Linear Regression

If you are in a drug store buying Tylenol. A 24 counts Tylenol costs $2.00.
\[ y = 2, \quad \text{if } x = 1 \quad (\text{Buy one cost } $2.00) \]

If you are in a drug store buying Tylenol. A 24 counts Tylenol costs $2.00.
\[ y = 4, \quad \text{if } x = 2 \quad (\text{Buy two cost } $4.00) \]

If you are in a drug store buying Tylenol. A 24 counts Tylenol costs $2.00.
\[ y = 6, \quad \text{if } x = 3 \quad (\text{Buy one cost } $6.00) \]

If the cost to get to the drug store is $3.00, then
\[ y = 3 + 2x \]
Can height information be used to find out the weight of an individual?

How long should you wait till next eruption?

How to predict “number of people killed by guns” by “number of handguns registered”?

<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>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>

Inter-eruption time: Response variable (Outcome, Dependent)
Duration of eruption: Explanatory variable (Predictor, Independent)
Correlation & Regression

Equation of a Straight Line
\[ y = \alpha + \beta \cdot x \]
- \( \beta \) is the slope
- \( \alpha \) is the \( y \)-intercept
- \( y \) response or dependent variable
- \( x \) explanatory, dependent, or predictor variable

Graph with a Fitted Line
\[ y = \hat{\alpha} + \hat{\beta} \cdot x \]

Least Square Principle
Find solution of \( \alpha \) and \( \beta \) of a straight line that minimizes the following variability measure:
\[ \sum [y_i - (\hat{\alpha} + \hat{\beta} \cdot x)]^2 \]
\[ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot x \]

The Equation of The Fitted Line
\[ y = \hat{\alpha} + \hat{\beta} \cdot x \]

The lease-squared estimate of \( \alpha \) and \( \beta \) are denoted as \( \hat{\alpha} \) and \( \hat{\beta} \) and they are
\[ \hat{\beta} = \frac{s_{xy}}{s_{xx}} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \]

Sum of Squares
\[ s_{xy} = \sum_{i=1}^{n} x_i y_i - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right) \left( \frac{\sum_{i=1}^{n} y_i}{n} \right) \]
\[ s_{xx} = \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \quad s_{yy} = \sum_{i=1}^{n} y_i^2 - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right)^2 \]
Correlation & Regression

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

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 \( x \).

Handgun Example

\[ x = 567.50, \quad y = 29.43, \quad s_x = 12.19, \quad s_y = 91.91, \quad r = .941 \]

\[ \hat{\beta} = .941 \cdot \frac{12.19}{91.91} = .124862 \]

\[ \hat{\alpha} = 29.43 - .124862 \times 567.50 = -41.430439 \]

The regression (prediction) equation:

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

\[ \hat{y} = -41.430369 + .124862 \cdot x \]

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.

\[ \hat{y} = -41.430439 + .124862 \times 700 \]

\[ = -41.430439 + .124862 \times 700 \]

\[ = 45.973 \]

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

Graph with a Fitted Line

Cautions

• Causality?
• Avoid unsure extrapolation.
**Problem of extrapolation**

- Extrapolated result for a value out of the scope of \( x \)
- A possible trend
- 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:

\[
y = \mu_{y|x} + \varepsilon
\]

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 \( y = -41.430439 + 0.124862 \cdot 460 = 16.01 \)

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

Residual Sum of Squares

\[
\text{Residual Sum of Squares (SSResid)} = \sum (y_i - \hat{y}_i)^2
\]

(or Error Sum of Squares, SSE)

Mean Square Error and Standard deviation for regression

Estimation of \( \sigma^2 \) :

\[
s^2 = s_{ij}^2 = \text{MSE} = \text{SSE} / (n - 2) = 18.287
\]

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

Estimated Standard Error of the regression model:

\[
s = \sqrt{s^2} = s_{ij} = 4.28
\]

Source of Variability

- Response variable
- Explanatory variables
- Error

Response Variable Variability

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

Error Variability

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

(or Error Sum of Squares, SSE)
Correlation & Regression

Regression Variability

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

\[ \bar{y} \]

\[ \hat{y} \]

Source of Variability

SSTo = SSR + 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

= SSTo - 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>MSR/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.

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

Evaluation of the Model

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of Estimate</th>
<th>Model Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.941</td>
<td>.886</td>
<td>.877</td>
<td>4.2764</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3711.978</td>
<td>12</td>
<td>310.165</td>
<td>18.287</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>1791.450</td>
<td>10</td>
<td>180.145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>13751.429</td>
<td>22</td>
<td>202.929</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Utility</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Square Error (MSE) = \sigma^2_{\hat{y}}</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Model Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Unstandardized Coefficients</td>
</tr>
<tr>
<td>(Constant)</td>
<td>41.430</td>
</tr>
<tr>
<td>Number of Handguns Registered</td>
<td>125</td>
</tr>
</tbody>
</table>

Regression coefficients

Equation of the regression line:

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

\[ \hat{y} = -41.430 + .125 \cdot x \]
Correlation & Regression

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

**Hypothesis:** $H_0: \beta = \beta_0$  v.s.  $H_a: \beta \neq \beta_0$

(It is often testing for $Ho: \beta = 0$ v.s. $Ha: \beta \neq 0$.)

**Test Statistic:**
$$ t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}(\hat{\beta})} $$
where
$$ \hat{\sigma}(\hat{\beta}) = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = 0.013 $$

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

Predicting Mean Response

The $(1-\alpha) 100\%$ confidence interval for predicting the mean response at $x$ is:
$$ \hat{y} \pm t_{\alpha/2} \cdot \hat{\sigma}(\hat{y}) $$
where
$$ \hat{\sigma}(\hat{y}) = s \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} $$

Predicted 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:
$$ \hat{y} \pm t_{\alpha/2} \cdot \hat{\sigma}(\hat{y}) $$
where
$$ \hat{\sigma}(\hat{y}) = s \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} $$

Predicted Number of People Killed at $x = 460$
$$ \Rightarrow 16.01 \pm 10.10 \Rightarrow (5.91, 26.11) $$

---

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

**Hypothesis:** $H_0: \alpha = \alpha_0$  v.s.  $H_a: \alpha \neq \alpha_0$

(It is often testing for $Ho: \alpha = 0$ v.s. $Ha: \alpha \neq 0$.)

**Test Statistic:**
$$ t = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}(\hat{\alpha})} $$
where
$$ \hat{\sigma}(\hat{\alpha}) = \frac{s}{\sqrt{n} + \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = 7.412 $$

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

---

**Equation of the regression line:**
$$ \hat{y} = \hat{\alpha} + \hat{\beta} \cdot x; \quad \hat{y} = -41.430 + .125 \cdot x $$

---

**Table:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ (Constant)</td>
<td>125</td>
<td>0.013</td>
<td>941</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Dependent Variable:** Number of People Killed

**Number of Handguns Registered**

<table>
<thead>
<tr>
<th>Model</th>
<th>Std. Error</th>
<th>Unstandardized Coefficients</th>
<th>Beta</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ (Constant)</td>
<td>7.412</td>
<td>-41.430</td>
<td>0.125</td>
<td>0.013</td>
</tr>
<tr>
<td>7.412</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.675</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Predicting Mean Response**

The $(1-\alpha) 100\%$ confidence interval for predicting the mean response at $x$ is:
$$ \hat{y} \pm t_{\alpha/2} \cdot \hat{\sigma}(\hat{y}) $$
where
$$ \hat{\sigma}(\hat{y}) = s \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} $$

Predicted 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:
$$ \hat{y} \pm t_{\alpha/2} \cdot \hat{\sigma}(\hat{y}) $$
where
$$ \hat{\sigma}(\hat{y}) = s \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} $$

Predicted Number of People Killed at $x = 460$
$$ \Rightarrow 16.01 \pm 10.10 \Rightarrow (5.91, 26.11) $$

---

**Confidence Interval Bands**

---

CR - 8
Residual Plot

A scatter plot of the residuals against the predicted values of the response variable to verify the assumptions behind the regression model.

- Homogeneity of variance
- Random normal error
- Appropriateness of the linear model

Graph with a Fitted Line

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

Before Transformation

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

After ln(GDP) Transformation
Correlation & Regression

**Example**: $y =$ female life expectancy  
$x =$ GDP (Gross domestic product)

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

**Transformation**

Circle of Powers: $x^p$ or $y^p$

- **For $x$ up or $y$ up**: 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 $x$ down or $y$ down**: try $p < 1$ for $x^p$ or $y^p$
  - Examples: $x^{-1/2}, y^{-1/2}, x^{-1}, y^{-1}, \ldots$ or $\ln(x), \ln(y)$