Linear Regression, Linearization, Linear Algebra and All That

Jim Fischer

Oregon Institute of Technology

jim.fischer@oit.edu

April 27, 2018

Chinook Juvenile Salmon

Length (mm)	Weight (g)
35	0.3
36	0.3
38	0.4
39	0.4
38	0.5
39	0.5
41	0.6
⋮	⋮

Table: Chinook Juvenile Salmon. Redwook Creek (CA)

Least Squares Linear Regression

- Find the best fit line to a collection of data (x_i, y_i), and determine a measure of the strength of the linear relationship.

Least Squares Linear Regression

- Find the best fit line to a collection of data (x_i, y_i), and determine a measure of the strength of the linear relationship.
- The equation of the line is:

$$\hat{y}_i = mx_i + b$$

We would like to determine the slope m and intercept b so that the total of the squared residuals is minimized.

Least Squares Linear Regression

- Find the best fit line to a collection of data (x_i, y_i), and determine a measure of the strength of the linear relationship.
- The equation of the line is:

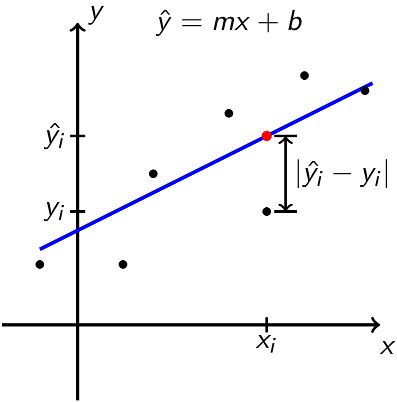
$$\hat{y}_i = mx_i + b$$

We would like to determine the slope m and intercept b so that the total of the squared residuals is minimized.

- $$\text{Minimize:} \quad E(m, b) = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

...

Linear Least Squares Regression



- Find m and b to minimize total least squares error:

$$E(m,b) = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Calculus Solution

- We compute the partial derivatives and set them equal to zero:

$$\begin{aligned} \frac{\partial E}{\partial m} &= -2\sum_{i=1}^{N} x_i(\hat{y}_i - y_i) = 0 \\ \frac{\partial E}{\partial b} &= -2\sum_{i=1}^{N} (\hat{y}_i - y_i) = 0 \end{aligned}$$

- By factoring out the m and b and rearranging terms the system looks like:

$$\begin{aligned} m\sum x_i^2 + b\sum x_i &= \sum x_i y_i \\ m\sum x_i + b\sum 1 &= \sum y_i \end{aligned}$$

- This linear system of equation can be solved, for example, using Cramer's rule:

$$
\begin{aligned}
m &= \quad = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \\
b &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum(x_i y_i)}{n\sum x_i^2 - (\sum x_i)^2}
\end{aligned}
$$

- The solution exists and is unique provided that

$$n\sum x_i^2 - \left(\sum x_i\right)^2 \neq 0$$

Calculus Solution Continued

Using the 2nd Derivative test for functions of two variables it is straight-forward to show that the unique solution corresponds to an absolute minimum

- $\dfrac{\partial^2 E}{\partial m^2} = 2\sum x_i^2 > 0$ and $D(m,b) = 4\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right) > 0$ implies the solution is a local minimum.
- E is continuous over R^2 and has one critical point, therefore the solution is an absolute minimum.

Linear Algebra Approach

- If the data (x_i, y_i) lie perfectly on a line, then:

$$\begin{aligned} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \\ &\vdots \\ y_n &= mx_n + b \end{aligned}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix}$$

Linear Algebra Approach

- If the data (x_i, y_i) lie perfectly on a line, then:

$$\begin{aligned} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \\ &\vdots \\ y_n &= mx_n + b \end{aligned} \qquad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix}$$

- Or more simply as:

$$\vec{\boldsymbol{y}} = A\boldsymbol{c}.$$

Linear Algebra Approach

- If the data (x_i, y_i) lie perfectly on a line, then:

$$\begin{aligned} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \\ &\vdots \\ y_n &= mx_n + b \end{aligned} \qquad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix}$$

- Or more simply as:

$$\vec{\boldsymbol{y}} = A\boldsymbol{c}.$$

- Note that $\vec{\boldsymbol{y}}$ is a vector in the column space of A.

Linear Algebra Approach Continued ...

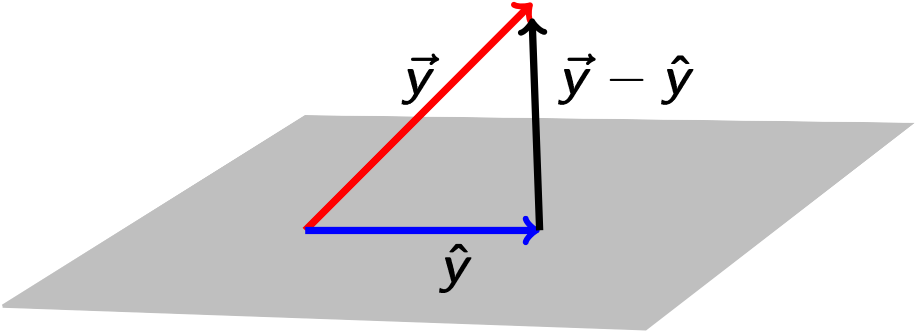
- Typically the data (x_i, y_i) does not lie perfectly on a line, and so we seek a solution $\hat{\boldsymbol{y}}$ (a vector in the column space of A) that is "closest" in some sense.

Linear Algebra Approach Continued ...

- Typically the data (x_i, y_i) does not lie perfectly on a line, and so we seek a solution $\hat{\boldsymbol{y}}$ (a vector in the column space of A) that is "closest" in some sense.
- When we choose "closest" to mean the least total squares error, then the solution turns out to be the orthogonal projection of $\boldsymbol{y}$ onto the column space of A where A is the matrix:

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$$

A Geometric Viewpoint



$$\hat{y} \in \text{Col}(A) = \text{span}\left(\begin{bmatrix} \vec{x} & \vec{1} \end{bmatrix}\right)$$

Linear Algebra Approach Continued ...

- Since $\hat{\boldsymbol{y}} \in \text{Col}(A)$, $\hat{\boldsymbol{y}}$ must be of the form $\hat{\boldsymbol{y}} = A\boldsymbol{c}$.

Linear Algebra Approach Continued ...

- Since $\hat{\boldsymbol{y}} \in \text{Col}(A)$, $\hat{\boldsymbol{y}}$ must be of the form $\hat{\boldsymbol{y}} = A\boldsymbol{c}$.
- Finding the vector $\hat{\boldsymbol{y}}$ such that $\vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}$ is orthogonal to the column space of A boils down to solving the "normal" equation(s):

$$\begin{aligned} A^T(\vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}) &= \vec{\mathbf{0}} \\ A^T A\boldsymbol{c} &= A^T\vec{\boldsymbol{y}} \end{aligned}$$

Linear Algebra Approach Continued ...

- Since $\hat{\boldsymbol{y}} \in \text{Col}(A)$, $\hat{\boldsymbol{y}}$ must be of the form $\hat{\boldsymbol{y}} = A\boldsymbol{c}$.
- Finding the vector $\hat{\boldsymbol{y}}$ such that $\vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}$ is orthogonal to the column space of A boils down to solving the "normal" equation(s):

$$\begin{aligned} A^T(\vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}) &= \vec{\boldsymbol{0}} \\ A^T A\boldsymbol{c} &= A^T\vec{\boldsymbol{y}} \end{aligned}$$

- Since $\hat{\boldsymbol{y}} \in \text{Col}(A)$, $\hat{\boldsymbol{y}}$ must be of the form $\hat{\boldsymbol{y}} = A\boldsymbol{c}$.
- Finding the vector $\hat{\boldsymbol{y}}$ such that $\vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}$ is orthogonal to the column space of A boils down to solving the "normal" equation(s):

$$\begin{aligned} A^T(\vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}) &= \vec{\boldsymbol{0}} \\ A^T A\boldsymbol{c} &= A^T\vec{\boldsymbol{y}} \end{aligned}$$

- Provided $A^T A$ is invertible, the solution is given by:

$$\boldsymbol{c} = (A^T A)^{-1} A^T \vec{\boldsymbol{y}}$$

Other Regression Models: Polynomials

- The Linear Algebra approach nicely lends itself to other models :

Other Regression Models: Polynomials

- The Linear Algebra approach nicely lends itself to other models :
- For example, to find the best (least squares) fit parabola, just change the matrix A by adding a column of squared values:

$$\hat{\boldsymbol{y}} = A\boldsymbol{c} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \\ x_N^2 & x_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Other Regression Models: Polynomials

- The Linear Algebra approach nicely lends itself to other models :
- For example, to find the best (least squares) fit parabola, just change the matrix A by adding a column of squared values:

$$\hat{\boldsymbol{y}} = A\boldsymbol{c} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \\ x_N^2 & x_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

- To find the best fit polynomial, simply add more columns. To fit a kth degree polynomial, the matrix A would have shape $N \times (k+1)$

Other Regression Models: Fourier Series

- To fit a finite fourier-sine series :

$$A = \begin{bmatrix} \sin(x_1) & \sin(2x_1) & +\cdots+ & \sin(kx_1) \\ \sin(x_2) & \sin(2x_2) & +\cdots+ & \sin(kx_2) \\ \vdots & \vdots & \vdots & \vdots \\ \sin(x_N) & \sin(2x_N) & +\cdots+ & \sin(kx_N) \end{bmatrix}$$

Other Regression Models

- In any case where the modeling function has coefficients that appear linearly, we can use the matrix setup.

Other Regression Models

- In any case where the modeling function has coefficients that appear linearly, we can use the matrix setup.
- The normal equations and the solution remain the same:

$$\begin{aligned} A^T A\mathbf{c} &= A^T \vec{\mathbf{y}} \\ \mathbf{c} &= (A^T A)^{-1} A^T \vec{\mathbf{y}} \end{aligned}$$

Other Regression Models

- In any case where the modeling function has coefficients that appear linearly, we can use the matrix setup.
- The normal equations and the solution remain the same:

$$\begin{aligned} A^T A\mathbf{c} &= A^T \vec{\mathbf{y}} \\ \mathbf{c} &= (A^T A)^{-1} A^T \vec{\mathbf{y}} \end{aligned}$$

- The vector of estimated values $\hat{\mathbf{y}}$ is given by:

$$\hat{\mathbf{y}} = A\mathbf{c}$$

A Pair of Nice Results from Linear Algebra:

Theorem (1)

For any matrix A, the null space of A is equal to the null space of $A^T A$.

A Pair of Nice Results from Linear Algebra:

Theorem (1)

For any matrix A, the null space of A is equal to the null space of $A^T A$.

Theorem (2)

For any matrix A, the matrix $A^T A$ is invertible (nonsingular) if and only if the columns of A form a linearly independent set of vectors.

Proof of Theorem 1: $\text{null}(A) = \text{null}(A^T A)$

($\Rightarrow$) Suppose that $\vec{x} \in \text{null}(A)$, so that $A\vec{x} = \vec{0}$. It follows that $A^T A\vec{x} = \vec{0}$ and so $\vec{x}$ is in the null space of $A^T A$.

($\Leftarrow$) Suppose that $\vec{x} \in \text{null}(A^T A)$, so that $A^T A\vec{x} = \vec{0}$. Multiply both sides of this equation by $\vec{x}^T$:

$$\begin{aligned} A^T A\vec{x} &= \vec{0} \\ \vec{x}^T A^T A\vec{x} &= \underset{(1\times n)}{\vec{x}^T} \cdot \underset{(n\times 1)}{\vec{0}} \\ (A\vec{x})^T (A\vec{x}) &= 0 \\ ||A\vec{x}||^2 &= 0 \\ A\vec{x} &= \vec{0} \end{aligned}$$

Therefore $\vec{x}$ is in the null space of A.

Proof of Theorem 2: $A^T A$ is nonsingular if and only if the columns of A form a L.I. set of a vectors

This theorem follows from Theorem 1 and the following two facts:

1. For any square matrix B, the matrix is invertible if and only if its null space is trivial: $\text{null}(B) = \left\{ \vec{\mathbf{0}} \right\}$.
2. For any matrix A, The null space of A is trivial if and only if the columns of A form a linearly independent set of vectors.

Existence and Uniqueness Continued ...

Theorem (The Fundamental Regression Theorem)

Let A be the matrix associated with a linear regression model. If the columns of A form a linearly independent set of $N \times k$ vectors, then there exists a unique solution $\mathbf{c}$ to the least squares problem. Where $\mathbf{c}$ is the column of coefficients for the modeling function (polynomial, fourier, etc.)

$$\mathbf{c} = (A^T A)^{-1} A^T \vec{\mathbf{y}} \quad \text{and} \quad \hat{\mathbf{y}} = A\mathbf{c}$$

Existence and Uniqueness Continued ...

Theorem (Orthogonal Projection)

Let V be a vector space in $\mathbb{R}^n$ and suppose W is a subspace of V with dimension less than n. If $\vec{\mathbf{y}} \in V$ and $\vec{\mathbf{y}} \notin W$, then there exists a unique $\hat{\mathbf{y}} \in W$ such that $\vec{\mathbf{y}} - \hat{\mathbf{y}} \perp W$.

Existence and Uniqueness Continued ...

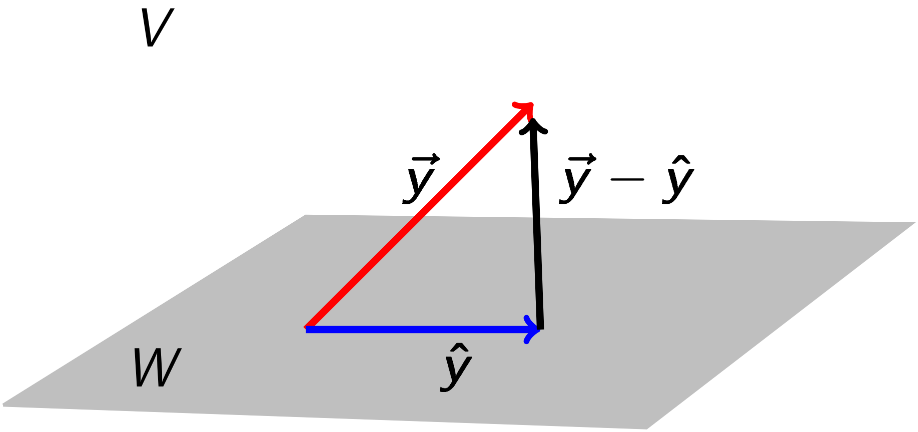
Theorem (Orthogonal Projection)

Let V be a vector space in $\mathbb{R}^n$ and suppose W is a subspace of V with dimension less than n. If $\vec{\mathbf{y}} \in V$ and $\vec{\mathbf{y}} \notin W$, then there exists a unique $\hat{\mathbf{y}} \in W$ such that $\vec{\mathbf{y}} - \hat{\mathbf{y}} \perp W$.

Theorem (Orthogonal $\Leftrightarrow$ Shortest distance)

The orthogonal condition is equivalent to: If $\vec{\mathbf{w}}$ is any vector in W, then $||\vec{\mathbf{y}} - \hat{\mathbf{y}}|| \leq ||\vec{\mathbf{y}} - \vec{\mathbf{w}}||$.

Orthogonal Projection



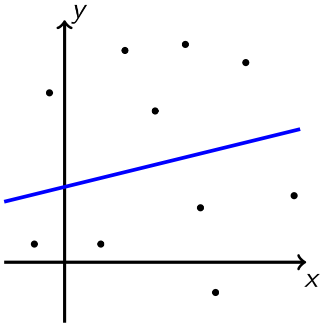
$$\hat{y} \in W = \text{span}\left(\begin{bmatrix} \vec{x_1} & \vec{x_2} & \ldots & \vec{x_m} \end{bmatrix}\right)$$

Correlation and Goodness of Fit

- The regression problem will almost always have a unique solution since the x data will usually be distinct.

Correlation and Goodness of Fit

- The regression problem will almost always have a unique solution since the x data will usually be distinct.
- However, the least squares error is not a good measure of how well the model approximates the dependent variable y. For example, we can find a linear model to approximate data that is clearly not linear:



Correlation and Goodness of Fit

- A good way to see if the model is a good fit is to compute the correlation coefficient r :

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

Correlation and Goodness of Fit

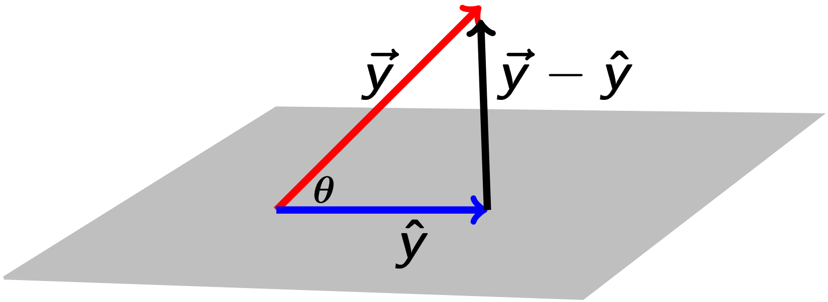
- A good way to see if the model is a good fit is to compute the correlation coefficient r :

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

- Note that this expression could be interpreted as the cosine of an angle by comparing the formula for r and the formula:

$$\boldsymbol{v} \cdot \boldsymbol{w} = ||\boldsymbol{v}||||\boldsymbol{w}|| \cos(\theta)$$

- While the angle shown below is not really the same angle as $\cos^{-1}(r)$, it is not unrelated and makes a reasonable interpretation.



$$\text{Col}(A) = \text{span}\left(\begin{bmatrix} \vec{x} & \vec{1} \end{bmatrix}\right)$$

Comparing r with $\cos(\theta)$

Data Set	r	$\cos^{-1}(r)$	$\cos(\theta)$	θ
Beam Project	0.9953273	5.5°	0.999996	0.2°
First Regression Slide	0.886456	27.6°	0.974923	12.9°
Low r Value Slide	0.017442	89.3°	0.012781	89.0°
Salmon Data (n=1940)	0.982955	10.6°	0.98317	10.5○

Table: Comparing r with the cosine of the angle between $\vec{y}$ and $\hat{y}$.

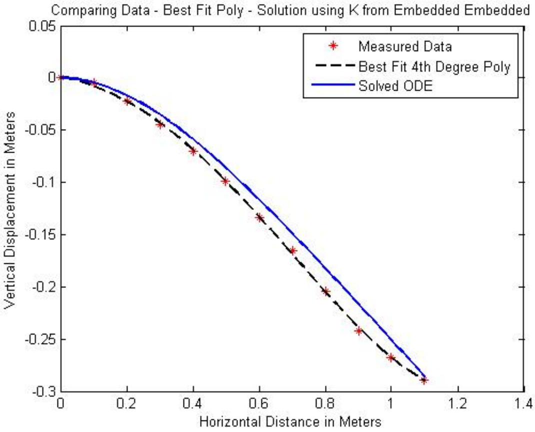
Differential Equations Group Project

- Recall the beam equation:

$$\frac{d^4y}{dx^4} = \frac{w}{EI}$$

- Students in a differential equations class measure deflections of a flexible beam. Two cases: embedded both ends, embedded one end and free on the other.
- They use software such as Excel to create scatter-plot and include a best fit 4th degree polynomial for each case.
- Students differentiate the polynomial 4 times to recover (estimate) the constant $K = \frac{w}{EI}$.
- Next they compare the constants K obtained for the two different set of boundary conditions. They should be close?

Typical Graph Submitted by Students



Sensitivity to Scale when using a Vandermonde matrix

- When completing the DE student project, students were instructed to measure in centimeters. (length of beam was about 200 cm, deflections from 0 to about 10 cm).

Sensitivity to Scale when using a Vandermonde matrix

- When completing the DE student project, students were instructed to measure in centimeters. (length of beam was about 200 cm, deflections from 0 to about 10 cm).
- When using centimeters, the coefficients for the best fit 4th degree polymomial appeared to be inaccurate. In some cases yielding negative values for K. The values obtained for K were extremely small and susceptable to error.

Sensitivity to Scale when using a Vandermonde matrix

- When completing the DE student project, students were instructed to measure in centimeters. (length of beam was about 200 cm, deflections from 0 to about 10 cm).
- When using centimeters, the coefficients for the best fit 4th degree polymomial appeared to be inaccurate. In some cases yielding negative values for K. The values obtained for K were extremely small and susceptable to error.
- When changing the data to meters by simply dividing all measurements by 100, the accuracy of regression results was greatly improved.

Sensitivity to Scale Continued ...

- Using the student data, the instructor created the matrices $A^T A$ and used Matlab to estimate the condition number of the matrices.

Sensitivity to Scale Continued ...

- Using the student data, the instructor created the matrices $A^T A$ and used Matlab to estimate the condition number of the matrices.
- When using the measurements in centimeters, the condition number was very large:

$$cond(A^T A) \approx 10^{18}$$

Sensitivity to Scale Continued ...

- Using the student data, the instructor created the matrices $A^T A$ and used Matlab to estimate the condition number of the matrices.
- When using the measurements in centimeters, the condition number was very large:

$$cond(A^T A) \approx 10^{18}$$

- When using the measurements in meters, the condition number improved:

$$cond(A^T A) \approx 10^5$$

Sensitivity to Scale Continued ...

- Using the student data, the instructor created the matrices $A^T A$ and used Matlab to estimate the condition number of the matrices.
- When using the measurements in centimeters, the condition number was very large:

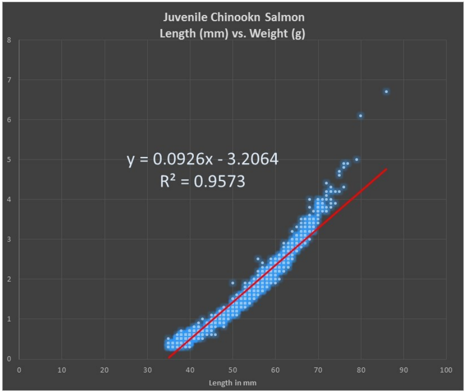
$$cond(A^T A) \approx 10^{18}$$

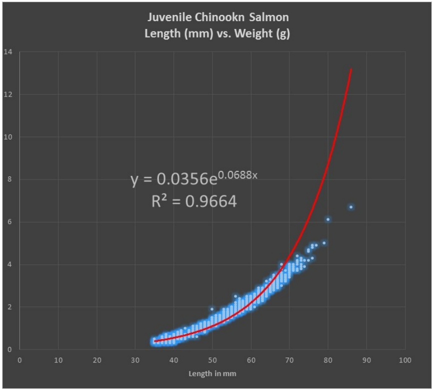
- When using the measurements in meters, the condition number improved:

$$cond(A^T A) \approx 10^{5}$$

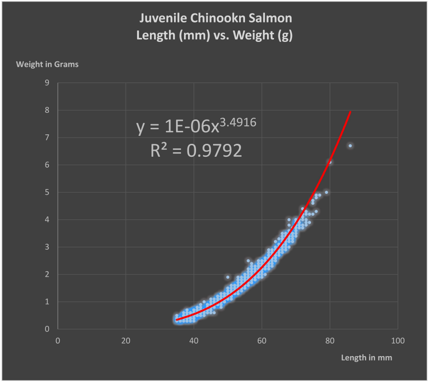
- It is known that computational error can result when the condition number exceeds $\approx 10^{16}$.

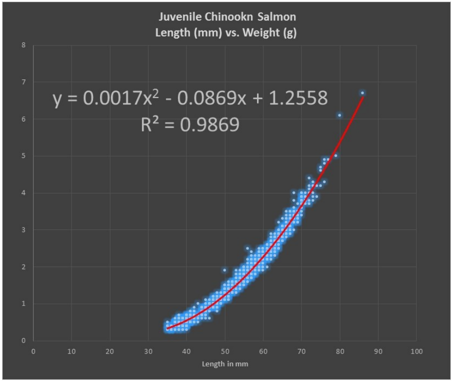
Chinook Juvenile Salmon: Linear Model and Exponential Model





Chinook Juvenile Salmon:
Power Model and Quadratic Model





Tranformations

Desired nonlinear model	Typical y_i	Transformation to linear model
Exponential	$y_i = me^{bx_i}$	$\ln(y_i) = \ln(m) + bx_i$
Power	$y_i = Ax^B$	$\ln(y_i) = ln(A) + B\ln(x_i)$

Table: Some common transformations

Summary

- Together with matrix software like MATLAB, Octave, Python,... the linear algebra approach to regression is an effective/efficient way to model data.
- The "general linear model" provides a framework for many types of curve fitting scenarios.
- The linear algebra approach gives a geometric view that is conceptually "pleasing".
- The linear algebra theorems can be generalized to more abstract settings, e.g. Hilbert Spaces.
- There is a rich amount of statistical analysis associated with linear regression.

References

1 Fogarty T, Waterman G., *Deflection of a Horizontal Beam*, SIMIODE, 2016

2 Michael Sparkman, Biologist, Calif. Dept. of Fish and Wildlife,

3 Professor Randall Paul, OIT Mathematics Dept.

4 Wikipedia: General Linear Model/ Generalized Linear Model

The End

Thanks for Listening!