Correlation & Regression

Examine Relation Between Variables

- **One quantitative and one qualitative** variables: Two-sample, paired sample, One-Way ANOVA, ...
- **Two qualitative variables**: Contingency tables, Chi-square test, McNemar’s test, ...

Examine Relation Between Two Quantitative Variables

Is there relation between “number of handguns registered” in the area and “number of people killed”?

<table>
<thead>
<tr>
<th>Year</th>
<th>NGR(x)</th>
<th>Nkill(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>447</td>
<td>13</td>
</tr>
<tr>
<td>78</td>
<td>460</td>
<td>21</td>
</tr>
<tr>
<td>79</td>
<td>481</td>
<td>24</td>
</tr>
<tr>
<td>80</td>
<td>498</td>
<td>16</td>
</tr>
<tr>
<td>81</td>
<td>513</td>
<td>24</td>
</tr>
<tr>
<td>82</td>
<td>512</td>
<td>20</td>
</tr>
<tr>
<td>83</td>
<td>526</td>
<td>15</td>
</tr>
<tr>
<td>84</td>
<td>559</td>
<td>34</td>
</tr>
<tr>
<td>85</td>
<td>585</td>
<td>33</td>
</tr>
<tr>
<td>86</td>
<td>614</td>
<td>33</td>
</tr>
<tr>
<td>87</td>
<td>648</td>
<td>39</td>
</tr>
<tr>
<td>88</td>
<td>675</td>
<td>43</td>
</tr>
<tr>
<td>89</td>
<td>711</td>
<td>50</td>
</tr>
<tr>
<td>90</td>
<td>719</td>
<td>47</td>
</tr>
</tbody>
</table>

Number of Handguns Registered

Number of People Killed

(447, 13)

(460, 21)

Scatter Plot

Pearson’s Sample Correlation

\[
 r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

\[
 = \frac{1}{n-1} \sum_{i=1}^{n} (z_{x_i} \cdot z_{y_i})
\]

\[ s_x: \text{Sample standard deviation of } x's \]

\[ s_y: \text{Sample standard deviation of } y's \]

\[ z_{x_i}: \text{Sample standard score of } x_i \]

\[ z_{y_i}: \text{Sample standard score of } y_i \]
Correlation & Regression

\[ \bar{x} = 567.50, \quad \bar{y} = 29.43 \]

\[ s_x = 91.91, \quad s_y = 12.19 \]

\[
\frac{447 - 567.5}{91.91} = -1.31 \\
\frac{13 - 29.43}{12.19} = -1.35
\]

\[ z_{x1} \cdot z_{y1} = (-1.31) \cdot (-1.35) \]

**Scatter Plot**

\[ \bar{x} = 567.5 \]

\[ \bar{y} = 29.43 \]

**Shortcut Formula**

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

\[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \]

\[ s_{xy} = \sum_{i=1}^{n} x_i y_i - \left( \frac{\sum x_i}{n} \right) \left( \frac{\sum y_i}{n} \right) \]

\[ s_{xx} = \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum x_i}{n} \right)^2 \]

\[ s_{yy} = \sum_{i=1}^{n} y_i^2 - \left( \frac{\sum y_i}{n} \right)^2 \]

**Interpretation of \( r \)**

- \(-1 \leq r \leq 1\)
- It measures the strength and direction of the linear relation between two quantitative variables.
- \( r = 1 \) if all points lie exactly on a straight line.
- \( \rho \) is the notation for population correlation coefficient.

**Main Table**

<table>
<thead>
<tr>
<th>Year</th>
<th>NGR (y)</th>
<th>Nkill (x)</th>
<th>( x - \bar{x} )</th>
<th>( y - \bar{y} )</th>
<th>( \frac{(x - \bar{x})(y - \bar{y})}{s_x s_y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>447</td>
<td>13</td>
<td>-1.31</td>
<td>-1.35</td>
<td>1.77</td>
</tr>
<tr>
<td>78</td>
<td>460</td>
<td>21</td>
<td>-1.17</td>
<td>-0.69</td>
<td>0.81</td>
</tr>
<tr>
<td>79</td>
<td>481</td>
<td>24</td>
<td>-0.94</td>
<td>-0.45</td>
<td>0.42</td>
</tr>
<tr>
<td>80</td>
<td>498</td>
<td>16</td>
<td>-0.76</td>
<td>-1.10</td>
<td>0.83</td>
</tr>
<tr>
<td>81</td>
<td>513</td>
<td>24</td>
<td>-0.59</td>
<td>-0.45</td>
<td>0.26</td>
</tr>
<tr>
<td>82</td>
<td>512</td>
<td>20</td>
<td>-0.66</td>
<td>-0.77</td>
<td>0.47</td>
</tr>
<tr>
<td>83</td>
<td>526</td>
<td>15</td>
<td>-0.45</td>
<td>-1.18</td>
<td>0.50</td>
</tr>
<tr>
<td>84</td>
<td>559</td>
<td>34</td>
<td>-0.09</td>
<td>0.38</td>
<td>-0.03</td>
</tr>
<tr>
<td>85</td>
<td>585</td>
<td>33</td>
<td>0.19</td>
<td>0.29</td>
<td>0.06</td>
</tr>
<tr>
<td>86</td>
<td>614</td>
<td>33</td>
<td>0.51</td>
<td>0.29</td>
<td>0.15</td>
</tr>
<tr>
<td>87</td>
<td>645</td>
<td>39</td>
<td>0.84</td>
<td>0.75</td>
<td>0.46</td>
</tr>
<tr>
<td>88</td>
<td>657</td>
<td>43</td>
<td>1.17</td>
<td>1.11</td>
<td>1.30</td>
</tr>
<tr>
<td>89</td>
<td>711</td>
<td>50</td>
<td>1.56</td>
<td>1.69</td>
<td>2.64</td>
</tr>
<tr>
<td>90</td>
<td>719</td>
<td>47</td>
<td>1.85</td>
<td>1.64</td>
<td>2.38</td>
</tr>
</tbody>
</table>

**Total**

\[ r = 0.941477289 \]

**Mean**

\[ r = 0.567.50, \quad r^2 = 29.43 \]
Correlation & Regression

**Correlation**
- Stronger linear trend

**Correlation Coefficient**
- \( r \) close to 1
- \( r \) close to 0
- \( r \) close to \(-1\)

**Births per 1000 population, 1992**
- 60
- 50
- 40
- 30
- 20
- 10

**Female life expectancy 1992**
- 90
- 80
- 70
- 60
- 50

**GDP per capita**
- 20000
- 10000
- 0
- (-10000)

**Deaths per 1000 people, 1992**
- 18
- 16
- 14
- 12
- 10
- 8
- 6
- 4

**Doctors per 10,000 people**
- 40
- 30
- 20
- 10
- 0
- (-10)

\[ r_1 = -0.968 \quad r_2 = 0.982 \quad r_3 = -0.165 \quad r_4 = 0.8 \]

**t-test for correlation**

**Hypothesis:** \( H_0: \rho = \rho_0 \), v.s. \( H_a: \rho \neq \rho_0 \)

**Test Statistic:**
\[ t = \frac{r - \rho_0}{\sqrt{\frac{1 - r^2}{n - 2}}} \sim t\text{-distribution d.f.} = n - 2 \]

**Decision rule:**
- C.V. approach: \( t < -t_{\alpha/2} \) or \( t > t_{\alpha/2} \)
- \( p \)-value approach: \( p \)-value < \( \alpha \)

**Is there a significant correlation?**

Example: (Handgun)
\[ H_0: \rho = 0, \quad \text{v.s.} \quad H_a: \rho \neq 0 \]
\[ t = .941 \cdot \frac{14 - 2}{\sqrt{1 - .941^2}} = 9.65 \]

\( \text{d.f.} = 14 - 2 = 12, \text{p-value} < .0005, \text{reject Ho, there is significant linear relation.} \)
Correlation & Regression

Correlation Does Not Imply Causation

The number of handguns registered may not be the direct cause for the number of people killed by guns.

Spearman’s Rank Correlation Coefficient

\[ r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]

Spearman’s correlation coefficient calculated based on the ranks of the observations, i.e., \( x_{ri} \) is rank of \( x_i \) among all \( x_i \)'s, and \( y_{ri} \) is rank of \( y_i \) among all \( y_i \)'s, where \( d_i = x_{ri} - y_{ri} \). (If ties, use average rank.)

\[ \sum d_i^2 = (1 - 1)^2 + (2 - 5)^2 + \ldots + (n - n)^2 \]

\[ t_s = \frac{1 - \frac{6 \cdot \sum d_i^2}{14(14^2 - 1)}}{\sqrt{1 - \frac{6 \cdot \sum d_i^2}{14(14^2 - 1)}}} = 6.26 \]

\[ r_s = 1 - \frac{6 \cdot 57}{14(14^2 - 1)} = .875 \]

There is significant correlation.

Test for rank correlation

- If sample size \( n \) is less than 10 use special table.
- If sample size large use \( t \)-distribution table.
- The \( t \)-test is the same as the test procedure for non-rank data.
- It is less sensitive to outliers with the disadvantage of loss of information.

Pearson’s Correlation

<table>
<thead>
<tr>
<th>Number of Handguns Registered</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>.941*</td>
<td>.000</td>
<td>14</td>
</tr>
</tbody>
</table>

Sample correlation

*p-value* for testing correlation equals 0. The p-value < .05 means that the correlation is statistically significantly different from 0.
Correlation & Regression

Spearman’s Correlation

<table>
<thead>
<tr>
<th>Number of Handguns Registered</th>
<th>Number of People Killed</th>
<th>Correlation Coefficient</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>.874**</td>
<td>.000</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).

Sample correlation

*p-value* for testing correlation equals 0. The *p*-value < .05 means that the correlation is statistically significantly different from 0.

Example: In an investigation, 122 countries were included to study the relation between female’s life expectancy and the birthrate.

Correlation is significant at the 0.01 level (2-tailed).

Linear Regression

Births per 1000 population, 1992

<table>
<thead>
<tr>
<th>Female life expectancy 1992</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Births per 1000 population, 1992</td>
<td>-.870*</td>
<td>.000</td>
<td>121</td>
<td>121</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).

Linear Relation

If you are in a drug store buying Tylenol. A 24 counts Tylenol costs $2.00. If the cost to get to the drug store is $3.00, then one can use a deterministic model

\[ y = 3 + 2x \]

Can height information be used to find out the weight of an individual?
How to predict “number of people killed by guns” by “number of handguns registered”?

<table>
<thead>
<tr>
<th>Year</th>
<th>NGR(x_i)</th>
<th>Nkill(y_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>447</td>
<td>13</td>
</tr>
<tr>
<td>78</td>
<td>460</td>
<td>21</td>
</tr>
<tr>
<td>79</td>
<td>481</td>
<td>24</td>
</tr>
<tr>
<td>80</td>
<td>498</td>
<td>16</td>
</tr>
<tr>
<td>81</td>
<td>513</td>
<td>24</td>
</tr>
<tr>
<td>82</td>
<td>512</td>
<td>20</td>
</tr>
<tr>
<td>83</td>
<td>526</td>
<td>15</td>
</tr>
<tr>
<td>84</td>
<td>559</td>
<td>34</td>
</tr>
<tr>
<td>85</td>
<td>585</td>
<td>33</td>
</tr>
<tr>
<td>86</td>
<td>614</td>
<td>33</td>
</tr>
<tr>
<td>87</td>
<td>645</td>
<td>39</td>
</tr>
<tr>
<td>88</td>
<td>675</td>
<td>43</td>
</tr>
<tr>
<td>89</td>
<td>711</td>
<td>50</td>
</tr>
<tr>
<td>90</td>
<td>719</td>
<td>47</td>
</tr>
</tbody>
</table>

number of people killed by guns  =>  Response variable
number of handguns registered =>  Explanatory variable

Equation of a Straight Line

\[ y = \alpha + \beta \cdot x \]

- \( y \) is the response or dependent variable
- \( x \) is the explanatory, dependent, or predictor variable

Least Square Principle

Find solution of \( \alpha \) and \( \beta \) of a straight line that minimizes the following variability measure:

\[ \sum (y_i - (\hat{\alpha} + \hat{\beta} \cdot x)^2) \]

\[ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot x \]
Correlation & Regression

The Equation of The Fitted Line

\[ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot x \]

The lease-squared estimate of \( \alpha \) , \( \beta \) are denoted as \( \hat{\alpha} \) and \( \hat{\beta} \) and they are

\[ \hat{\beta} = \frac{s_{xy}}{s_{xx}} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \]

Handgun Example

- \( \bar{x} = 567.50 \), \( \bar{y} = 29.43 \), \( s_x = 12.19 \), \( s_y = 91.91 \), \( r = .941 \)
- \( \hat{\beta} = \frac{.941 \cdot 12.19}{91.91} = .124862 \)
- \( \hat{\alpha} = 29.43 - .124862 \times 567.50 = -41.430439 \)

The regression (prediction) equation:

\[ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot x \]

The Equation of a Fitted Line

\[ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot x \]

• Can be used for estimation or prediction.
• Give the estimate of location of mean response for various \( x \).

Sum of Squares

\[ s_{xy} = \sum_{i=1}^{n} x_i y_i - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right) \left( \frac{\sum_{i=1}^{n} y_i}{n} \right) \]

\[ s_{xx} = \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \]

Other formula

\[ \hat{\beta} = r \cdot \frac{s_y}{s_x} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \]

\( s_x \) is the sample standard deviation of \( x_i \)’s
\( s_y \) is the sample standard deviation of \( y_i \)’s

Correlation & Regression

minimize \( q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta \cdot x_i)]^2 \)

\[ \frac{\partial q}{\partial \alpha} = \sum_{i=1}^{n} (-2)[y_i - (\alpha + \beta \cdot x_i)] = 0 \]

\[ \frac{\partial q}{\partial \beta} = \sum_{i=1}^{n} (-2)x_i[y_i - (\alpha + \beta \cdot x_i)] = 0 \]

\[ \sum_{i=1}^{n} y_i = \alpha \cdot n + \beta \cdot \sum_{i=1}^{n} x_i \]

\[ \sum_{i=1}^{n} x_i y_i = \alpha \cdot \sum_{i=1}^{n} x_i + \beta \cdot \sum_{i=1}^{n} x_i^2 \]

\[ \alpha = ? \]
\[ \beta = ? \]
An Estimation

If at a certain year the number of handguns registered is 700,000, estimate how many people on average would be killed by guns.

\[ \hat{y} = -41.430439 + .124862 \cdot x \]
\[ = -41.430439 + .124862 \cdot 700 \]
\[ = 45.973 \]

The average response at \( x = 700 \) is 45.973.

Cautions

- Causality?
- Avoid unsure extrapolation.

Problem of extrapolation

Error

Extrapolated result for a value out of the scope of \( x \)

A possible trend

Scope of \( x \) data

Estimate \( y \) at \( x \)

Regression and Causality

Regression itself provides no information about causal patterns and must be supplemented by additional analysis (with designed and controlled experiments) to obtain insight about causal relationship.

Regression Model

Regression Model:

\[ y = \mu_{\hat{y}} + \varepsilon \]

with errors, \( \varepsilon \), independent, identically and normally distributed as \( N(0, \sigma^2) \).
First Order Simple Linear Regression Model

Model assumptions:

\[ y = \alpha + \beta x + \varepsilon \]

with errors, \( \varepsilon \), independent, identically and normally distributed as \( N(0, \sigma^2) \), and mean of \( y \) at \( x \) is \( \mu_{y|x} = \alpha + \beta x \).

Model Assumptions

- Equal variances
- Normal errors

Residual

Residual:

\[ e_i = y_i - \hat{y}_i \]
\[ = y_i - (\hat{\alpha} + \hat{\beta} \cdot x_i) \]

Example: Find the residual at \( x = 460 \) and the observed \( y = 21 \).

Predicted \( y = -41.430439 + 0.124862 \cdot 460 \)
\[ = 16.01 = \hat{y} \]

The residual = \( 21 - 16.01 = 4.99 \).

Residual Sum of Squares

Residual Sum of Squares (SSResid) = \( \sum \left( y_i - \hat{y}_i \right)^2 \)

(or Error Sum of Squares, SSE)

Mean Square Error and Standard deviation for regression

Estimation of \( \sigma^2 \):

\[ s^2 = s_{\varepsilon}^2 = MSE = SSE / (n - 2) = \frac{18.287}{18} \]

(Degrees of freedom = \( n - 2 \))

Estimated Standard Error of the regression model:

\[ s = s_{\varepsilon} = \sqrt{s^2} = \sqrt{18.287} = 4.28 \]
Source of Variability

- Response variable
- Explanatory variables
- Error

Response Variable Variability

Total Sum of Squares (SS To) = \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \)

Error Variability

Residual Sum of Squares (SSResid) = \( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) (or Error Sum of Squares, SSE)

Regression Variability

Regression Sum of Squares (SSR) = \( \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \)

Source of Variability

\[ \text{SSTo} = \text{SSR} + \text{SSE} \]

Total Sum of Squares (SSTo) = \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \)

Regression Sum of Squares (SSR) = \( \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \)

Error Variability

\[ \text{SSTo} - \text{SSE} \]

ANOVA Table

<table>
<thead>
<tr>
<th>Source of Var.</th>
<th>S.S.</th>
<th>d.f.</th>
<th>M.S.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>SSR</td>
<td>1</td>
<td>SSR/1=MSR</td>
<td>MSR/MSE</td>
</tr>
<tr>
<td>Error</td>
<td>SSE</td>
<td>n - 2</td>
<td>SSE/(n - 2)=MSE</td>
<td></td>
</tr>
<tr>
<td>Total (corrected)</td>
<td>SSTo</td>
<td>n - 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation of the Model

Coefficient of Determination: It is the proportion of variation in observed $y$ that can be explained by the variable $x$ with the linear regression model.

$$r^2 = 1 - \frac{SSE}{SSTo} = \frac{SSR}{SSTo}$$

Regression coefficients

Equation of the regression line:

$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot x; \quad \hat{y} = -41.430 + .125 \cdot x$$

Inference for Regression Coefficients $\beta$ (t-Test)

Hypothesis: $H_0: \beta = \beta_0$ vs. $H_1: \beta \neq \beta_0$

(It is often testing for $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$)

Test Statistic:

$$t = \frac{\hat{\beta} - \beta_0}{\hat{se}(\hat{\beta})} \sim t\text{-distribution d.f.} = n - 2,$$

where

$$\hat{se}(\hat{\beta}) = \frac{s_{yx}}{\sqrt{\sum(x_i - \bar{x})^2}} = .013$$

Decision rule: Reject $H_0$, if

C.V. approach: $|t| > t_{df}$ or $t > t_{df}$

p-value approach: $p$-value < $\alpha$

Inference for Regression Coefficients $\alpha$ (t-Test)

Hypothesis: $H_0: \alpha = \alpha_0$ vs. $H_1: \alpha \neq \alpha_0$

(It is often testing for $H_0: \alpha = 0$ vs. $H_1: \alpha \neq 0$)

Test Statistic:

$$t = \frac{\hat{\alpha} - \alpha_0}{\hat{se}(\hat{\alpha})} \sim t\text{-distribution d.f.} = n - 2,$$

where

$$\hat{se}(\hat{\alpha}) = s_{y|x} \sqrt{\frac{1}{n} \sum(x_i - \bar{x})^2} = 7.412$$

Decision rule: Reject $H_0$, if

C.V. approach: $|t| > t_{df}$ or $t > t_{df}$

p-value approach: $p$-value < $\alpha$
Predicting Mean Response

The (1 - \(\alpha\)) 100% confidence interval for predicting the mean response at \(x\) is:

\[
\hat{y} \pm t_{\alpha/2} \cdot \hat{s}\varepsilon(\hat{y})
\]

where \(\hat{s}\varepsilon(\hat{y}) = s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}\) d.f. = \(n - 2\)

Predicted Average Number of People Killed at \(x = 460\)

\[
\Rightarrow \ 16.01 \pm 3.91 \Rightarrow (12.09, 19.92)
\]

Predicting a Single New Response

The (1 - \(\alpha\)) 100% confidence interval for predicting an individual outcome at \(x\) is:

\[
\hat{y} \pm t_{\alpha/2} \cdot \hat{s}\varepsilon(\hat{y})
\]

where \(\hat{s}\varepsilon(\hat{y}) = s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}\) d.f. = \(n - 2\)

Predicted Number of People Killed at \(x = 460\)

\[
\Rightarrow \ 16.01 \pm 10.10 \Rightarrow (5.91, 26.11)
\]
Correlation & Regression

Residual Plot

No a good linear model
Variances are homogeneous

Example: \( y = \text{female life expectancy} \)
\( x = \text{GDP (Gross domestic product)} \)

Before Transformation

After ln(GDP) Transformation

\[ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot \ln(x) \]

Transformation

Circle of Powers: \( x^p \) or \( y^p \)

- For \( x \uparrow \) or \( y \uparrow \): try \( p > 1 \) for \( x^p \) or \( y^p \)
  Examples: \( x^2, y^2, x^3, y^3, \ldots \) or \( e^x, e^y \)
- For \( x \downarrow \) or \( y \downarrow \): try \( p < 1 \) for \( x^p \) or \( y^p \)
  Examples: \( x^{-1/2}, y^{-1/2}, x^{-1}, y^{-1}, \ldots \) or \( \ln(x), \ln(y) \)