Regression Analysis

Regression: *Methodology for studying the relationship among two or more variables*

Two major aims:

- Determine an appropriate **model** for the relationship between the variables.

- **Predict** the value of one variable given known values of the other variables.

**Simple** regression refers to the case of only two variables. The variable to be predicted is called the *response* or *dependent* variable. The other variable goes by many names: *explanatory* or *independent* variable and also *covariate* or *predictor*. 
The response is usually denoted $y$ and the predictor $x$.

In *deterministic* situations, $y$ and $x$ are related by an exact functional relationship:

$$y = f(x).$$

Example of deterministic relationship:

- $x = \text{height of an object above earth's surface}$
- $y = \text{time it takes object to reach earth when dropped from height } x$

$$y = \sqrt{\frac{2x}{g}}$$
Examples of *nondeterministic* relationships:

- $x = \text{age}, \ y = \text{blood pressure}$
- $x = \text{SAT score}, \ y = \text{freshman gpr}$

Suppose we have data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The goal is to determine the approximate relationship between $x$ and $y$. A useful tool in doing so is the scatter plot.

*Scatterplot for fish data*
Ultimately, we will study multiple regression, the case in which there is one response variable and at least two independent variables.

We start with a review of the straight line model in simple regression:

\[(\text{Average of } y) = ax + b\]

**Statistical model**

\(x_1, \ldots, x_n\) are values of an independent variable. These values are fixed by the experimenter.

\(y_1, \ldots, y_n\) are corresponding values of response or dependent variable. These are random variables.

We assume that

\[y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.\]
The following assumptions are made about the preceding model:

(1) $\epsilon_1, \ldots, \epsilon_n$ are unobserved random variables.

(2) $\epsilon_1, \ldots, \epsilon_n$ are independent of each other.

(3) Each $\epsilon_i$ has a $N(0, \sigma^2)$ distribution.

(4) $\beta_0$, $\beta_1$ and $\sigma^2$ are unknown, but fixed, parameters.

The statistical problem is to infer the unknown parameters $\beta_0$, $\beta_1$ and $\sigma^2$ using the data $(x_1, y_1)$, $(x_2, y_2)$, $\ldots$, $(x_n, y_n)$. 
The parameter $\sigma^2$ determines how tightly the data will cluster about the line $y = \beta_0 + \beta_1 x$. If $\sigma^2 = 0$, the relationship is deterministic, and all points lie exactly on the straight line.

An analysis of the relationship between $x$ and $y$ is referred to as a regression analysis.

Given a set of data, how do we estimate the parameters $\beta_0$ and $\beta_1$? One method of doing so is based on the principle of least squares.

Given the data $(x_1, y_1), \ldots, (x_n, y_n)$, choose $b_0$ and $b_1$ to minimize the sum of the following values:

$$[y_i - (b_0 + b_1 x_i)]^2, \quad i = 1, \ldots, n.$$
Illustration of Least Squares Principle

Two candidate lines and the vertical deviations of the data from each line
Form of Least Squares Estimates

The least squares estimates are as follows:

$$\hat{\beta}_1 = \frac{\text{sum of } (x_i - \bar{x})(y_i - \bar{y})}{\text{sum of } (x_i - \bar{x})^2}$$

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

The line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the least squares line.

The $i$th predicted value is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n,$$

and the $i$th residual is $y_i - \hat{y}_i$. 
The error sum of squares, or $SSE$, is

$$SSE = \text{sum of } (y_i - \hat{y}_i)^2.$$ 

We estimate the error variance $\sigma^2$ by

$$\hat{\sigma}^2 = \frac{SSE}{n - 2}.$$ 

The total sum of squares, or $SST$, is

$$SST = \text{sum of } (y_i - \bar{y})^2.$$ 

The total sum of squares satisfies the following important relationship:

$$SST = SSE + SSR,$$

where

$$SSR = \text{sum of } (\hat{y}_i - \bar{y})^2.$$
$SSR$ is called the sum of squares due to regression. $SSR$ measures how much the least squares line differs from the flat line $y = \bar{y}$. The bigger $SSR$ is, the steeper the least squares line.

- $SST$: total variation
- $SSE$: variation due to error, i.e., unexplained variation
- $SSR$: variation due to linear relationship, i.e., explained or systematic variation

Practical interpretation of $SST = SSE + SSR$:

“The total variation in the data is the sum of variation due to error and variation due to linear relationship between $x$ and $y$. “
Coefficient of determination:

\[ R^2 = \frac{SSR}{SST} \]

\( R^2 \) represents the proportion of variation in the data due to the linear relationship of \( x \) and \( y \)

Note: \( 0 \leq R^2 \leq 1 \)

Example

- \( y = \) ozone level (ppm) at given location
- \( x = \) secondary carbon pollutants (millionths of grams per cubic meter) at same location
### Ozone/Carbon-Pollutant Data

<table>
<thead>
<tr>
<th>Ozone</th>
<th>Carbon</th>
<th>Ozone</th>
<th>Carbon</th>
</tr>
</thead>
<tbody>
<tr>
<td>.066</td>
<td>4.61</td>
<td>.071</td>
<td>2.87</td>
</tr>
<tr>
<td>.088</td>
<td>11.62</td>
<td>.110</td>
<td>13.00</td>
</tr>
<tr>
<td>.120</td>
<td>9.51</td>
<td>.081</td>
<td>5.79</td>
</tr>
<tr>
<td>.162</td>
<td>13.88</td>
<td>.115</td>
<td>5.83</td>
</tr>
<tr>
<td>.050</td>
<td>6.35</td>
<td>.143</td>
<td>10.21</td>
</tr>
<tr>
<td>.186</td>
<td>15.44</td>
<td>.122</td>
<td>14.18</td>
</tr>
<tr>
<td>.057</td>
<td>2.53</td>
<td>.157</td>
<td>16.58</td>
</tr>
<tr>
<td>.100</td>
<td>11.83</td>
<td>.138</td>
<td>9.62</td>
</tr>
<tr>
<td>.112</td>
<td>8.08</td>
<td>.094</td>
<td>9.25</td>
</tr>
<tr>
<td>.055</td>
<td>7.06</td>
<td>.063</td>
<td>8.99</td>
</tr>
<tr>
<td>.154</td>
<td>20.69</td>
<td>.141</td>
<td>11.39</td>
</tr>
<tr>
<td>.074</td>
<td>16.64</td>
<td>.169</td>
<td>13.52</td>
</tr>
<tr>
<td>.111</td>
<td>9.28</td>
<td>.106</td>
<td>7.78</td>
</tr>
<tr>
<td>.140</td>
<td>14.90</td>
<td>.118</td>
<td>9.55</td>
</tr>
</tbody>
</table>
Scatterplot of ozone data
Computation of least squares line:

\[ \text{sum of } (x_i - \bar{x})(y_i - \bar{y}) = 2.985411 \]

\[ \text{sum of } (x_i - \bar{x})^2 = 516.8985 \]

\[ n = 28, \quad \bar{x} = 10.39214, \quad \bar{y} = 0.1108214 \]

\[ \hat{\beta}_1 = \frac{2.985411}{516.8985} = 0.005775623 \]

\[ \hat{\beta}_0 = 0.1108214 - \hat{\beta}_1(10.39214) \]

\[ = 0.05080033 \]
So, the least squares line is

\[ y = 0.0508 + 0.00578x, \]

which is plotted below.
\[ SST = \text{sum of } (y_i - \bar{y})^2 \]
\[ = 0.03825211 \]

\[ SSR = \text{sum of } (\hat{y}_i - \bar{y})^2 \]
\[ = 0.01724261 \]

\[ SSE = SST - SSR = 0.0210095 \]

\[ R^2 = \frac{0.01724261}{0.03825211} = .4508 \]

About 45% of the variation in ozone levels is explained by the relationship between ozone level and secondary carbon pollutants.
ANOVA Table

The analysis of variance, or ANOVA, table summarizes some of the calculations in a regression analysis. Here is the ANOVA table for the ozone data.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>0.01724</td>
<td>1</td>
<td>0.01724</td>
<td>21.335</td>
</tr>
<tr>
<td>Error</td>
<td>0.02101</td>
<td>26</td>
<td>0.00081</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.03825</td>
<td>27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In general:

\[
MSR = \frac{SSR}{(\text{model df})} = \frac{SSR}{1}
\]

\[
MSE = \frac{SSE}{(\text{error df})} = \frac{SSE}{(n - 2)}
\]

\[
F = \frac{MSR}{MSE}.
\]

The \( F \)-statistic is used to test the hypothesis \( H_0 : \beta_1 = 0 \). More on this later.
Residuals

The residuals are $e_1, \ldots, e_n$, where

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, \ldots, n.$$  

These approximate the error terms, i.e.,

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i), \quad i = 1, \ldots, n.$$  

We may use the residuals to check our various assumptions about the model. Examining plots of the residuals is the most common way to check assumptions.

Methods of plotting residuals

1) Construct a box plot of the residuals. This will reveal any outliers, which are worrisome in regression since they can inflate $SSE$. This makes a test of $H_0: \beta_1 = 0$ less powerful, possibly causing one to miss a significant linear relationship.
2) Construct a histogram of the residuals. This allows one to check the normality assumption.

3) Plot $e_i$ vs. $x_i$, which may point out the presence of a nonlinear relationship.

4) Plot $e_i$ vs. $\hat{y}_i$. This is a good plot for checking the assumption of equal variances.
   
   **Note:** Don’t plot $e_i$ vs. $y_i$ since a pattern may be expected in this plot even when the errors are independent.
Consider the following plot of log-transformed gene expression data along with the least squares line.

**Boxplot of residuals**
Smooth histogram of residuals and normal curve with same standard deviation

The dashed line is the normal curve. Note the slightly heavier right-hand tail for the histogram. This agrees with the boxplot.
Residual plot along with a “running average”

There is some evidence of curvature in the running average. This suggests that a straight line model is not the best model for curve of averages.