Introduction to Regression

Brad Jones\textsuperscript{1}

\textsuperscript{1}Department of Political Science
University of California, Davis

January 15, 2010
Today: Preliminaries
Preliminaries

- Regression Model and Conceptual Regression
- More preliminaries . . . not all of which is review.
Regression

- The basic model (in scalar form):

\[ Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + \epsilon_j \]  \hspace{1cm} (1)

- Or

\[ Y_j = \beta_0 + \sum_{i=1}^{k} \beta_j X_{ij} + \epsilon_j \]  \hspace{1cm} (2)

- Estimation problem arises because of the error term.

- Error? It summarizes our ignorance … the inherent unpredictability of the world.

- Basic idea here: find a way to minimize this quantity.
Regression

- Implications of the model.
- Fitted value is given by $\hat{Y}$.
- $Y$ is a function of $X_k$.
- The “residual”:

$$ r_j = Y_j - \beta_0 + \sum_{i=1}^{k} \beta_j X_{ij} $$  \hspace{1cm} (3) $$

- Or . . .

$$ r_j = Y_j - \hat{Y}_j $$  \hspace{1cm} (4) $$

- “Observed minus predicted.”
Consider bivariate setting: 1 regressor and 1 response variable.

Working example: Percentage for Obama as a function of vote on Prop. 8

NB: a pedagogical example only.

Begin with a scatterplot:
Artwork
Regression

- Suppose we wanted to predict Obama county vote and the only information we had was county vote on prop. 8?
- What does the scatterplot reveal?
- The question is, how do we derive a prediction . . . that is, come up with some concrete number.
- Let’s start with linear regression:

\[ \hat{O}_c = \beta_0 + \beta_1 P8_c \]  

(5)

- A simple fit might be to “draw a line” through the cloud of points.
- `lm(obamapercent ~ proportionforprop8)`
Artwork
Regression

- Red line is the fitted regression function and the red dots correspond to the point prediction.
- San Francisco county: Obama vote: 84 percent; Prop. 8: 23.5 percent.
- The fitted line gives a prediction of 81.98 or 82 percent.
- Simple linear estimator:
  \[ \hat{O}_c = 102.33 - .866(P8_c) \]
- Substitute the value of \( P8_c \) and you obtain the predicted value for San Francisco.
Essential Geometry

\[ r_j = Y_j - \hat{Y}_j \]  \hfill (6)

- The residual is the “signed vertical distance between the point and the line.” (Fox, p. 78)
- For SF, the residual is about 2 (84 – 82).
- The fitted line “underpredicts” the true value for this observation.
- Thus, the sign of the residual informs us about its location relative to the fitted line.
- The question is: where are you going to put the line?
If you think of the residual as “error”, one might want to minimize $\sum r_j$.

Obvious problem: if a line passes through the point of means $(\bar{X}, \bar{Y})$, $\sum r_j = 0$.

The fairly obvious proof of this is given by Fox in equation 5.2.

Simple minimization won’t work. So what about:

$$\sum |r_j|$$

Gives rise to a “least absolute value” estimator or LAV regression.
Essential Geometry

- A more tractable approach:
  \[
  \min \sum r_j^2
  \]

- This gives rise to a “least squares” estimator or LS regression.

- The criterion:
  \[
  \sum_{i=1}^{n} r_j^2 = \sigma(Y_j - \beta_0 - \beta X_i)^2
  \]

- The LS solution minimizes the sum of the squared residuals over all observations.
S(β₀, β₁) = \sum_{i=1}^{n} r_j^2

Taking the partial derivatives of the sum-of-squares with respect to the regression parameters gives:

\frac{\partial S(β₀, β₁)}{\partial β₀} = \sum (-1)(2)(Y_j - β₀ - β₁ X_i)

\frac{\partial S(β₀, β₁)}{\partial β₁} = \sum (-X_i)(2)(Y_j - β₀ - β₁ X_i)

Set these to 0 gives the *normal equations* (a system of simultaneous linear equations):

β₀ n + β₁ \sum X_i = \sum Y_j

β₀ \sum X_i + β₁ \sum X_i^2 = \sum X_i Y_i
Least Squares Solution

- Solving the normal equations gives:

\[
\beta_0 = \overline{Y} - \beta_1 \overline{X} \\
\beta_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\
= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}
\]

- Note the regression function passes through the point of means. Also note that since \( \sum r_j = 0 \), then \( \sum X_ir_j = 0 \) (see Fox, p. 80; work through this yourself).

- It is also the case that \( \sum \hat{Y}r_j = 0 \).

- This implies the LS residuals are *uncorrelated with the X and \( \hat{Y} \).*
Least Squares Solution

- The system is identified so long as: $X$ is not a constant and the model is of full rank.

- Interpretation of the regression coefficient (from previous example): a 1 percent change in Prop 8 vote is associated with a .87 decrease in *expected* vote share for Obama. The sign is negative.

- Suppose the coefficient was 1?

- Multiple regression is a straightforward extension.
Multiple Regression

Regression model:

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \]

For two variable regression, differentiate

\[ \sum \hat{r}_j^2 = \sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 \]

partially with respect to the three unknown parameter estimates.
Multiple Regression

This gives

\[
\frac{\partial \sum \hat{r}_i^2}{\partial \hat{\beta}_0} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-1),
\]

\[
\frac{\partial \sum \hat{r}_i^2}{\partial \hat{\beta}_1} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{1i}),
\]

\[
\frac{\partial \sum \hat{e}_i^2}{\partial \hat{\beta}_2} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{2i})
\]
Regression

When set to 0 and rearranging terms produces the normal equations:

\[
\begin{align*}
\bar{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \\
\Sigma Y_i X_{1i} &= \hat{\beta}_0 \Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i} \\
\Sigma Y_i X_{2i} &= \hat{\beta}_0 \Sigma X_{2i} + \hat{\beta}_1 \Sigma X_{2i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i}.
\end{align*}
\]
These equations can be rewritten yet again in terms of the parameter estimates:

\[
\hat{\beta}_1 = \frac{\sum(X_1 - \bar{X}_1)(Y_i - \bar{Y})(\sum(X_2 - \bar{X}_2)^2) - \sum(X_2 - \bar{X}_2)(Y_i - \bar{Y})(\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\sum(X_1 - \bar{X}_1)^2\sum(X_2 - \bar{X}_2)^2 - (\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2}
\]

\[
\hat{\beta}_2 = \frac{\sum(X_2 - \bar{X}_2)(Y_i - \bar{Y})(\sum(X_1 - \bar{X}_1)^2) - \sum(X_1 - \bar{X}_1)(Y_i - \bar{Y})(\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\sum(X_1 - \bar{X}_1)^2\sum(X_2 - \bar{X}_2)^2 - (\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2}
\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2
\]
Several things are worth noting.

First, this system is identified because there are three equations and three unknowns.

Under pretty general conditions, you will get estimates using the least squares approach.

Indeed, the only conditions under which you will not get estimates are when $X_1$ and $X_2$ are perfectly correlated. To see this, note that if

$$\Sigma(X_1 - \bar{X})^2 \Sigma(X_2 - \bar{X})^2 = (\Sigma(X_1 - \bar{X})(X_2 - \bar{X}))^2,$$

then it must be the case that one variable is a linear function of another variable.
The previous expression can be rewritten in terms of the correlation coefficient,

\[ r_{X_1,X_2} = \frac{\sum (X_1 - \bar{X})(X_2 - \bar{X})}{\sqrt{\sum (X_1 - \bar{X})^2 \sum (X_2 - \bar{X})^2}}, \]

This makes it easy to see the condition under which \( r = 1 \) (the product mean deviations of \( xy \) is equal to the product of the standard deviations of \( x \) and \( y \).

What would produce this?

Any time \( X_1 \) is a perfect linear combination of \( X_2 \) (or vice versa), this will occur. Example: \( X_2 = X_1 \times K \) (where \( K \) is a constant).
Regression

- Another situation where nonestimation of the parameters will occur is when the standard deviation of one of the regressors is 0—that is, one of the $X_k$ is constant.
- This is trivial: seldom will you run into this problem.
- However, suppose that the variance on one of the variables is very small?
- The implication is that there will be a limited amount of information upon which the coefficient estimate will be based.
- So even though you can estimate it, the coefficient will be highly unreliable.
Pedagogical example: simulated data.

The correlation between the two independent variables is 0.0002 (nearly 0, but not quite).

The normal equations use information regarding the variances and covariances to compute the regression coefficients.

The quantities on the next slide are necessary in order to solve for the regression estimates:
The quantities:  
\[ \bar{Y} = .4918 \]  
\[ \bar{X}_1 = .4969 \]  
\[ \bar{X}_2 = .5095 \]  
\[ \Sigma (X_1 - \bar{X})^2 = 8.7018 \]  
\[ \Sigma (X_2 - \bar{X})^2 = 8.2850 \]  
\[ \Sigma (X_1 - \bar{X})(X_2 - \bar{X}) = .0018 \]  
\[ \Sigma (X_1 - \bar{X})(Y - \bar{Y}) = -.1432 \]  
\[ \Sigma (X_2 - \bar{X})(Y - \bar{Y}) = 1.1040 \]  
\[ (\Sigma (X_1 - \bar{X})(X_2 - \bar{X}))^2 = .0000033 \]
Regression

Compute $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{(-.1432)(8.2850) - (1.1040)(.0018)}{(8.7018)(8.2850) - (.0000033)}$$

$$\hat{\beta}_1 = -.01649;$$

Now, $\hat{\beta}_2$:

$$\hat{\beta}_2 = \frac{(1.1040)(8.7018) - (.1432)(.0018)}{(8.7018)(8.2850) - (.0000033)}$$

$$\hat{\beta}_2 = .1333;$$

Now $\hat{\beta}_0$:

$$\hat{\beta}_0 = .4918 + .01649(.4969) - .1333(.5095)$$

$$\hat{\beta}_0 = .43208.$$
Regression

- Taken together, the three parameter estimates gives us the following model (rounding to the third decimal):

\[
\hat{Y} = 0.43 - 0.016X_1 + 0.133X_2.
\]

- Computation is not magical, mystical, or mysterious.
- It only is a function of the means of $X_1$, $X_2$, and $Y$, deviations from the mean, and products of deviations from the mean.
- The regression coefficients are a function of all the pieces of information (all of the data).
- The coefficients may also be referred to as partial regression coefficients because they give the direct “effect” of $X_1$ on $Y$, net any effects from $X_2$.
- Note what is happening with $X_2$ in the estimation of $\hat{\beta}_1$: its influence is being “swept” out or partialed out of the estimate so what is left is the direct effect of $X_1$ on $Y$. Similar remarks apply about $\hat{\beta}_2$. 
Back to the simple case now.
Why least squares?
Understand what “you get” from the least squares solution.
In the bivariate setting, you get a straight line with slope $\beta_1$.
Implications of this?
Alternatives?
Regression

- Naive regression and smoothers.
- Is a straight-line function always the best way to go?
- Consider the plot on the next slide.
Artwork

Lowess smoother

bandwidth = .8
This plot is based on a nonparametric estimate of the relationship between Obama vote share and Prop. 8 vote share.

We will cover nonparametric regression at the end of the quarter.

But note there are many ways to fit a line through a cloud of points.
We deal with stochastic relationships—everything is probabilistic—so we have to recognize there is error around our estimate of the slope coefficient.

Omitting the mathematical details that gets us to this uncertainty, consider the following:

\[ s_{\beta_1} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (X_{1i} - \bar{X})^2(1 - R_i^2)(n - k - 1)}}. \]

What do we have here?

You might “prefer” a standard error to be small.

Why? More precise inferences are permitted.
“Conceptual” Uncertainty and Research Design

- What makes this “thing” small?
- When the stuff in the denominator is big relative to the stuff in the numerator, then what happens?
- When the stuff in the numerator is big relative to the stuff in the denominator, the s.e. increases.
- What does this say about research design and quantitative methods?
First, what is the term in the numerator representing?

It represents deviations from a model’s predictions and the observed values of $y$. Smaller deviations, better predictions.

The denominator?

The first term denotes the variance of $x$.

What is the implication here? More variance on $x$, the better off you will be. Research design issues: collect more data on $x$!
“Conceptual” Uncertainty and Research Design

- Second term: known as auxiliary regression. Think of it, for now, as an indication of how highly correlated two independent variables are.

- The extent to which your independent variables are not measuring unique concepts effects your ability to make inference.

- Third term: cases. If $n$ denotes the sample size, then what happens as $n$ increases? The S.E. decreases (a good thing).