Association/relationship between quantitative variables

Objective

(1) Learn how to explore association between 2 quantitative variables.

(2) Learn basic concepts related to making and interpreting scatterplots.

(3) Understand the concept of correlation coefficient.

Concept briefs:

* Scatterplot = Basic x-y plot showing one quant. variable on the x-axis, and the other on the y-axis.

* Explanatory variable = The variable whose effect on the other we want to study.

* Response variable = The other variable.

* Shape, strength, direction (of association) = Key items we investigate in scatterplots.

* Linear association = Scatter pattern with approximately "straight line" trend.

* Correlation coefficient = Numerical measure of strength of linear association.

Scatterplots: What do they indicate about "association"?

Three key things to look for:

[1] Strength [2] Direction [3] Form (or, general shape)

[1] Strength of association

- Low scatter ⇒ strong association
- High scatter ⇒ weak association

[2] Direction of association

Does the response variable go ↑ or ↓ as the explanatory goes ↑?

- Response ↑ ⇒ positive association
- Response ↓ ⇒ negative association

[3] Form/shape of association

Is the overall scatter pattern roughly straight or curved?

- Approximately straight ⇒ linear association
- Not straight ⇒ not linear association

Correlation coefficient

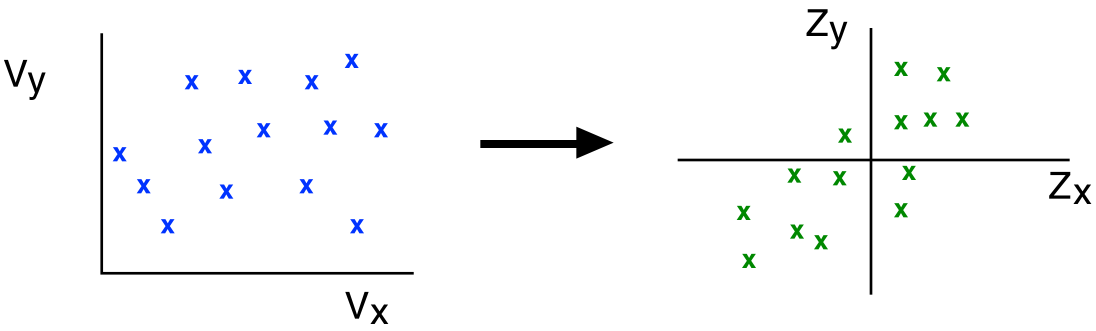
What is it?

Numerical measure of the strength of linear association between 2 quantitative variables.

* Notation: r
* Value of r always between -1 and +1
* Strong association ⇒ large magnitudes of r [i.e., close to ±1]
* Weak association ⇒ small magnitude of r
* Sign tells us whether association is positive or negative

How to calculate r?

Easiest to do this via z-scores: Rescale & standardize both quant. variables.



* Multiply each pair of Z_x, Z_y values and add together

* Divide by the total number of cases minus 1 (i.e., n-1)

Correlation coefficient, r = $\frac{\sum Z_x Z_y}{n-1}$

Warning: This concept only applies to linear associations, with no outliers.
Do not use it if scatterplot looks curved, or has outliers.

Association vs. Correlation vs. Causation

* Each term has a very specific (different) meaning in statistics.

* It is important to understand the difference & to learn to use them correctly.

Association: Means there is a pattern that seems to relate changes in one variable to changes in the other.

E.g., (1) As a student's study time increases, so does the GPA. (2) When we compare internet usage & infant mortality for different countries, higher internet usage is associated with lower infant mortality rates.

Correlation: Is a numerical measure of the strength of linear association between 2 variables.

E.g., (1) The correlation between GPA and study time is 0.68. (2) The correlation between internet usage and infant mortality rate is −0.52.

Causation: Says that changes in one variable causes changes in the other.

E.g., (1) Studying longer causes one's GPA to increase. (2) Using the internet causes infant mortality to decrease.

Association is NOT Causation

* But it is often misinterpreted that way.

Lurking variables: Often hide behind associations, making them look like "causation."

E.g., Internet usage across different countries is associated with infant mortality rates because of lurking variables (poverty levels, income, education).

Linear regression

Objective

(1) Learn how to model association bet. 2 variables using a straight line (called "linear regression").

(2) Learn to assess the quality of regression models.

(3) Understand & avoid common mistakes made in regression modeling.

Concept briefs:

* General straight line equation (Y = m X + b)

* Slope of a straight line (m) = How much Y changes when X increases by 1.

* Intercept of straight line (b) = Value of Y when X=0.

* Linear regression = Process of finding the "best fit" straight line to approximately model linear associations.

* Predicted response = Y-value obtained by plugging in any X into regr. model.

* Residual = Error between predicted Y-value & true Y-value (if known).

* Least-squares best fit = The particular straight line that minimizes the sum of the square of the residuals.

"Best-fit" straight line model

Problem statement

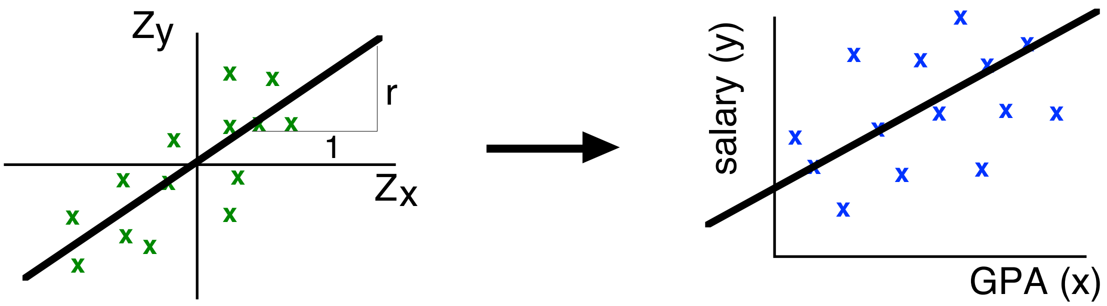
- You have a scatterplot that shows linear association bet. 2 variables. [e.g., Starting salary vs GPA]
- You want "to model" the association with a straight line equation. For example:

$$\boxed{\text{Predicted\_starting\_salary} = m * \text{GPA} + b}$$

where "m" and "b" are to be determined so that we get the best line.

How to find best fit?

- Develop concept using the Z_x - Z_y scatterplot.
- Then rescale & apply it in any other setting



Fact: Best-fit line has

Slope = r, Intercept = 0

(r = correlation coefficient)

Best Fit line: Slope = $r \dfrac{S_y}{S_x}$

Intercept = solve for it using mean values of x, y variables.

S_x & S_y denote the standard deviations

E.g., Predicted_starting_salary = 1000 * GPA + 28,800 ($)

Errors, residuals and R-squared

What is an error or residual?

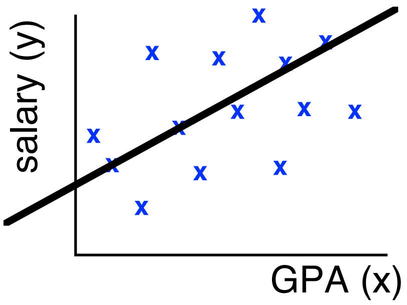
* Recap: The goal is to predict the response (y-value)
for any given x-value.

* We have some known data values that can be used to judge quality.

* Residual = (error) = True y - Predicted y

$$\boxed{e = y - \hat{y}}$$

X	True Y ($)	Pred. Y ($)	Error ($)
2.6	33,400	31,400	2,000
2.7	31,800	31,500	300
2.8	29,200	31,600	-2,400



How to assess quality of regression model?

There are 3 things one should always look at:

(1) Graphical: Scatterplot of e vs x-data value (Or, y-data value).

(2) Numerical: Calculate standard deviation, or variance, in e.

(3) Numerical: Calculate R^2.

* Scatterplot should look random, with no patterns or outliers.

* Variance (or SD) of e should be small relative to variance of y.

What is R-squared?

This is a key numerical measure of quality:

* It is related to both r and to the variance in residuals (e).

* Simplest interpretation: $R^2 = r^2$ x 100 (expressed as %)

E.g., If r = -0.8, then $R^2 = 100 \times (-0.8)^2 = 64\%$

Very important connection to residuals and 'e':

* R^2 tells how well the regression model accounts for the variance in the data. Large R^2 says most of the data's variance is modeled by the regression equation.

* 100 - R^2 is the percent of data variance <u>not</u> accounted for by the model. This is equal to the variance in e.

In what way is our line the "best-fit"?

* This is based on the <u>least-squares</u> principle:

→ This particular line has the minimum sum of the square of the residuals.

→ In other words, it gives the least possible variance in e.

Ex.69 (see last page for copy of exercise)

(a) Use the given R^2 to find r: Since $R^2 = r^2 \times 100$,

$$r = \sqrt{84/100} = \quad \sqrt{0.84} = 0.9165$$

(b) 84 % of the variability in mean temperature is accounted for by the variability in CO_2 levels.

(c) Must learn to interpret typical regression table (comes from software)

* "Intercept" gives the "b" value in the straight line equation.

* Variable CO_2's coefficient is the slope of the straight line.

Thus, the regression equation here would be

Predicted_mean_temp = 11.0276 + 0.0089$*$$CO_2$ (in degrees C)

(d, e) Reword the question to: "Interpret the slope and intercept in this context."

Slope: "The model predicts that for each 1 ppm increase in CO_2 level, the mean temperature will increase by 0.0089 ^oC."

Intercept: [Makes no sense here, but you can still blindly interpret!]

"According to the model, when there is no CO_2 in the atmosphere the mean temperature will be 11.0276 ^oC."
Not very meaningful in this context, since there is always some CO_2 in the atmosphere.

(f) The scatterplot of the residuals looks fairly random, with no specific pattern. Thus it seems appropriate to use a linear model here.

(g) Predicted_mean_temp = 11.0276 + 0.0089*CO_2
= 11.0276 + 0.0089*(400)
= 14.59 ^oC

The model predicts the mean temperature will be 14.59 ^oC when the CO_2 level in the atmosphere is 400 ppm.

Ex.60 (see last page for copy of exercise)

(a) Yes, because the scatterplot shows an association that looks fairly strong and linear.

(b) $R^2 = 87.3\%$ says: 87.3% of variability in the use of other drugs (Y-variable) is accounted for by variability in the use of Marijuana (X-variable).

(c) Regression equation is of the form: $Y = m X + b$.

* First, figure out what our Y and X are: $Y = \widehat{\text{Other drug use}}$ (predicted)

$X = \text{Marijuana use } \%$

* Next, find slope, $m = r \dfrac{S_y}{S_x}$.

The exercise has given us $S_y = 10.2\%$, $S_x = 15.6\%$

We can calculate r from R^2: $r = \sqrt{87.3/100} = \sqrt{0.873} = 0.934$

Thus, slope $m = 0.934 *(10.2 / 15.6) = 0.611$

So our equation is: $\widehat{\text{Other drug use}} = 0.611 \times \text{Marijuana\_use} + b\ (\%)$

* Next, find b by plugging in the mean values and solving for it:

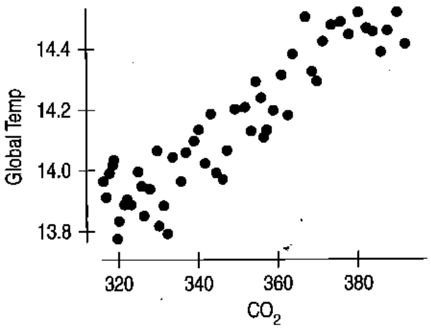
$$11.6\ (\%) = 0.611 \times 23.9\ (\%) + b \longrightarrow b = -3.0\ \%$$

* Final model: $\boxed{\widehat{\text{Other drug use}} = 0.611 \times \text{Marijuana\_use} - 3\ \ (\%)}$

(d) Slope means: "For each 1% increase in teens using Marijuana, the model predicts there is a 0.611% increase in teens using other drugs."

(e) No, this doesn't confirm that Marijuana use leads to the use of other drugs. It only shows there is an association between use of these two drugs.

69. Climate change 2011 The earth's climate is getting warmer. The most common theory attributes the increa to an increase in atmospheric levels of carbon dioxide (CO_2), a greenhouse gas. Here is a scatterplot showin the mean annual CO_2 concentration in the atmosphere measured in parts per million (ppm) at the top of Mau Loa in Hawaii, and the mean annual air temperature o both land and sea across the globe, in degrees Celsius (°C) for the years 1959 to 2011.



A regression predicting *Temperature* from (produces the following output table (in part):

```
Dependent variable is Global Temperature (°C)
R-squared = 84.0%

Variable        Coefficient
Intercept         11.0276
CO2 (ppm)          0.0089
```

a) What is the correlation between CO_2 and *Temperature*?

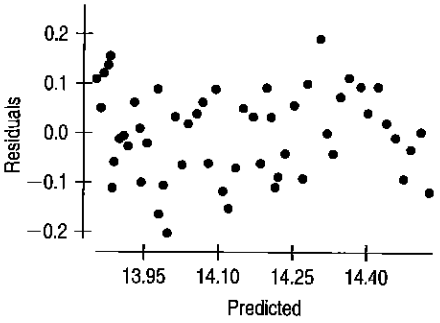
b) Explain the meaning of R-squared in this context.

c) Give the regression equation.

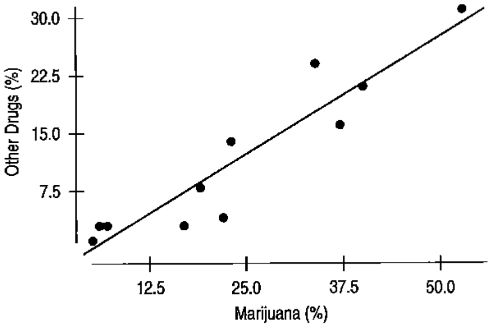
d) What is the meaning of the slope in this equation?

e) What is the meaning of the y-intercept of this equation?

f) Here is a scatterplot of the residuals vs. CO_2. Does this plot show evidence of the violation of any assumptions behind the regression? If so, which ones?



60. Drug abuse In the exercises of the last chapter, you examined results of a survey conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. Below is the scatterplot. Summary statistics showed that the mean percent that had used marijuana was 23.9%, with a standard deviation of 15.6%. An average of 11.6% of teens had used other drugs, with a standard deviation of 10.2%.



a) Do you think a linear model is appropriate? Explain.

b) For this regression, R^2 is 87.3%. Interpret this statistic in this context.

c) Write the equation you would use to estimate the percentage of teens who use other drugs from the percentage who have used marijuana.

d) Explain in context what the slope of this line means.

e) Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs?