Data Analysis, Statistics, Machine Learning

Leland Wilkinson
Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai
leland.wilkinson@gmail.com
Predicting

Most statistical prediction models take one of two forms

\[ y = \Sigma_j(\beta_j x_j) + \varepsilon \]  \hspace{1cm} \text{(additive function)}
\[ y = f(x_j, \varepsilon) \]  \hspace{1cm} \text{(nonlinear function)}

The distinction is important

The first form is called an additive model
The second form is called a nonlinear model
Additive models can be curvilinear (if terms are nonlinear)
Nonlinear models cannot be transformed to linear

Examples of linear or linearizable models are

\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \]
\[ y = \alpha e^{\beta x} + \varepsilon \]

Examples of nonlinear models are

\[ y = \beta_1 x_1 / \beta_2 x_2 + \varepsilon \]
\[ y = \log(\beta_1 x_1) + \varepsilon \]
Predicting

Regression predicts a set of values on a variable $y$ from values on one or more $x$ variables.

This is done by fitting a mathematical function that, for any value(s) on the $x$ variable(s), yields the most probable value of $y$.

Simple linear model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Estimates are

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

The badness or loss of this prediction in a sample of values is represented by the discrepancies between the $y$ values and their corresponding predicted values.

$$\text{loss} = \sum_{i=1}^{n} (y - \hat{y})^2$$
Predicting

Ordinary Least Squares (OLS)

Legendre (1805)

“Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of making the sum of the squares of the errors a minimum. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.” (translation by Stigler, 1986)
Predicting

Regression

Francis Galton (1887)

He noticed that the blue lines were the same length as the red ones

That is, the line best predicting $Y$ from $X$ “regressed” away from the major axis
Predicting

Regression

Francis Galton (1887)

His data weren’t as clean and linear as he imagined, but that didn’t matter
Predicting

Regression

Francis Galton (1887)

That’s because he aggregated over different sources

Wachsmuth, Wilkinson & Dallal, 2003
Predicting

Estimating via Ordinary Least Squares
Predicting

Estimating via Ordinary Least Squares

For intercept \((b_0)\) and slope \((b_1)\), we could use calculus

The way we did when we used maximum likelihood to estimate mean and sd

In this case, we want to minimize the sum of squared residuals \(SSE\)

Given
\[
y_i = b_0 - b_1 x_i + e_i
\]

We sum the \(e_i\) to get \(SSE\)

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

Compute the partial derivatives with respect to \(b_0\) and \(b_1\)
Set these derivatives to zero (where the minimum \(SSE\) exists)
Solve the resulting simultaneous equations
This is what Legendre originally did

But there is an easier way
Predicting

Estimating via OLS (we’ll use matrices)

\[
Y = XB + E
\]

\[
XB \perp E
\]

\[
X'Y = X'XB + X'E
\]

\[
(X'X)^{-1}X'Y = (X'X)^{-1}(X'X)B
\]

\[
(X'X)^{-1}X'Y = B
\]

\[
E = Y - XB
\]

\[
Y = \begin{bmatrix}
3 \\
2 \\
2 \\
4 \\
4 \\
6 \\
5
\end{bmatrix}
\]

\[
X = \begin{bmatrix}
1 & 2 \\
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4 \\
1 & 5 \\
1 & 6
\end{bmatrix}
\]

\[
B = \begin{bmatrix}
1.25 \\
0.75
\end{bmatrix}
\]

\[
E = \begin{bmatrix}
0.25 \\
0.00 \\
-0.75 \\
0.50 \\
-0.25 \\
1.00 \\
-0.75
\end{bmatrix}
\]
Predicting

Estimating the regression model parameters

How do we know we minimized the error sum of squares?

\[ X \cdot Y = \|X\|\|Y\| \cos \theta \]

\[ \cos \theta = \frac{X \cdot Y}{\|X\|\|Y\|} \quad \text{Pearson correlation coefficient (if data are centered)} \]

length of \( E \) = \( \|E\| \)

Y and X are fixed (because of \( \theta \) and \( \|X\| \) and \( \|Y\| \))

Shortest distance from point Y to line X is \( \|E\| \) \quad QED
Predicting

Estimation (assuming we sampled from a population)

\[
B = \begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1
\end{bmatrix}
\]

is a least-squares estimator
  unbiased (assuming residuals are homogeneous)
  smallest variance among unbiased estimators
  Best Linear Unbiased Estimator (BLUE)

If the residuals are normally distributed
  \( B \) is a maximum likelihood estimator
  We can do classical statistical tests on estimates of parameters
Predicting

Estimating via OLS

Two predictors (same formulas)

\[ Y = XB + E \]
\[ XB \perp E \]
\[ X'Y = X'XB + X'E \]
\[ (X'X)^{-1}X'Y = (X'X)^{-1}(X'X)B \]
\[ (X'X)^{-1}X'Y = B \]
\[ E = Y - XB \]

\[
Y = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 4 \\ 4 \\ 6 \\ 5 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 4 \\ 1 & 2 & 2 \\ 1 & 3 & 1 \\ 1 & 4 & 2 \\ 1 & 5 & 4 \\ 1 & 6 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 0.897 \\ 0.746 \\ 0.135 \end{bmatrix} \quad E = \begin{bmatrix} 0.206 \\ -0.183 \\ -0.659 \\ 0.730 \\ -0.151 \\ 0.833 \\ -0.778 \end{bmatrix}
\]
Predicting

The two-predictor vector space

\[ Y = XB + E \]
The two-predictor vector space

It is possible for $y$ not to be correlated much with either $X_1$ or $X_2$ yet be highly correlated with the linear combination of $X_1$ and $X_2$. So don’t throw out predictors by looking at their correlations with the dependent variable.
Predicting

The two-predictor vector space

Z is not significantly related to X
Z is not significantly related to Y
But the multiple correlation of Z with X and Y is almost 1!

Copyright © 2016 Leland Wilkinson
Predicting

Standardized Regression Coefficients (*Beta* Weights)
Standardize variables, then compute regression
Removes scales from consideration of size of coefficients
Social scientists love this stuff

“Why then are correlation coefficients so attractive? Only bad reasons seem to come to mind. Worst of all, probably, is the absence of any need to think about units for either variable. Given two perfectly meaningless variables, one is reminded of their meaninglessness when a regression coefficient is given, since one wonders how to interpret its value.”

Tukey (1969)
Predicting

Sums of Squares

\[ SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 \]
regression sum of squares (explained)

\[ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]
error sum of squares (unexplained)

\[ SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \]
total sum of squares
Predicting

Goodness of fit

Pearson correlation

\[ \rho_{X,Y} = \frac{COV(X,Y)}{\sqrt{VAR(X)VAR(Y)}} \]

\[ \hat{\rho}_{X,Y} = r_{X,Y} = \frac{X \cdot Y}{\|X\|\|Y\|} \]

(if \( X \) and \( Y \) are centered)

Multiple correlation (sqrt of coefficient of determination)

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

regression sum of squares / total sum of squares

\[
R_{xx} = \begin{bmatrix}
1 & r_{x1x2} & \cdots & r_{x1xp} \\
r_{x2x1} & 1 & \cdots & r_{x2xp} \\
\vdots & \vdots & \ddots & \vdots \\
r_{xp1} & r_{xp2} & \cdots & 1
\end{bmatrix}
\]

\[
R_{yx} = \begin{bmatrix}
r_{yx1} & r_{yx2} & \cdots & r_{yxp}
\end{bmatrix}
\]

\[ R^2 = R_{yx} R_{xx}^{-1} R_{xy} \]

≈ sum of squared correlations with \( Y \) / correlations among \( X \)
Predicting

Inference

F-statistic (test of significance for overall prediction)

\[ F_{p,n-p-1} = \frac{R^2/p}{(1 - R^2)/(n - p - 1)} \]

Confidence intervals on regression coefficients

\[ c_j = \text{diag}(X'X)_j^{-1} \]

\[ s = \sqrt{\frac{SSE}{n - p - 1}} \]

\[ CI = (\hat{\beta} \pm t_{n-p-1}^{\alpha/2} s \sqrt{c_j}) \]
Predicting

An interesting application of Ordinary Least Squares

We have

- $Z_{n \times p}$: $n$ fixed points in $p$ dimensions (cell towers, MDS or other configuration, ...)
- $d_n$: a new point $y$'s estimated distance to each point in $Z$

Solve for vector $b_p$ specifying location of new point
Predicting

An interesting application of Ordinary Least Squares

We have

$$Z_{n \times p} : n \text{ fixed points in } p \text{ dimensions (cell towers, MDS or other configuration, ...)}$$

$$d_n : \text{a new point } y \text{'s estimated distance to each point in } Z$$

Solve for vector $$b_p$$ specifying location of new point

$$s = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \sum_{k=1}^{p} (Z_{ik} - Z_{jk})^2 \quad \text{(a scaling constant)}$$

$$y_m = d_i^2 - d_j^2 + s, \quad j = 1, \cdots, i - 1, \quad i = 2, \cdots, n, \quad m = (i - 1)(i - 2)/2 + j$$

$$X_{mk} = 2(Z_{ik} - Z_{jk}), \quad j = 1, \cdots, i - 1, \quad i = 2, \cdots, n, \quad m = (i - 1)(i - 2)/2 + j$$

$$b = (X'X)^{-1}X'y$$
Predicting

OLS Assumptions

The true model is linear in the parameters
The $X$ variables are not random and are measured without error
The $X$ variables are not collinear
Residuals are uncorrelated/independent
The residuals have constant variance (homoscedasticity)
If $t$ and $F$ test statistics are computed
  Residuals must be Normally distributed
  Random sample from a population
Predicting

Evaluating OLS assumptions

The true model is linear in the parameters

Use LOESS to put a smooth through residual plot
Predicting

Transformations

Dealing with nonlinearity

Wilkinson, Blank, & Gruber (1996)
Predicting

Transformations

Dealing with nonlinearity
Tukey-Mosteller bulging rule

\[ x^* = x^p \]
\[ y^* = y^p \]
Predicting

Evaluating OLS assumptions

The X variables are not random and are measured without error
There is no simple test or graphic for this
Know the source of your measurements
If there is measurement error, you can use errors-in-the-variables methods
Or, you can just forgeddaboutit, which is what most of the world does

Who cares if your coefficient estimates are biased?
They are usually biased downward, so no harm done
Just don't show them to a social scientist
Social scientists love latent variable models
They are Platonists
Predicting

Evaluating OLS assumptions

The $X$ variables are not collinear

Values of Variance Inflation Factor (VIF) above 10 are worrisome

\[ VIF = \frac{1}{1 - R_i^2} \]

$R^2 = 0.995$

$R^2 = 0.97$
Predicting

Evaluating OLS assumptions

Residuals are uncorrelated/independent

An ACF plot can help spot violations, but there are other types of dependencies

- Economists spend all their time worrying about this
- And for good reason – serial dependence is more toxic than outliers
- Don’t even THINK of using ordinary linear regression (trend line) on time series data
Predicting

Evaluating OLS assumptions

The residuals have constant variance (homoscedasticity)

Try transformations (usually logging works for a power model)

Or use one of the heteroscedasticity corrections (MacKinnon-White, etc.)
Predicting

Transformations
Dealing with heteroscedasticity
log-log transformation
Predicting

Evaluating OLS assumptions

If $t$ and $F$ test statistics are computed
- Residuals must be Normally distributed
- Random sample from a population
- Forget about tests for normality; they are worthless

![Residuals vs. Normal Order Statistics](chart1.png)

![Histogram of Residuals](chart2.png)
Predicting

Evaluating OLS assumptions

If $t$ and $F$ test statistics are computed

Residuals must be Normally distributed
Predicting

Evaluating OLS assumptions

Leverage

Diagonal of the hat matrix (puts the hat on the parameters)

\[ H = X(X'X)^{-1}X' \]

Cook’s D

Measures the decrement in prediction by removing an observation

A type of leverage measure that is transformable to an approximate \( F \)
Predicting

Evaluating OLS assumptions

Leverage

Deletions
- Yes
- No
Predicting

Model Selection

Partial Residual Plots

Each plot consists of sets of residuals plotted against each other.

One set, on the vertical axis, consists of the residuals from regressing $y$ on all the $X$ (predictor) variables except for the predictor on the horizontal axis.

The other set, on the horizontal axis, consists of the residuals from regressing $X_i$ on all the other $X$ variables.

The result is a plot which shows you how each $X$ variable is related to the $Y$ variable when all the other $X$ variables are taken into account.

Partial residual plots have several useful features:

- The slope of the line in the plot is the partial regression regression coefficient corresponding to the predictor in the plot.
- A line with a steep slope and nice looking residuals is a sign that a predictor belongs in the model.
- The residuals from this line are the same as the residuals from regressing $Y$ on all the predictors.
- The plot helps you to judge whether the conditional relationship is linear or nonlinear.
- Extreme values on the horizontal axis help you to identify high leverage points.
Predicting

Model Selection

Partial Residual Plots

\[ \text{time} = -10.679 + 6.697 \times \text{distance} + 0.008 \times \text{climb} \]
Predicting

Model Selection

- Forward Selection Regression
- Backward Elimination Regression
- Stepwise Regression
- All Possible Subsets Regression
Predicting

Model Selection

Wilkinson (1979) has discussed the case in which a subset of \( k \) predictors is to be chosen, where \( 1 < k < m \), and has provided tables of the upper 95\(^{th} \) and 99\(^{th} \) percentage points of the sample \( R^2 \) distribution in forward selection based on simulations (other tables and discussions of this problem can be found in Hocking, 1983; Rencher & Pun, 1980; and Wilkinson & Dallal, 1982). These tables are more conservative than the usual \( F \) tables. For example, with \( N = 35 \) and \( \alpha = .05 \), if all four members of a set of predictor variables are to be included in the regression equation, it is appropriate to use the standard \( F \) test to test \( R^2 \) for significance. When this is done, it is found that the sample \( R^2 \) has to exceed .26 in order to reject the hypothesis that the population multiple correlation coefficient is 0. However, if the four predictors are to be selected from a larger set of 20 predictors by a forward selection procedure, according to Wilkinson’s tables, the sample \( R^2 \) must exceed .51 in order to reject the null hypothesis. Many researchers do not seem to be aware of this problem; for a sample of 66 published papers that reported significant forward selection analyses according to the usual \( F \) tests, Wilkinson found that 19 were not significant when his tables were used.

### Predicting

#### Model Selection

Akaike Information Criterion (AIC)

Longley Data: Predicting TOTAL (RSQ = .995)

<table>
<thead>
<tr>
<th></th>
<th>Constant</th>
<th>Deflator</th>
<th>GNP</th>
<th>Unemployment</th>
<th>Armforce</th>
<th>Population</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3,482,258.635</td>
<td>890,420.384</td>
<td>-5,496,529.488</td>
<td>-1,467,987.781</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEFLATOR</td>
<td>15.062</td>
<td>84.915</td>
<td>-177.029</td>
<td>207.153</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNP</td>
<td>-0.036</td>
<td>0.033</td>
<td>-0.112</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNEMPLOY</td>
<td>-2.020</td>
<td>0.488</td>
<td>-3.125</td>
<td>-0.915</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARMFORCE</td>
<td>-1.033</td>
<td>0.214</td>
<td>-1.518</td>
<td>-0.549</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POPULATN</td>
<td>-0.051</td>
<td>0.226</td>
<td>-0.563</td>
<td>0.460</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIME</td>
<td>1,829.151</td>
<td>455.478</td>
<td>798.788</td>
<td>2,859.515</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predicting

Model Selection ($m$ submodels)

$$AIC = 2(p + 2) - 2 \log(L)$$

$$\Delta AIC_k = AIC_k - AIC_{\text{min}}$$

$$L_k = \exp(-\Delta AIC / 2)$$

$$W_k = L_k / \sum_{i=1}^{m} L_i$$

<table>
<thead>
<tr>
<th>RSQ</th>
<th>Increment</th>
<th>AIC Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNP&amp;UNEMPLOY&amp;ARMFORCE&amp;TIME 0.995</td>
<td>0.051</td>
<td>0.503</td>
</tr>
<tr>
<td>UNEMPLOY&amp;ARMFORCE&amp;TIME 0.993</td>
<td>0.013</td>
<td>0.192</td>
</tr>
<tr>
<td>UNEMPLOY&amp;ARMFORCE&amp;POPULATN&amp;TIME 0.995</td>
<td>0.027</td>
<td>0.167</td>
</tr>
<tr>
<td>GNP&amp;UNEMPLOY&amp;ARMFORCE&amp;POPULATN&amp;TIME 0.995</td>
<td>0.053</td>
<td>0.050</td>
</tr>
<tr>
<td>DEFLATOR&amp;GNP&amp;UNEMPLOY&amp;ARMFORCE&amp;TIME 0.995</td>
<td>0.073</td>
<td>0.049</td>
</tr>
<tr>
<td>DEFLATOR&amp;UNEMPLOY&amp;ARMFORCE&amp;POPULATN&amp;TIME 0.995</td>
<td>0.028</td>
<td>0.020</td>
</tr>
<tr>
<td>DEFLATOR&amp;UNEMPLOY&amp;ARMFORCE&amp;TIME 0.993</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>DEFLATOR&amp;GNP&amp;UNEMPLOY&amp;ARMFORCE&amp;POPULATN&amp;TIME 0.995</td>
<td>0.995</td>
<td>0.002</td>
</tr>
<tr>
<td>UNEMPLOY&amp;TIME 0.982</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>GNP&amp;UNEMPLOY&amp;ARMFORCE 0.985</td>
<td>0.048</td>
<td>0.001</td>
</tr>
<tr>
<td>GNP&amp;UNEMPLOY 0.981</td>
<td>0.047</td>
<td>0.000</td>
</tr>
<tr>
<td>GNP&amp;ARMFORCE&amp;POPULATN 0.984</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>DEFLATOR&amp;UNEMPLOY&amp;TIME 0.983</td>
<td>0.012</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Predicting

Model Selection ($m$ submodels)
Predicting

Other approaches to multicollinearity

Regularization

Ridge regression (Hoerl, 1962)

Minimize

$$\sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

So, \( \hat{\beta} = (X'X + \lambda I)^{-1} X'y \)

We inflate the diagonal to reduce the influence of the off-diagonal covariances

Ridge regression shrinks the estimates toward zero, introducing bias

But this reduces the variance of the estimates
Predicting

Other approaches to multicollinearity

Regularization

LASSO (Least Absolute Value Shrinkage and Selection Operator)
Minimize

\[ \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} | \beta_j | \]

Solution requires iteration

Allows some coefficients to go to zero with others nonzero

Efron and Tibshirani use Least Angle Regression with Shrinkage (LARS)

Similar to stepwise regression
Predicting

Logistic Regression

OLS estimates
Predicting

Logistic Regression

LR estimates

Copyright © 2016 Leland Wilkinson
Predicting

Logistic Regression
Model and estimation

\[ p(x; \beta) = \frac{e^{x\beta}}{1 + e^{x\beta}} = \frac{1}{1 + e^{x\beta}} \]

\[ \text{logit}(p) = \log \frac{p(x)}{1 - p(x)} = x\beta \]

\[ L(\beta; x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \]

\[ l(\beta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \text{logit}_i - \log(1 + p_i) \]

Must use iterative optimization to find maximum
Different model for more than two categories

\[ x = x_0, \ldots, x_p \]
\[ \beta = \beta_0, \ldots, \beta_p \]

log-odds

likelihood (Binomial)

log-likelihood
Predicting

Poisson Regression (log-linear model)

Model and estimation

\[ E[Y|\mathbf{x}] = e^{\mathbf{x} \beta} \]

Poisson distribution has only 1 parameter

\[ p(y; \mathbf{x}, \beta) = \frac{e^{y \mathbf{x} \beta} e^{-e^{\mathbf{x} \beta}}}{y!} \]

\( y \) is integer valued

\[ L(\beta; y, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \frac{e^{y_i \mathbf{x}_i \beta} e^{-e^{\mathbf{x}_i \beta}}}{y_i!} \]

likelihood

\[ l(\beta; y, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \left\{ (y_i \mathbf{x}_i \beta) - e^{\mathbf{x}_i \beta} - \log(y_i!) \right\} \]

log-likelihood

(don’t need the yellow term
In order to maximize)

Must use iterative optimization to find maximum

Copyright © 2016 Leland Wilkinson
Predicting

Poisson Regression

OLS

Poisson
Predicting

Generalized Linear Models (GLM)
Nelder & Wedderburn (1972)
Not to be confused with
  General Linear Model (GLM) for OLS with or without dummy variables
  Generalized Least Squares (GLS) for dealing with heteroscedasticity

Estimation done through Iteratively Reweighted Least Squares (IRWLS)

\[ x\beta = g(E[Y]) \quad \text{link function} \]
\[ E[Y] = \mu = g^{-1}(x\beta) \quad \text{modeling mean of } Y \text{ through inverse of link function} \]

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Link Function Name</th>
<th>Link Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
<td>( x\beta = \mu )</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit</td>
<td>( x\beta = \log \left( \frac{\mu}{1 - \mu} \right) )</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
<td>( x\beta = \log(\mu) )</td>
</tr>
</tbody>
</table>
Predicting

Nonlinear Regression

Functions of the form $y = f(x + \epsilon)$

Left is linearizable by transformation, right is intrinsically nonlinear

$y = e^{\beta_0 + \beta_1 x + \epsilon}$

$y = e^{\beta_0 + \beta_1 x + \epsilon}$
Predicting

Nonlinear Regression

\[ E[Y] = e^{a + b/x + c \log(x)} \]

Minimize SSE (Newton, Quasi-Newton, Metropolis, ...)
Magic, right?

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Lower95%</th>
<th>Upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>12.547</td>
<td>1.134</td>
<td>10.324</td>
</tr>
<tr>
<td>c</td>
<td>-1.934</td>
<td>0.227</td>
<td>-2.379</td>
</tr>
</tbody>
</table>
Predicting

Nonlinear Regression

Things to worry about

Don’t waste your time with $R^2$
  There are all sorts of definitions
  All are nonsense
  The bad ones are ridiculously large
  Look at sum of squared errors instead

Don’t hunt around for nonlinear equations that work
  That’s a fishing expedition
  The equation you choose should be driven by theory and the domain it applies to

Half the time your iterative fitting method will croak
  That’s a sign that there’s a local minimum
  Or your equation is wrong
  Or your starting values are wildly off-target
  Don’t waste your time looking for another computer program

Nonlinear programs are notoriously finicky
  Your data are probably crap or your model is ridiculous

Look at the fit before you examine any statistics
  Most of the time you have one predictor and one dependent variable
  So look at the fitted equation
  Ignore the fit statistics and tests of significance until it looks good

EVERYTHING fits
  Keep in mind that almost any bad model will look pretty good
  If it doesn’t make theoretical sense, don’t trust it
Predicting

References
