EAD 115

Numerical Solution of Engineering and Scientific Problems

David M. Rocke
Department of Applied Science
Curve Fitting

• Given a set of n points \((x_i, y_i)\), find a fitted curve that provides a fitted value \(y = f(x)\) for each value of \(x\) in a range.

• The curve may interpolate the points (go through each one), either linearly or nonlinearly, or may approximate the points without going through each one, as in least-squares regression.
Simple Linear Regression

• We have a set of \( n \) data points, each of which has a measured predictor \( x \) and a measured response \( y \).
• We wish to develop a prediction function \( f(x) \) for \( y \).
• In the simplest case, we take \( f(x) \) to be a linear function of \( x \), as in \( f(x) = a_0 + a_1 x \).
Criteria and Estimation

• If we have one point, way (1,1), then many lines fit perfectly:
  – $f(x) = x$
  – $f(x) = 2x-1$
  – $f(x) = -x+2$

• If there are two points, say (1,1) and (2,3), then in general there is exactly one line going through the points: $f(x) = 2x-1$. 
• If there are more than two points, then in general there is no straight line through all of them.

• These problems are, respectively, underdetermined, determined, and overdetermined.

• Reasonable criteria for choosing the coefficients $a$ and $b$ in $f(x) = a_0 + a_1 x$ lie in minimizing the size of the residuals: $r_i = y_i - f(x_i) = y_i - (a_0 + a_1 x_i)$, but how to combine different residuals?
• The least-squares criterion minimizes

\[ SS = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \]

• There are many other possible criteria. Use of the least-squares criterion does not imply any beliefs about the data.
• Use of the linear form for \( f(x) \) assumes that this straight-line relationship is reasonable.
• Assumptions are needed for inference about the predictions or about the relationship itself.
Minimize Sum of Residuals

Minimize Sum of Absolute Values of Residuals

Minimize Max Residual
Computing the Least-Squares Solution

• We wish to minimize the sum of squares of deviations from the regression line by choosing the coefficients $a_0$ and $a_1$ accordingly.

• Since this is a continuous, quadratic function of the coefficients, one can simply set the partial derivatives equal to zero.
\[ SS(a_0, a_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \]

\[ 0 = \frac{\partial SS(a_0, a_1)}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = -2 \left[ \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_0 - \sum_{i=1}^{n} a_1 x_i \right] \]

\[ 0 = \frac{\partial SS(a_0, a_1)}{\partial a_1} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) x_i = -2 \left[ \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} a_0 x_i - \sum_{i=1}^{n} a_1 x_i^2 \right] \]

\[ n a_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \]

\[ a_0 \sum_{i=1}^{n} x_i - a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \]
• These *normal equations* have a unique solution as two equations in two unknowns.

• The straight line that is calculated in this way is used in practice to see if there is a relationship between $x$ and $y$.

• It is also used to predict $y$ from $x$.

• It can also be used to predict $x$ from $y$ by inverting the equation.

• We now look at some practical uses of least squares.
Quantitative Prediction

• Regression analysis is the statistical name for the prediction of one quantitative variable (fasting blood glucose level) from another (body mass index)

• Items of interest include whether there is in fact a relationship and what the expected change is in one variable when the other changes
Assumptions

• Inference about whether there is a real relationship or not is dependent on a number of assumptions, many of which can be checked.
• When these assumptions are substantially incorrect, alterations in method can rescue the analysis.
• No assumption is ever exactly correct.
Linearity

• This is the most important assumption
• If $x$ is the predictor, and $y$ is the response, then we assume that the average response for a given value of $x$ is a linear function of $x$
  
  • $E(y) = a + bx$
  
  • $y = a + bx + \varepsilon$
  
  • $\varepsilon$ is the error or variability
Regression when the Assumptions are Satisfied
Regression with nonlinearity
• In general, it is important to get the model right, and the most important of these issues is that the mean function looks like it is specified

• If a linear function does not fit, various types of curves can be used, but what is used should fit the data

• Otherwise predictions are biased
Independence

• It is assumed that different observations are statistically independent
• If this is not the case inference and prediction can be completely wrong
• There may appear to be a relationship even though there is not
• Randomization and control prevents this in general
Lack of Independence

![Graph showing a lack of independence between y1 and y2 with a linear trend line.](image-url)
Note no relationship between x and y

These data were generated as follows:

\[ x_1 = y_1 = 0 \]
\[ x_{i+1} = 0.95x_i + \varepsilon_i \]
\[ y_{i+1} = 0.95y_i + \eta_i \]
Constant Variance

- Constant variance, or homoscedasticity, means that the variability is the same in all parts of the prediction function.
- If this is not the case, the predictions may be on the average correct, but the uncertainties associated with the predictions will be wrong.
- Heteroscedasticity is non-constant variance.
Consequences of Heteroscedasticity

- Predictions may be unbiased (correct on the average)
- Prediction uncertainties are not correct; too small sometimes, too large others
- Inferences are incorrect (is there any relationship or is it random)
Normality of Errors

• Mostly this is not particularly important
• Very large outliers can be problematic
• Graphing data often helps
• If in a gene expression array experiment, we do 40,000 regressions, graphical analysis is not possible
• Significant relationships should be examined in detail
Consequences of Outliers
Example Analysis

• Standard aqueous solutions of fluorescein (in pg/ml) are examined in a fluorescence spectrometer and the intensity (arbitrary units) is recorded

• What is the relationship of intensity to concentration?

• Use later to infer concentration of labeled analyte
<table>
<thead>
<tr>
<th>concen~n</th>
<th>intens~y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>12.6</td>
</tr>
<tr>
<td>8</td>
<td>17.3</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>24.7</td>
</tr>
</tbody>
</table>
. regress intensity concentration

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
</tr>
<tr>
<td>Residual</td>
<td>.936784731</td>
<td>5</td>
<td>.187356946</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
</tr>
</tbody>
</table>

Number of obs = 7
F( 1,  5) = 2227.53
Prob > F = 0.0000
R-squared = 0.9978
Adj R-squared = 0.9973
Root MSE = .43285

| intensity | Coef.      | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|-----------|------------|-----------|-------|--------|----------------------|
| concentration | 1.930357   | .0409002  | 47.20 | 0.000  | 1.82522 2.035495    |
| _cons     | 1.517857   | .2949358  | 5.15  | 0.004  | .7597003 2.276014   |
. regress intensity concentration

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
</tr>
<tr>
<td>Residual</td>
<td>.936784731</td>
<td>5</td>
<td>.187356946</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
</tr>
</tbody>
</table>

Number of obs = 7
F( 1, 5) = 2227.53
Prob > F = 0.0000
R-squared = 0.9978
Adj R-squared = 0.9973
Root MSE = 0.43285

| intensity | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----------|--------|-----------|-------|------|-----------------------|
| concentration | 1.930357 | 0.0409002  | 47.20 | 0.000 | 1.825222 - 2.035495  |
| _cons      | 1.517857 | 0.2949358  | 5.15  | 0.004 | .7597003 - 2.276014  |

Slope
. regress intensity concentration

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
<td>F(  1,     5) = 2227.53</td>
</tr>
<tr>
<td>Residual</td>
<td>.936784731</td>
<td>5</td>
<td>.187356946</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
<td>R-squared = 0.9978</td>
</tr>
</tbody>
</table>

| intensity | Coef.       | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----------|-------------|-----------|-------|------|----------------------|
| concentration | 1.930357    | .0409002  | 47.20 | 0.000 | 1.82522   2.035495    |
| _cons     | 1.517857    | .2949358  | 5.15  | 0.004 | .7597003  2.276014    |

Intercept = intensity at zero concentration
. regress intensity concentration

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
<td>F(1, 5) = 2227.53</td>
</tr>
<tr>
<td>Residual</td>
<td>0.936784731</td>
<td>5</td>
<td>0.187356946</td>
<td>Pr(&gt;F) = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
<td>R-squared = 0.9978</td>
</tr>
</tbody>
</table>

| intensity | Coef.    | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----------|----------|-----------|-------|------|-------------------|
| concentration | 1.930357 | 0.0409002 | 47.20 | 0.000 | 1.82522 2.035495  |
| _cons     | 1.517857 | 0.2949358 | 5.15  | 0.004 | 0.7597003 2.276014 |
```
. regress intensity concentration

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
</tr>
<tr>
<td>Residual</td>
<td>.936784731</td>
<td>5</td>
<td>.187356946</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
</tr>
</tbody>
</table>

Number of obs = 7
F( 1, 5) = 2227.53
Prob > F = 0.0000
R-squared = 0.9978
Adj R-squared = 0.9973
Root MSE = 0.43285

<table>
<thead>
<tr>
<th>intensity</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>t</th>
<th>[t]</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>concentration</td>
<td>1.930357</td>
<td>0.0409002</td>
<td>47.20</td>
<td>0.000</td>
<td>1.82522  2.035495</td>
</tr>
<tr>
<td>_cons</td>
<td>1.517857</td>
<td>0.2949358</td>
<td>5.15</td>
<td>0.004</td>
<td>.7597003 2.276014</td>
</tr>
</tbody>
</table>
```

Test of overall model
. regress intensity concentration

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
<td>F( 1, 5) = 2227.53</td>
</tr>
<tr>
<td>Residual</td>
<td>.936784731</td>
<td>5</td>
<td>.187356946</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
<td>R-squared = 0.9978</td>
</tr>
</tbody>
</table>

| intensity | Coef.       | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-----------|-------------|-----------|------|------|----------------------|
| concentration |             |           |      |      |                      |
|             |             |           |      |      |                      |

Variability around the regression line

<table>
<thead>
<tr>
<th>corr</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.035495</td>
<td>2.035495</td>
</tr>
<tr>
<td>3</td>
<td>2.276014</td>
<td>2.276014</td>
</tr>
</tbody>
</table>
Use of the calibration curve

\[ \hat{y} = 1.52 + 1.93x \]

\( \hat{y} \) is the predicted average intensity
\( x \) is the true concentration

\[ \hat{x} = \frac{y - 1.52}{1.93} \]

\( y \) is the observed intensity
\( \hat{x} \) is the estimated concentration
Measurement and Calibration

• Essentially all things we measure are indirect

• The thing we wish to measure produces an observed transduced value that is related to the quantity of interest but is not itself directly the quantity of interest

• Calibration takes known quantities, observes the transduced values, and uses the inferred relationship to quantitate unknowns
Measurement Examples

• Weight is observed via deflection of a spring (calibrated)
• Concentration of an analyte in mass spec is observed through the electrical current integrated over a peak (possibly calibrated)
• Gene expression is observed via fluorescence of a spot to which the analyte has bound (usually not calibrated)
Measuring Variation

• If we do not use any predictor, the variability of $y$ is its variance, or mean square difference between $y$ and the mean of all the $y$’s.

• If we use a predictor, then the variability is the mean square difference between $y$ and its prediction.
\[ MST = (n - 1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

\[ MSE = (n - 2)^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

\[ = (n - 2)^{-1} \sum_{i=1}^{n} (y_i - (a_0 - a_1 x_i))^2 \]

\[ MSR = \left( SST - SSE \right) / 1 \]
concentration vs. intensity Fitted values
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>417.343228</td>
<td>1</td>
<td>417.343228</td>
</tr>
<tr>
<td>Residual</td>
<td>.936784731</td>
<td>5</td>
<td>.187356946</td>
</tr>
<tr>
<td>Total</td>
<td>418.280013</td>
<td>6</td>
<td>69.7133355</td>
</tr>
</tbody>
</table>
Multiple Regression

• If we have more than one predictor, we can still fit the least-squares equations so long as we don’t have more coefficients than data points

• This involves solving the normal equations as a matrix equation
\[ y = xB + \varepsilon \]
\[ Y = XB + E \]

\( n = \) number of data points

\( p = \) number of predictors including constant

\( y \) is \( 1 \times 1 \)

\( x \) is \( 1 \times p \)

\( B \) is \( p \times 1 \)

\( \varepsilon \) is \( 1 \times 1 \)

\( Y \) is \( n \times 1 \)

\( X \) is \( n \times p \)

\( E \) is \( n \times 1 \)
\[ Y = XB + E \]

\( n = \) number of data points

\( p = \) number of predictors including constant

\( B \text{ is } p \times 1 \)

\( Y \text{ is } n \times 1 \)

\( X \text{ is } n \times p \)

\( E \text{ is } n \times 1 \)

\( (Y - XB) \text{ is } n \times 1 \)

\( (Y - XB)'(Y - XB) \text{ is } 1 \times 1, \text{ the SSE} \)
\[(Y - X\hat{B})'(Y - X\hat{B})\text{ is }1 \times 1\text{, the SSE}\]

To minimize this over choices of \(B\), solve

\[
(X'X)\hat{B} = (X'Y)
\]

\[
\hat{B} = (X'X)^{-1}(X'Y)
\]
Linearization of Nonlinear Relationships

• We can fit a curved relationship with a polynomial
• The relationship $f(x) = a_0 + a_1x + a_2x^2$ can be treated as a problem with two predictors
• This can then be dealt with as any multiple regression problem
- Sometimes a nonlinear relationship can be linearized by a transformation of the response and the predictor.
- Often this involves logarithms, but there are many possibilities.
\[ y = \alpha e^{\beta x} \]
\[ \ln y = \ln \alpha + \beta x \]
\[ y = a + \frac{d - a}{1 + (x / c)^b} \]
\[ \frac{d - a}{y - a} = 1 + (x / c)^b \]
\[ \frac{d - y}{y - a} = (x / c)^b \]
\[ \ln \left( \frac{d - y}{y - a} \right) = -b \ln c + b \ln x \]
Intrinsic nonlinearity

• We can still solve the least-squares problem even if $f(x)$ is not linear in the parameters

• We do this by approximate linearization at each step = Gauss-Newton

• There are other, more effective methods, but this is beyond our scope
\[y_i = f(x_i; a_0, a_1, \ldots a_n) + \varepsilon\]
\[y_i = f(x_i; a_{0,j}, a_{1,j}, \ldots a_{n,j}) + f_0(x_i; a_{0,j}, a_{1,j}, \ldots a_{n,j})(a_0 - a_{0,j}) + \ldots\]
\[\quad + f_n(x_i; a_{0,j}, a_{1,j}, \ldots a_{n,j})(a_n - a_{n,j}) + \varepsilon\]
\[y = f(x) + \varepsilon = a_0 \left(1 - e^{-a_1 x}\right) + \varepsilon\]
\[f_0(x_j) = \left(1 - e^{-a_{1,j} x_j}\right)\]
\[f_1(x_j) = a_{0,j} x_j e^{-a_{1,j} x_j}\]
\[y \doteq f_0(x_j) + f_0(x_j)(a_0 - a_{0,j}) + f_1(x_j)(a_1 - a_{1,j})\]
\[y - f_0(x_j) \doteq f_0(x_j) \Delta a_0 + f_1(x_j) \Delta a_1\]
Solve for \(\Delta a_0\) and \(\Delta a_1\) by linear least squares.
Repeat until convergence.