ICEF Probability theory and Statistics

Home assignments 9-10
23.05.2025

This extended home assignment will have double weight
and be marked according to 0-200 scale.

Problem 1. A school principal wanted to analyze the dependence between students' scores for home assignment in science and their exam scores. He recorded the results of 30 randomly selected students and computed the following statistics

	Mean	Standard deviation
HA	60	12
Exam	38	11

Correlation coefficient: 0.7

(a) Is there statistically significant correlation between the scores for the home assignments and the exam? Use the 10% significance level.

(b) Find the equation of the regression line of the exam score y on the home assignment score x.

(c) If a student gets 55 points for home assignments, what will his expected exam score?

Problem 2. Suppose that you are given independent observations y_1, y_2, y_3 such that

$$\begin{aligned} y_1 &= \alpha + 3\beta + \varepsilon_1, \\ y_2 &= \alpha + 2\beta + \varepsilon_2, \\ y_3 &= \alpha + \beta + \varepsilon_3. \end{aligned}$$

The random variables $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are normally distributed with mean of 0 and a variance of 2.

(a) Find the least squares estimators of the parameters α and β, and verify that they are unbiased estimators.

(b) Calculate the variance of estimator $\widehat{\beta}$ of β.

Problem 3. The following pairs of values (x, y) were obtained as a result of sampling.

x	106	136	101	75	124	92	115	86	50	91	75	141	85	82
y	87	84	71	68	52	74	80	54	66	99	60	89	83	78

(a) Estimate the correlation coefficient of X and Y.

(b) At 5% significance level test whether regression $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$ is significant. Assume that conditions needed to apply the test are satisfied.

(c) Represent the data graphically, find the equation of the regression line, and draw it on the plot.

Problem 4. In a study of dependence between variables x and y, it was computed that their correlation coefficient was $r_{xy} = 0.4$, their sample means were $\overline{x} = 12$, $\overline{y} = 18$, and the sample standard deviations were $s_x = 2.5$, $s_y = 1.5$. The size of the sample was $n = 52$.

(a) Can the null hypothesis of the absence of correlation between x and y be rejected at the 1% significance level?

(b) Find the coefficients $\widehat{\alpha}, \widehat{\beta}$ of the regression equation $y = \widehat{\alpha} + \widehat{\beta}x$.

(c) What will be the predicted value of y corresponding to $x = 11$?

(d) What is the residual corresponding to the observation $x = 9$, $y = 14$?

Problem 5. Below is data from a study on how a number of railcars in a train affects the fuel consumption.

Number of Railcars	30	25	37	31	47	43	39	63	40	28	34
Fuel Consumption (units/mile)	57	53	98	80	106	105	101	103	106	60	73

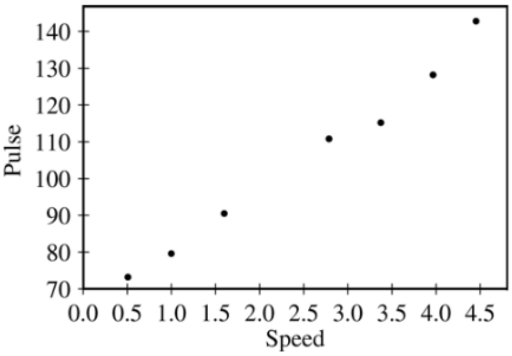
(a) Find the OLS estimates of the linear regression coefficients.

(b) Find residuals e_i of the model and plot the points (x_i, e_i). By looking at the graph, can you say that a linear model is appropriate for these data?

(c) Construct a two-sided 90% confidence interval for the slope coefficient and test the hypothesis that it is zero at the 10% significance level. Assume that the necessary conditions are satisfied.

(d) Find the coefficient of determination of the model and provide an interpretation for its value.

Problem 6. The graph below shows a relation between a person's walking speed (in miles per hour) and his pulse rate (in beats per minute). A regression output is also provided.



Predictor	Coef	SE	t	P
Constant	60.1	1.7	35.353	0.000
Speed	15.3	0.84	18.214	0.000

s = 2.03 $\quad$ $R^2 = 87.2\%$

(a) Provide an interpretation of the slope and intercept coefficients in this model.

(b) Provide an interpretation of the p-values and R^2 shown in the above output.

(c) Construct 99% confidence interval for the slope.

Problem 7. Based on a sample of 25 observations, the regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid was estimated. The least squares estimates obtained were $\widehat{\alpha} = 15.6$, $\widehat{\beta} = 1.3$.

The total and error sums of squares were $SST = 268$, $SSE = 204$.

(a) Find and interpret the coefficient of determination R^2.

(b) At 5% significance level test null hypothesis that the slope of regression line is 0 against two-sided alternative.

(c) Find 95% confidence interval for β.

Problem 8. A simple linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid is estimated. The following results and statistics are available:

$R^2 = 0.8$, $\widehat{\beta} = 1.6$, $n = 20$, $\overline{x} = 10$, $\overline{y} = 12$, $\sum_{i=1}^{20} x_i^2 = 2500$.

(a) Find $\sum_{i=1}^{20} y_i^2$.

(b) At 10% significance level test null hypothesis $H_0\colon \beta = 1.3$ against two-sided alternative.

Problem 9. Consider the following model: $y_i \ = \alpha + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid, $i = 1, \dots, n$. Derive OLS estimator for parameter α and find it's expectation and variance.

Problem 10. Show that for linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid

$$SST = S_{yy},$$
$$SSR = \widehat{\beta}^2 S_{xx},$$
$$SSE = S_{yy} - \widehat{\beta} S_{xy}$$

Additional problems

Problem 11. Following concerns about the appropriateness of using the license to kill by secret service agents, M, the Head of the Secret Intelligence Service, has to report information on the number of people killed during special missions. To prepare the report, he has collected data shown in the table below. One column lists the number of people killed directly by James Bond, the second column lists the number of people killed by others.

Mission	Bond killed	Others killed	Mission	Bond killed	Others killed
The Man with the Golden Gun	1	5	Octopussy	15	43
Dr. No	4	8	Quantum of Solace	16	15
On Her Majesty's Secret Service	5	37	Skyfall	17	23
A View to a Kill	5	57	For Your Eyes Only	18	36
Diamonds Are Forever	7	42	Thunderball	20	90
Live and Let Die	8	5	You Only Live Twice	21	175
Goldfinger	9	68	The World is Not Enough	27	43
Licence to Kill	10	13	Tomorrow Never Dies	30	24
Casino Royale	11	11	Die Another Day	31	20
From Russia With Love	11	16	The Spy Who Loved Me	31	116
Moonraker	12	69	GoldenEye	47	25
The Living Daylights	13	29			

(a) To report the data in a concise form, M decides to draw a graph which will summarize the number of people James Bond kills in one mission. Choose an appropriate graphical method to do that and draw the graph.
Answer: Boxplot with outlier 47.

(b) M has noticed that there is probably a dependence between the number of people killed by Bond and killed by others. Find the equation of the least squares regression line that shows the relation between these variables and represent this relation graphically.
Answer: $\widehat{y} = 29.115 + 0.814x$ regression line.

(c) Bond has reported that in the last mission, Spectre, he killed 32 people. What prediction can M make about the number of killed people by others using the regression line?
Answer: $x_{24} = 32$, $y_{24} = 35.627$.

(d) Is the relation between the variables in c) statistically significant? Use the p-value approach and assume the conditions for applying the appropriate statistical test are met.
Answer: $T_{st} = 1.048$, $p - val = 0.306$ – not enough evidence to reject $H_0\colon \beta = 0$.

Problem 12. Prove that

(a) $R^2 = r_{xy}^2$.

(b)

$$\frac{r_{xy}}{\sqrt{(1 - r_{xy}^2)/(n-2)}} = \frac{\widehat{\beta}}{s/\sqrt{s_{xx}}}$$

Problem 13. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be datapoints. Model $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid, $\widehat{y} = \widehat{\alpha} + \widehat{\beta} x$ regression line, $e_i = y_i - \widehat{y}_i$ residuals. Prove that

$$\sum_{i=1}^{n} e_i = 0, \tag{1}$$

$$\sum_{i=1}^{n} e_i x_i = 0 \tag{2}$$

This is intuitive explanation why we need to divide $SSE = \sum_{i=1}^{n} e_i^2$ by $n-2$ (instead of $n-1$) to get an unbiased estimator of σ^2.

Problem 14. Consider the following model (**linear regression without intercept**)

$y_i \ = \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid, $i = 1, \ldots, n$. Derive OLS estimator for parameter β and find it's expectation and variance.

Answer: $\widehat{\beta} = \dfrac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$, $\mathrm{E}(\widehat{\beta}) = \beta$, $\mathrm{Var}(\widehat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^n x_i^2}$.

Problem 15. For linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ iid find $\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta})$.

Answer: $\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = -\dfrac{\overline{x} \cdot \sigma^2}{S_{xx}}$.