Overview: Correlation & Regression

- Pearson correlation coefficient
- Simple Linear Regression.
  - What and why?
  - How (interpretation, estimation & diagnostics).
  - Statistical Inference.
  - Comments regarding interpretation.
- Bi-variate regression
- Multiple regression
- General Linear Model
Outline: Pearson Correlation Coefficient

- Definition & Properties.

- Statistical Inference
  - $t$-test that correlation equals 0.
  - Fisher’s $Z$-Transformation.
  - Confidence intervals for $\rho$.
  - Test of $H_o : \rho = K$.
  - Test of $H_o : \rho_1 = \rho_2$ (2 independent populations).
Correlation: Definition & Properties

- “Pearson Product Moment Correlation”

- Two numerical variables measured on same individual,

\[(X_i, Y_i) \text{ for } i = 1, \ldots, n. \text{ e.g.,}\]

- Height and weight.
- Math and science scores.
- Salary and merit.
- High school GPA and college GPA.
- Cost of wine and annual rainfall.
- Conservative Party donors and people who buy garden bulbs by mail.
Scatter Diagram & Summary Statistics

\[ N = 1000, \ r = 0.72 \]

\[ x_{\text{bar}} = 10 \text{ and } y_{\text{bar}} = 0 \]

\[ s_x = 1.0 \text{ and } s_y = 2.75 \]
Definition: Correlation coefficient

- $\rho$ (Greek “rho”) = population correlation.
- $r$ = sample correlation.

Formal definition

$$r = \frac{\text{cov}(X, Y)}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

- It measures the extent to which two random variables are linearly related.
How $r$ Works

\[ r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \cdot \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n} \]

Formula and Scatter Plot

$\begin{align*}
    r &= .80 \\
    r &= -.80
\end{align*}$
Examples of Different $r$'s

Postive Correlations

$r = 0$

$r = 0.60$

$r = 0.90$

$r = 0.80$

$r = 0.55$

$r = 0.90$
Examples of Different $r$'s

Negative Correlations

- $r = 0$
- $r = -0.6$
- $r = -0.2$
- $r = -0.8$
- $r = -0.9$
- $r = -0.95$
- $r = -0.96$
- $r = -0.99$

Fisher's $Z$-Transformation
Non-Linear Relationships

\[ r = 0 \]
Properties: Correlation Coefficient

-1 ≤ r ≤ +1
- 1 ≤ r < 0 → small values of X go with large values of Y and large values of X go with small values of Y.
- 0 < r ≤ +1 → large values of X go with large values of Y and small values of X go with small values of Y.
- r = 0 → No linear relationship.

r measures the strength of the relationship (magnitude) between two variables and the direction of the relationship (sign).
Properties: Correlation Coefficient

- $r$ measures linear relationship.
- Linear transformations of $X$ and/or $Y$ do not change the size (magnitude) of $r$. Linear transformations do not change the direction (sign) as long as

$$X^* = aX + b$$

where $a > 0$ (e.g., $z$ scores).
- In a scatter plot, a linear transformation(s) (where $a > 0$) simply corresponds to relabelling axis (axes).
Inference & the Correlation Coefficient

- **Preliminaries:** bivariate normal distribution.
- This is a generalization of the normal distribution for two random variables (say $X$ and $Y$).
- The parameters of the bivariate normal distribution are: $\mu_x, \sigma^2_x, \mu_y, \sigma^2_y$, and $\rho_{xy}$
- It looks like a bell or a little hill.
- MatLab program.
The Bivariate Normal Distribution

- If $X$ and $Y$ have a bivariate normal distribution, then
  - $X \sim \mathcal{N}(\mu_x, \sigma^2_x)$
  - $Y \sim \mathcal{N}(\mu_y, \sigma^2_y)$
  - $\rho_{xy}$ measures how related $X$ and $Y$ are.

- If $X$ and $Y$ are bivariate normal and $\rho_{xy} = 0$, then $X$ and $Y$ are statistically independent.

- If $X$ and $Y$ are statistically independent, then $\rho_{xy} = 0$.

- The case where $\rho_{xy} = 0$ and the (joint) distribution of $X$ and $Y$ is not bivariate normal does not imply that $X$ and $Y$ are statistically independent.
Example: $\rho = 0$ & dependent

$r = 0$

Marginal distributions of $X$ and $Y$ are not normal:
Hypothesis Testing

- **Statistical Hypotheses**: The most common case,  
  \[ H_0 : \rho = 0 \quad \text{versus} \quad H_a : \rho \neq 0 \]

- **Assumptions**:  
  - \( X \) and \( Y \) are random variables whose joint distribution is bivariate normal.*** qualification.  
  - Observations are independent.
Hypothesis Testing

- **Test Statistic:** Given the assumptions above and $H_0 : \rho = 0$, 
  \[ t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \]

- **Sampling Distribution** of the test statistic is Student’s $t$ with 
  \( \nu = n - 2 \).

- **Note:** the test statistic depends on both $r$ and the sample size $n$. So for a given $\alpha$-level, you do not have to compute the test statistic... just find the “critical” value for $r$. 
Example Hypothesis Testing for $\rho$

- High School & Beyond: Reading scores and Motivation
- $H_0 : \rho_{\text{read,mot}} = 0$ vs $H_a : \rho_{\text{read,mot}} \neq 0$.
- Test statistic
  \[ t = \frac{.21061}{\sqrt{\frac{(1-.21061^2)}{600-2}}} = \frac{.21061}{\sqrt{.9556/598}} = 5.269 \]

- For $\nu = 600 - 2 = 598$, $p$ value $= P(|t| \geq 5.269) < .001$; therefore, Reject $H_0$.
- Conclusion: The data provide evidence that there is a linear relationship between reading and motivation.
Alternative Method

- Find the critical $r$ and compare to the observed $r$.
- Will reject $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$ whenever
  
  
  \[
  \text{observed}\ t_{n-2} \leq 0.025\ t_{n-2}\ \text{or}\ \text{observed}\ t_{n-2} \geq 0.975\ t_{n-2}
  \]

- Take
  
  \[
  t = \frac{r}{\sqrt{1 - r^2}} = r \frac{\sqrt{n - 2}}{\sqrt{1 - r^2}}
  \]

  and $r$ as a function of $t$. 
Alternative Method

\[ t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \]

- Square both sides and solve for \( r \):

\[ t^2 = \frac{r^2(n-2)}{1-r^2} \]

\[ \frac{t^2(1-r^2)}{(n-2)} = r^2 \]

\[ \frac{t^2}{(n-2)} = r^2 \left( 1 + \frac{t^2}{n-2} \right) \]

\[ r^2 = \frac{t^2}{(n-2)(1 + \frac{t^2}{n-2})} \]
Alternative Method

- So
  \[ r_{crit} = \frac{t_{crit}}{\sqrt{(n - 2)(1 + t_{crit}^2/(n - 2))}} \]

- For our HSB example:
  \[ r_{crit} = \frac{1.9639}{\sqrt{598 \left(1 + \frac{(1.9639)^2}{598}\right)}} = \frac{1.9639}{\sqrt{601.85}} = .08 \]

- Any correlation > .08 (or < -.08) would be “significant” for \( n = 600 \).

- Note: “Statistical significance” does not imply “importance”.

Note: “Statistical significance” does not imply “importance”.
More correlations from the HSB data where \( N = 600 \), \( \alpha = .05 \), and \( r_{crit} = .0877 \):

<table>
<thead>
<tr>
<th></th>
<th>Locus of control</th>
<th>Self concept</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading ((p\text{-value}))</td>
<td>.38 ((&lt; .01))</td>
<td>.06 (.14)</td>
<td>.21 ((&lt; .01))</td>
</tr>
<tr>
<td>Science ((p\text{-value}))</td>
<td>.32 ((&lt; .01))</td>
<td>.07 (.09)</td>
<td>.12 ((&lt; .01))</td>
</tr>
</tbody>
</table>

Note \( p\text{-value} = \text{Prob}(|r| \geq r \text{ given } \rho = 0) \), i.e., \( H_0 : \rho = 0 \).
Computing Correlations: SAS

- SAS program command:
  
  ```sas
  PROC CORR;
  VAR rdg sci;
  With locus concpt mot;
  
  or
  PROC CORR;
  VAR rdg sci locus concpt mot;
  
  ASSIST
  ANALYST
  Interactive data analysis
  ```
Fisher’s $Z$-Transformation

- **Why?** When $\rho \neq 0$, the sampling distribution for $r$ is skewed.

- Fisher’s $Z$-Transformation is a function of $r$ whose sampling distribution of the transformed value is close to normal.

- Can compute confidence intervals and a variety of tests using Fisher’s $Z$.

- **Requirement:** For the distribution of $Z$ to be approximately normal,
  - Variables from a bivariate normal distribution.
  - Sample size should be $n \geq 10$ (and larger if question the bivariate normal assumption).
Fisher’s \( Z \)-Transformation:

\[
Z = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right),
\]

where

- \( r \) is the sample correlation
- \( Z \) is the transformed value of \( r \)
- \( \ln \) is the natural logarithm.

Note: the natural logarithm has base equal to \( \exp = e = 2.718281828 \); that is,

\[
\text{if} \quad \exp^a = x \quad \text{then} \quad \ln(x) = a
\]
Fisher’s $Z$-Transformation

- Taking the logarithm of numbers has the effect of “compressing” the differences or space between the larger values and “stretching” the space between smaller values.

- If a distribution is positively skewed, the taking the logarithm has the effect of making the distribution more symmetric.

- How Fisher’s $Z$-Transformation works...
Simulated Sampling Distributions
rho = .95, n = 200 per sample (lots of replications)
Simmulated Sampling Distributions

$\rho = .95$, $n = 10$ per sample (lots of replications)
Sampling Distribution of Fisher’s $Z$

- **IF Observations**
  - Are from a bivariate normal distribution.
  - Are independent across individuals.
  - $n \geq 10$

- **THEN** the sampling distribution of $Z$ is $\approx \mathcal{N}(\mu_Z, \sigma^2_Z)$ where

$$E(Z) = \mu_Z = Z_\rho = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right) + \frac{\rho}{2(n - 1)}$$

$$\sigma^2_Z = \frac{1}{n - 3}$$

- The value $\frac{\rho}{2(n - 1)}$ is the bias factor, which in SAS you can request that a bias adjustment be used (in confidence intervals).

- $\mu_Z$ and $\sigma^2_Z$ are independent of each other.

- The transformation of $r$ is known as the “inverse of the hyperbolic tangent of $r$”.
Using Fisher’s $Z$

- HSB data: Are there relationships between psychological variables and achievement: motivation and reading?

- Observed correlation, $r = .21061$.

- If the true population correlation coefficient is $\rho > 0$, then the sampling distribution of $r$ will be skewed.

- Use Fisher’s $Z$ transformation,

$$Z = \frac{1}{2} \ln \left( \frac{1 + .21061}{1 - .21061} \right) = \frac{1}{2} \ln(1.53360) = \frac{1}{2}(0.4276) = .2138$$

- The standard deviation,

$$\sigma_Z = \frac{1}{\sqrt{600 - 3}} = .04093$$
Example Using Fisher’s $Z$

- Suppose want to test $H_0 : \rho = .25$ vs $H_a : \rho \neq .25$.

- Need the value of $Z$ for $\rho = .25$,

$$Z_{.25} = \frac{1}{2} \ln \left( \frac{1 + .25}{1 - .25} = .2554 \right)$$

- Test statistic is

$$z = \frac{Z_{obs} - Z_{null}}{\sigma_Z} = \frac{.2138 - .2554}{.04093} = -1.016$$

- Retain $H_0$.

- Note: a lower case $z$ is used for the test statistic and upper case $Z$ is denotes Fisher’s Z-transformed value of $r$.
Confidence Interval for $\rho$

- Another use for Fisher’s $Z$-transformation.

- Suppose we want a 95% CI for correlation between motivation and reading scores.

- Steps:
  1. Transform the sample correlation: $Z_{\text{obs}} = .2138$.
  2. Compute the $(1 - \alpha)\%$ CI for $Z_{\rho}$

\[
Z_{\text{obs}} \pm \frac{z_{\alpha/2} \sigma_Z}{2}
\]

\[
.2138 \pm 1.96(.04093) \implies (.13, .29)
\]

3. Un-transform the end points of the CI above.
Confidence Interval for $\rho$

- Reversing the Fisher $Z$ transformation...a little algebra gives

\[ r = \frac{e^{2Z} - 1}{e^{2Z} + 1} \]

- Our example

\[
\begin{align*}
r_{\text{lower}} &= \frac{e^{2(.1336)} - 1}{e^{2(.1336)} + 1} = \frac{2.71828^{.2672} - 1}{2.71828^{.2672} + 1} = \frac{.3063}{2.3063} = .1328 \\
r_{\text{upper}} &= \frac{e^{2(.2940)} - 1}{e^{2(.2940)} + 1} = \frac{2.71828^{.5880} - 1}{2.71828^{.5880} + 1} = \frac{.8004}{2.8004} = .2858
\end{align*}
\]

- The 95% confidence interval for $\rho$ between motivation and reading scores is (.13, .29).
Fisher’s Z in SAS

```sas
TITLE ’Testing Ho: rho=0 using Fisher-Z transformation’;
proc corr data=hsb fisher;
    var mot rdg;
RUN;

TITLE ’Ho: rho= .25 , No bias adjustment’;
proc corr data=hsb fisher(rho0=.25 biasadj=no alpha=.05);
    var mot rdg;
RUN;

TITLE ’Ho: rho= .25 , With bias adjustment’;
proc corr data=hsb fisher(rho0=.25 biasadj=yes alpha=.05);
    var mot rdg;
RUN;
```
Two Independent Group Test

- Test whether the correlation from 2 independent groups are the same or different.

- The same procedure that we used for testing difference between mean for large samples.

- Statistical hypotheses:
  \[ H_0 : \rho_1 = \rho_2 \quad \text{vs} \quad H_a : \rho_1 \neq \rho_2 \]

- Assumptions:
  - Observations are independent within and between populations
  - The joint distribution of the two variables in each population is bivariate normal.
Two Independent Group Test

Test Statistic:

\[ z = \frac{Z_1 - Z_2}{\sigma_{Z_1 - Z_2}} \]

where

- \( Z_1 \) and \( Z_2 \) are Fisher \( Z \)-transformations of the sample correlations, \( r_1 \) and \( r_2 \), from the two groups.

- Standard deviation,

\[ \sigma_{Z_1 - Z_2} = \sqrt{\sigma^2_{Z_1} + \sigma^2_{Z_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \]

Why?

- Sampling distribution of the test statistic is \( \mathcal{N}(0, 1) \).
Is the relationship between writing scores and locus of control the same or different for male and female high school students?

The data: \( n_{male} = 327 \) and \( r_{male} = .40196 \)
\( n_{female} = 273 \) and \( r_{female} = .28250 \)

Statistical hypotheses:
\[ H_0 : \rho_{male} = \rho_{female} \text{ vs } H_a : \rho_{male} \neq \rho_{female} \]

Assumptions:
- Scores come from bivariate normal populations.
- Independence within and between groups.

...so what's dependent?
Example Two Independent Group Test

Test Statistic:

\[
Z_{female} = \frac{1}{2} \ln \left( \frac{1 + .28250}{1 - .28250} \right) = \frac{1}{2}(.5808) = .29040
\]

\[
Z_{male} = \frac{1}{2} \ln \left( \frac{1 + .40196}{1 - .40196} \right) = \frac{1}{2}(.85197) = .425980
\]

\[
\sigma(Z_m - Z_f) = \sqrt{\frac{1}{n_{male} - 3} + \frac{1}{n_{female} - 3}} = \sqrt{\frac{1}{324} + \frac{1}{270}} = .08240
\]

\[
z = \frac{.42598 - .29040}{.08240} = 1.645
\]

Conclusion: Retain \( H_o \) for \( \alpha = .05 \). The difference between the correlations more likely to be due to chance than reflect real a difference.