L3. Correlation and Regression

Example 1: Plastic Bags

A Canadian grocery store's plan to shame customers out of using plastic bags backfired spectacularly. Vancouver's East West Market printed embarrassing phrases like "Dr. Toews' Wart Ointment Wholesale", hoping to deter use—but instead, people loved them and flocked to collect them.If guilt and humiliation aren't enough to change consumer behaviour, then Kenya's especially draconian ban on plastic bags might help; anyone caught selling, producing, or even carrying a plastic bag faces a $38000 fine or four years in prison.

The decomposition time for biodegradable plastic bags depends on moisture levels. The table below shows how long they take to break down based on environmental humidity.

Moisture in Environment (%)	2	3	4	7	8	9	12	13	14	16	17
Time to Decompose (days)	39	36	36	34	33	31	28	27	26	25	23

a. Make a scatter plot of the data.

b. Calculate the sample covariance between the moisture level in the environment and the number of days it takes for the bag to decompose.

c. Calculate the coefficient of correlation, r, for the given data. Comment on the strength and direction of the relationship.

d. Calculate the coefficient of determination, R^2, for the given data. Comment on the amount of variation that is accounted for by this data.

e. Calculate the slope of the least squares line and interpret it in the context of the problem.

f. Calculate the y-intercept of the least squares line and interpret it in the context of the problem.

g. Estimate how long it would take for a bag to decompose if there is 18% moisture in the air. Comment on the reliability of this estimate.

h. Predict how long it will take for a bag to fully decompose in an environment that contains 5% moisture. Comment on the reliability of this estimate.

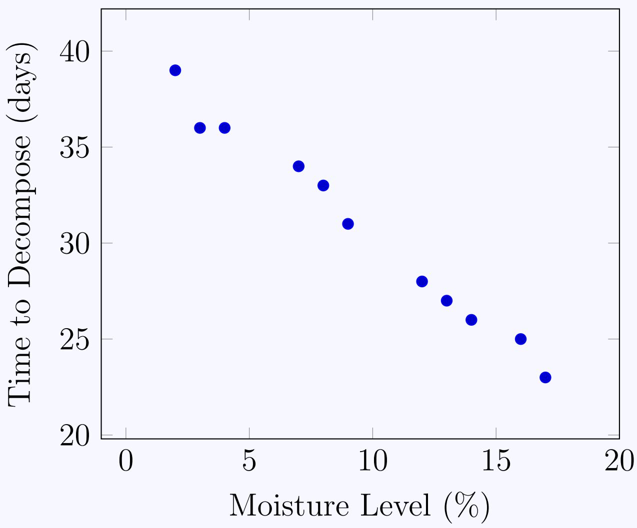
i. Scientists observed that it took 33 days for a bag to fully decompose in an environment that had 8.5% moisture in it. Calculate the residue, and comment on the performance of the model

Solution

Here is the table augmented with additional columns to facilitate our calculations.

Moisture Level (%) x	Time ($days$) y	x^2	y^2	xy
2	39	4	1521	78
3	36	9	1296	108
4	36	16	1296	144
7	34	49	1156	238
8	33	64	1089	264
9	31	81	961	279
12	28	144	784	336
13	27	169	729	351
14	26	196	676	364
16	25	256	625	400
17	23	289	529	391
105	338	1277	10662	2953

a. Here is a scatter plot of the given data.



b. Sample covariance

$$
\begin{aligned}
s_{xy} = \frac{1}{n-1}\left[\sum xy - \frac{\sum x \sum y}{n}\right] &= \frac{1}{11-1}\left[2953 - \frac{(105)(338)}{10}\right] \\
&= \frac{1}{10}[-273.6363] \\
&= -27.3364
\end{aligned}
$$

c. Coefficient of correlation.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{11(2953) - (105)(338)}{\sqrt{11(1277) - (105)^2}\sqrt{11(10662) - (338)^2}}$$
$$= -\frac{3007}{\sqrt{3022}\sqrt{3038}}$$
$$= -0.9924$$

Comment: $r = -0.9924$ indicates a very strong negative correlation between environmental moisture and the number of days required for a bag to decompose. The higher the level of moisture in the environment, the fewer it days it takes for the bag to decompose.

d. Coefficient of determination.

$$R^2 = r^2 = (-.9924)^2$$
$$= 0.9894$$

$$1 - R^2 = 1 - (-0.9924)^2$$
$$= 0.0151$$

Interpretation: 98.49% of the variability in decomposition time can be explained by the moisture levels in the environment. The remaining 1.51% of the variation in decomposition time is due to other factors not accounted for by moisture. This suggests that other influences (such as temperature, microbial activity, or bag material) play only a minor role in comparison to moisture. Thus, we can confidently use moisture levels to predict decomposition time with a high degree of accuracy.

e. Slope of regression line.

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{11(2953) - (105)(338)}{11(1277) - (105)^2}$$
$$= -\frac{3007}{3022}$$
$$= -0.995$$

Interpretation: For every 1% increase in moisture levels, the number of days for the bag to decompose decreases by 0.995 days.

f. Intercept of regression line.

$$\begin{aligned} a = \frac{\sum y - b\sum x}{n} &= \frac{10662 - (-0.995)(105)}{11} \\ &= 40.2253 \end{aligned}$$

Interpretation: With 0% moisture in the environment, it would take 40.2253 days for the bag to decompose completely.

g. $\hat{y} = 40.2253 - 0.995x$

$$\hat{y} = 40.2253 - 0.995(18) = 22.1347 \text{ days}$$

Comment: With 18% moisture in the environment, it would take approximately 22.1247 days to decompose. This estimate uses **extrapolation** and unreliable, since $18 \notin [2, 17]$

h. $\hat{y} = 40.2253 - 0.995x$

$$\hat{y} = 40.2253 - 0.995(5) = 35.2502 \text{ days}$$

Comment: With 5% moisture in the environment, it would take approximately 35.2502 days to decompose. This estimate uses **interpolation** and reliable, since $18 \in [2, 17]$

i. Residual error

$$\hat{y} = 40.2253 - 0.995(8.5) = 31.7675 \text{ days}$$

$$\begin{aligned} e &= y - \hat{y} \\ &= 33 - 31.7675 \\ &= 1.2325 \text{ days} \end{aligned}$$

Comment: The residual error is 1.2325 days; and the model underestimates the number of days for the bag to decompose.

Example 2: Stealthy Starbucks

With over 29,000 locations across the U.S., Starbucks is everywhere—including inside the CIA headquarters in Langley, Virginia. This ultra-secret café, known as "Stealthy Starbucks", is so classified that it doesn't show up on GPS or on Google Maps. But the secrecy doesn't stop there. Receipts are titled 'Store Number 1' and no names—real or fake—are written on cups. Despite its clandestine nature, Stealthy Starbucks is the busiest in the world, as CIA agents rarely leave the premises. Its top-selling items? Lemon pound cake and Frappuccinos.

The amount of sugar and caloric count for some of Starbucks' more popular drinks are shown in the table below:

Sugar (g)	42	35	40	18	47	54	30	36
Calories (kcal)	290	250	260	180	300	290	190	230

a. Make a scatter plot of the data.

b. Calculate the coefficient of correlation, r, for the given data. Comment on the strength and direction of the relationship.

c. Calculate the coefficient of determination, R^2, for the given data. Comment on the amount of variation that is accounted for by this data.

d. Calculate the slope of the least squares line and interpret it in the context of the problem.

e. Calculate the y-intercept of the least squares line and interpret it in the context of the problem.

f. Make a prediction for a drink that contains 60 g of sugar. Is this estimate trustworthy?

g. Make a prediction for a drink that contains 10 g of sugar. Is this estimate reliable?

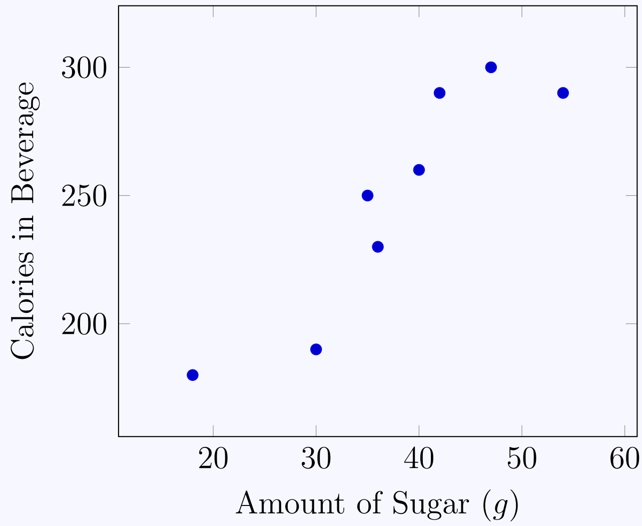
h. A Starbucks coffee beverage that has $37g$ of sugar, actually contains 250 calories. Calculate the residual and interpret the result.

Solution

Here is the table augmented with additional columns to facilitate our calculations.

Sugar (g) x	Calories y	x^2	y^2	xy
42	290	1764	84100	12180
35	250	1225	62500	8750
40	260	1600	67600	10400
18	180	324	32400	3240
47	300	2209	90000	14100
54	290	2916	84100	15660
30	190	900	36100	5700
36	230	1296	52900	8280
302	1990	12234	509700	78310

a. Here is a scatter plot of the given data.



b. Coefficient of correlation.

$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{8(78310) - (302)(1990)}{\sqrt{8(12234) - (302)^2}\sqrt{8(509700) - (1990)^2}} \\ &= \frac{25500}{\sqrt{6668}\sqrt{117500}} \\ &= 0.911 \end{aligned}$$

Comment: $r = 0.911$ indicates a very strong positive correlation between the amount of sugar and the number of calories in the beverage. This means that as sugar content increases, the calorie count also increases in a very consistent manner. In other words, beverages with higher sugar content tend to be associated with higher calories, and beverages with lower sugar content tend to have fewer calories.

c. Coefficient of determination.

$$\begin{aligned} R^2 = r^2 &= (0.911)^2 \\ &= 0.8299 \end{aligned}$$

$$\begin{aligned} 1 - R^2 &= 1 - (0.8299)^2 \\ &= 0.1701 \end{aligned}$$

Interpretation: 82.99% of the variation in calorie content is directly related to sugar content. This suggests that sugar is a major factor influencing calories, but not the only one. The remaining 17%of the variability in calories is due to other factors, such as fats, protein, etc.

d. Slope of regression line.

$$b = \frac{\sum xy - n\sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{8(78310) - (302)(1990)}{8(12234) - (302)^2} = \frac{2550}{6668}$$
$$= 3.8242$$

Interpretation: For every $1\,g$ increase in sugar, the number of calories in the beverage increases by 3.8242 calories.

e. Intercept of regression line.

$$a = \frac{\sum y - b\sum x}{n} = \frac{1990 - (3.8242)(302)}{8} = 104.386$$

Interpretation: A beverage with $0g$ of sugar will contain 104.356 calories.

f. $\hat{y} = 104.386 + 3.824x$

$$\hat{y} = 104.386 + 3.824(60) = 333.8384 \text{ calories}$$

Comment: With $60g$ of sugar, the beverage would contain 333.8384 calories. This estimate uses **extrapolation** and unreliable, since $60 \notin [18, 54]$

g. $\hat{y} = 104.386 + 3.824x$

$$\hat{y} = 104.386 + 3.824(10) = 142.628 \text{ calories}$$

Comment: With $10g$ of sugar, the beverage would contain 142.628 calories. This estimate uses **extrapolation** and unreliable, since $60 \notin [18, 54]$

h. Residual error

$$\hat{y} = 104.386 + 3.824(37) = 245.8814 \text{ calories}$$

$$\begin{aligned} e &= y - \hat{y} \\ &= 250 - 245.8814 \\ &= 4.1186 \text{ calories} \end{aligned}$$

Comment: The residual error is 4.1186 calories; and the model underestimates the number of calories in the beverage.

Example 3: Hot vs. Cold

You're more likely to devour a meal if it's cold—because, apparently, hot food tricks your brain into thinking you're full faster. So if you've ever wondered why ice cream disappears faster than a bowl of oatmeal, science has your answer. To put this to the test, researchers served participants identical portions of soup at different temperatures and measured how much they actually ate. The results are shown in the table below.

Temperature (C)	75	65	55	45	35	25	15	5
Average Consumption (g)	250	270	300	340	390	450	510	580

a. Make a scatter plot of the data.

b. Calculate the coefficient of correlation, r, for the given data. Comment on the strength and direction of the relationship.

c. Calculate the coefficient of determination, R^2, for the given data. Comment on the amount of variation that is accounted for by this data.

d. Calculate the slope of the least squares line and interpret it in the context of the problem.

e. Calculate the y-intercept of the least squares line and interpret it in the context of the problem.

f. Predict how much soup would be consumed if it were served at $10°C$. Comment on the reliability of the estimate.

g. Predict how much soup would be consumed if it were served at $85°C$. Comment on the reliability of the estimate.

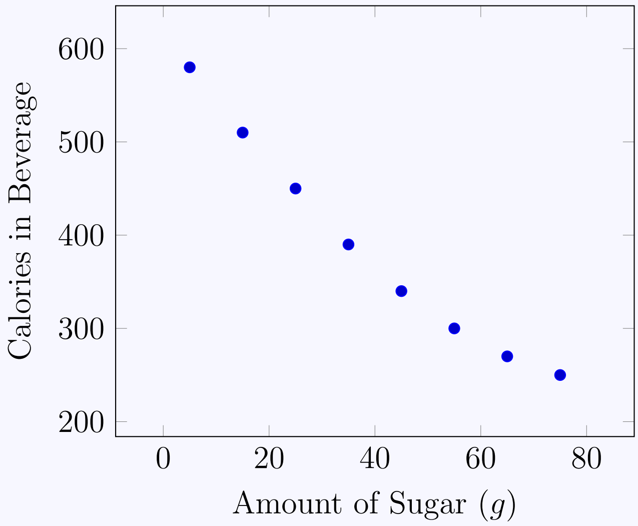
h. Suppose that soup served at $20°C$ resulted in an average consumption of $475\,g$ of soup, calculate the residual and comment on the result.

Solution

Here is the table augmented with additional columns to facilitate our calculations.

Temperature ($^\circ C$) x	Average Consumption (g) y	x^2	y^2	xy
75	250	5625	62500	18750
65	270	4225	72900	17550
55	300	3025	90000	16500
45	340	2025	115600	15300
35	390	1225	152100	13650
25	450	625	202500	11250
15	510	225	260100	7650
5	580	25	336400	2900
320	3090	17000	1292100	103550

a. Here is a scatter plot of the given data.



b. Coefficient of correlation.

$$
\begin{aligned}
r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} &= \frac{8(103550) - (320)(3090)}{\sqrt{8(17000) - (320)^2}\sqrt{8(1292100) - (3090)^2}} \\
&= -\frac{160400}{\sqrt{33600}\sqrt{788700}} \\
&= -0.9853
\end{aligned}
$$

Comment: $r = -0.9853$ indicates a very strong negative correlation between the temperature of soup and the average amount consumed. This means that as the temperature increases, the average consumption decreases in a very consistent manner. In other words, people consume less soup when it is hotter and more soup when it is cooler.

c. Coefficient of determination.

$$
\begin{aligned}
R^2 = r^2 &= (-0.9853)^2 \\
&= 0.9709
\end{aligned}
$$

$$
\begin{aligned}
1 - r^2 &= 1 - (0.9853)^2 \\
&= 0.0291
\end{aligned}
$$

Interpretation: 97.09% of the variability in the amount of soup consumed can be explained by the temperature of the soup (ie. soup consumption is directly tied to the temperature. This suggests that soup temperature is the main determining factor for how much people eat.). Meanwhile, the remaining 2.91% of the variation could be due to other factors (e.g. individual preferences, hunger levels, etc).

d. Slope of regression line.

$$b = \frac{\sum xy - n\sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{8(103550) - (320)(3090)}{8(17000) - (320)^2} = -\frac{160400}{33600}$$
$$= -4.7738$$

Interpretation: For every $1°C$ increase in soup temperature, the average amount consumed decreases by approximately 4.7738 grams.

e. Intercept of regression line.

$$\begin{aligned} a &= \frac{\sum y - b\sum x}{n} \\ &= \frac{17000 - (-4.7738)(320)}{8} \\ &= 577.2024 \end{aligned}$$

Interpretation: If soup temperature was $0°C$, the predicted amount of soup consumed would be 577.2 grams. However, this is likely outside the realistic range of serving temperatures, so while it provides a mathematical reference point, it may not have practical significance.

f. $\hat{y} = 577.2024 - 4.7738x$

$$\hat{y} = 577.2024 - 4.7738(10) = 529.4643\,g$$

Comment: If the soup was served at $10^{\circ C}$, the predicted average consumption would be $529.4643\,g$. This is an interpolation, and is a reliable estimate since $10 \in [5, 75]$

g. $\hat{y} = 577.2024 - 4.7738x$

$$\hat{y} = 577.2024 - 4.7738(85) = 171.4286\,g$$

Comment: If the soup was served at $85^{\circ C}$, the predicted average consumption would be $171.4286\,g$. This estimate uses extrapolation, and is a unreliable since $85 \notin [5, 75]$

h. Residual error

$$\hat{y} = 577.2024 - 4.7738(20) = 481.7262\,g$$

$$\begin{aligned} e &= y - \hat{y} \\ &= 475 - 481.7262 \\ &= -6.7262\,g \end{aligned}$$

Comment: The residual error is $6.7262\,g$; and the model overestimates the average amount of soup consumed.