ANOVA in simple linear regression

Statistics

Anton Afanasev

Higher School of Economics

DSBA 211

March 11, 2023

Seminar Overview

Quiz

1. Explain the difference between following terms in simple linear regression:
   - y_i,
   - $\widehat{y}_i$,
   - $\mathsf{E}(y_i)$.
2. Sketch a simple linear regression with
   - confidence intervals for $\mathsf{E}(y)$,
   - prediction intervals for y,

   of arbitrary confidence level in chosen interval. What's the difference between confidence and prediction intervals?

Problem 1

A car insurance company would like to examine the relationship between driving experience and insurance premium. For this reason, a random sample of ten drivers is taken and the years of driving experience (x) as well as the monthly insurance premium (y, in £) is recorded. The data are shown in the table below:

Driver	№1	№2	№3	№4	№5	№6	№7	№8	№9	№10
Driving experience (x)	6	3	11	10	15	6	25	16	15	20
Insurance premium (y)	66	88	51	70	44	56	42	60	45	40

The summary statistics for these data are:

Sum of x data: 127	Sum of the squares of x data: 2033
Sum of y data: 562	Sum of the squares of y data: 33662
Sum of the products of x and y data: 6402	

Problem 1

1. Draw a scatter diagram of these data. Label the diagram carefully.
2. Calculate the sample correlation coefficient. Interpret your findings.
3. Calculate the least squares line of y on x and draw the line on the scatter diagram.
4. Based on the regression equation in part 3, what will be the predicted monthly insurance premium for a driver with 10 years of experience? Will you trust this value? Justify your answer.

Sources of variance in simple linear regression

- Model of simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \mathsf{E}(\varepsilon_i) = 0, \mathsf{V}(\varepsilon_i) = \sigma^2.$$

Regressand y can change because of:

$$\begin{aligned} \varepsilon_i &- \text{error/noise}, \\ \beta_1 x_i &- \text{change in regressor } x. \end{aligned}$$

- We can compare variances, created by both sources, to find a significant evidence of presence of dependency between x and y.

$$\text{Total SS} = \underbrace{\text{Regression SS}}_{\beta_1 x_i} + \underbrace{\text{Residual SS}}_{\varepsilon_i}$$

Variation, produced by error

- Estimate of ε_i:

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i,$$

where $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ are sample regression line instants (estimates of $\mathsf{E}(y_i)$).

- Variation, created by ε_i (residual, error):

$$RSS = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$$

with $n-2$ degrees of freedom.

- Correspondence of estimates to a model:

$$\begin{array}{rccccc} \text{model:} & y_i & = & \beta_0 + \beta_1 x_i & + & \varepsilon_i \\ \text{estimates:} & y_i & = & \widehat{y}_i & + & (y_i - \widehat{y}_i) \end{array}$$

Variation, produced by regression

- Sample regression line with OLS estimate of β_0:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$
$$\widehat{y}_i = \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i.$$

- Deviation of regressand estimate from the mean is in direct ratio to the deviation of regressor:

$$\widehat{y}_i - \overline{y} = \widehat{\beta}_1 (x_i - \overline{x}).$$

This difference will produce required variation (from regression):

$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i),$$
$$(y_i - \overline{y}) = \widehat{\beta}_1 (x_i - \overline{x}) + (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i).$$

ANOVA decomposition

- Total variation decomposition:

$$
\begin{array}{ccccc}
\text{Total SS} & = & \text{Regression SS} & + & \text{Residual SS} \\
TSS & = & ESS & + & RSS \\
SST & = & SSR & + & SSE \\
\sum_{i=1}^{n}(y_i - \overline{y})^2 & = & \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 & + & \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \\
SS_{yy} & = & \widehat{\beta}_1^2 \cdot SS_{xx} & + & \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2
\end{array}
$$

with degrees of freedom

$$
\begin{array}{ccccc}
n-1 & = & 1 & + & n-2
\end{array}
$$

Pivot function

- Under assumptions of normal regression ($\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$) and truth of null hypothesis of ANOVA in simple linear regression

$$\mathrm{H}_0 : \beta_1 = 0,$$

the following is true:

$$SST/\sigma^2 \sim \chi^2_{n-1},$$
$$SSR/\sigma^2 \sim \chi^2_{1},$$
$$SSE/\sigma^2 \sim \chi^2_{n-2}.$$

- Pivot function:

$$F\Big|_{\mathrm{H}_0} = \frac{SSR/1}{SSE/(n-2)} = \frac{\widehat{\beta}_1^2 \cdot SS_{xx}}{MSE} \sim F_{1;\, n-2}.$$

H_0 is rejected in favor of $\mathrm{H}_1 : \beta_1 \neq 0$ if $F > F_{1;\, n-2;\, \alpha}$.

Coefficient of determination

- Coefficient of determination is a goodness-of-fit metric:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST},$$

  proportion of the total variation of y, explained by x.
  The closer R^2 to 1 – the better explanatory power of the regression model.

- For OLS estimates in simple linear regression:

$$R^2 = \frac{\widehat{\beta}_1^2 \cdot SS_{xx}}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}^2} \cdot \frac{SS_{xx}}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx} \cdot SS_{yy}} = \widehat{\rho}^2.$$

Problem 2

A simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is estimated. The following results and statistics are available:

$$R^2 = 0.80; \qquad \widehat{\beta}_1 = 1.6; \qquad n = 20;$$

$$\overline{x} = 10; \qquad \overline{y} = 12; \qquad \sum_{i=1}^{20} x_i^2 = 2500.$$

1. Find $\sum\limits_{i=1}^{20} y_i^2$.
2. At 10% significance level test $\mathrm{H}_0 : \beta_1 = 1.3$ against two-sided alternative.

Problem 3

The table shows, for eight vintages of select wine, purchases per buyer (y) and the wine buyer's rating in a year (x).

x	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
y	24	21	22	22	18	13	9	6

1. Estimate the regression of purchases per buyer on the buyer's rating.
2. Interpret the slope of the estimated regression line.
3. Find and interpret the coefficient of determination.
4. Find and interpret a 90% confidence interval for the slope of the population regression line.
5. Find a 90% confidence interval for expected purchases per buyer for a vintage for which the buyer's rating is 2.0.

Problem 4

A random sample of 15 less-developed countries showed the following relation between population density X and economic growth rate Y (*Simon*, 1981).

Country	Population density per km^2 (X)	Percent annual change in per capita income (Y)
A	27	3.3
B	32	0.8
C	118	1.4
D	270	5.4
E	10	1.4
...	...	...
Average	54	2.4
Total SS	80920	46.9
MS (variance)	5780	3.35
St. dev.	76.0	1.83
Correlation, $\widehat{\rho}$	0.54	

Problem 4

1. Calculate the regression line of Y on X. Graph the regression line, along with the first 5 points.
2. Carry out the ANOVA table as far as the p-value for $\mathrm{H}_0 : \beta_1 = 0$. Can you reject H_0 at the 5% error level?
3. Using the slope in part 1 and the residual variance in part 2, calculate the 95% confidence interval for β_1. Can you reject H_0 at 5% error level?
4. Test $\mathrm{H}_0 : \rho = 0$ against $\mathrm{H}_1 : \rho \neq 0$. Can you reject H_0 at the 5% error level?
5. Do you get consistent answers in parts 2, 3, and 4 for the question "Are X and Y linearly related?"
6. From the ANOVA table in 2, find the proportion of the SS that is explained by the regression. Does it agree with $\widehat{\rho}^2$? Also find the proportion left unexplained. Does it agree with $\left(1 - \widehat{\rho}^2\right)$?

Look at the time!