Stats for Staffers Presents: Regression Analysis

Michael Costello
RTI International
Washington Statistical Society
American Statistical Association

What we will cover

1. What is regression?
2. Simple Linear Regression
3. General Statistics Aside
4. Multiple Regression
5. Logistic Regression

What is regression and why do we use it?

- Focuses on the relationship between a dependent variable (Y) and one or more independent variables (X).
- Estimates the conditional expectation of the dependent variable, given the independent variables.
- To determine the strength of a relationship between variables.

Simple Linear Regression Takeaways

- Knowing about more than just one variable can help us more accurately predict future events.
- The Least Squares Regression Line (LSRL or OLS) is the linear best fit model for a dataset.
- Math: $y = mx + b$

  Statistics: $y = a + bx$

  $y = b_0 + b_1x_1 + b_2x_2 + \ldots$

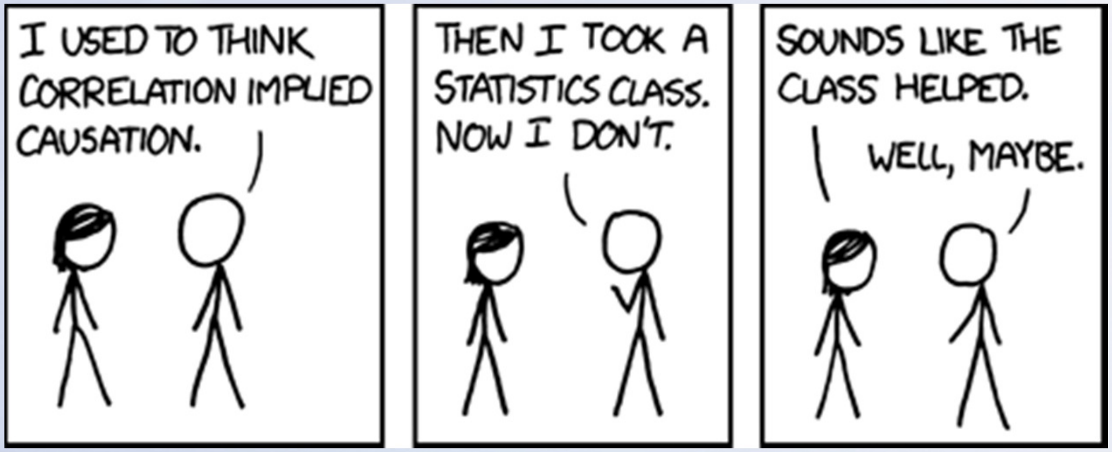
Knowing when your model is good

- Look at the correlation coefficient (r) and the coefficient of determination (R^2)
- Examine the slope of the line
- Examine the residuals

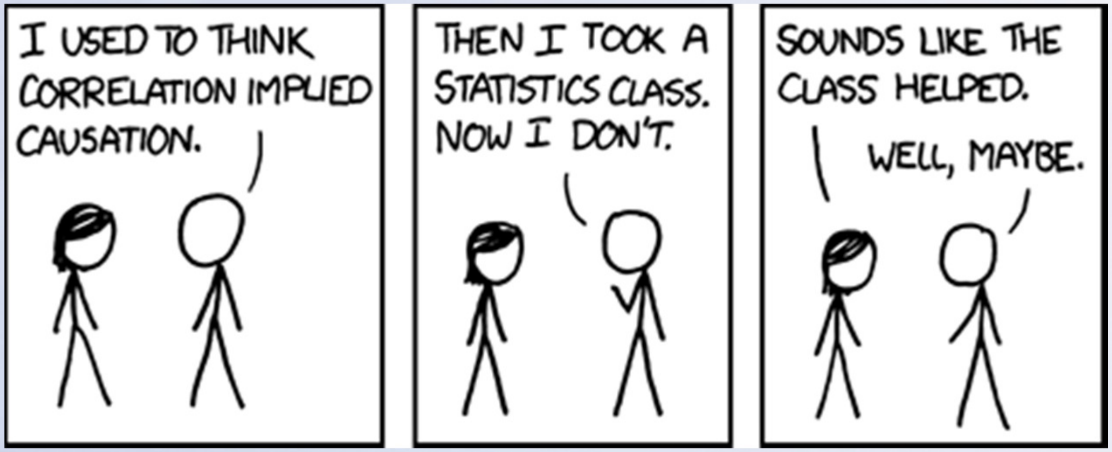
Knowing when your model is good

- Look at the correlation coefficient (r) and the coefficient of determination (R^2)
- Examine the slope of the line
- Examine the residuals

Knowing when your model is good



Knowing when your model is good



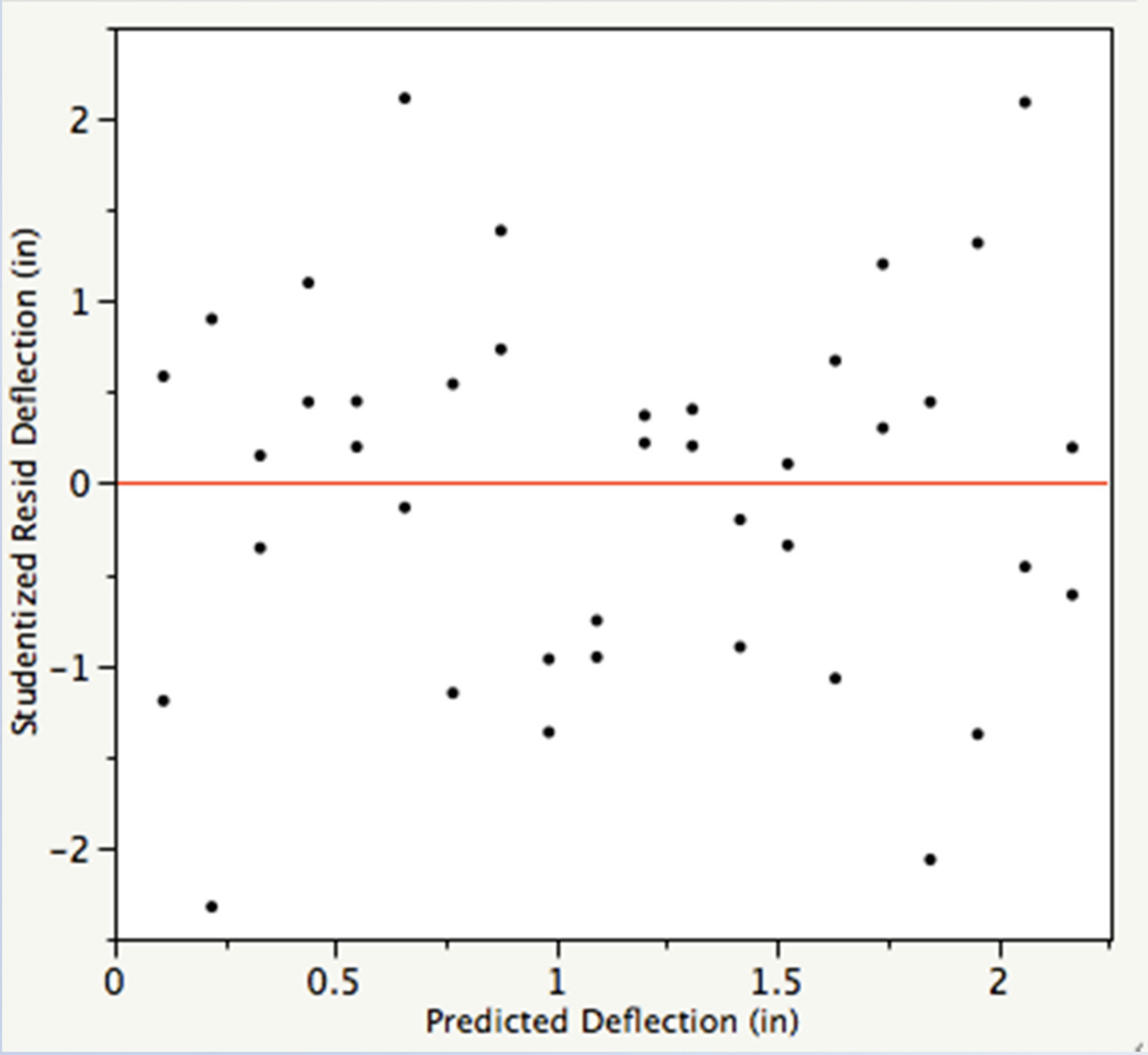
Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there.'

Regression Formulas

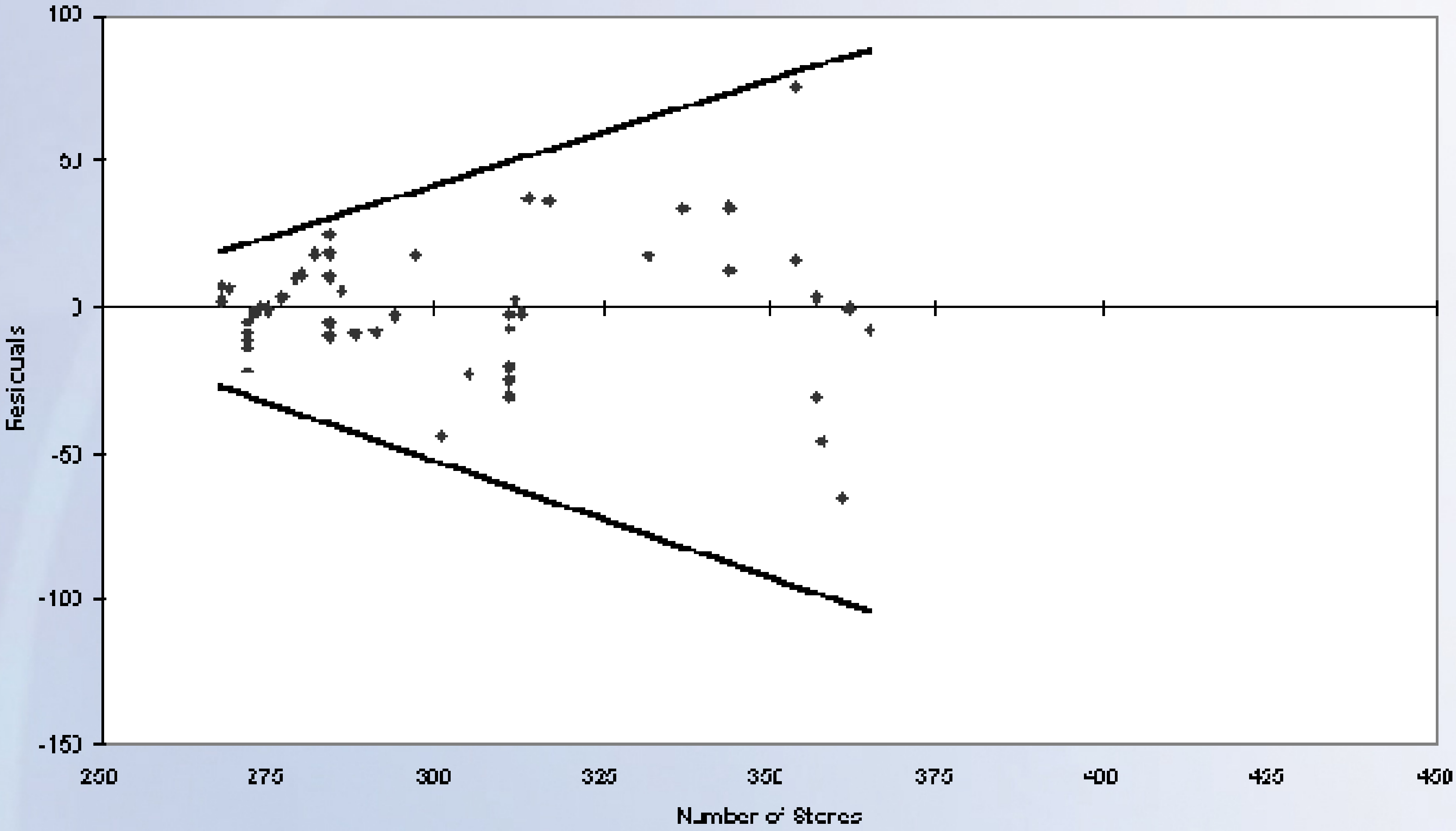
- Slope: b = $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r\,\frac{s_y}{s_x}$

- Y-intercept: a = $\bar{y} - b\bar{x}$

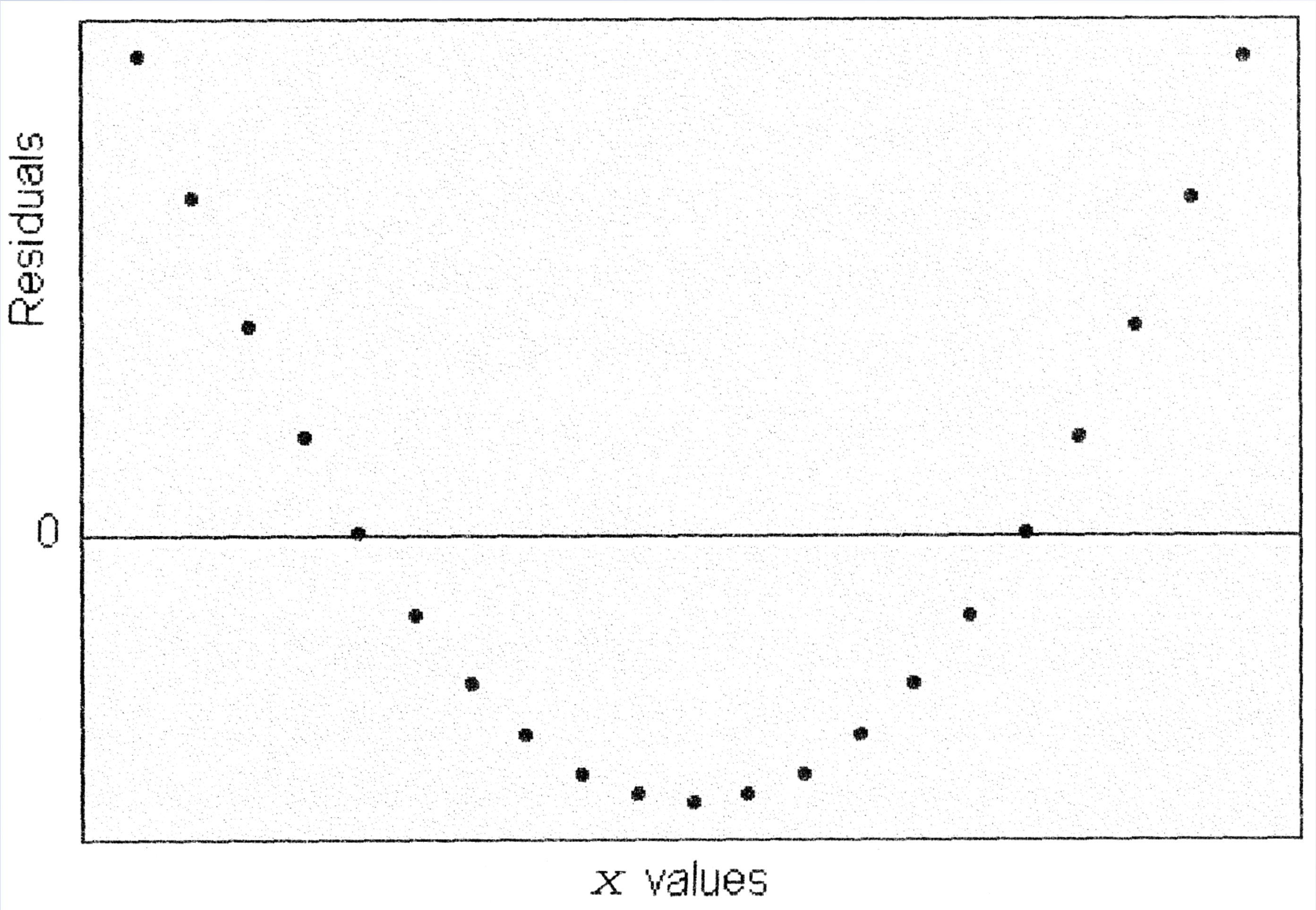
Looking at Residuals



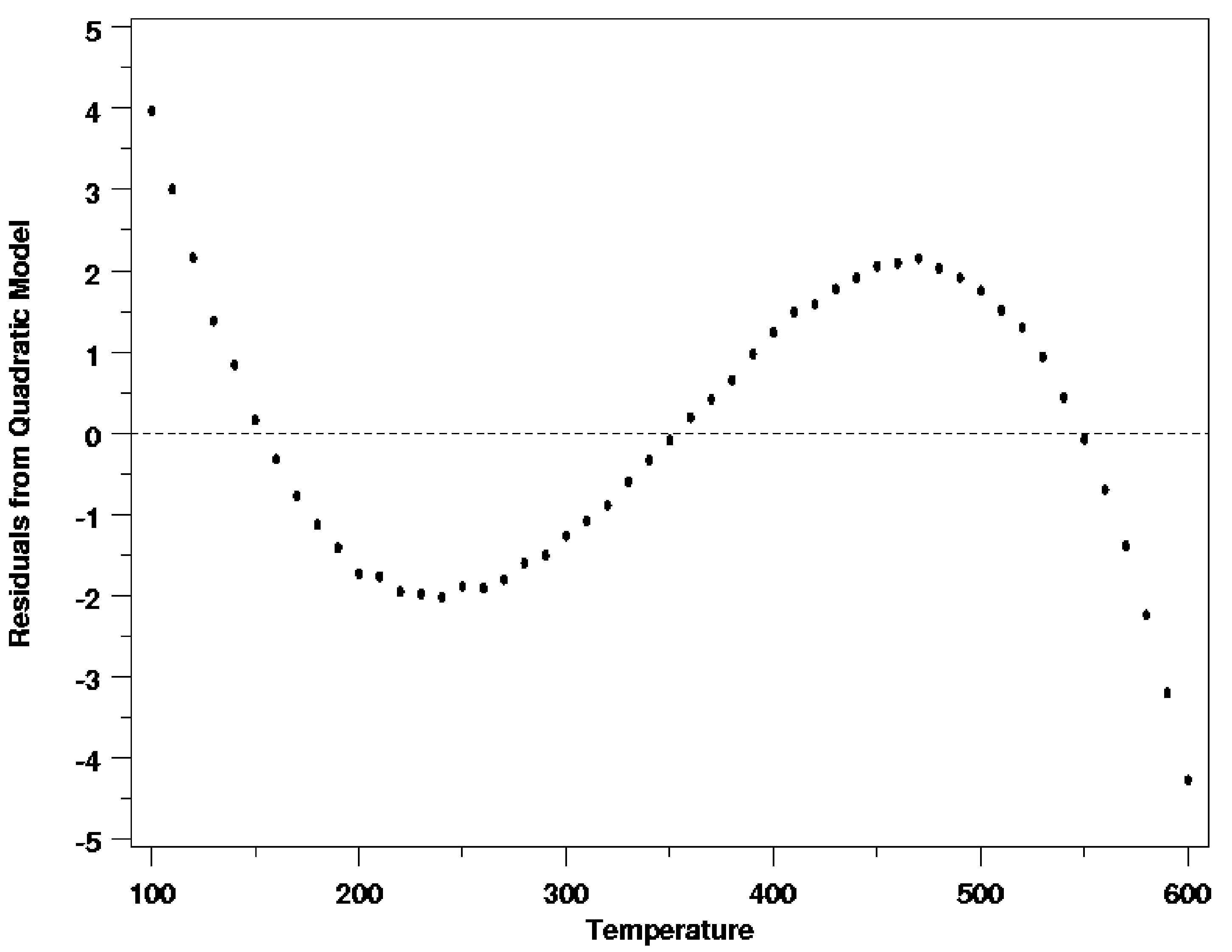
Looking at Residuals



Looking at Residuals



Looking at Residuals



General Statistical Aside

- Parameter vs. Statistic –

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information
- Standard Deviation –

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information
- Standard Deviation – Deviation of the data from the mean.

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information
- Standard Deviation – Deviation of the data from the mean.
- Standard Error –

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information
- Standard Deviation – Deviation of the data from the mean.
- Standard Error – Deviation of possible means from the current mean we see in our sample.

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information
- Standard Deviation – Deviation of the data from the mean.
- Standard Error – Deviation of possible means from the current mean we see in <u>our sample</u>.
- t-Statistic –

General Statistical Aside

- Parameter vs. Statistic – Population vs. Sample Information
- Standard Deviation – Deviation of the data from the mean.
- Standard Error – Deviation of possible means from the current mean we see in our sample.
- t-Statistic – The number of standard deviations / standard errors a dataset's mean is from the expected mean.

General Statistical Aside

- P-Value –

General Statistical Aside

- P-Value – The probability of observing our dataset mean (or one more extreme) when the expected mean is actually true.

General Statistical Aside

- P-Value – The probability of observing our dataset mean (or one more extreme) when the expected mean is actually true.
- Confidence Interval –

General Statistical Aside

- P-Value – The probability of observing our dataset mean (or one more extreme) when the expected mean is actually true.
- Confidence Interval – Estimate of a parameter, used to indicate an estimate's reliability.

Regression Tables in Excel, Stata, etc.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.931183427
R Square	0.867102575
Adjusted R Square	0.864743449
Standard Error	6.172971999
Observations	173

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	42017.40498	14005.8	367.5525	8.40189E-74
Residual	169	6439.843577	38.10558		
Total	172	48457.24855			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	16.17957577	2.986473768	5.417619	2.05E-07	10.28397657	22.07517497
democA	2.374906018	1.156002059	2.054413	0.041475	0.092841837	4.656970199
prtystrA	0.215889774	0.057858766	3.731323	0.00026	0.101670758	0.33010879
shareA	0.435885557	0.016375789	26.61768	2.43E-62	0.403558105	0.468213009

Regression Tables in Excel, Stata, etc.

```
. regress voteA democA prtystrA shareA

      Source |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  3,   169) =  367.55
       Model |  42017.405     3  14005.8017           Prob > F      =  0.0000
    Residual | 6439.84358   169  38.1055833           R-squared     =  0.8671
-------------+------------------------------           Adj R-squared =  0.8647
       Total | 48457.2486   172  281.728189           Root MSE      =   6.173

------------------------------------------------------------------------------
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      democA |   2.374906   1.156002     2.05   0.041     .0928418     4.65697
    prtystrA |   .2158898   .0578588     3.73   0.000     .1016708    .3301088
      shareA |   .4358856   .0163758    26.62   0.000     .4035581     .468213
       _cons |   16.17958   2.986474     5.42   0.000     10.28398    22.07517
------------------------------------------------------------------------------
```

Multiple Regression

- When we have more than one x-variable
- Dummy or Binary X Variables
- Examples

Logistic Regression

- When our Independent variable (Y) is binary
- Produces an odds ratio
- Examples

Logistic Regression

- π = probability of success
- Logit form: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$
- Probit Form: $\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$

Logistic Regression

```
. logit oral_read_score_zero grade female grade_size
Logistic regression                                 Number of obs   =       3012
                                                    LR chi2(3)      =    1099.06
                                                    Prob > chi2     =     0.0000
Log likelihood =  -1294.945                         Pseudo R2       =     0.2979
------------------------------------------------------------------------------------
oral_read_score_zero |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
               grade |  -1.666301   .0677756   -24.59   0.000    -1.799139   -1.533463
              female |   .1271608   .0964876     1.32   0.188    -.0619515     .316273
               _cons |    6.35269   .2712615    23.42   0.000     5.821027    6.884353
------------------------------------------------------------------------------------
. logistic oral_read_score_zero grade female grade_size
Logistic regression                                 Number of obs   =       3012
                                                    LR chi2(3)      =    1099.06
                                                    Prob > chi2     =     0.0000
Log likelihood =  -1294.945                         Pseudo R2       =     0.2979
------------------------------------------------------------------------------------
oral_read_score_zero | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
               grade |   .1889447   .0128058   -24.59   0.000     .1654413     .215787
              female |     1.1356   .1095713     1.32   0.188     .9399285    1.372005
               _cons |   574.0349   155.7136    23.42   0.000     337.3184    976.8693
------------------------------------------------------------------------------------
```

More Information

Michael Costello

202.728.2487

mcostello@rti.org