University Introductory Statistics · Practice Worksheet · tutorinmontreal.ca

Simple Linear Regression and Correlation

Name: ______________________ Date: ____________

Concepts

- Scatterplots and the Correlation Coefficient
- The Least-Squares Regression Line
- Interpreting Slope, Intercept and Predictions (theory)
- The Coefficient of Determination
- Residuals and the Conditions for Regression (theory)
- Inference for the Slope
- Confidence and Prediction Intervals for a Response (theory)

Scatterplots and the Correlation Coefficient

1. A community-garden cooperative adds a different amount of compost to five plots and records the tomato harvest of each.

Compost, x (kg)	2	4	6	8	10
Harvest, y (kg)	16	22	20	23	24

Using $S_{xx} = \sum(x_i - \bar{x})^2$, $S_{yy} = \sum(y_i - \bar{y})^2$, $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$ and $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, compute the correlation coefficient r and describe the direction and strength of the linear association.

2. For 40 greenhouses, the correlation between the mean daily temperature x (in °C) and the daily water use y (in litres) is $r = 0.62$. Recall that $r = \frac{1}{n-1}\sum z_{x_i} z_{y_i}$, where $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$ and $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$ are standard scores. Give the new correlation, with a one-line reason, if

a) the temperatures are converted to degrees Fahrenheit, $F = 1.8x + 32$;

b) the water use is recorded in millilitres;

c) y is replaced by the water left in a full 500 L tank at the end of the day, $500 - y$;

d) the roles of the two variables are swapped.

The Least-Squares Regression Line

3. A bicycle courier records the distance x (km) and the delivery time y (min) for six deliveries:

$$(2, 10),\ (3, 13),\ (5, 20),\ (6, 24),\ (9, 38),\ (11, 39).$$

The least-squares line $\hat{y} = b_0 + b_1 x$ has $b_1 = \frac{S_{xy}}{S_{xx}}$ and $b_0 = \bar{y} - b_1\bar{x}$, with $S_{xx} = \sum(x_i - \bar{x})^2$ and $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$. Find the line, and use it to predict the time for a 7 km delivery.

4. An agronomist records June rainfall x (mm) and hay yield y (bales per hectare) on 25 farms and reports only the sums

$$n = 25, \quad \sum x = 2000, \quad \sum y = 1050, \quad \sum x^2 = 165\,000, \quad \sum xy = 85\,500.$$

Using $S_{xx} = \sum x^2 - \dfrac{(\sum x)^2}{n}$ and $S_{xy} = \sum xy - \dfrac{(\sum x)(\sum y)}{n}$, find the least-squares line and predict the yield after 100 mm of June rain.

Interpreting Slope, Intercept and Predictions

5. For 30 households with 1 to 6 occupants, the least-squares line relating the monthly water bill y ($) to the number of occupants x is $\hat{y} = 18.40 + 11.25x$.

a) Predict the monthly bill of a household of 4.

b) By how much do the predicted bills of two households differ if one has 2 more occupants than the other?

c) Compute the prediction for a household of 12, and say whether it should be trusted.

The Coefficient of Determination

6. A transit analyst regresses the percentage of on-time trips y on route length x (km) for a sample of bus routes. The slope is $b_1 = -1.8$ and the analysis-of-variance table reads:

Source	df	Sum of squares
Regression	1	540
Residual (error)	18	360
Total	19	

The residual degrees of freedom are $n-2$, $r^2 = \dfrac{\text{SSR}}{\text{SST}}$ and $s = \sqrt{\dfrac{\text{SSE}}{n-2}}$. Find the number of routes n, the total sum of squares, r^2, s and r.

7. For 12 cafés on the same street, daily foot traffic x (hundreds of passers-by) and daily sales y (hundreds of dollars) give

$$\bar{x} = 7.5, \quad \bar{y} = 24, \quad S_{xx} = 80, \quad S_{yy} = 500, \quad S_{xy} = 180.$$

With $b_1 = \dfrac{S_{xy}}{S_{xx}}$, $\text{SSE} = S_{yy} - b_1 S_{xy}$ and $r^2 = 1 - \dfrac{\text{SSE}}{S_{yy}}$, find the least-squares line, SSE and r^2. Check r^2 against $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$.

Residuals and the Conditions for Regression

8. A garden centre prices six clay pots by their diameter. The least-squares line relating price y (\$) to diameter x (cm) is $\hat{y} = 4 + 0.8x$. The data are

$$(10, 13),\ (15, 15),\ (20, 18),\ (25, 26),\ (30, 29),\ (35, 31).$$

Compute every residual $e_i = y_i - \hat{y}_i$, check that the residuals sum to 0, and name the pot whose price is furthest below the line.

Inference for the Slope

9. For 20 sourdough loaves, a baker regresses the rise y (cm) on the proofing time x (h) and finds

$$b_1 = 0.45, \qquad s = 1.2, \qquad S_{xx} = 64.$$

The standard error of the slope is $\text{SE}(b_1) = \dfrac{s}{\sqrt{S_{xx}}}$ and the test statistic $t = \dfrac{b_1}{\text{SE}(b_1)}$ has $n-2$ degrees of freedom. Use $t_{0.025,\,18} = 2.101$. Assuming the regression conditions hold, test at the 5% level whether proofing time and rise are linearly related.

10. A housing researcher regresses the monthly rent y (\$) of 14 apartments on their distance x (km) from the nearest metro station. The software prints

Predictor	Coef	SE Coef	T
Constant	1620.0	85.0	19.06
Distance	−42.5	12.0	−3.54

The interval is $b_1 \pm t_{\alpha/2,\,n-2}\,\mathrm{SE}(b_1)$; use $t_{0.025,\,12} = 2.179$. Compute a 95% confidence interval for β_1, and say what it implies about a two-sided test of $H_0\colon \beta_1 = 0$ at the 5% level.

Confidence and Prediction Intervals for a Response

11. A maple-syrup producer records, for 12 batches, the sap boiled x (hundreds of litres, between 10 and 30) and the syrup obtained y (litres). The least-squares line is $\hat{y} = 0.6 + 2.4x$, with $\bar{x} = 20$, $S_{xx} = 250$ and $s = 3.0$. At $x = x^*$, with $\hat{y}^* = b_0 + b_1 x^*$,

$$\text{mean response:}\quad \hat{y}^* \pm t_{\alpha/2,\,n-2}\, s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}},$$

$$\text{new observation:}\quad \hat{y}^* \pm t_{\alpha/2,\,n-2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.$$

Use $t_{0.025,\,10} = 2.228$. For batches of 2500 L of sap, compute a 95% confidence interval for the mean syrup yield and a 95% prediction interval for one batch.

Synthesis — drawing on several topics in this unit

12. A bakery records the number of staff x on five morning shifts and the trays of pastries y finished by 9 a.m.:

$$(1, 9),\ (2, 8),\ (3, 12),\ (4, 16),\ (5, 15).$$

Use $b_1 = \dfrac{S_{xy}}{S_{xx}}$, $b_0 = \bar{y} - b_1\bar{x}$, $s = \sqrt{\dfrac{\text{SSE}}{n-2}}$, $\text{SE}(b_1) = \dfrac{s}{\sqrt{S_{xx}}}$, and $t_{0.025,\,3} = 3.182$.

a) Find the least-squares line.

b) Compute the residuals, SSE and $r^2 = 1 - \dfrac{\text{SSE}}{S_{yy}}$.

c) Assuming the regression conditions hold, test $H_0\colon \beta_1 = 0$ against $H_a\colon \beta_1 \neq 0$ at the 5% level.

Want the harder problems and the worked solutions?

The *Challenge Problems* set and the *Answer Keys* for Simple Linear Regression and Correlation — with the reasoning written out in full — are available at **tutorinmontreal.ca**.