Multiple Regression

Kathleen M. Saul

May 4, 2011

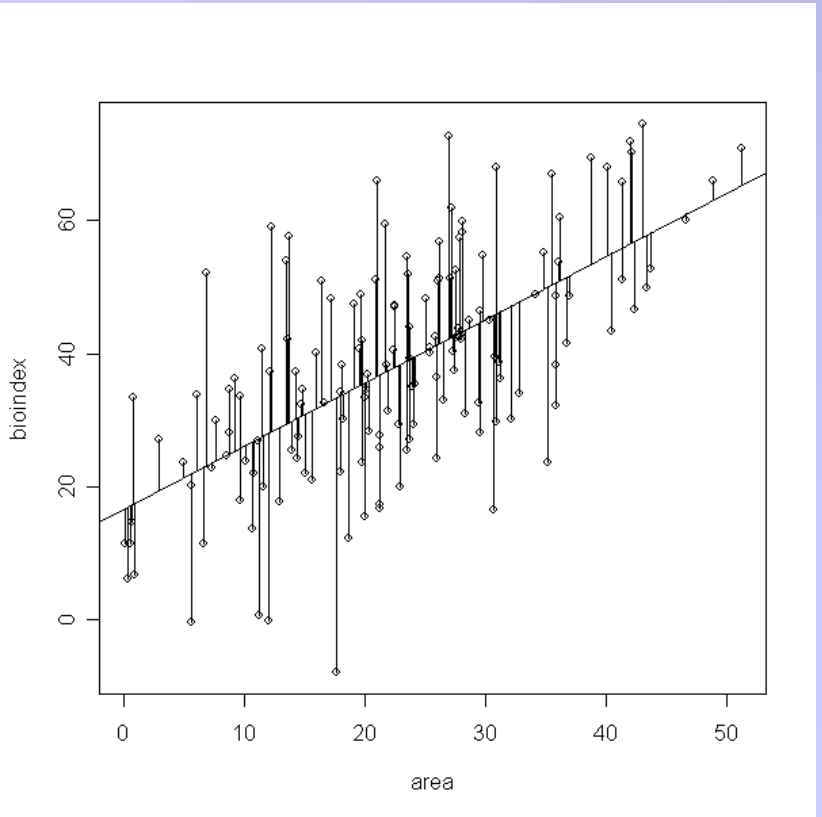
For Simple Linear Regression:

- Underlying assumption: Causation
- A linear model: $y = a + b x + error$
  - A = y intercept
  - B = slope
- Residual = The vertical distance between the point and the estimated line.

Understanding the Results

- The p-value associated with each coefficient tells us whether we can reject the null hypotheses that those coefficients equal 0.
- The r^2 value tells us the proportion of the variation in y that is explained by the variation in x. The large the r^2 value, the better the model.
- A plot of the residuals vs the predicted value of y should show no pattern.
- A histogram of the residuals should approximate a normal distribution.

- In R: lm(y~x)



Multiple Regression

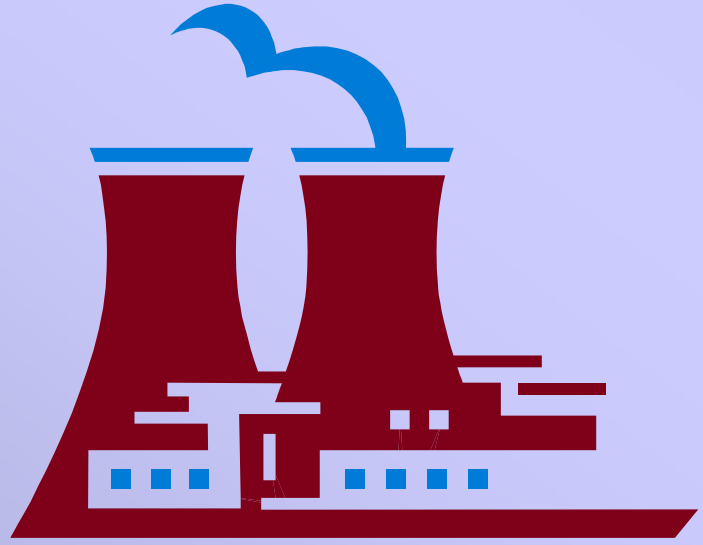
- Model: $y = z + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$
- Generates a multidimensional surface, rather than a line
- Assumes an additive effect of the x variables
- Null hypothesis:
  - Ho: $B_1 = B_2 = \ldots = B_k = 0$
- Alternative hypothesis:
  - Ha: At least one of the B is not equal to 0

- We can reject the null hypothesis ($B = 0$) for a given coefficient if the associated p-value is smaller than our alpha.
  - p-values in multiple regression assess the significance of each x variable, assuming that all other x variables are included in the regression equation (a situational/conditional value for p)
- Finding the best model to fit the data involves a lot of trial and error

- New issues to address:
  - *Multicollinearity*, or a correlation between two or more of the x variables
    - Increases the standard errors, results in incorrect sign and magnitude of coefficient estimates
    - Evaluate the correlation matrix of the variables first and eliminate one of any pair that appears highly correlated
  - *Interactions* between the x variables
    - Interaction implies a combined effect of the variables on the y variable: The impact of one x variable depends on the level of the other(s)
    - Run regression models that include interaction terms and evaluate the significance of those terms

Example: Nuclear Power Plant Construction Times

Scientific Hypothesis: The length of time it took to construct nuclear power plants in the U.S. depended on who supplied the reactor and the size (capacity) of that reactor.



- Data: Construction time (numeric: days from construction start to going on line), Suppler (categorical variable), Capacity (numeric: MW)
- Approach:
  - First, generate a correlation matrix
  - > nuke<-read.csv(file="Nuclear.csv",sep=",",head=T)
  - > cor(nuke)

	Supplier	Capacity	Constr
Supplier	1.00000000	0.06534536	0.05384731
Capacity	0.06534536	1.00000000	0.68937442
Constr	0.05384731	0.68937442	1.00000000

  - No apparent strong correlation between the x variables

- Next, run the “everything and the kitchen sink” model, including all x variables and the interaction term

- > summary(lm(nuke$Constr~nuke$Supplier+nuke$Capacity+nuke$Supplier*nuke$Capacity))

- Call:
- lm(formula = nuke$Constr ~ nuke$Supplier + nuke$Capacity + nuke$Supplier *
- nuke$Capacity)

- Residuals:
- Min 1Q Median 3Q Max
- -1628.2 -647.1 -188.9 364.7 4495.1
- Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	568.2856	1991.3451	0.285	0.776
nuke$Supplier	-507.2821	619.7776	-0.818	0.415
nuke$Capacity	2.6725	2.1588	1.238	0.219
nuke$Supplier:nuke$Capacity	0.5738	0.6696	0.857	0.394

- Residual standard error: 1021 on 100 degrees of freedom
- Multiple R-squared: 0.4791, Adjusted R-squared: 0.4635
- F-statistic: 30.66 on 3 and 100 DF, p-value: 3.855e-14

- Try again!
- > summary(lm(nuke$Constr~nuke$Supplier+nuke$Capacity))
- Call:
- lm(formula = nuke$Constr ~ nuke$Supplier + nuke$Capacity)

- Residuals:
- Min 1Q Median 3Q Max
- -1533.6 -636.3 -194.5 339.9 4472.4

- Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1068.1579	564.0495	-1.894	0.0611
nuke$Supplier	14.2973	116.8518	0.122	0.9029
nuke$Capacity	4.4780	0.4696	9.536	9.44e-16 ***

- ---
- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
- Residual standard error: 1020 on 101 degrees of freedom
- Multiple R-squared: 0.4753, Adjusted R-squared: 0.4649
- F-statistic: 45.75 on 2 and 101 DF, p-value: 7.16e-15

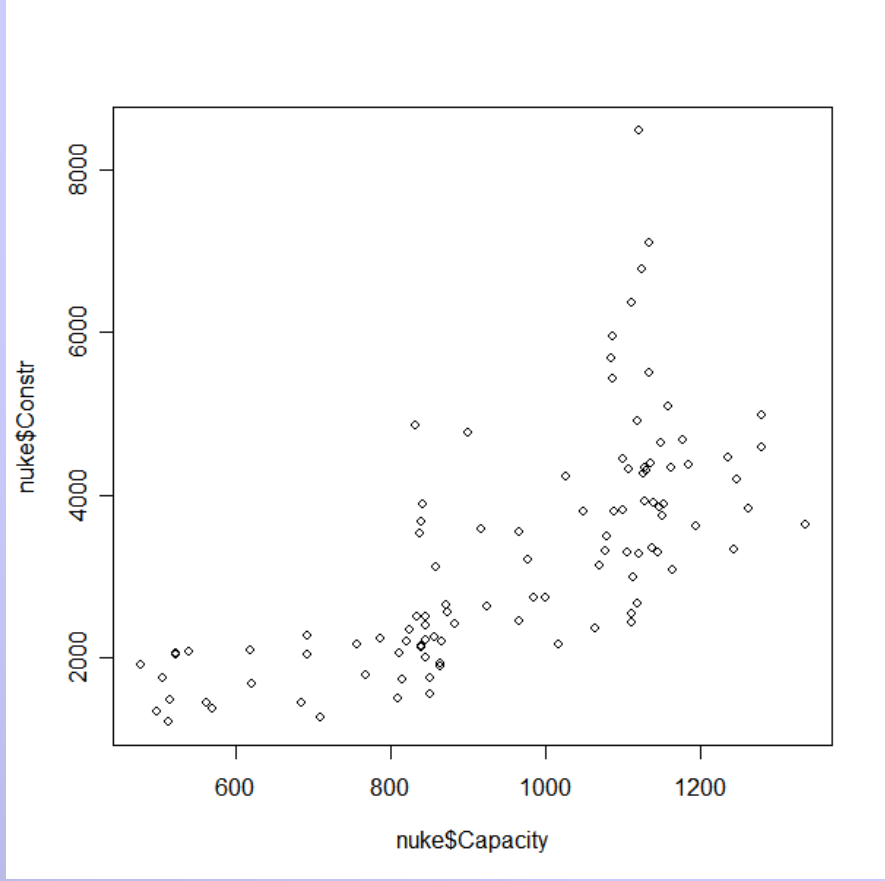
- And again!
- > summary(lm(nuke$Constr~nuke$Capacity))
- Call:
- lm(formula = nuke$Constr ~ nuke$Capacity)

- Residuals:
- Min 1Q Median 3Q Max
- -1520.9 -645.1 -188.4 354.0 4470.8
- Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1027.8853	455.8364	-2.255	0.0263 *
nuke$Capacity	4.4818	0.4663	9.611	5.91e-16 ***

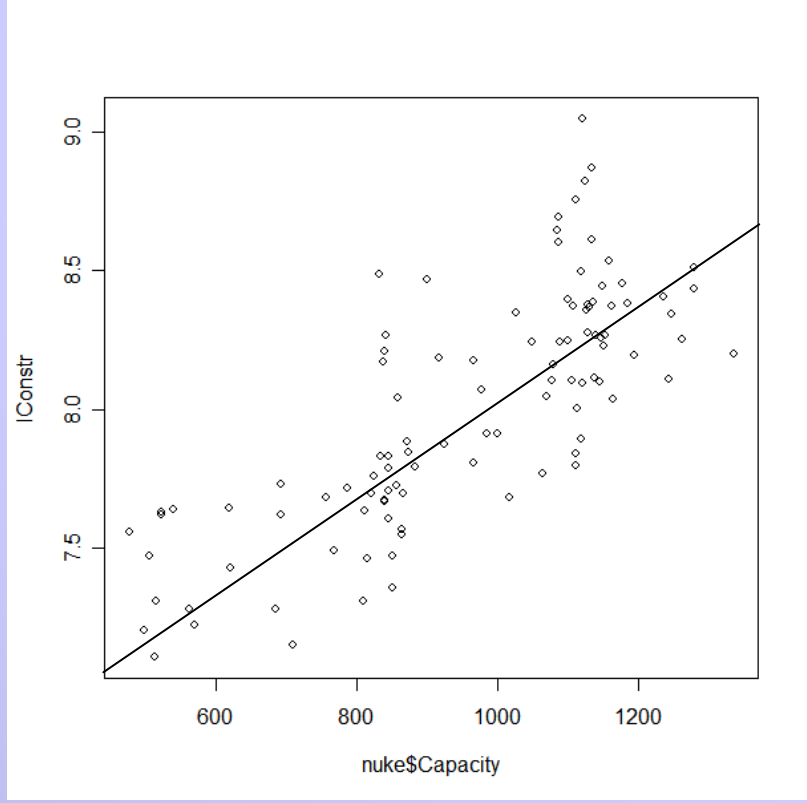
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- Residual standard error: 1015 on 102 degrees of freedom
- Multiple R-squared: 0.4752, Adjusted R-squared: 0.4701
- F-statistic: 92.37 on 1 and 102 DF, p-value: 5.912e-16

Plotting Construction Time vs. Capacity:



Not exactly linear!

Transform Constr by taking the log:



Much better!

- Rerun the model using lConstr:
- > summary(lm(lConstr~nuke$Capacity))
- Call:
- lm(formula = lConstr ~ nuke$Capacity)
- Residuals:
- Min 1Q Median 3Q Max
- -0.4835 -0.1648 -0.0299 0.1481 0.7903
- Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.5339881	0.1210338	53.98	<2e-16 ***
nuke$Capacity	0.0015340	0.0001238	12.39	<2e-16 ***

- ---
- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
- Residual standard error: 0.2695 on 102 degrees of freedom
- Multiple R-squared: 0.6008, Adjusted R-squared: 0.5969
- F-statistic: 153.5 on 1 and 102 DF, p-value: < 2.2e-16

- Results in the model for U.S. nuclear reactor construction time:
  - Log(Construction Time) = 688.14 + 1.001 * Capacity (MW), or
  - Construction Time (days) = 6.534 * exp (0.0015 * Capacity)

Recap

- Check for correlations between x variables
- Run the full model
- Eliminate variables that do not make a significant contribution to explaining y
- Rerun the model
- Continue until you have a model with a relatively high r^2 value and statistically significant x variables