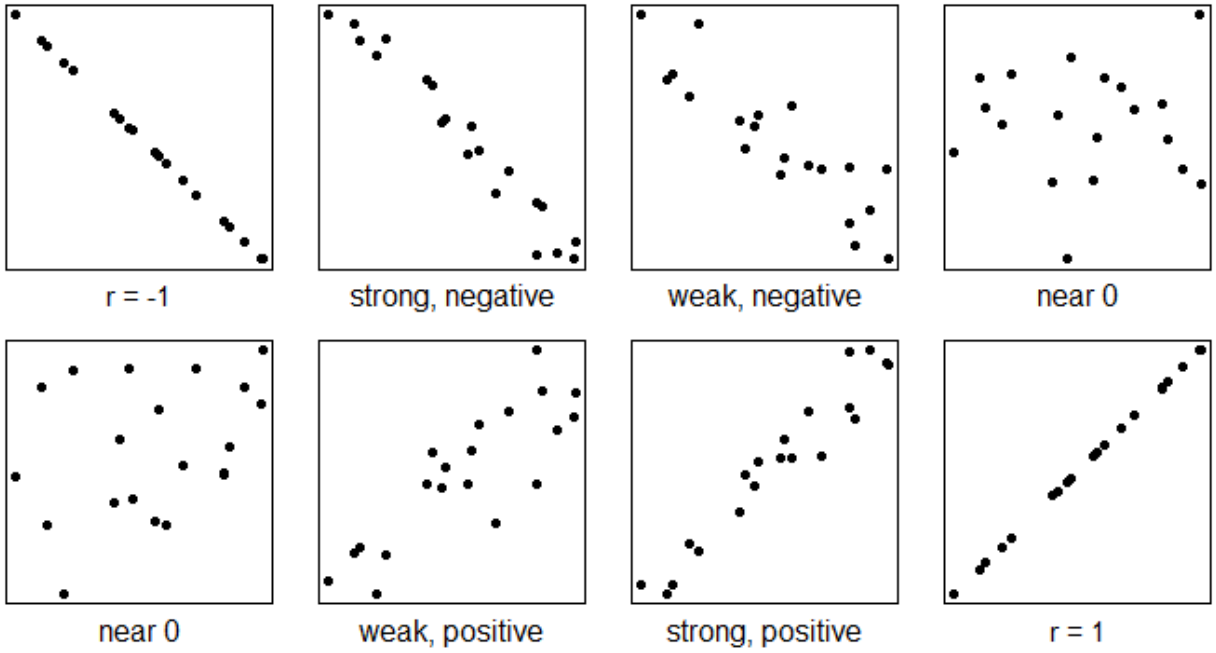
Simple Linear Regression
Part 8: Correlation Analysis

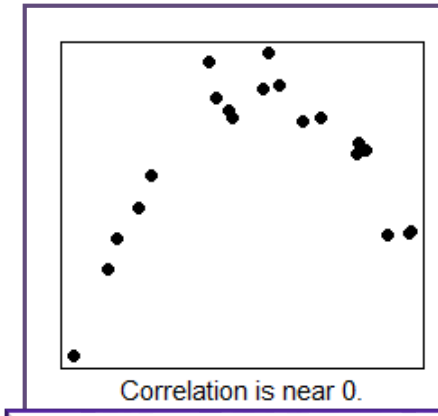
STAT 705: Regression and Analysis of Variance

Correlation Coefficient

- Denoted by r
- Official name: "Pearson's product moment coefficient of correlation"
- Measures the strength of the linear association between X and Y
- r is ALWAYS between -1 and 1
  - Closer to 1 or -1 ⇨ strong linear association
  - Close to 0 ⇨ weak (or no) linear association

Visualize Correlation





Correlation measures the LINEAR association between X and Y.

Top row:

Negative correlation, from perfect (r = -1) on the left, to nonexistent ($r \approx 0$) on the right.

Bottom row:

Positive correlation, from nonexistent ($r \approx 0$) on the left to perfect (r =1) on the right.

Correlation and Related Measures

- Three ways to measure linear relationship between X and Y
  (1) coefficient of determination, (2) correlation, and
  (3) the estimated slope from simple linear regression
- These are all related

Coefficient of determination (R^2)	Correlation coefficient (r)
$R^2 = \frac{SSTot - SSE}{SSTot} = \frac{SSReg}{SSTot}$	$r = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}} = \hat{\beta}_1 \sqrt{\frac{SS_{XX}}{SS_{YY}}}$

- Recall: R^2 measures the proportion of the variability in Y that is explained by the regression on X
- For simple linear regression, $R^2 = r^2$

Hypothesis Test for Correlation

- Sample correlation (r) is a point estimate for the population correlation (ρ) [this is "rho"]
- We want to test if the population correlation is 0:

  $H_0: \rho = 0$ vs. $H_a: \rho \neq 0$

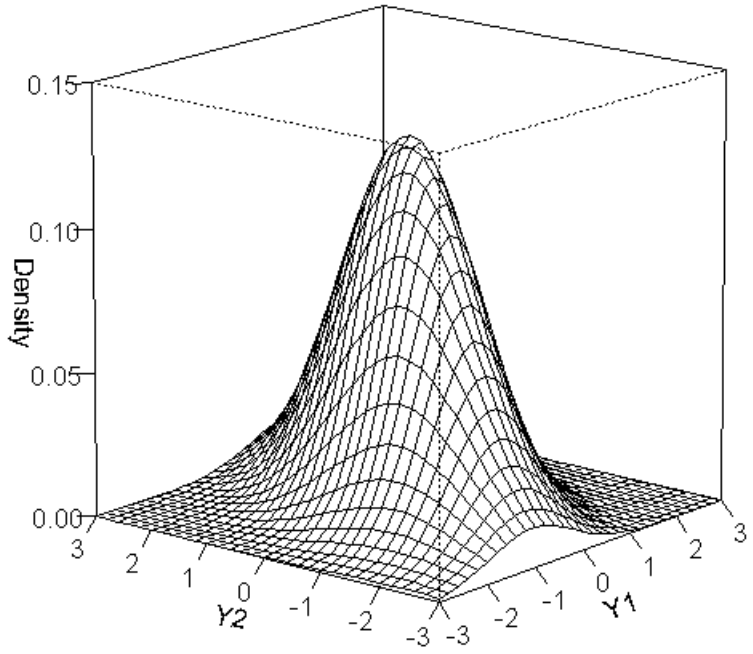
- In words:

  H_0: In the population, there is <u>not</u> a linear association between X and Y.

  H_a: In the population, there <u>is</u> a linear association between X and Y.

Assumption for the Correlation Test

- Both X and Y are random variables
- Jointly, (X, Y) follows a bivariate normal distribution



- Density curve becomes a density surface
- Make any slice parallel to an axis, and the result is a normal curve

Test for Correlation

- Null hypotheses: $H_0: \rho = 0$
- Alternative can be left-tailed, right-tailed or two-tailed
- Test statistic: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- Reference distribution: t, with df = $n - 2$
- Rejection criteria
  - Left-tailed test: Reject H_0 if $t < -t_\alpha$
  - Right-tailed test: Reject H_0 if $t > t_\alpha$
  - Two-tailed test: Reject H_0 if $|t| > t_{\alpha/2}$

Correlation Test: Lead vs. Traffic

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} = \frac{22,996.23}{\sqrt{643.56 \times 892,522.67}} = 0.9595$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9595\sqrt{9}}{\sqrt{1-0.9595^2}} = 10.218$$

$$\text{Critical value} = t_{\alpha/2, df=9} = 2.262$$

- Since 10.218 > 2.262, we reject H_0. The sample provides convincing evidence that there is a linear relationship between X and Y.
- Limitation: This test can be used only when the test $\rho = 0$
- To test against values other than 0, use Fisher's Z (later)

SAS Code: Correlation

(Append this to previous SAS code.)

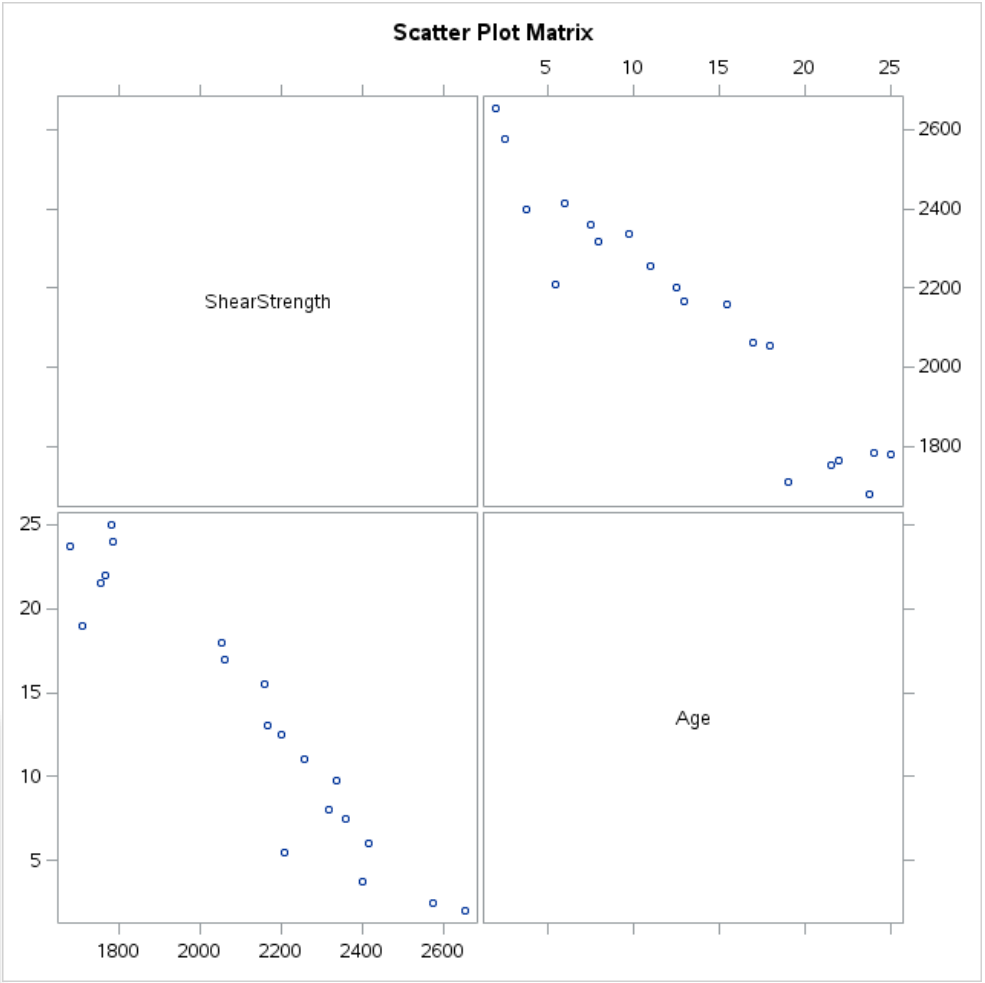
```
ods html;              /* Saves output in a html file */
ods graphics on;    /* In order to generate graphs */
proc corr data=nasa plots=matrix;
  var ShearStrength Age    ;
run;
ods graphics off;
ods html close;
```

SAS Output from PROC CORR

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	ShearStrength	**Age**
ShearStrength	1.00000 20	-0.94965 <.0001 20
Age	-0.94965 <.0001 20	1.00000 21

Three entries in each cell:

- Sample correlation
- p-value for testing $\rho = 0$
- Number of observations



Fisher's Z: Another test for ρ

- Earlier test for correlation is not flexible; restricted to H_0: $\rho = 0$
- Suppose we want to test H_0: $\rho = \rho_0$ vs. H_a: $\rho \neq \rho_0$
  - ρ_0 is some constant between -1 and 1
  - Let r denote the sample correlation
  - We transform both r and ρ_0, then use a regular two-sided Z test
- Calculate $r' = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$ and $\rho_0' = \frac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right)$
- The test statistic $Z = \sqrt{n-3}\left(r' - \rho_0'\right)$ is approx. N(0, 1)
- Reject H_0 if $|Z| > z_{\alpha/2}$

Example: Fisher's Z test

The body weights of 100 fathers and their first-born sons are measured, resulting in a sample correlation r of 0.38. Is this compatible with an underlying correlation of 0.5 that might be expected under genetic theory (i.e. Mendelian sampling)?

$$r' = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\log_e\left(\frac{1+.38}{1-.38}\right) = 0.40$$

$$\rho_0' = \frac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = \frac{1}{2}\log_e\left(\frac{1+0.5}{1-0.5}\right) = 0.55$$

$$Z = \sqrt{100-3}\left(0.40-0.549\right) = -1.47$$

- Hypotheses:
  H_0: $\rho = 0.5$ vs. H_a: $\rho \neq 0.5$
- α=.05 critical value:
  $z_{.025} = 1.96$
- |-1.47| is not > 1.96
- Do not reject H_0

Conclusion: The sample does not provide enough evidence to suspect the body weights of fathers and first-born sons do not follow genetic theory

Correlation: Final Comments

- **CORRELATION IS NOT CAUSATION**.

  Example: The per capita consumption of soft drinks in the U.S. has increased steadily since 1950. The rate of obesity in the U.S. has also increased steadily since 1950. These two variables are highly correlated.

  - Does obesity 'cause' soft drink consumption?
  - Does soft drink consumption 'cause' obesity?
  - To answer either of these questions, we need a randomized controlled EXPERIMENT, not an OBSERVATIONAL STUDY.

- If the bivariate normal assumption is not appropriate, we can use nonparametric rank correlation (Spearman's)

Things You Should Know

- Interpret the sample correlation
- Know the relationship between sample correlation, coefficient of determination, and the estimated slope from simple linear regression
- Write and interpret SAS code to perform correlation analysis
- Perform Fisher's Z test 'by hand'