1 Single Factor Analysis of Variance

Analysis of Variance (ANOVA), refers to a broad collection of experimental situations and statistical procedures for the analysis of quantitative responses from experimental units.

The simplest one of these is called Single Factor ANOVA and it essentially an extension of the pooled t-test. Remember that in the pooled t-test we compared two population means to each to each other:

\[ H_0 : \mu_1 = \mu_2 \]

In a Single Factor ANOVA we compare multiple population means to each other:

\[ H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_I \]

These \( I \) population means are from a factor: a characteristic that differentiate the treatments or populations from each other. The number of different treatments or populations is referred to as the levels of the factor.
What are the factors, response and levels?

- An experiment to test the tensile strength of four different mixing techniques of portland cement.

- A manufacturer of television sets is interested in the effect on tube conductivity of four different types of coating for color picture tubes.

- The response time in milliseconds was determined for three different types of circuits used in an automatic valve shutoff mechanism.

- Four different digital computer circuits are being studied in order to compare the amount of noise present.

- Three brands of batteries are under study. It is suspected that the life (in weeks) of the three brands is different. Five batteries of each brand are tested for longevity.

- Four chemists are asked to determine the percentage of methyl alcohol in a certain chemical compound. Each chemist makes three determinations.
In Single Factor ANOVA we have one factor with two or more levels:

- \( I \) = number of treatments compared (levels)
- \( N \) = total number of observations
- \( \mu_i \) = mean of population \( i \) = mean of treatment \( i \)

The hypothesis of interest is:

\[
H_0 : \quad \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_I \\
H_a : \quad \text{At least two } \mu_i \text{'s are different.}
\]

\( x_{ij} \) is the \( j \)'s observed value of the \( i \)'th treatment (level).

*Example:* The response time in milliseconds was determined for three different types of circuits use in an automatic value shutoff mechanism. Does the data suggest at level .05 that all three circuits had the same response time? The results where:

<table>
<thead>
<tr>
<th>Circuit Type</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9 12 10 8 15</td>
</tr>
<tr>
<td>2</td>
<td>20 21 23 17 30</td>
</tr>
<tr>
<td>3</td>
<td>6 5 8 16 7</td>
</tr>
</tbody>
</table>
The requirements for a Single Factor ANOVA are an extensions of the requirements for the pooled t-test:

The I population or treatment distributions are all Normal with the same variance \( \sigma^2 \). That is, each \( X_{ij} \) is normally distributed with

\[
E[X_{ij}] = \mu_i \quad V[X_{ij}] = \sigma^2
\]

**Example:** In our circuit example we want to test:

\[
H_0 : \quad \mu_1 = \mu_2 = \mu_3 \\
H_a : \quad \text{At least two } \mu_i \text{'s are different.}
\]

The summary data:

\[
\begin{align*}
\bar{x}_1 &= 10.8 \quad s_1^2 = 7.7 \quad s_1 = 2.77 \\
\bar{x}_2 &= 22.2 \quad s_2^2 = 23.7 \quad s_2 = 4.87 \\
\bar{x}_3 &= 8.4 \quad s_3^2 = 19.3 \quad s_3 = 4.39
\end{align*}
\]

Is our equal variance assumption violated? A very rough rule of thumb is that if the largest \( s \) is not too much more than two times the smallest \( s \) then it is reasonable to assume equal \( \sigma^2 \)'s.

It is hard to check for normality when we only have 5 observations, but we can examine the raw data or boxplots for obvious violations.
We will not present the formulas for ANOVA. We will explain the basic principles behind ANOVA and show how to perform the analysis.

The term Analysis of Variance is derived from a partitioning of total variability into its component parts.

- The Total Sum of Squares ($SS_{\text{tot}}$) is our raw measure of variability in the data:

$$SS_{\text{tot}} = \sum_{\text{all obs}} (\text{obs}_{i,j} - \text{grand mean})^2$$

The grand mean is the mean of all observations.

- The Treatment Sum of Squares ($SS_{\text{trt}}$) is the raw variability between treatment:

$$SS_{\text{trt}} = \sum_{\text{treatments}} (\text{treatment mean}_i - \text{grand mean})^2$$

- The Error Sum of Squares ($SS_{\text{err}}$) is the raw variability within treatment:

$$SS_{\text{err}} = \sum_{\text{all obs}} (\text{obs}_{i,j} - \text{treatment mean}_i)^2$$

- The component parts add up to the total part:

$$SS_{\text{tot}} = SS_{\text{trt}} + SS_{\text{err}}$$
Once we have partitioned the variability into its two component parts we scale our raw sum of squares:

- The *Mean Square Treatment* is:
  \[
  \text{MS}_{\text{trt}} = \frac{\text{SS}_{\text{trt}}}{I - 1} = \frac{\text{SS}_{\text{trt}}}{\text{DF}_{\text{trt}}}
  \]
  where \(\text{DF}_{\text{trt}}\) is the *Treatment Degrees of Freedom* \(\text{DF}_{\text{trt}} = I - 1\).

- The *Mean Square Error* is:
  \[
  \text{MS}_{\text{err}} = \frac{\text{SS}_{\text{err}}}{N - I} = \frac{\text{SS}_{\text{err}}}{\text{DF}_{\text{err}}}
  \]
  where \(\text{DF}_{\text{err}}\) is the *Error Degrees of Freedom* \(\text{DF}_{\text{err}} = N - I\).

- Note that:
  \[
  \text{DF}_{\text{tot}} = \text{DF}_{\text{trt}} + \text{DF}_{\text{err}} = N - 1
  \]

The \(\text{MS}_{\text{err}}\) is the common variance \(\sigma^2\) of our treatments. It is an extension of the pooled variance we discussed earlier.
It turns out (distribution theory) that under the null hypothesis:

$$E[\text{MS}_{\text{trt}}] = E[\text{MS}_{\text{err}}] = \sigma^2$$

If $H_0$ is false then:

$$E[\text{MS}_{\text{trt}}] > E[\text{MS}_{\text{err}}] = \sigma^2$$

so under $H_0$ we would expect $\text{MS}_{\text{trt}} \approx \text{MS}_{\text{err}}$.

Expressing this in terms of a ratio:

$$f = \frac{\text{MS}_{\text{trt}}}{\text{MS}_{\text{err}}}$$

under $H_0$ should be 1. The more $f$ exceeds 1 the more likely it is we reject $H_0$ in favor of $H_a$. 
The final step in performing any ANOVA is to know the distribution of

\[ f = \frac{MS_{trt}}{MS_{err}} \sim F^{DF_{trt},DF_{err}} \]

therefore we reject \( H_0 \) at level \( \alpha \) if

\[ f > F_{\alpha,DF_{trt},DF_{err}} \]
1.1 ANOVA Table

To recap, our test for equality of population means is:

\[ H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_I \]
\[ H_a : \text{At least two } \mu_i \text{'s are different.} \]

We summarize the data in an ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>f statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>( I - 1 )</td>
<td>SS_{trt}</td>
<td>MS_{trt} = \frac{SS_{trt}}{DF_{trt}}</td>
<td>( f = \frac{MS_{trt}}{MS_{err}} )</td>
</tr>
<tr>
<td>Error</td>
<td>( N - I )</td>
<td>SS_{err}</td>
<td>MS_{err} = \frac{SS_{err}}{DF_{err}}</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( N - 1 )</td>
<td>SS_{tot}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We reject \( H_0 \) if \( f \geq F_{\alpha,DF_{trt},DF_{err}} \).
Example: Circuit example continued:

\[
H_0 : \mu_1 = \mu_2 = \mu_3 \\
H_a : \text{At least two } \mu_i \text{'s are different.}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>f statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>2</td>
<td>543.6</td>
<td>271.8</td>
<td>16.08</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>202.8</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>746.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Since \( f = 16.08 \geq F_{.05,2,12} = 3.89 \) we reject \( H_0 \) and conclude the response time for the three circuits is not the same.
### 1.1.1 SAS System Output

The SAS System 1
14:59 Monday, April 26, 1999

Analysis of Variance Procedure
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIRCUIT</td>
<td>3</td>
<td>1 2 3</td>
</tr>
</tbody>
</table>

Number of observations in data set = 15

The SAS System 2
14:58 Monday, April 26, 1999

Analysis of Variance Procedure

<table>
<thead>
<tr>
<th>source</th>
<th>DF</th>
<th>sum of squares</th>
<th>mean square</th>
<th>F value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>2</td>
<td>543.600000000</td>
<td>271.8000000</td>
<td>16.08</td>
<td>0.0004</td>
</tr>
<tr>
<td>error</td>
<td>12</td>
<td>202.800000000</td>
<td>16.9000000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>corrected total</td>
<td>14</td>
<td>746.400000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square  
C.V. Root MSE RESPONSE Mean
0.728296 29.78957 4.1109610 13.800000

<table>
<thead>
<tr>
<th>source</th>
<th>DF</th>
<th>anova ss</th>
<th>mean square</th>
<th>F value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>circuit</td>
<td>2</td>
<td>543.600000000</td>
<td>271.8000000</td>
<td>16.08</td>
<td>0.0004</td>
</tr>
</tbody>
</table>
1.1.2 Splus System Output

```
> aov.circuit <- aov(response ~ circuit, circuit.frame)
> summary(aov.circuit)

        Df Sum of Sq Mean Sq F value   Pr(>F)
 circuit  2   543.6  271.80 16.0828 0.0004023
 Residuals 12  202.8  16.90
```

1.1.3 Stata System Output

```
. anova response circuit, cat(circuit)

Number of obs = 15  R-squared = 0.7283
Root MSE = 4.11096  Adj R-squared = 0.6830

Source | Partial SS  df       MS     F         Prob > F
-----------------------------------------------------------------------------------
Model | 543.60      2  271.80   16.08      0.0004
     |            |          |          |            |
circuit | 543.60     2  271.80   16.08      0.0004
     |            |          |          |            |
Residual | 202.80     12  16.90
     |            |          |          |            |
Total | 746.40      14 53.3142857
```

2 Multiple Comparisons in ANOVA

If we do not reject $H_0$ in an ANOVA, the analysis is finished—there are no differences among the means. If we, however, reject $H_0$ then we want to know which of the $\mu_i$’s are different from each other.

Any method for carrying out this further analysis is called a *multiple comparison procedure*. There a number of such procedure in statistical literature. The naive way of doing this would be to construct pooled t-tests for each pair of $\mu_i$’s. There is a problem with this...

The significance level is no longer valid. We cannot conduct multiple pooled t-test at level $\alpha$ and have the overall significance level maintained at $\alpha$. Multiple tests at level $\alpha$ result in an overall $\alpha' > \alpha$.

The *Tukey’s Procedure* is a multiple comparison procedure that make *simultaneous confidence statements* about the true value of all differences $\mu_i - \mu_j$. Each interval that does not include zero yields the conclusion that $\mu_i$ and $\mu_j$ differ significantly at level $\alpha$. 
The Tukey’s procedure can be performed through the following series of steps:

- Select $\alpha$ and find $Q_{\alpha, I, \text{DF}_{\text{err}}}$ from Table A8.
- Determine $w = Q_{\alpha, I, \text{DF}_{\text{err}}} \cdot \sqrt{\text{MSE}_{\text{err}}/J}$
- List the sample mean in increasing order and underline those pairs that differ by less than $w$. Any pair of sample means not underscored by the same line corresponds to a pair of true treatment means that are judged to be significantly different.

Note: $J$ is the number of observations per treatment. The above procedure will only work if all the treatments contain the same number of observations. A more general Tukey’s procedure is also available.
Example: For the circuit example, do Tukey’s multiple comparison procedure at $\alpha = .05$ to determine which circuits have different response times. The summary data:

$$\bar{x}_1 = 10.8 \quad \bar{x}_2 = 22.2 \quad \bar{x}_3 = 8.4$$

First we find $w$:

$$Q_{\alpha,I,DF_{err}} = Q_{\alpha,I,DF_{err}} = Q_{.05,3,12} = 3.77$$

$$w = Q_{.05,3,12} \sqrt{MS_{err}/J} = 3.77 \sqrt{16.9/5} = 6.931$$

We sort our sample means from smallest to largest:

$$8.4 \quad 10.8 \quad 22.2$$

Therefore we conclude that circuits 1 and 3 are not significantly different from each other, but that they are significantly different from circuit 2.
Example: The tensile strength of synthetic fiber used to make cloth for men's shirts is of interest to a manufacturer. It is suspected that the strength is affected by the percentage of cotton in the fiber. Five levels of cotton percentages are of interest, 15, 20, 25, 30 and 35 percent.

<table>
<thead>
<tr>
<th>Percentages of Cotton</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

Test at $\alpha = .05$ if the tensile strength is different for the different cotton percentages and, if they differ find which ones differ from each other.

Sample means:

$$\bar{x}_1 = 9.8 \quad \bar{x}_2 = 15.4 \quad \bar{x}_3 = 17.6 \quad \bar{x}_4 = 21.6 \quad \bar{x}_5 = 10.8$$
Chapter 10: Single Factor Analysis of Variance
$H_0$ : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H_a$ : At least two $\mu_i$’s are different.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>f statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4</td>
<td>474.76</td>
<td>118.94</td>
<td>14.76</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>161.20</td>
<td>8.06</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>636.96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We reject $H_0$ if

$$f = 14.76 \geq F_{0.05,4,20} = 2.87.$$  

We reject $H_0$ and conclude the tensile strength for the 5 cotton percentages are different.

Tukey’s multiple comparison:

$$Q_{\alpha,I,DF_{err}} = Q_{\alpha,I,DF_{err}} = Q_{0.05,5,20} = 4.23$$

$$w = Q_{0.05,4,20} \sqrt{MS_{err}/J} = 4.23 \sqrt{8.06/5} = 5.37$$

We sort our sample means from smallest to largest:

<table>
<thead>
<tr>
<th>1</th>
<th>5</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.8</td>
<td>10.8</td>
<td>15.4</td>
<td>17.6</td>
<td>21.6</td>
</tr>
</tbody>
</table>

We conclude that a cotton percentage of 30% yields the highest tensile strength.
### Analysis of Variance Procedure

#### Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERCEN</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>35</td>
</tr>
</tbody>
</table>

Number of observations in data set = 25

#### Analysis of Variance Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>475.76000000</td>
<td>118.9400000</td>
<td>14.76</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>161.20000000</td>
<td>8.0600000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>24</td>
<td>636.96000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>C.V.</th>
<th>Root MSE</th>
<th>STRENGTH Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.746923</td>
<td>18.87642</td>
<td>2.8390139</td>
<td>15.0400000</td>
</tr>
</tbody>
</table>

#### Source

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Anova SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERCEN</td>
<td>4</td>
<td>475.76000000</td>
<td>118.94000000</td>
<td>14.76</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
The SAS System

16:29 Monday, April 26, 1999

Analysis of Variance Procedure

Tukey’s Studentized Range (HSD) Test for variable: STRENGTH

NOTE: This test controls the type I experimentwise error rate, but
generally has a higher type II error rate than REGWQ.

Alpha = 0.05  df = 20  MSE = 8.06
Critical Value of Studentized Range = 4.232
Minimum Significant Difference = 5.373

Means with the same letter are not significantly different.

Tukey Grouping  Mean  N  PERCENT

    A           21.600   5  30
    A
    B  A      17.600   5  25
    B
    B  C      15.400   5  20
    C
    D  C      10.800   5  35
    D
    D          9.800   5  15
1 Simple Linear Regression

1.1 Deterministic and Nondeterministic Models

In regression analysis we are interested in studying variables that are related to each other in a non-deterministic way.

- **Deterministic:** If we say that $x$ and $y$ are deterministically related, we are saying that knowing $x$ completely specifies the value of $y$ (and vice-versa). Thus, we can write

  $$ y = f(x). $$

  **Example:** Suppose the cost of leasing a computer ($x$) is a $200 initial payment plus $50 per month then the cost of leasing the computer $y$ is

  $$ y = 200 + 50x. $$

- **Nondeterministic:** A nondeterministic or probabilistic relationship is less precise than a deterministic one. We typically assume that there is some underlying deterministic relationship with an error term added on:

  $$ y = f(x) + \epsilon $$

  **Example:** People who score higher on SAT score tend to have higher GPA in college, but knowing the SAT score does not tell you what GPA a student will have in college.

  $$ \text{GPA} = f(\text{SAT}) + \text{error} $$
Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a nondeterministic fashion.

### 1.2 Simple Linear Regression Model

A simple deterministic relationship between two variables $x$ and $y$ is a linear relationship:

$$ y = \beta_0 + \beta_1 x $$

The $y$-intercept is $\beta_0$ and the slope of this line is $\beta_1$.

Generally, in a regression scenario, the $x$ variable is fixed by the experimenter and it is called the **independent variable or predictor**. In other settings, the $x$ variable is an observational variable that we wish to use as a basis for predicting the other variable. The $y$ variable depends on the value of the predictor, and we refer to it as the **dependent variable or response**.

\[
\begin{align*}
x & = \text{predictor or independent variable} \\
y & = \text{response or dependent variable}
\end{align*}
\]
Model for Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n. \]

\( Y_1, \ldots, Y_n \) – observed values of response

\( x_1, \ldots, x_n \) – observed values of predictor

\( \beta_0, \beta_1 \) – unknown parameters to be estimated from data

\( \epsilon_1, \ldots, \epsilon_n \) – unknown random error terms, usually assumed to be independent \( N(0, \sigma^2) \) random variables

Example: In our SAT versus GPA example, the independent variable \( x \) is SAT score, the dependent variable \( y \) is GPA. Since GPA and SAT scores are nondeterministically related, we use the nondeterministic model

\[ Y = \beta_0 + \beta_1 x + \epsilon \]

Questions:

1. How do we estimate \( \beta_0, \beta_1, \) and \( \sigma^2 \)?

2. Does the proposed model fit the data adequately?

3. Are the assumptions satisfied?
Implication of Model:

For each $x$ value, the observed $Y$ will fall above or below the line $y = \beta_0 + \beta_1 x$ according to the value of the error term $\epsilon$. For each fixed $x$

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

Notation: $\mu_{Y|x} = E(Y|x) = \beta_0 + \beta_1 x$

$$\sigma^2_{Y|x} = \text{Var}(\beta_0 + \beta_1 x + \epsilon) = \text{Var}(\epsilon) = \sigma^2$$

1.3 Plotting the Data

A scatter plot of the data is useful to see whether a linear relationship is plausible. We can also look for constant variance and possible outliers.

Examples: Sulphur Dioxide and Ozone.
1.4 Estimating $\beta_0$ and $\beta_1$

Consider an arbitrary line $y = b_0 + b_1 x$ drawn through a scatter plot. We want the line to be as close to the points in the scatter plot as possible. The vertical distance from a point $(x, y)$ to the corresponding point on the line $(x, b_0 + b_1 x)$ is $y - (b_0 + b_1 x)$.

The least squares estimates minimize the sum of the squared deviations:

$$D = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The values of $b_0$ and $b_1$ are called the least squares estimates of $\beta_0$ and $\beta_1$. We write them as $\hat{\beta}_0$ and $\hat{\beta}_1$. 
Formulas for Least Squares Estimates

\[ \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \]

and

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

where

\[ S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \]

and

\[ S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \]

Estimating \( \sigma^2 \)

The minimum value of \( D \) is

\[ D = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = SSE \]

To estimate \( \sigma^2 \), we divide this SS by its degrees of freedom \((n - 2)\) to get

\[ \hat{\sigma}^2 = s_e^2 = \frac{SSE}{n - 2} \]
Example: Ozone Data

<table>
<thead>
<tr>
<th>x</th>
<th>0.066</th>
<th>0.088</th>
<th>0.120</th>
<th>0.162</th>
<th>0.050</th>
<th>0.186</th>
<th>0.057</th>
<th>0.100</th>
<th>0.112</th>
<th>0.055</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>4.61</td>
<td>11.62</td>
<td>9.51</td>
<td>13.88</td>
<td>6.35</td>
<td>15.44</td>
<td>2.53</td>
<td>11.83</td>
<td>8.08</td>
<td>7.06</td>
</tr>
<tr>
<td>x</td>
<td>0.154</td>
<td>0.074</td>
<td>0.111</td>
<td>0.140</td>
<td>0.071</td>
<td>0.110</td>
<td>0.081</td>
<td>0.115</td>
<td>0.143</td>
<td>0.122</td>
</tr>
<tr>
<td>y</td>
<td>20.69</td>
<td>16.64</td>
<td>9.28</td>
<td>14.90</td>
<td>2.87</td>
<td>13.00</td>
<td>5.79</td>
<td>5.83</td>
<td>10.21</td>
<td>14.18</td>
</tr>
<tr>
<td>x</td>
<td>0.157</td>
<td>0.138</td>
<td>0.094</td>
<td>0.063</td>
<td>0.141</td>
<td>0.169</td>
<td>0.106</td>
<td>0.118</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The summary statistics are

\[ \sum x_i = 290.98 \quad \sum x_i^2 = 3540.8042 \]

\[ \sum y_i = 3.103 \quad \sum y_i^2 = 0.382131 \]

\[ \sum x_i y_i = 35.23223 \quad n = 28 \]

Compute the estimates:
1.5 Examining the Overall Fit of the Model

After we obtain the least squares line, we can obtain the predicted value corresponding to $x_i$ from

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$ 

In computing $SSE$, we looked at the deviation of the observed value from that predicted by the least squares line. This deviation is the residual,

$$e_i = y_i - \hat{y}_i.$$ 

We will again break up the total variation of the response into two parts: one explainable by the model and one due to error.

**Total Sum of Squares:**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$$

**Error Sum of Squares:**

$$SSE \quad = \quad \sum_{i=1}^{n} e_i^2 \quad = \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad = \quad \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \quad = \quad \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$
### Regression Sum of Squares:

\[
SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SST - SSE
\]

We obtain the decomposition of \( SST \):

\[
SST = SSR + SSE
\]

The corresponding degrees of freedom are

\[
n - 1 = 1 + (n - 2)
\]

We summarize our calculations in an ANOVA table for regression:

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( SSR )</td>
<td>1</td>
<td>( MSR )</td>
<td>( MSR/MSE )</td>
</tr>
<tr>
<td>Error</td>
<td>( SSE )</td>
<td>( n - 2 )</td>
<td>( MSE )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( TSS )</td>
<td>( n - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### An Overall Measure of Fit

A useful measure of fit is the coefficient of determination:

\[
r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}
\]
Remarks:

1. $0 \leq r^2 \leq 1$

2. $r^2 = 1$ if all the data lie on a straight line.

3. $r^2$ near zero indicates no linear relationship. However, there may be some strong nonlinear relationship.

4. $r^2$ is the proportion of variation of $y$ “explained” by the linear relationship with $x$.

1.6 Testing for a Linear Relationship

Consider testing the hypotheses:

$H_0 : \beta_1 = 0$ (no linear relationship)

$H_a : \beta_1 \neq 0$ (some linear relationship)

Test Statistic:

$$F = \frac{MSR}{MSE} = (n - 2) \frac{r^2}{1 - r^2}$$

Rejection Region:

$$F > F_\alpha \text{ where } df = (1, n - 2)$$
1.7 Inference Concerning $\beta_1$

Mean and Variance of $\hat{\beta}_1$

When the assumptions of the regression model hold, one can find the mean and variance of $\hat{\beta}_1$:

$$E(\hat{\beta}_1) = \beta_1$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Then one can form the statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

has a $t$ distribution with $n - 2$ d.f. Here

$$s_{\hat{\beta}_1} = \frac{s_{\epsilon}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

Confidence Interval for $\beta_1$

A $100(1 - \alpha)$ confidence interval for $\beta_1$ is given by

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$$
Tests Concerning $\beta_1$

We can test $H_0 : \beta_1 = \beta_{10}$ using the test statistic

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$$

Remarks

1. Inference concerning $\beta_1$ is more important than that concerning $\beta_0$. $\beta_1$ measures the effect on $E(Y)$ of changing $x$ by one unit. $\beta_0$ is a special case of $E(Y)$, being the mean response when $x = 0$.

2. To estimate $\beta_1$ well, we need small $\sigma^2_{\beta_1}$. This occurs if $\sum (x_i - \bar{x})^2$ is large. Thus, we want the $x$ values to be as scattered as possible. A drawback is that linearity occurs over only a limited range of $x$ values.
1.8 Inference on the Mean Response

Let \( x^* \) be a specified value of the predictor \( x \). The mean response when \( x = x^* \) is \( E(y) = \beta_0 + \beta_1 x^* \). Better notation is \( \mu_{y,x^*} \) or \( E(Y|x^*) \) showing the dependence on \( x^* \).

**Point Estimate:**

\[
\hat{\mu}_{y,x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*
\]

**Note:** The estimated mean falls on the line of best fit or the least squares line.

**Mean and Variance of \( \hat{\mu}_{y,x^*} \)**

\[
E(\hat{\mu}_{y,x^*}) = \beta_0 + \beta_1 x^*
\]

\[
\text{Var}(\hat{\mu}_{y,x^*}) = \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)
\]

Again we will use a \( t \) variable for inference by replacing \( \sigma^2 \) by its estimate \( s^2_e \).
Confidence Interval for the Mean Response

A 100(1 − α) confidence interval for 𝜇_{y|x*} is given by

\[ \hat{β}_0 + \hat{β}_1 x^* \pm t_{α/2} s_ε \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} , \]

with \( n - 2 \) degrees of freedom.

Remark: One can also carry out tests concerning 𝜇_{y|x} using a 𝑡 test as described on p. 517. However, confidence intervals are typically used in connection with the mean response.
1.9 Prediction of a Single Response

Often we would like to predict a new $y$ at a given $x$ value. Our best guess for the new $y$ is the estimated mean response,

$$ \hat{y} = \hat{\mu}_{y \cdot x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*. $$

The error in prediction is $y - \hat{y}$. The variance of prediction is

$$ \text{Var}(y - \hat{y}) = \text{Var}(y) + \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) $$

$$ = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right) $$

A $100(1 - \alpha)$ prediction interval for $y$ is given by

$$ \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} s_{\epsilon} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} $$

Remarks:

1. The best estimation of $\mu_{y \cdot x^*}$ and prediction of a new $y$ occurs at $x^* = \bar{x}$. The variance increases as $x^*$ moves away from $\bar{x}$.

2. The use of regression for estimating mean response and predicting a new response is justifiable only for $x$ values within the range of the data. We do not have information on whether the linear relationship hold outside this region.

3. The above procedures give a confidence interval for a single $x^*$. If we want intervals that hold simultaneously for several $x^*$ values, we would need to use other methodology.