different days to determine the relationship between wind speed in mi/h (x) and current in kA (y). The data are presented in the following table.

Day	Wind Speed	Current	Day	Wind Speed	Current	Day	Wind Speed	Current
1	4.2	1.9	16	3.7	2.1	31	2.6	1.4
2	1.4	0.7	17	5.9	2.2	32	7.7	2.8
3	6.6	2.2	18	6.0	2.6	33	6.1	2.4
4	4.7	2.0	19	10.7	3.2	34	5.5	2.2
5	2.6	1.1	20	5.3	2.3	35	4.7	2.3
6	5.8	2.6	21	5.1	1.9	36	4.0	2.0
7	1.8	0.3	22	4.9	2.3	37	2.3	1.2
8	5.8	2.3	23	8.3	3.1	38	11.9	3.0
9	7.3	2.6	24	7.1	2.3	39	8.6	2.5
10	7.1	2.7	25	9.2	2.9	40	5.6	2.1
11	6.4	2.4	26	4.4	1.8	41	4.2	1.7
12	4.6	2.2	27	8.0	2.6	42	6.2	2.3
13	1.6	1.1	28	10.5	3.0	43	7.7	2.6
14	2.3	1.5	29	5.1	2.1	44	6.6	2.9
15	4.2	1.5	30	5.8	2.5	45	6.9	2.6

a. Compute the least-squares line for predicting y from x. Make a plot of residuals versus fitted values.
b. Compute the least-squares line for predicting y from $ln\ x$. Make a plot of residuals versus fitted values.
c. Compute the least-squares line for predicting $ln\ y$ from x. Make a plot of residuals versus fitted values.
d. Compute the least-squares line for predicting $\sqrt{y}$ from x. Make a plot of residuals versus fitted values.
e. Which of the four models (a) through (d) fits best? Explain.
f. For the model that fits best, plot the residuals versus the order in which the observations were made. Do the residuals seem to vary with time?
g. Using the best model, predict the current when wind speed is 5.0 mi/h. Navidi-1820036 book November 16, 2009 8:31 7.4 Checking Assumptions and Transforming Data 581
h. Using the best model, find a 95% prediction interval for the current on a given day when the wind speed is 5.0 mi/h.

10. Two radon detectors were placed in different locations in the basement of a home. Each provided an hourly measurement of the radon concentration, in units of PC_i/L. The data are presented in the following table.

R_1	R_2	R_1	R_2	R_1	R_2	R_1	R_2
1.2	1.2	3.4	2.0	4.0	2.6	5.5	3.6
1.3	1.5	3.5	2.0	4.0	2.7	5.8	3.6
1.3	1.6	3.6	2.1	4.3	2.7	5.9	3.9
1.3	1.7	3.6	2.1	4.3	2.8	6.0	4.0
1.5	1.7	3.7	2.1	4.4	2.9	6.0	4.2
1.5	1.7	3.8	2.2	4.4	3.0	6.1	4.4
1.6	1.8	3.8	2.2	4.7	3.1	6.2	4.4
2.0	1.8	3.8	2.3	4.7	3.2	6.5	4.4
2.0	1.9	3.9	2.3	4.8	3.2	6.6	4.4
2.4	1.9	3.9	2.4	4.8	3.5	6.9	4.7
2.9	1.9	3.9	2.4	4.9	3.5	7.0	4.8
3.0	2.0	3.9	2.4	5.4	3.5		

a. Compute the least-squares line for predicting the radon concentration at location 2 from the concentration at location 1.
b. Plot the residuals versus the fitted values. Does the linear model seem appropriate?
c. Divide the data into two groups: points where $R1 < 4$ in one group, points where $R1 \geq 4$ in the other. Compute the least-squares line and the residual plot for each group. Does the line describe either group well? Which one?
d. Explain why it might be a good idea to fit a linear model to part of these data, and a nonlinear model to the other.

11. The article "The Equilibrium Partitioning of Titanium Between T_i^{3+} and T_i^{4+} Valency States in $CaO - SiO_2 - TiO_x$ Slags" (G. Tranell, O. Ostrovski, and S. Jahanshahi, Metallurgical and Materials Transactions B, 2002:61–66) discusses the relationship between the redox ratio T_i^{3+} / T_i^{4+} and oxygen partial pressure pO_2 in $CaO - SiO_2 - TiO_x$ melts. Several independent measurements of the redox ratio were made at each of five different partial pressures: $10^{-7}, 10^{-8}, 10^{-9}, 10,^{-10}$ and 10^{-12} atmospheres. The results for the runs at 14 mass percent TiO_x are presented in the following table.

Oxygen Partial Pressure	Redox Ratio Measurements
10^{-7}	0.011, 0.017, 0.034, 0.039
10^{-8}	0.018, 0.011, 0.026, 0.050, 0.034, 0.068, 0.061
10^{-9}	0.027, 0.038, 0.076, 0.088
10^{-10}	0.047, 0.069, 0.123, 0.162
10^{-12}	0.160, 0.220, 0.399, 0.469

a. Denoting the redox ratio by y and the partial pressure by x, theory states that y should be proportional to x^{β} for some β. Express this theoretical relationship as a linear model.
b. Compute the least-squares line for this linear model. Plot the residuals versus the fitted values. Does the linear model hold?
c. Further theoretical considerations suggest that under the conditions of this experiment, y should be proportional to $x^{-1/4}$Are the data in the preceding table consistent with this theory? Explain.

12. The article "The Selection of Yeast Strains for the Production of Premium Quality South African Brandy Base Products" (C. Steger and M. Lambrechts, Journal of Industrial Microbiology and Biotechnology, 2000:431–440) presents detailed information on the volatile compound composition of base wines made from each of 16 selected yeast strains. Below are the concentrations of total esters and total volatile acids (in mg/L) in each of the wines

Esters	Acids	Esters	Acids	Esters	Acids	Esters	Acids
284.34	445.70	173.01	265.43	229.55	210.58	312.95	203.62
215.34	332.59	188.72	166.73	144.39	254.82	172.79	342.21
139.38	356.88	197.81	291.72	303.28	215.83	256.02	152.38
658.38	192.59	105.14	412.42	295.24	442.55	170.41	391.30

a. Construct a scatterplot of acid concentration versus ester concentration. Indicate the outlier. b. Compute the coefficients of the least-squares line for predicting acid level (y) from ester level (x), along with their estimated standard deviations.

b. Compute the $P - value$ of the test of the null hypothesis $H_0 : \beta_1 = 0$.
c. Delete the outlier, and recompute the coefficients of the least-squares line, along with their estimated standard deviations.
d. Compute the $P - value$ of the test of the null hypothesis $H_0 : \beta_1 = 0$ for the data with the outlier deleted.
e. Does a linear model appear to be useful for predicting acid concentration from ester concentration? Explain.

13. The article "Mathematical Modeling of the Argon-Oxygen Decarburization Refining Process of Stainless Steel: Part II. Application of the Model to Industrial Practice" (J. Wei and D. Zhu, Metallurgical and Materials Transactions B, 2001:212–217) presents the carbon content (in mass %) and bath temperature (in K) for 32 heats of austenitic stainless steel. These data are shown in the following table.

Carbon %	Temp.	Carbon %	Temp.	Carbon %	Temp.	Carbon %	Temp.
19	1975	17	1984	18	1962	17	1983
23	1947	20	1991	19	1985	20	1966
22	1954	19	1965	19	1946	21	1972
16	1992	22	1963	15	1986	17	1989
17	1965	18	1949	20	1946	18	1984
18	1971	22	1960	22	1950	23	1967
12	2046	20	1960	15	1979	13	1954
24	1945	19	1953	15	1989	15	1977

a. Compute the least-squares line for predicting bath temperature (y) from carbon content (x).
b. Identify two outliers. Compute the two least-squares lines that result from the deletion of each outlier individually, and the least-squares line that results from the deletion of both outliers.
c. Are the least-squares lines computed in parts (a) and (b) similar? If so, report the line that was fit to the full data set, along with 95% confidence intervals for the slope and intercept. If not, report the range of slopes, without a confidence interval.

14. The article "Characteristics and Trends of River Discharge into Hudson, James, and Ungava Bays, 1964–2000" (S. D´ery, M. Stieglitz, et al., Journal of Climate, 2005:2540–2557) presents measurements of discharge rate x

(in km3 /yr) and peak flow y (in m3 /s) for 42 rivers that drain into the Hudson, James, and Ungava Bays. The data are shown in the following table:

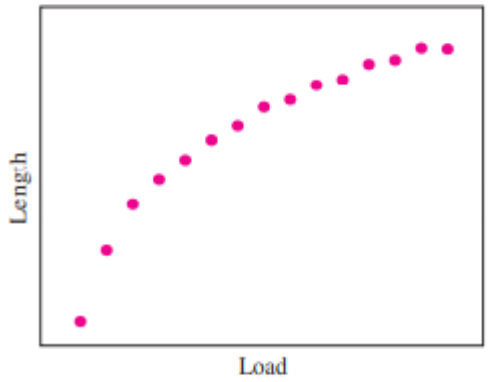
Discharge	Peak Flow	Discharge	Peak Flow	Discharge	Peak Flow
94.24	4110.3	17.96	3420.2	3.98	551.8
66.57	4961.7	17.84	2655.3	3.74	288.9
59.79	10275.5	16.06	3470.3	3.25	295.2
48.52	6616.9	14.69	1561.6	3.15	500.1
40.00	7459.5	11.63	869.8	2.76	611.0
32.30	2784.4	11.19	936.8	2.64	1311.5
31.20	3266.7	11.08	1315.7	2.59	413.8
30.69	4368.7	10.92	1727.1	2.25	263.2
26.65	1328.5	9.94	768.1	2.23	490.7
22.75	4437.6	7.86	483.3	0.99	204.2
21.20	1983.0	6.92	334.5	0.84	491.7
20.57	1320.1	6.17	1049.9	0.64	74.2
19.77	1735.7	4.88	485.1	0.52	240.6
18.62	1944.1	4.49	289.6	0.30	56.6

a. Compute the least-squares line for predicting y from x. Make a plot of residuals versus fitted values.
b. Compute the least-squares line for predicting y from $ln\ x$. Make a plot of residuals versus fitted values.
c. Compute the least-squares line for predicting $ln\ y$ from $ln\ x$. Make a plot of residuals versus fitted values.
d. Which of the three models (a) through (c) fits best? Explain.
e. Using the best model, predict the peak flow when the discharge is 50.0 km3 /yr.
f. Using the best model, find a 95% prediction interval for the peak flow when the discharge is 50.0 km3 /yr.

15. The article "Some Parameters of the Population Biology of Spotted Flounder (Ciutharus linguatula Linnaeus, 1758) in Edremit Bay (North Aegean Sea)" (D. T¨urker, B. Bayhan, et al., Turkish Journal of Veterinary and Animal Science, 2005:1013–1018) models the relationship between weight W and length L of spotted flounder as $W = aL^b$ where a and b are constants to be estimated from data. Transform this equation to produce a linear model.

16. The article "Mechanistic-Empirical Design of Bituminous Roads: An Indian Perspective" (A. Das and B. Pandey, Journal of Transportation Engineering, 1999:463–471) presents an equation of the form $y = a(1/x_1)^b(1/x_2)^c$ for predicting the number of repetitions for laboratory fatigue failure (y) in terms of the tensile strain at the bottom of the bituminous beam(x_1) and the resilient modulus(x_2). Transform this equation into a linear model, and express the linear model coefficients in terms of a, b, and c.

17. An engineer wants to determine the spring constant for a particular spring. She hangs various weights on one end of the spring and measures the length of the spring each time. A scatterplot of length (y) versus load (x) is depicted in the following figure.



a. Is the model $y = \beta_0 + \beta_1 x$ an empirical model or a physical law? Should she transform the variables to try to make the relationship more linear, or would it be better to redo the experiment? Explain.