Bivariate Data: Graphical Display

The scatterplot is the basic tool of graphically displaying bivariate quantitative data.

Example:

Some investors think that the performance of the stock market in January is a good predictor of its performance for the entire year. To see if this is true, consider the following data on Standard & Poor's 500 stock index (found in SASDATA.SANDP).

Year	Percent January Gain	Percent 12 Month Gain
1985	7.4	26.3
1986	0.2	14.6
1987	13.2	2.0
1988	4.0	12.4
1989	7.1	27.3
1990	-6.9	-6.6
1991	4.2	26.3
1992	-2.0	4.5
1993	0.7	7.1
1994	3.3	-1.5

The plot your instructor is about to show you is a scatterplot of the percent gain in the S&P index over the year (vertical axis) versus the percent gain in January (horizontal axis).

How to analyze a scatterplot

The scatterplot of the S&P data can illustrate the general analysis of scatterplots. You should look for:

- Association. This is a pattern in the scatterplot.

- Type of Association. If there is association, is it:

  - Linear.

  - Nonlinear.

- Direction of Association.

For the S&P data, there is association. This shows up as a general positive relation (Larger % gain in January is generally associated with larger % yearly gain.) It is hard to tell if the association is linear, since the spread of the data is increasing with larger January % gain. This is due primarily to the 1987 datum in the lower right corner of plot, and to some extent the 1994 datum. Eliminate those two points, and the association is strong linear and positive, as the second plot shows.

There is some justification for considering the 1987 datum atypical. That was the year of the October stock market crash. The 1994 datum is a mystery to me.

Data Smoothers

Data smoothers can help identify and simplify patterns in large sets of bivariate data. You have already met one data smoother: the moving average.

Another is the median trace. Here's how it works:

Correlation

Pearson Correlation

Suppose n measurements, $(X_i,\ Y_i),\ i = 1, \ldots, n$ are taken on the variables X and Y. Then the Pearson correlation between X and Y computed from these data is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} X_i' Y_i',$$

where

$$X_i' = \frac{X_i - \overline{X}}{S_X} \text{ and } Y_i' = \frac{Y_i - \overline{Y}}{S_Y}$$

are the standardized data.

The following illustrate what Pearson correlation measures.

Good Things to Know About Pearson Correlation

- Pearson correlation is always between -1 and 1. Values near 1 signify strong positive linear association. Values near -1 signify strong negative linear association. Values near 0 signify weak linear association.

- Correlation between X and Y is the same as the correlation between Y and X.

- Correlation can never by itself adequately summarize a set of bivariate data. Only when used in conjunction with $\overline{X}$, $\overline{Y}$, S_X, and S_Y **and a scatterplot** can an adequate summary be obtained.

- The meaningfulness of a correlation can only be judged with respect to the sample size.

Example:

Back to the S&P data, the SAS macro CORR gives a 95% confidence interval for ρ as (-0.2775, 0.8345).

We can also test

$$\begin{array}{rccc} H_0: & \rho & = & 0 \\ H_{a\pm}: & \rho & \neq & 0 \end{array}$$

by computing

$$t^* = 0.4295\sqrt{\frac{8}{(1-0.4295^2)}} = 1.3452,$$

and comparing this with a t_8 distribution. The resulting values are

$$p^+ = P(t_8 > 1.3452) = 0.1077,$$

and

$$p_- = P(t_8 < 1.3452) = 0.8923,$$

so that

$$p = 2\min(0.1077, 0.8923) = 0.2154.$$

QUESTION: What is ρ? Does this make sense?

Spearman's Correlation

An alternative to Pearson correlation which is resistant to outliers and which can pick up some kinds of nonlinear association is Spearman's correlation. Spearman's correlation is just Pearson correlation computed from the ranks of the data.

Example:

Back to the S&P data, we compute the ranks

Year	Percent January Gain	Percent 12 Month Gain	Rank Percent January Gain	Rank Percent 12 Month Gain
1985	7.4	26.3	9	8.5
1986	0.2	14.6	3	7
1987	13.2	2.0	10	3
1988	4.0	12.4	6	6
1989	7.1	27.3	8	10
1990	-6.9	-6.6	1	1
1991	4.2	26.3	7	8.5
1992	-2.0	4.5	2	4
1993	0.7	7.1	4	5
1994	3.3	-1.5	5	2

The Pearson correlation between the January gain and the 12 month gain is 0.4295. The Spearman correlation, which is just the Pearson correlation between the ranks, is 0.4802. We can test

H_0 : JANGAIN and YEARGAIN are independent.

versus

$H_{a\pm}$: JANGAIN and YEARGAIN are not independent.

by using Table B.8. There we see that if $r_s^* = 0.454$, $p^+ = 0.095$, which implies $p = 2(0.095) = 0.190$. We also see that if $r_s^* = 0.551$, $p^+ = 0.052$, which implies $p = 2(0.052) = 0.104$.

Final note: If the 1987 and 1994 data are omitted, the Pearson and Spearman correlations are 0.9360 and 0.8862, respectively. These are significantly different from 0.

Simple Linear Regression

The SLR model assumes a predictor variable X and a response variable Y are related by

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is a random error term.

We want to fit the model to a set of data (X_i, Y_i), $i = 1, \ldots, n$. As with the measurement model, two options are least absolute errors, which finds values b_0 and b_1 to minimize

$$\text{SAE}(b_0, b_1) = \sum_{i=1}^{n} \mid Y_i - (b_0 + b_1 X_i) \mid,$$

or least squares, which finds values b_0 and b_1 to minimize

$$\text{SSE}(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2.$$

We'll concentrate on least squares. Using calculus, we find the least squares estimators of β_0 and β_1 to be

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

and

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

The relevant SAS/INSIGHT output for the regression of YEARGAIN on JANGAIN looks like this:

And the relevant SAS/INSIGHT output for the regression of YEARGAIN on JANGAIN, with the years 1987 and 1994 removed, looks like this:

Residuals, Predicted and Fitted Values

- The **predicted value** of Y at X is

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

- For $X = X_i$, one of the values in the data set, the predicted value is called a **fitted value** and is written

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$

- The **residuals**, $e_i, i = 1, \ldots, n$ are the differences between the observed and fitted values for each data value:

$$e_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i).$$

Tools to Assess the Quality of the Fit

- Residuals. Residuals should exhibit no patterns when plotted versus the X_i, $\widehat{Y}_i$ or other variables, such as time order. Studentized residuals should be plotted on a normal quantile plot.

- Coefficient of Determination. The coefficient of determination, r^2, is a measure of (take your pick):

  - How much of the variation in the response is "explained" by the predictor.

  - How much of the variation in the response is reduced by knowing the predictor.

The notation r^2 comes from the fact that the coefficient of determination is the square of the Pearson correlation. Check out the quality of the two fits for the S&P data:

Model Interpretation

- The Fitted Slope. The fitted slope may be interpreted as the estimated change in the mean response per unit increase in the predictor.

- The Fitted Intercept. The fitted intercept is the estimate of the response when the predictor equals 0, provided this makes sense.

- The Mean Square Error. The mean square error or MSE, is an estimator of the variance of the error terms ϵ, in the simple linear regression model. Its formula is

$$\text{MSE} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

It measures the "average prediction error" when using the regression.

Classical Inference for the SLR Model

Estimation of Slope and Intercept

Level L confidence intervals for β_0 and β_1 are

$$(\hat{\beta}_0 - \hat{\sigma}(\hat{\beta}_0)t_{n-2,\frac{1+L}{2}}, \hat{\beta}_0 + \hat{\sigma}(\hat{\beta}_0)t_{n-2,\frac{1+L}{2}}),$$

and

$$(\hat{\beta}_1 - \hat{\sigma}(\hat{\beta}_1)t_{n-2,\frac{1+L}{2}}, \hat{\beta}_1 + \hat{\sigma}(\hat{\beta}_1)t_{n-2,\frac{1+L}{2}}),$$

respectively, where

$$\hat{\sigma}(\hat{\beta}_0) = \sqrt{\text{MSE} \Big/ \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]},$$

and

$$\hat{\sigma}(\hat{\beta}_1) = \sqrt{\text{MSE} \Big/ \sum_{i=1}^n (X_i - \overline{X})^2}$$

Estimation of The Mean Response

The mean response at $X = x_0$ is

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The point estimator of μ_0 is

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

A level L confidence interval for μ_0 is

$$(\widehat{Y}_0 - \widehat{\sigma}(\widehat{Y}_0) t_{n-2,\frac{1+L}{2}}, \widehat{Y}_0 + \widehat{\sigma}(\widehat{Y}_0) t_{n-2,\frac{1+L}{2}}),$$

where

$$\widehat{\sigma}(\widehat{Y}_0) = \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]}.$$

Prediction of a Future Observation

A level L prediction interval for a future observation at $X = x_0$ is

$$(\hat{Y}_{new} - \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-2,\frac{1+L}{2}},$$

$$\hat{Y}_{new} + \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-2,\frac{1+L}{2}})),$$

where

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_0,$$

and

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = \sqrt{\text{MSE}\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]}.$$

The macro REGPRED will compute confidence intervals for a mean response and prediction intervals for future observations for each data value and for other user-chosen X values.

The Relation Between Correlation and Regression

If the standardized responses and predictors are

$$Y_i' = \frac{Y_i - \overline{Y}}{S_Y},$$

and

$$X_i' = \frac{X_i - \overline{X}}{S_X},$$

Then the regression equation fitted by least squares can be written as

$$\widehat{Y}' = r \cdot X',$$

Where X' is any value of a predictor variable standardized as described above.

The Regression Effect refers to the phenomenon of the standardized predicted value being closer to 0 than the standardized predictor. Equivalently, the unstandardized predicted value is fewer Y standard deviations from the response mean than the predictor value is in X standard deviations from the predictor mean.

For the S&P data $r = 0.4295$, so for a January gain X' standard deviations (S_X) from $\overline{X}$, the regression equation estimates a gain for the year of

$$\widehat{Y}' = 0.4295 \cdot X'$$

standard deviations (S_Y) from $\overline{Y}$.

With 1987 and 1994 removed, the estimate is

$$\widehat{Y}' = 0.9360 \cdot X',$$

which reflects the stronger relation.

The Relationship Between Two Categorical Variables

Analysis of categorical data is based on counts, proportions or percentages of data that fall into the various categories defined by the variables.

Some tools used to analyze bivariate categorical data are:

- Mosaic Plots.

- Two-Way Tables.

Example:

A survey on academic dishonesty was conducted among WPI students in 1993 and again in 1996. One question asked students to respond to the statement "Under some circumstances academic dishonesty is justified." Possible responses were "Strongly agree", "Agree", "Disagree" and "Strongly disagree". Here are the results:

Association is NOT Causation

Two variables may be associated due to a number of reasons, such as:

1. X **could cause** Y**.**

2. Y **could cause** X**.**

3. X **and** Y **could cause each other.**

4. X **and** Y **could be caused by a third (lurking) variable Z.**

5. X **and** Y **could be related by chance.**

6. **Bad (or good) luck.**

The Issue of Stability

- When assessing the stability of a process in terms of bivariate measurements X and Y, always consider the evolution of the relationship between X and Y, as well as the individual distribution of the X and Y values, over time or order.

- Suppose we have a model relating a measurement from a process to time or order. If, as more data are taken the pattern relating the measurement to time or order remains the same, we say that the process is stable **relative to the model**.