Tutorial 6 Regression

1. The following statistics were obtained from a sample of size n = 75:
   – the predictor variable X has mean 32.2, variance 6.4;
   – the response variable Y has mean 8.4, variance 2.8; and
   – the sample covariance between X and Y is 3.6.

   a. Estimate the linear regression equation predicting Y based on X. [$\hat{y} = -9.71 + 0.56x$]

   b. Complete the ANOVA table. What portion of the total variation of Y is explained by variable X? [199.85, S57.35, 207.2, 0.7232]

   c. Construct a 99% confidence interval for the regression slope. Is the slope significant? [(0.45, 0.679, yes]

2. The data below represent investments, in 1000s, in the development of new software by some computer company over an 11 year period.

Year, X	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Investment, Y	17	23	31	29	33	39	39	40	41	44	47

   a. In the regression model with Y as a dependent variable, estimate the variance of Y. [7.39]

   b. Test whether the investment increases by more than 1,800 every year, on the average. [yes]

   c. Give a 95% prediction interval for the investment in new-product development in the year 2017. [50.36, 66.98]

(Subtracting 2000 from each year will certainly simplify the calculations.)

3. The following data set shows population of the United States (in million) since 1790,

Year	1790	1800	1810	1820	1830	1840	1850	1860
Population	3.9	5.3	7.2	9.6	12.9	17.1	23.2	31.4

Year	1870	1880	1890	1900	1910	1920	1930	1940
Population	38.6	50.2	63	76.2	92.2	106	123.2	132.2

Year	1950	1960	1970	1980	1990	2000	2010
Population	151.3	179.3	203.3	226.5	248.7	281.4	308.7

   a. Fit a linear regression model estimating the time trend of the U.S. population. For simplicity, subtract 1800 from each year and let x = year – 1800 serve as a predictor. [$\hat{y} = -32.0261 + 1.36(x - 1800)$]

   b. Complete the ANOVA table and compute R-square. [187277, 16434, 203711, 0.919]

   c. According to the linear model, what population would you predict for years 2015 and 2020? Is this a reasonable prediction? Comment on the prediction ability of this linear model. [260.4, 267.2]

5. Eight students, randomly selected from a large class, were asked to keep a record of the hours they spent studying before the midterm examination. The following table gives the number of hours these eight students studied before the midterm and their scores on the midterm.

Hours studied	15	7	12	8	18	6	9	11
Midterm score	97	78	87	92	89	57	74	69

a. Do the midterm scores depend on hours studied or do hours studied depend on the midterm scores? Do you expect a positive or a negative relationship between these two variables?

b. Taking hours studied as an independent variable and midterm scores as a dependent variable, compute S_{XX}, S_{YY} and S_{XY}. [119.5, 1251.875, 237.75]

c. Find the least squares regression line. [$\hat{y} = 58.9874 + 1.9896x$]

d. Interpret the meaning of the values of *intercept* and *slope* calculated in *part (c)*.

e. Plot the scatter diagram and the regression line.

f. Calculate *r* and also the coefficient of determination. [0.5714, 0.3778]

6. A sample of 10 pairs of values (x_i, y_i) which will be represented by the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

The summaries of the data are:

$\sum_{i=1}^{10}(x_i - \bar{x})^2 = 400$; $\sum_{i=1}^{10}(y_i - \bar{y})^2 = 425$; $\sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2 = 225$

Test at the 5% significance level whether the slope of the regression line is significant. [$reject\ H_0$]

7. A sample of 20 pairs of values (x_i, y_i) which will be represented by the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

F – ratio for testing the hypothesis $H_0 : \beta_1 = 0$ is equal to 12. Calculate R^2. [0.4]

8. A sample of 20 pairs of values (x_i, y_i) which will be represented by the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

The summaries of the data are:

$\hat{\beta}_0 = 138.561$; $\hat{\beta}_1 = -1.104$; $\sum(x_i - \bar{x})^2 = 10.668$; $\sum(y_i - \bar{y})^2 = 20.838$;

$\sum e_i^2 = \sum(y_i - \hat{y})^2 = 7.832$; $\bar{x} = 2$

a. Construct a 99% confidence interval for the mean per capita consumption of natural gas, $E(\hat{y}_h)$ for all gas stations with the prices of natural gas at \$2.11. [135.8, 136.7]

b. Construct a 99% prediction interval for the per capita consumption of natural gas, Y_h for a gas stations with the prices of natural gas at \$2.11. [134.3, 138.2]

9. In a medical survey, the weight X (in pounds) and systolic blood pressure Y of 26 randomly selected males in the age group 25 – 30 are recorded and the summaries are given below.

$$\sum_{i=1}^{26} x_i = 4743;\ \sum_{i=1}^{26} y_i = 3786;\ \sum_{i=1}^{26} x_i y_i = 697076;\ \sum_{i=1}^{26} x_i^2 = 880545;\ \sum_{i=1}^{26} y_i^2 = 555802;$$

$$\sum_{i=1}^{26} (x_i - \bar{x})^2 = 15312.35;\ \sum_{i=1}^{26} (y_i - \bar{y})^2 = 4502.15;\quad \sum_{i=1}^{26} (y_i - \hat{y}_i)^2 = 1808.5726$$

a. Fit a simple regression model to the data. [$\hat{y} = 69.1071 + 0.4194x$]

b. Find a 99% confidence interval on the slope. Interpret the interval obtained. [0.2231, 0.6157]

c. Construct the ANOVA table and test for significance of regression by using the significance level, $\alpha = 0.05$. [reject H_0]

d. Construct a 95% confidence interval on the mean response with $x_h = 180\,lb$. Interpret the interval obtained. [141.07, 148.13]

e. Find the coefficient of determination and explain what it means. [0.5983]

10. The number of pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature. Last year's usage Y (in 1000 pounds) and temperatures X (in $^\circ F$) were studied, the summaries of the data is given below.

$$\sum_{i=1}^{12} x_i = 558;\ \sum_{i=1}^{12} y_i = 5062.34;\ \sum_{i=1}^{12} x_i y_i = 265869.63;\ \sum_{i=1}^{12} x_i^2 = 29256;\ \sum_{i=1}^{12} y_i^2 = 2416234.61;$$

$$\sum_{i=1}^{12} (x_i - \bar{x})^2 = 3309;\ \sum_{i=1}^{12} (y_i - \bar{y})^2 = 280627.4204;\quad \sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = 37.8547.$$

a. Fit a simple regression model to the data. State one of the assumptions made in this model. Interpret the meaning of $\hat{\beta}_1$ found. [$\hat{y} = -6.3336 + 9.2085x$]

b. State the mean and variance for the least square estimator of β_1.

c. State the relationship between the correlation coefficient and the slope, and hence find the correlation coefficient and interpret its meaning. [$r \approx 1$]

d. Plant management believes that an increase in average ambient temperature of $1^\circ F$ will increase average monthly steam usage by $9100\,lb$. Do the data support this statement? Use the significance level, $\alpha = 0.01$. [reject H_0]

e. Construct a 99% prediction interval on Y_h with $x_h = 58^\circ F$. Interpret the interval obtained. [521.22, 534.29]