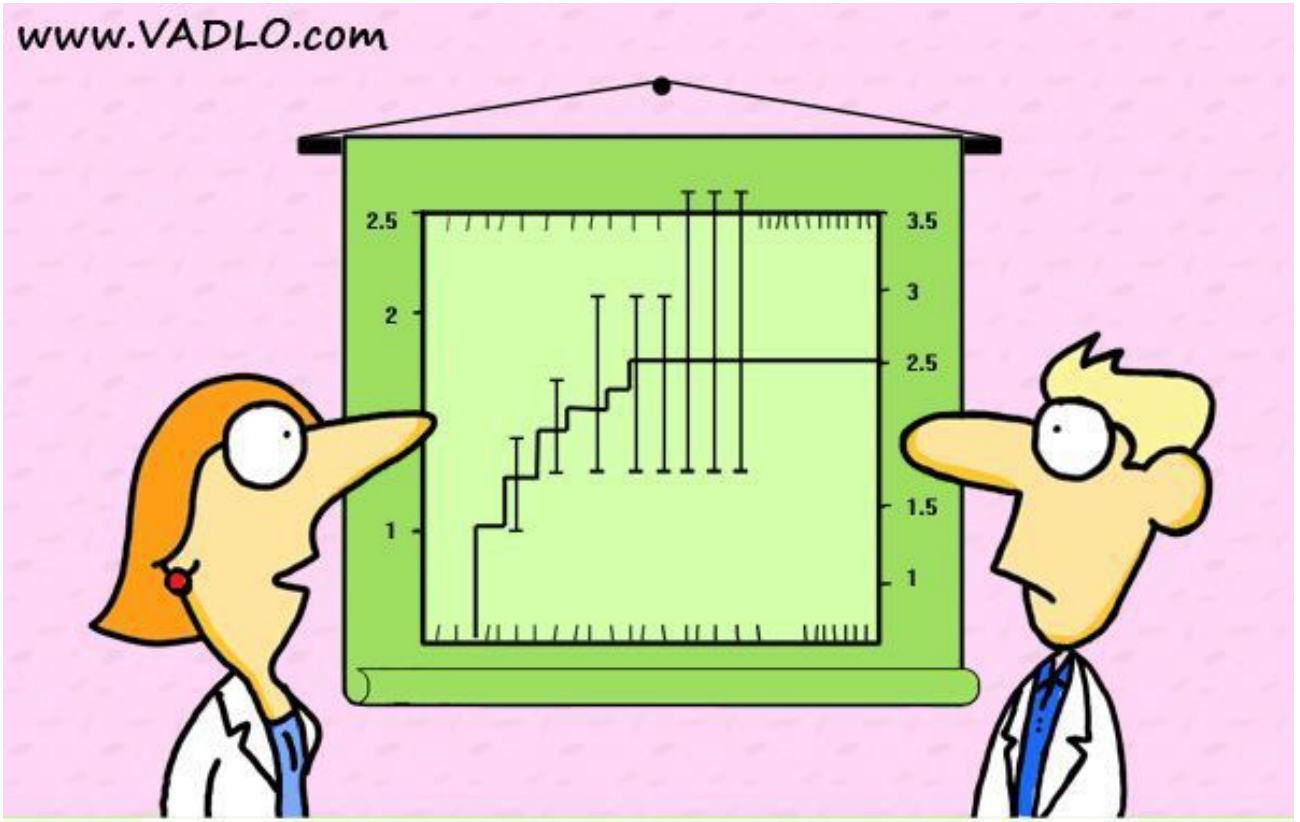
BST 210
Applied Regression Analysis



"Did you really have to show the error bars?"

High Level Recap

- Variability in world – regression modeling helps capture that; provides logic amidst uncertainty
- Regression explores associations, prediction
- Assumptions must be met
- Testing can be done and inferences drawn
- Can be generalized, extended to accommodate many underlying data processes (GLM, survival)

- Today's topics are equally exciting and useful!

Begin to develop general framework for approaching analyses

First -

- Learn the topic/study well, really well
- Collaborate to define motivating questions of interest, check PubMed
- What techniques might help to achieve answers? Which do the data warrant? (develop intuition, read literature)
- Possible Confounding or Effect Modification to account for?
- Keep an open mind, and the larger picture – there is no recipe

Begin to develop framework for analysis

Next -

- Diagnostics/Checking Assumptions:
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing (example: Lowess)
- Hypothesis testing/modeling:
  - t-test?
  - Correlation (r)?
  - ANOVA useful?
  - Nonparametric approach better?
  - Linear regression or extensions?
  - Generalizations

Begin to develop framework for analysis

Then -

- Assess Model Fit:
  - R^2
  - MSE
  - Confidence Intervals
  - Residual Analysis
- Interpretation and Inference, constant collaboration around data and meaning
- Backup and regroup as needed, delve deeper, use caution, stay organized

The New England
Journal of Medicine

Volume 329 OCTOBER 7, 1993 Number 15

A COMPARISON OF THE PERIOPERATIVE NEUROLOGIC EFFECTS OF HYPOTHERMIC CIRCULATORY ARREST VERSUS LOW-FLOW CARDIOPULMONARY BYPASS IN INFANT HEART SURGERY

JANE W. NEWBURGER, M.D., M.P.H., RICHARD A. JONAS, M.D., GIL WERNOVSKY, M.D., DAVID WYPIJ, PH.D., PAUL R. HICKEY, M.D., KARL C.K. KUBAN, M.D., S.M., DAVID M. FARRELL, M.A., C.C.P., GREGORY L. HOLMES, M.D., SANDRA L. HELMERS, M.D., JULES CONSTANTINOU, F.R.A.C.P., ENRIQUE CARRAZANA, M.D., JOHN K. BARLOW, M.D.,* AMY Z. WALSH, R.N., B.S.N., KRISTIN C. LUCIUS, R.N., M.S., JANE C. SHARE, M.D., DAVID L. WESSEL, M.D., FRANK L. HANLEY, M.D., JOHN E. MAYER, JR., M.D., ALDO R. CASTANEDA, M.D., AND JAMES H. WARE, PH.D.

(At the time of this study, hypothermic circulatory arrest (HCA) and low-flow cardiopulmonary bypass (CPB) were accepted techniques for the operative management of complex cardiovascular pathology, with the potential for neurologic sequelae being a concern.)

Circulatory Arrest Study

- Bellinger *et al*. (NEJM, 1995) report on a clinical trial comparing CA (Circulatory Arrest) vs. LFB (Low Flow Bypass) for repair of transposition of the great arteries

- Primary outcome at one year was PDI (Psychomotor Development Index), a continuous measure of motor skills (scaled similar to IQ)

Circulatory Arrest Study

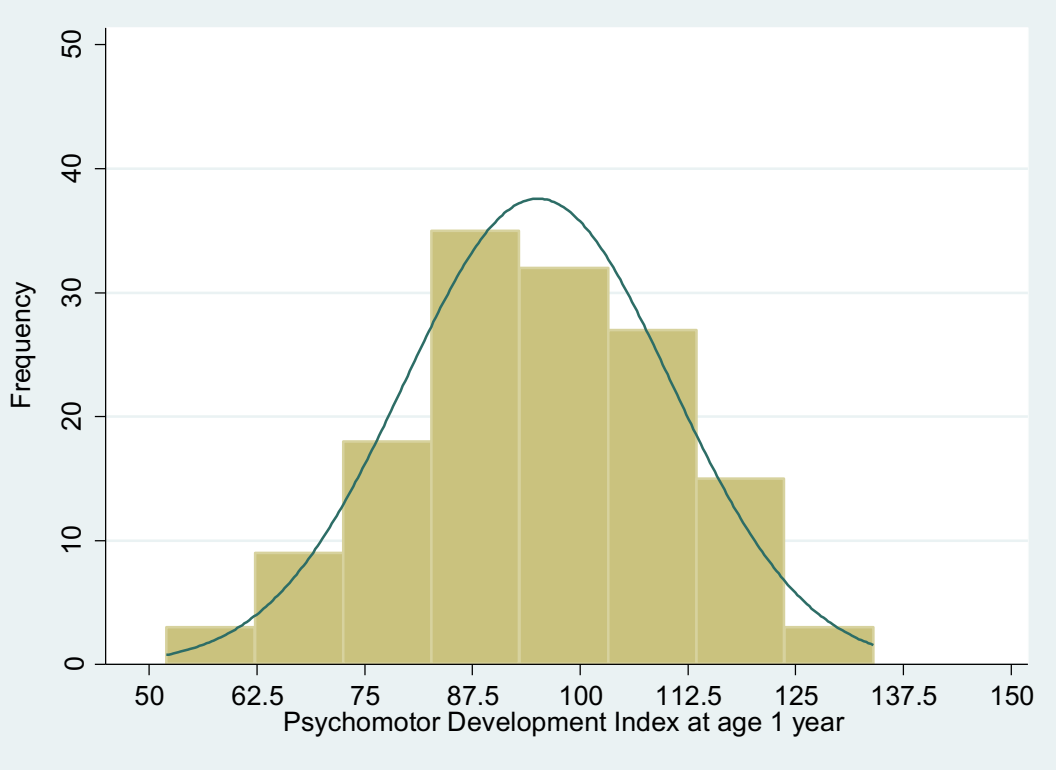
- Predictor variables include:
  - treatment group (CA vs. LFB) - nominal
  - duration of circulatory arrest (minutes) - continuous
  - diagnosis group (IVS, Intact Ventricular Septum, vs. VSD, Ventricular Septal Defect) - nominal
  - age at surgery, birth weight, etc.

Motivating Questions

- Is PDI related to the (continuous) duration of CA?
- Is PDI related to the (categorical) treatment group (CA vs. LFB)?
- Is PDI related to treatment group, after adjusting for diagnosis group (IVS and VSD)?
- Other predictor variables?

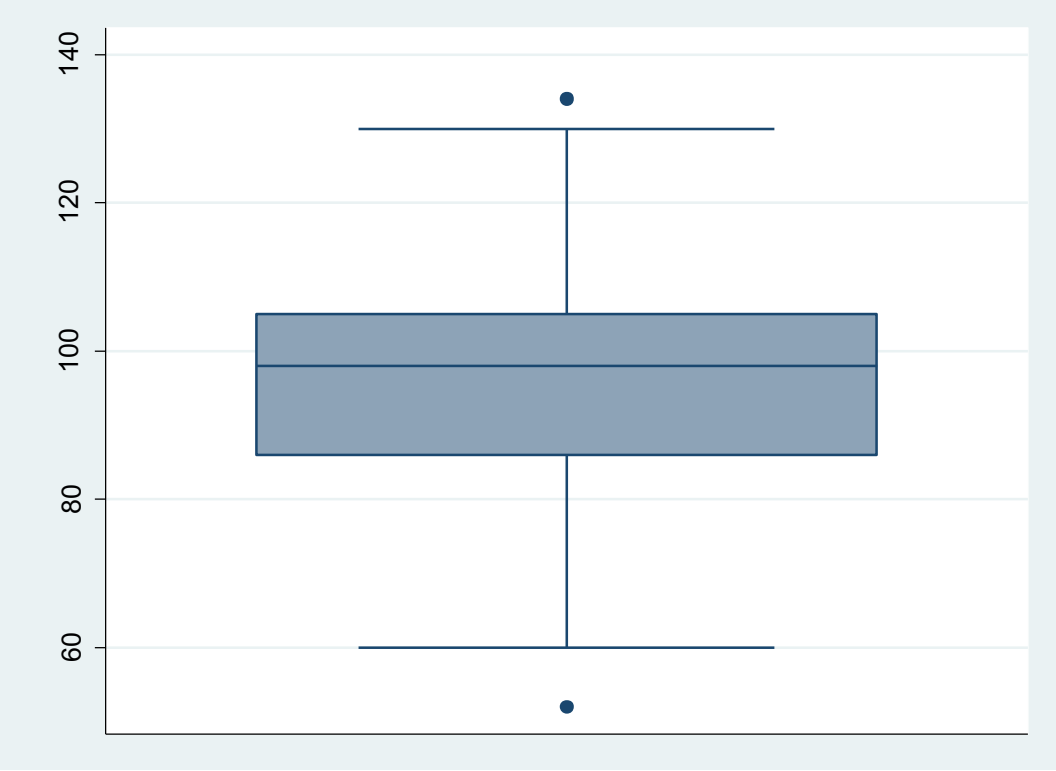
Diagnostics

```
histogram pdi, bin(8) freq xlabel(50(12.5)150)
ylabel(0(10)50) normal
```



Diagnostics

```
graph box pdi
```

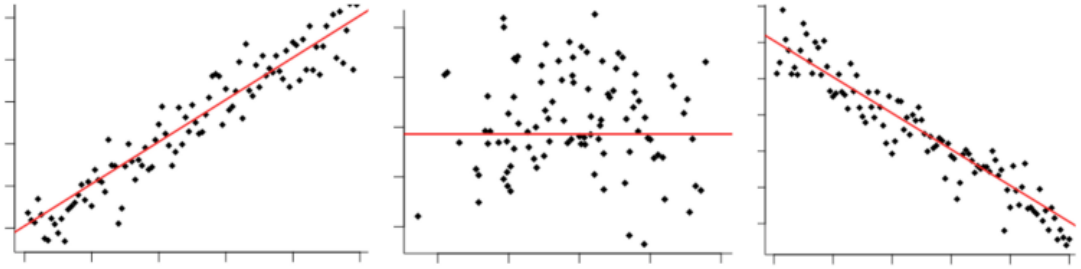


Correlation Analysis

- Quantifies the degree to which two continuous variables are *linearly* related
- Pearson correlation coefficient, often denoted by *r*
- Correlation coefficients are dimensionless (no units)

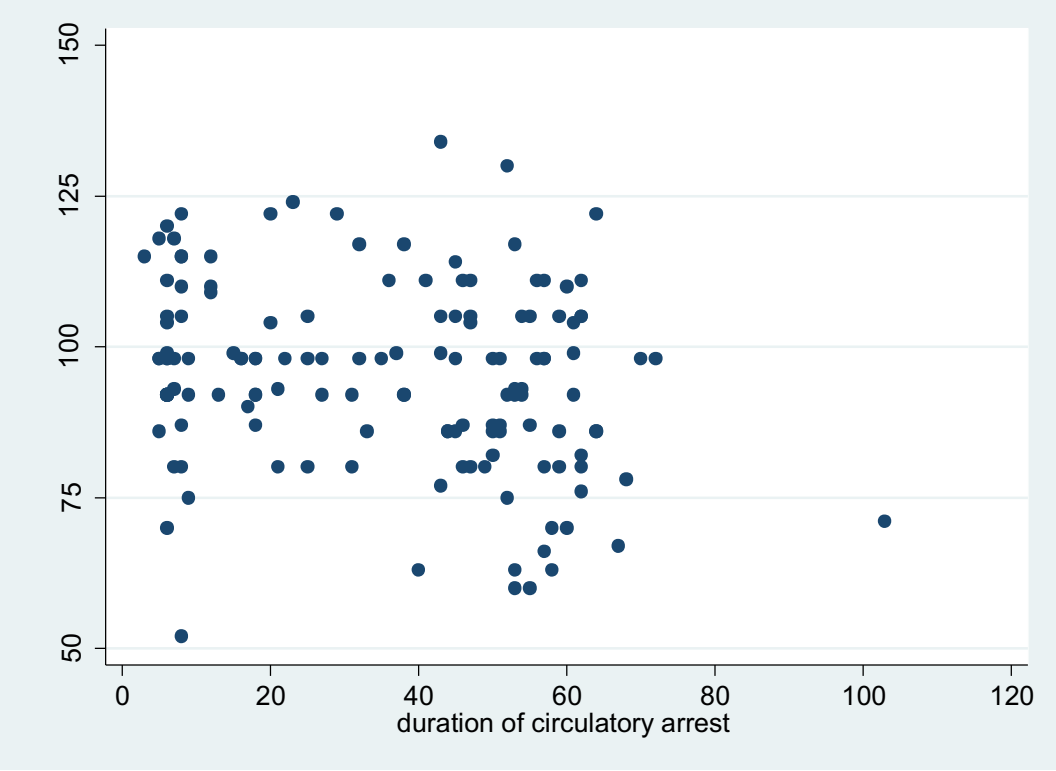
Correlation Analysis

- The variables are
  - positively correlated ($r > 0$) if one tends to increase as the other does
  - negatively correlated ($r < 0$) if one tends to decrease as the other increases
  - uncorrelated ($r = 0$) if no linear association
- Correlations can range between –1 to +1



Results

```
scatter pdi minutes, xlabel(0(20)120)
ylabel(50(25)150)
```



Results

```
pwcorr pdi minutes, sig obs
```

	pdi	minutes
pdi	1.0000	
	142	
minutes	-0.2198	1.0000
	0.0086	
	142	171

Results

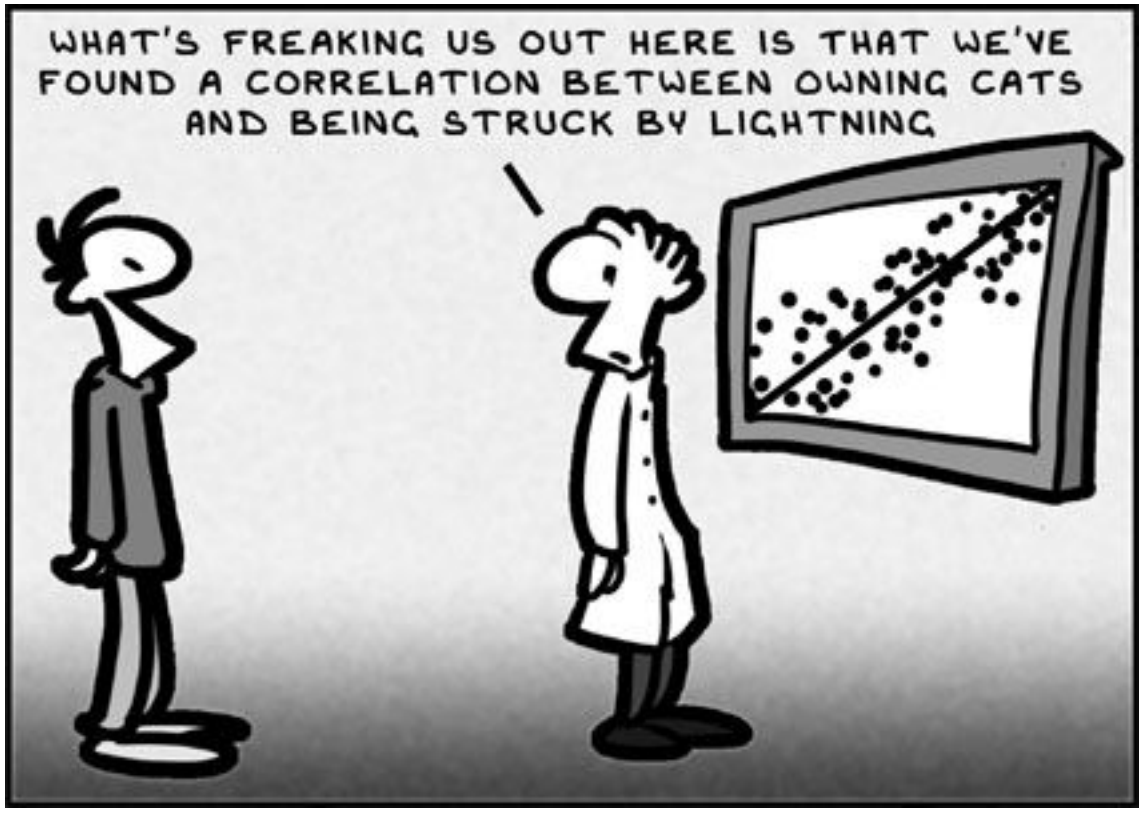
- Pearson correlation coefficient measuring the association between PDI and minutes of CA is $r = -0.22$ ($P = 0.009$)

- Strong (?) evidence of an inverse relationship between these variables

Limitations of Correlations

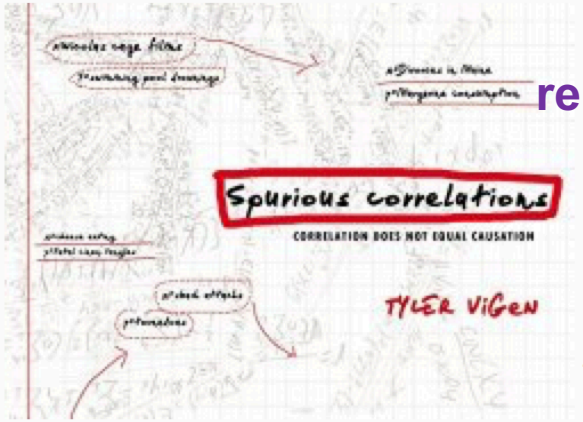
- Only quantifies the strength of a **linear relationship**; not a valid measure of a nonlinear relationship
- Highly sensitive to outliers
- Cannot be extrapolated beyond the observed ranges of the variables
- Doesn't provide any sort of 'slope' or strength of association estimate
- …and…

** A strong correlation does not necessarily imply a cause-and-effect relationship or causation **



Use Caution!

Spurious correlations

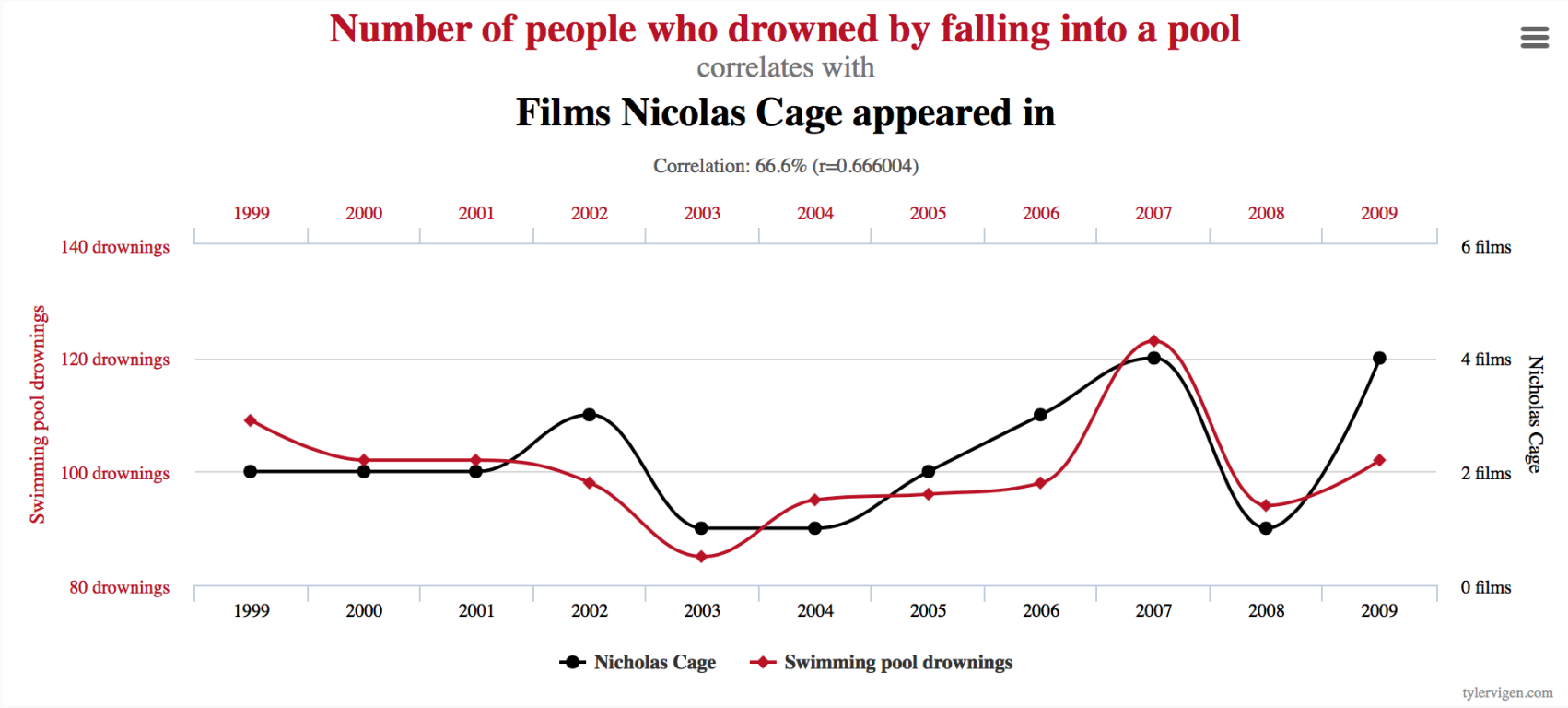


Now a ridiculous book!

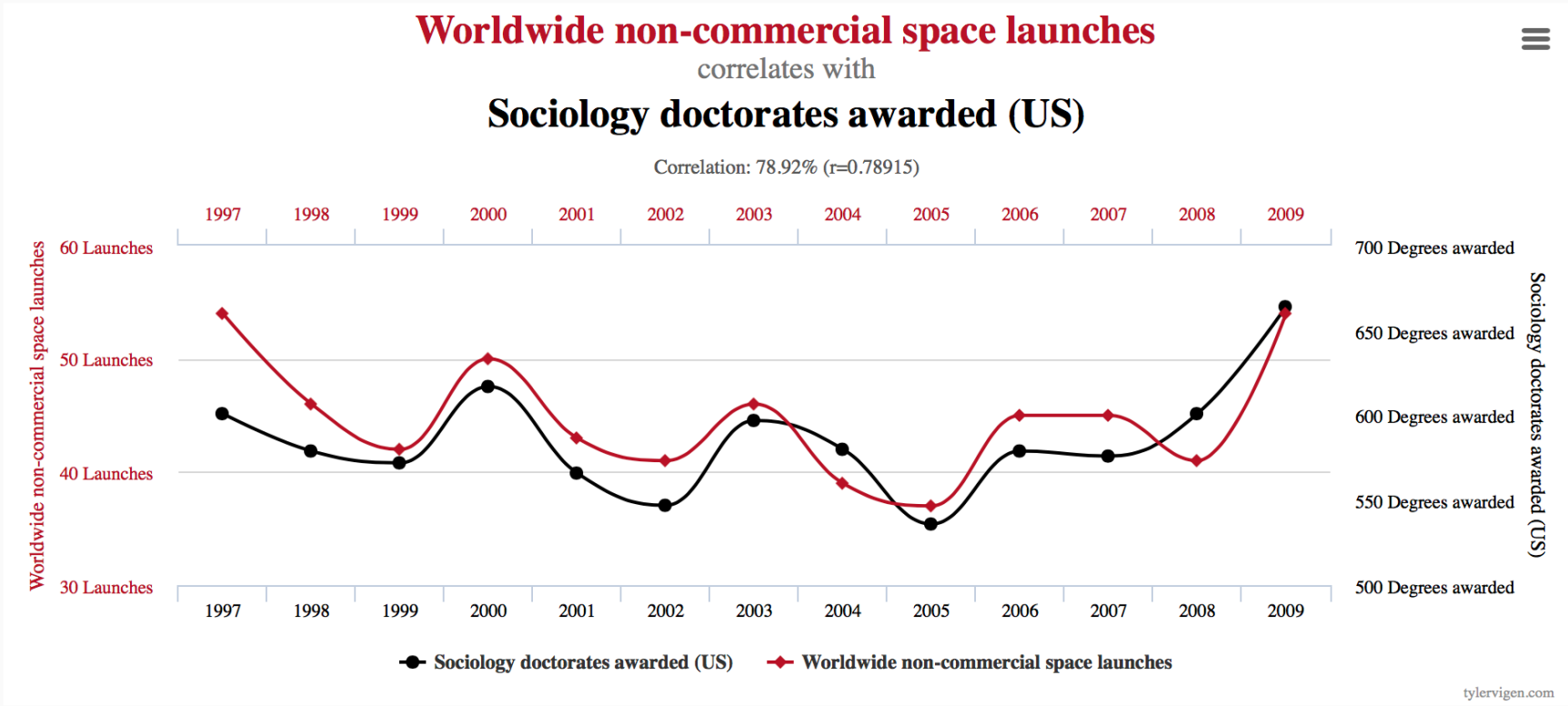
relati

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

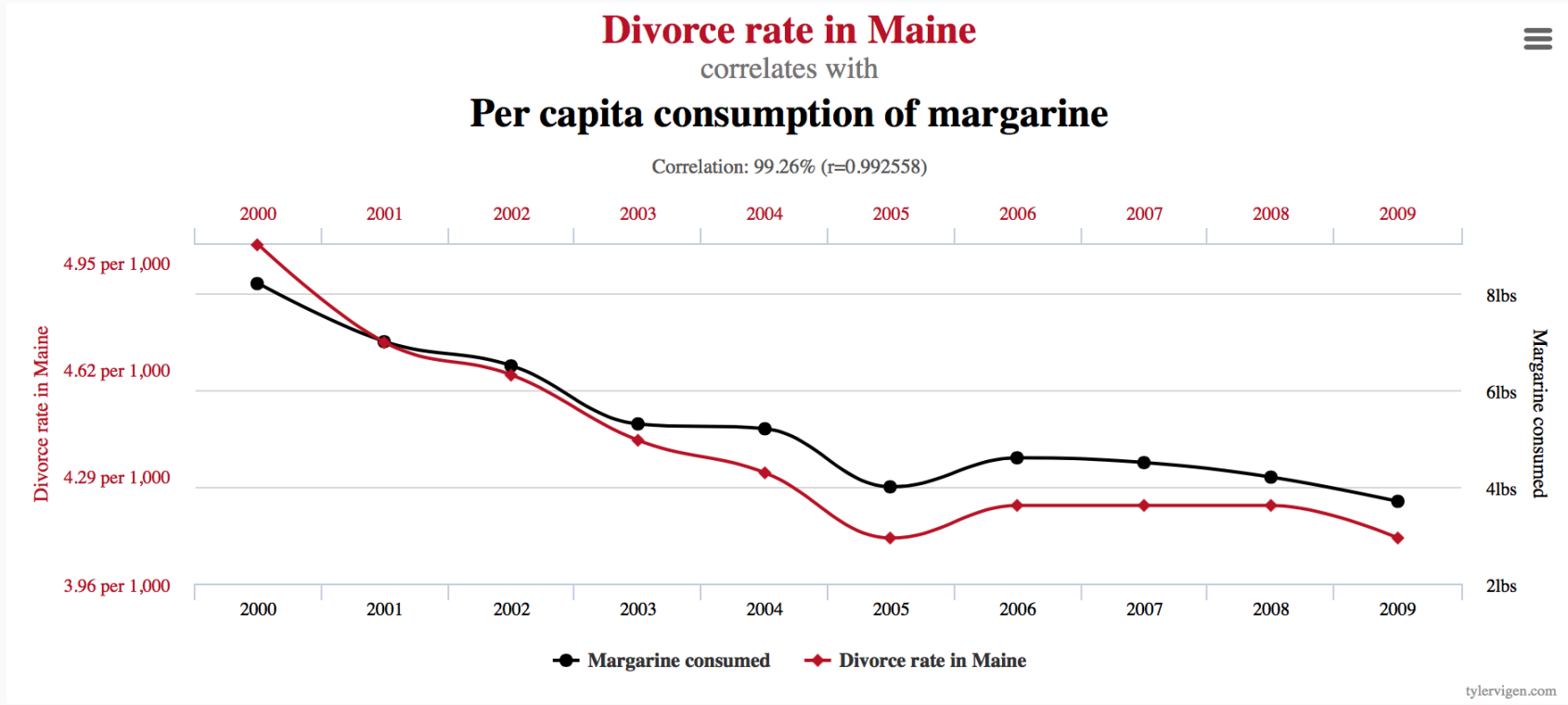
Amazon | Barnes & Noble | Indie Bound



Data sources: Centers for Disease Control & Prevention and Internet Movie Database



Data sources: Federal Aviation Administration and National Science Foundation



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Next let's consider *t*-tests

- The unpaired *t*-test compares the means of two independent groups of normally distributed observations, testing the null hypothesis

$$H_0: \mu_1 = \mu_2 \quad \text{(or } H_0: \mu_1 - \mu_2 = 0\text{)}$$

- Two versions: equal variance (assumes $\sigma_1^2 = \sigma_2^2$) or unequal variance (assumes σ_1^2 may be different than σ_2^2)

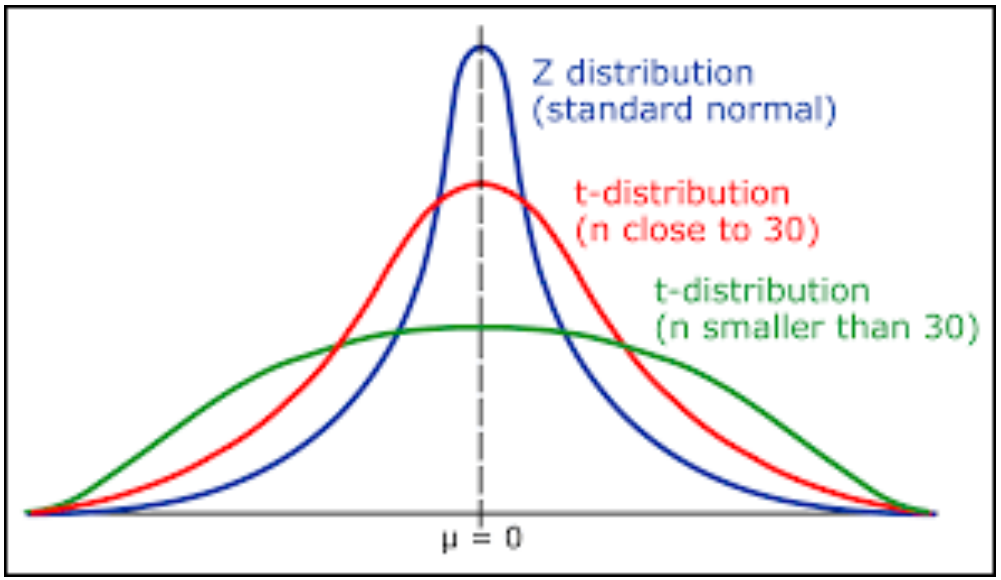
Comparing Groups: *t*-tests

- *t*-tests depend on degrees of freedom, here $n_1 + n_2 - 2$ (related to sample sizes)

- The assumption of normality is important, although the *t*-test tends to work alright for non-normal distributions if the sample sizes are large

Review the t-distribution

- t-distribution (df related to sample size, and we don't know σ^2)



Results

- PDI scores for a healthy, normal population of children (newer versions have mean 100):

$$110 \pm 15 \text{ (mean} \pm \text{SD)}$$

- PDI scores for the combined cardiac population:

$$95.1 \pm 15.5$$

- The dispersion seems similar, but the mean seems lower for the cardiac population ($P < 0.001$, one sample *t*-test)

Results

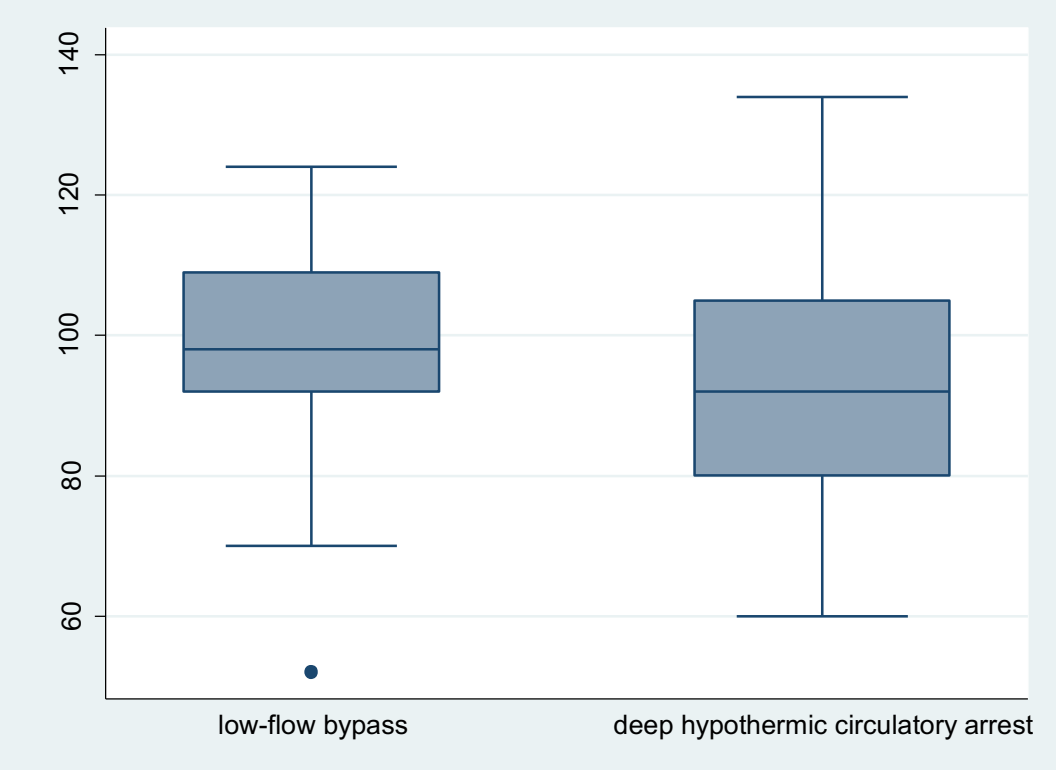
```
ttest pdi=110, level(95)

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pdi |     142    95.09859    1.296565    15.45036    92.53537    97.66181
------------------------------------------------------------------------------
Degrees of freedom: 141

                     Ho: mean(pdi) = 110

    Ha: mean < 110               Ha: mean ~= 110               Ha: mean > 110
       t = -11.4930                 t = -11.4930                  t = -11.4930
   P < t =   0.0000           P > |t| =   0.0000           P > t =   1.0000
```

Comments on Computer Output

- As for any computer output, be sure to accurately input your data and interpret your results
- For very small *P*-values, it is better to write $P < 0.001$ or $P < 0.0001$ rather than $P = 0$, even though the output sometimes says that (similarly $P > 0.99$)
- Use an appropriate number of decimal places when reporting numeric results, e.g., write mean = 95.1 rather than 95.09859

Results

```
graph box pdi, over (dhca)
```



Results

```
ttest pdi, by(dhca) level(95)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
low-flow |      69    98.46377    1.636459    13.59345    95.19827    101.7293
deep hyp |      73    91.91781    1.929775      16.488    88.07087    95.76474
---------+--------------------------------------------------------------------
combined |     142    95.09859    1.296565    15.45036    92.53537    97.66181
---------+--------------------------------------------------------------------
    diff |             6.54596    2.543947                 1.51644    11.57548
------------------------------------------------------------------------------
Degrees of freedom: 140

                 Ho: mean(low-flow) - mean(deep hyp) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   2.5732                  t =   2.5732                  t =   2.5732
   P < t =   0.9944              P > |t| =   0.0111              P > t =   0.0056
```

Results

- The test statistic comparing the PDI scores of subjects randomized to DHCA vs. LFB is T = 2.57 on 140 degrees of freedom, *P* = 0.011 (two-sided test)

- Strong evidence of a difference in mean values between these two groups

Results

- Comparing the two treatment groups, we find that the LFB group has a mean PDI score of 6.5 points higher than the CA group (95% CI, 1.5-11.6)

- Even with a statistically significant difference between groups, a confidence interval for the mean difference is useful

Other forms of *t*-test

- In addition to two-sample *t*-tests, there are also:
  - One sample *t*-tests (e.g., testing $H_0 : \mu = 110$ which would use $n - 1$ degrees of freedom)
  - Paired *t*-tests (e.g., when you have pre- and post-measurements on subjects, or right and left measurements on subjects) – equivalent to using a one sample *t*-test on the paired differences

Limitations of *t*-tests

- Can be sensitive to non-normality, outliers, unequal variances

- Need regression or ANOVA methods for adjustment for additional factors/levels

Analysis of Variance

- In a **t-test**, we compare values for a continuos variable across two levels of a categorical one.

- In **analysis of variance (ANOVA)**, we extend this to explain the variance of the outcome variable as a function of at least two *categorical* factors

One-way ANOVA

- In **one-way ANOVA**, we have *one* predictive factor that takes on *two or more levels*
- Examples: treatment group, diagnostic group, age group
- Thus, one-way ANOVA serves as the extension of the two-sample *t*-test to three or more samples

Family of approaches based on same model

- A linear regression model uses one or more *categorical or continuous* covariates to predict the outcome variable
- In fact, the model for one-way ANOVA is basically equivalent to that of linear regression using indicator variables
  - use one indicator variable for a factor with two levels, use two indicator variables for a factor with three levels, etc.
- Similarly, a one-way ANOVA serves as the extension of the two-sample *t*-test to three or more samples

One-way ANOVA

- In one-way ANOVA, we are interested in testing the null hypothesis that all population means are identical
- If we reject this null hypothesis, we use multiple comparisons procedures to test pairwise group differences (for $\geq$ 2 groups)

Two-way ANOVA

- In **two-way ANOVA**, we have *two* predictive factors that takes on two or more levels
- Two-way ANOVA allows us to assess the effects of one factor, controlling for the other (confounding)
- We can also assess two-way interactions (effect modification)
- In general, ANOVA reduces the probability of Type I error (which would occur if we did multiple t-tests rather than one single ANOVA).

Two-way ANOVA

- In fact, the model for two-way ANOVA can be described by linear regression using two sets of indicator variables, with product variables if the interaction is included
- An example could be using diagnosis group and treatment group to predict PDI (plus possibly their interaction)
- Or, using treatment group (two categories, one indicator variable) and age group (three categories, two indicator variables)

ANOVA Extensions

- Many extensions exist
  - three-way ANOVA
  - repeated measures ANOVA
  - nested models, etc.

- These generally can also be written as multiple linear regression models

Limitations of ANOVA

- ANOVA methods require normality and equal variances across groups, and can be affected by outliers or nonnormality (similar to linear regression)
- ANOVA methods don't take the ordering of categories into account (e.g., they use nominal, not ordinal categories, so don't test for trends)
- Often you need to move back to linear regression to allow continuous covariates

We see mounting evidence toward flexibility of the Regression Model!

But first a bit on several useful

- Nonparametric and
- Smoothing approaches

Remember where we are in the 'framework'

- Diagnostics/Checking Assumptions:
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing (example: Lowess)
- Hypothesis testing/modeling:
  - t-test?
  - Correlation (r)?
  - ANOVA useful?
  - Nonparametric approach better?
  - Linear regression or extensions?
  - Generalizations

Nonparametric Methods

- If the data are normally distributed, or if the sample sizes are relatively large and the data are approximately normally distributed, then normality-based methods work well
- ‘Parametric’ means that the researcher or analyst assumes in advance that the data fits some type of distribution (i.e. the normal distribution)
- But, if the data are far from normal (skewed) or the sample sizes are small so you are not sure of normality, nonparametric methods may be better to use

Nonparametric Methods

- Often nonparametric methods work on the **ranks** (or midranks, in the case of ties) of the data, rather than the data values themselves

- The **Spearman rank correlation coefficient** is basically the Pearson correlation coefficient applied to the ranks of the two variables being looked at

Nonparametric Methods

- The **Wilcoxon rank sum test** (or the Mann-Whitney *U* test) is a two group comparison of the ranks of the data (when the two groups are combined)
- When there is uncertainty about the normality assumption, a nonparametric method may be more appropriate, as they more robust to outliers

What is Smoothing and how is it related?

- Tool for summarizing the trend in a response variable as a function of one or more predictor variables
- Is less variable than the original data, hence the name smoother
- Is nonparametric – doesn't force a rigid form of the dependence (e.g., linear, quadratic)

A Nonparametric Smoothing Method

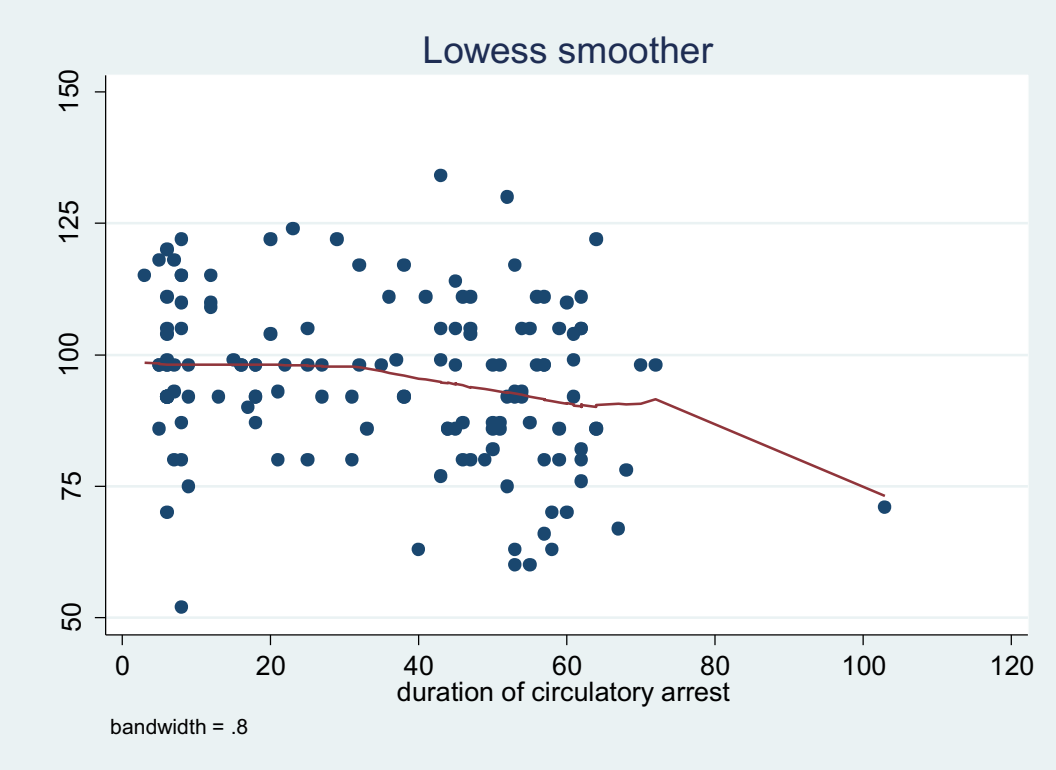
- **Lowess**: Locally weighted running line smoother of *Y* over a (moving) neighborhood of *x* (higher weights the closer you are to the middle of the neighborhood)
  - makes a smooth fit through a scatterplot of data
  - A sophisticated but very useful approach to smoothing
  - Available in Stata, SAS, and R, though exact implementation may vary slightly

Lowess Smoothing

- Use it to help assess for linearity or nonlinearity in either outcome variable or continuous covariates
- Do so either at diagnostic stage, or during testing
- Upside: Provides flexible approach to presenting ‘gist’ of trend
- Downside: can’t be used to obtain simple equation for set of data

Results

```
lowess pdi minutes, xlabel(0(20)120)
ylabel(50(25)150)
```



Finally back to Simple Linear Regression

- Investigates the *linear relationship* between a continuous response or outcome variable as a function of an explanatory or predictor variable or covariate
- Can also estimate the predicted change in the response for a given change in the covariate

With Simple Linear Regression, we are looking at models of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where the Y_i are our outcomes and the X_i are our explanatory variables (covariates).

Simple Linear Regression

- Data: (x_i, Y_i), $i = 1, \ldots, N$

  $x_i = i^{th}$ predictor variable, measured without error

  $Y_i = i^{th}$ outcome variable (random, continuous, may have error)

  N = number of subjects

Simple Linear Regression

- Assumptions:
  - the mean of *Y* is an unknown, but linear, function of *x*
  - the variability of *Y* about its mean value is equal for all *x* values (*homoscedasticity*)
  - the distribution of *Y* about its mean value is *normally distributed*
  - all responses are *independent*

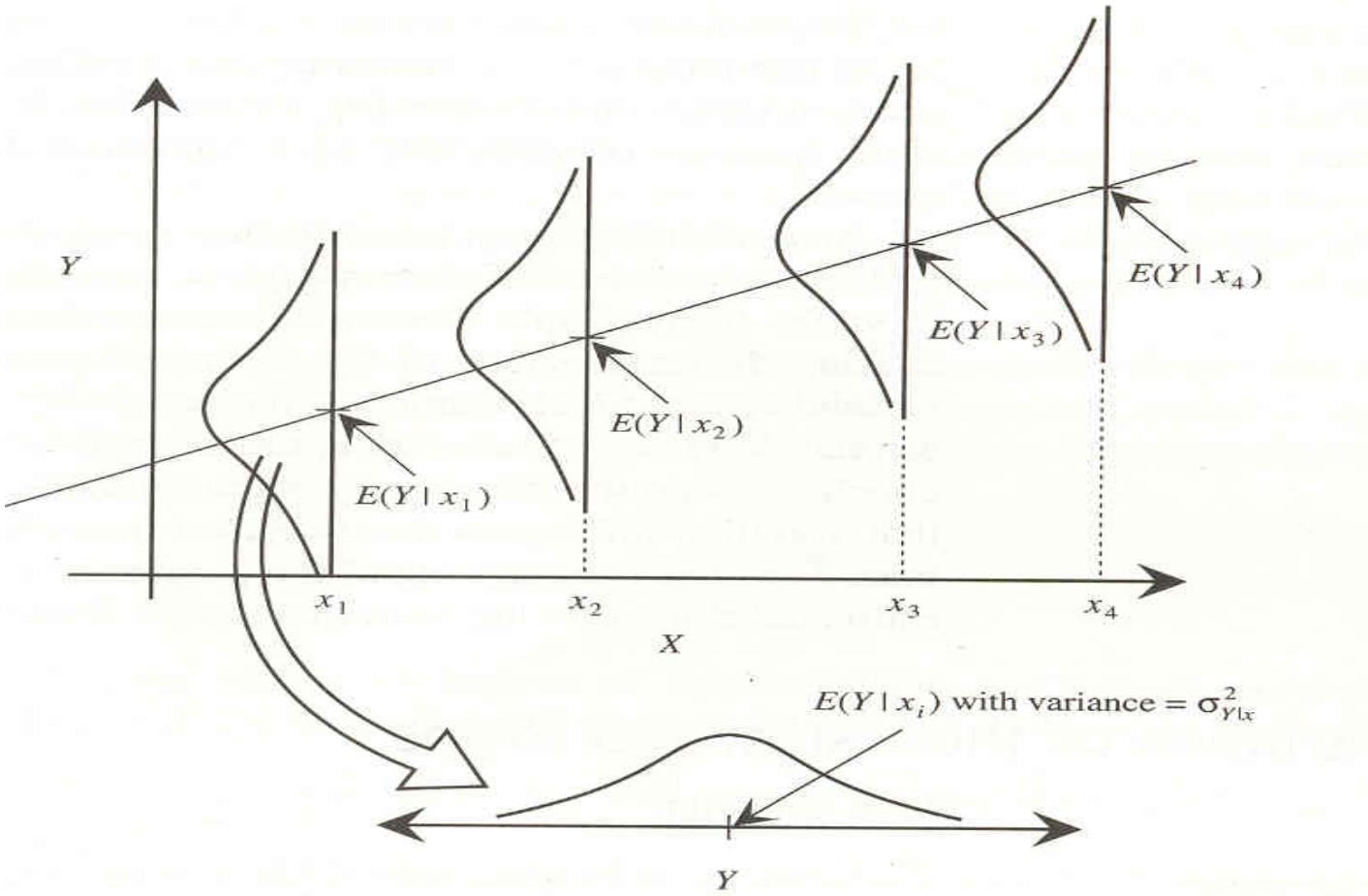
Simple Linear Regression

- Assumptions:
  - **(L)** the mean of *Y* is an unknown, but linear, function of *x*
  - **(I)** all responses are *independent*
  - **(N)** the distribution of *Y* about its mean value is *normally distributed*
  - **(E)** the variability of *Y* about its mean value is equal for all *x* values (*homoscedasticity*)

Simple Linear Regression

- Model: $E(Y) = \beta_0 + \beta_1 \cdot x$
  - $E(Y)$ = expected value of Y for a given x
  - β_0 = intercept, or constant term, corresponding to the mean value of Y when $x = 0$
  - β_1 = slope, or the change in Y corresponding to a 1 unit increase in x

Simple Linear Regression



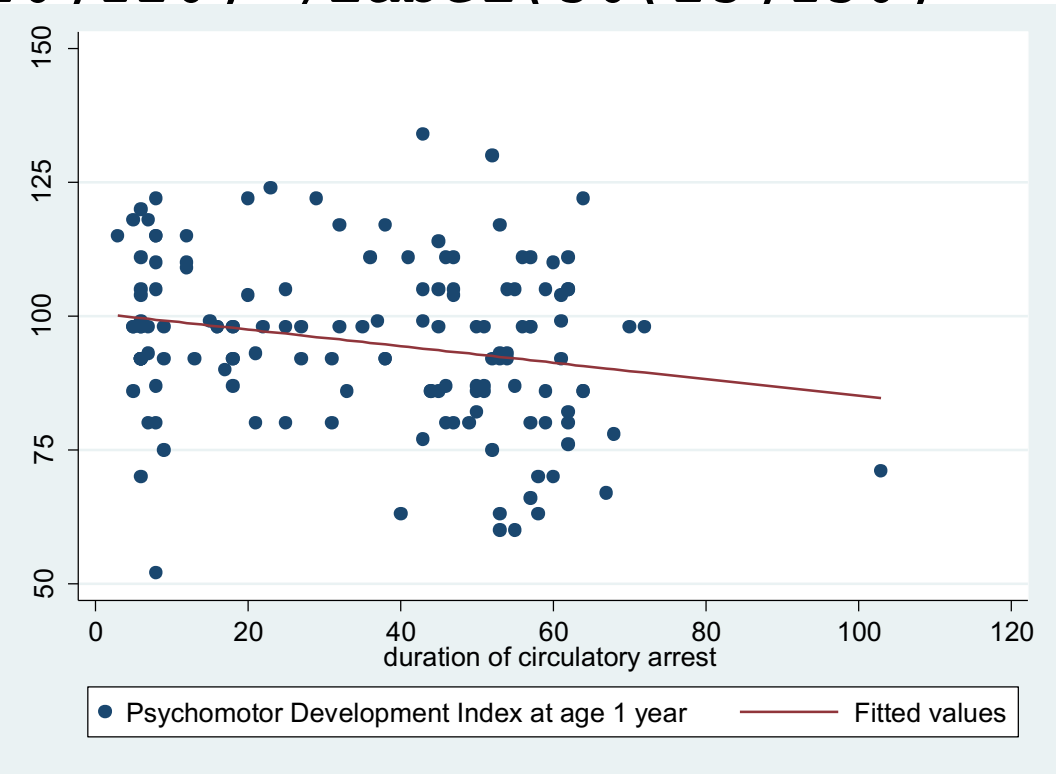
Results

```
regress pdi minutes
```

```
      Source |       SS       df       MS              Number of obs =     142
-------------+------------------------------           F(  1,   140) =    7.10
       Model |  1625.47748     1  1625.47748            Prob > F      =  0.0086
    Residual |  32033.1422   140  228.808159            R-squared     =  0.0483
-------------+------------------------------           Adj R-squared =  0.0415
       Total |  33658.6197   141  238.713615            Root MSE      =  15.126

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     minutes |  -.1545158    .057972    -2.67   0.009    -.2691295   -.0399021
       _cons |   100.5708   2.413829    41.66   0.000     95.79857    105.3431
```

Results

```
predict yhat

scatter pdi yhat minutes, ms(o i) connect(. l)
xlabel(0(20)120) ylabel(50(25)150)
```



Results

- To predict PDI as a function of minutes of CA, the following model was fit:

$$PDI = 100.57 - 0.155 \text{ minutes}$$

- The intercept is 100.57, which is the predicted value of PDI when duration of CA is 0 minutes (that never occurred, but still leave the intercept in the model!)

Results

PDI = 100.57 – 0.155 minutes

- The slope is – 0.155, so each 1 minute increase of CA leads to a 0.155 point decrease in PDI ($P = 0.009$)
- A 60 minute increase of CA leads to a $0.155 \times 60 = 9.3$ point decrease in PDI

Simple Linear Regression

- Because it quantifies the change in *Y* corresponding to a 1 unit increase in *x*, the slope is usually the most important regression model coefficient
- Standard errors, *P*-values, and confidence intervals for β_0 and β_1 can be estimated easily in Stata, SAS, or R

Evaluation of the Model

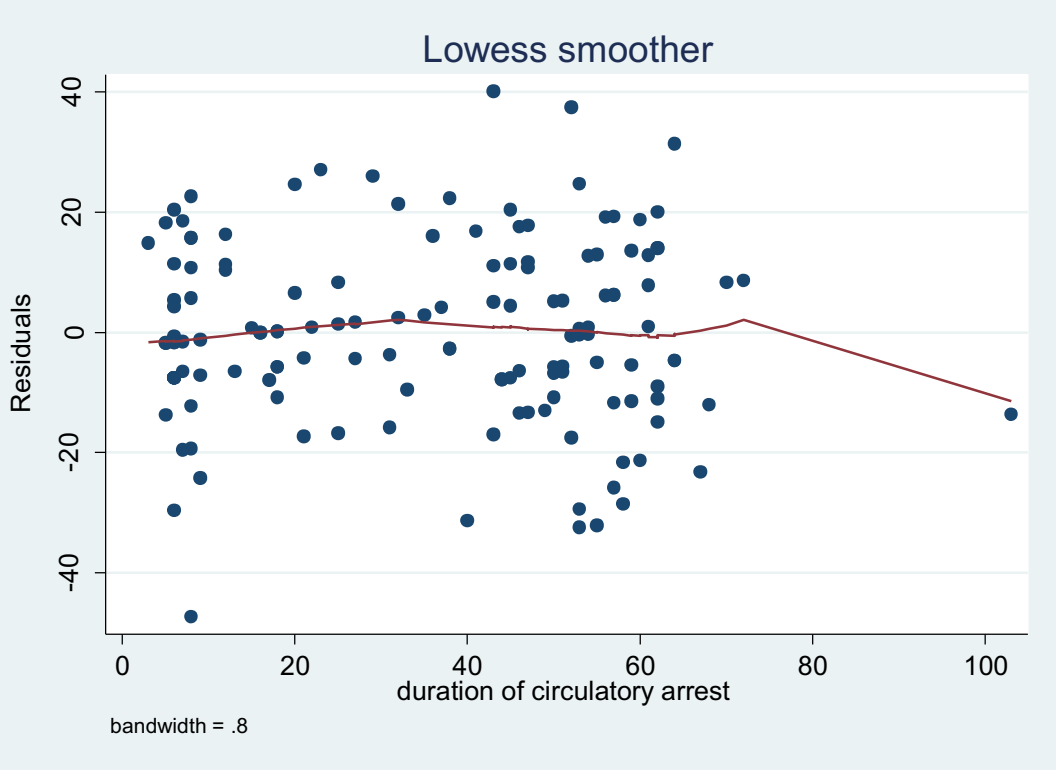
- **Coefficient of determination (R^2)**, the proportion of variability among the observed *Y* values that can be explained by the linear regression of *Y* on *x*
- **Mean Square Error (MSE)**, the estimate of the variance of *Y* about the linear regression of *Y* on *x*

Evaluation of the Model

- **Residual analysis** (differences between the observed data and the fitted values)
  - least squares -- we're minimizing the sum of the squares of the residuals
  - plotting residuals can help detect outliers, assist with checking for normality, and help to add or transform variables in the model

Results

```
predict resids, residuals
lowess resids minutes
```



Limitations of Simple Linear Regression

- Only valid for linear relationships
- Highly sensitive to outliers
- Cannot be extrapolated beyond the observed ranges of the variables
- A strong slope does not necessarily imply a cause-and-effect relationship
- May need to add in other covariates

PDI data come from subjects in this trial

The New England
Journal of Medicine

Volume 329 OCTOBER 7, 1993 Number 15

A COMPARISON OF THE PERIOPERATIVE NEUROLOGIC EFFECTS OF HYPOTHERMIC CIRCULATORY ARREST VERSUS LOW-FLOW CARDIOPULMONARY BYPASS IN INFANT HEART SURGERY

JANE W. NEWBURGER, M.D., M.P.H., RICHARD A. JONAS, M.D., GIL WERNOVSKY, M.D., DAVID WYPIJ, PH.D., PAUL R. HICKEY, M.D., KARL C.K. KUBAN, M.D., S.M., DAVID M. FARRELL, M.A., C.C.P., GREGORY L. HOLMES, M.D., SANDRA L. HELMERS, M.D., JULES CONSTANTINOU, F.R.A.C.P., ENRIQUE CARRAZANA, M.D., JOHN K. BARLOW, M.D.,* AMY Z. WALSH, R.N., B.S.N., KRISTIN C. LUCIUS, R.N., M.S., JANE C. SHARE, M.D., DAVID L. WESSEL, M.D., FRANK L. HANLEY, M.D., JOHN E. MAYER, JR., M.D., ALDO R. CASTANEDA, M.D., AND JAMES H. WARE, PH.D.

PDI data are analyzed in this paper

The New England
Journal of Medicine

Volume 332 MARCH 2, 1995 Number 9

DEVELOPMENTAL AND NEUROLOGIC STATUS OF CHILDREN AFTER HEART SURGERY WITH HYPOTHERMIC CIRCULATORY ARREST OR LOW-FLOW CARDIOPULMONARY BYPASS

DAVID C. BELLINGER, PH.D., RICHARD A. JONAS, M.D., LEONARD A. RAPPAPORT, M.D., DAVID WYPIJ, PH.D., GIL WERNOVSKY, M.D., KARL C.K. KUBAN, M.D., PATRICK D. BARNES, M.D., GREGORY L. HOLMES, M.D., PAUL R. HICKEY, M.D., ROY D. STRAND, M.D., AMY Z. WALSH, R.N., B.S.N., SANDRA L. HELMERS, M.D., JULES E. CONSTANTINOU, F.R.A.C.P., ENRIQUE J. CARRAZANA, M.D., JOHN E. MAYER, M.D., FRANK L. HANLEY, M.D., ALDO R. CASTANEDA, M.D., JAMES H. WARE, PH.D., AND JANE W. NEWBURGER, M.D., M.P.H.

Descriptive Statistics

Table 1. Characteristics of Infants with D-Transposition of the Great Arteries, According to Ventricular Septal Status and Treatment Group.*

VARIABLE	INTACT VENTRICULAR SEPTUM		VENTRICULAR SEPTAL DEFECT	
	CIRCULATORY ARREST (N = 61)	LOW-FLOW BYPASS (N = 59)	CIRCULATORY ARREST (N = 18)	LOW-FLOW BYPASS (N = 17)
	mean ±SD			
Preoperative characteristics				
Birth weight (g)	3601±470	3480±414	3436±270	3564±344
Gestational age (wk)	39.8±1.3	39.7±1.1	40.1±1.2	39.4±1.0
Apgar score at 5 min	8.1±1.0	8.4±0.7	8.7±0.5	8.6±0.6
Age at surgery (days)	7.6±5.7	7.0±3.9	24.1±21.4	13.8±17.5
Surgical data				
Circulatory arrest (min)	52±13	14±12	54±8	33±16
Total bypass time (min)	81±28	127±26	114±26	124±16
Total support time (min)	134±31	141±31	168±27	157±20
	no. with abnormality/total no. (%)			
Postoperative neurologic abnormalities				
Clinical seizures within 7 days	3/61 (5)	0/59	5/18 (28)	1/17 (6)
Ictal activity within 48 hr†	7/54 (13)	6/46 (13)	10/15 (67)	2/11 (18)
	mean ±SD			
Follow-up data at 1 year				
Age (mo)	12.4±0.8	12.5±0.8	12.1±0.4	12.4±0.9
Weight (z score)	−0.30±1.04	−0.41±0.87	−0.62±0.94	−0.18±0.77
Length (z score)	0.17±1.68	−0.37±1.90	−1.11±2.17	−0.74±1.15
Social class‡	41±15	45±14	43±13	40±15
Maternal IQ§	95±14	98±13	95±8	98±14

*Only the 155 children who returned for evaluation at one year of age are included.

†Rhythmic epileptiform activity continuing for more than five seconds on continuous video EEG monitoring.

‡Score on the Hollingshead Four Factor Index of Social Status, with higher scores indicating higher social status.[5]

§Score on the Peabody Picture Vocabulary Test (revised).[6]

Are means and SD's appropriate for all variables?

Outcome Variables

Table 2. Scores on Developmental Tests, According to Ventricular Septal Status and Treatment Group.

Test	Intact Ventricular Septum		Ventricular Septal Defect		P Value*
	Circulatory Arrest	Low Flow Bypass	Circulatory Arrest	Low Flow Bypass	
	mean ±SD				
Psychomotor Development Index	94.1±15.3	99.1±14.2	84.2±18.7	96.0±11.1	0.01
Mental Development Index	106.0±15.6	108.0±12.4	92.8±15.7	104.3±15.1	0.10
Fagan Test of Infant Intelligence†	59.4±7.1	59.8±6.2	54.6±10.7	57.5±9.0	0.49
	no. with low score/total no. (%)				
Psychomotor Development Index ≤80	12/57 (21)	5/54 (9)	8/16 (50)	3/15 (20)	0.02
Mental Development Index ≤80	3/58 (5)	1/54 (2)	3/16 (19)	1/15 (7)	0.27
Fagan Test <53†	10/42 (24)	6/41 (15)	4/13 (31)	3/11 (27)	0.35

*P values are for differences between treatment groups, with adjustment for diagnosis; P values were determined by linear regression for continuous outcome variables and by stratified exact tests for dichotomous outcome variables.

†Restricted to infants between 11 and 13 months of age at the time of examination.

Graphical Display

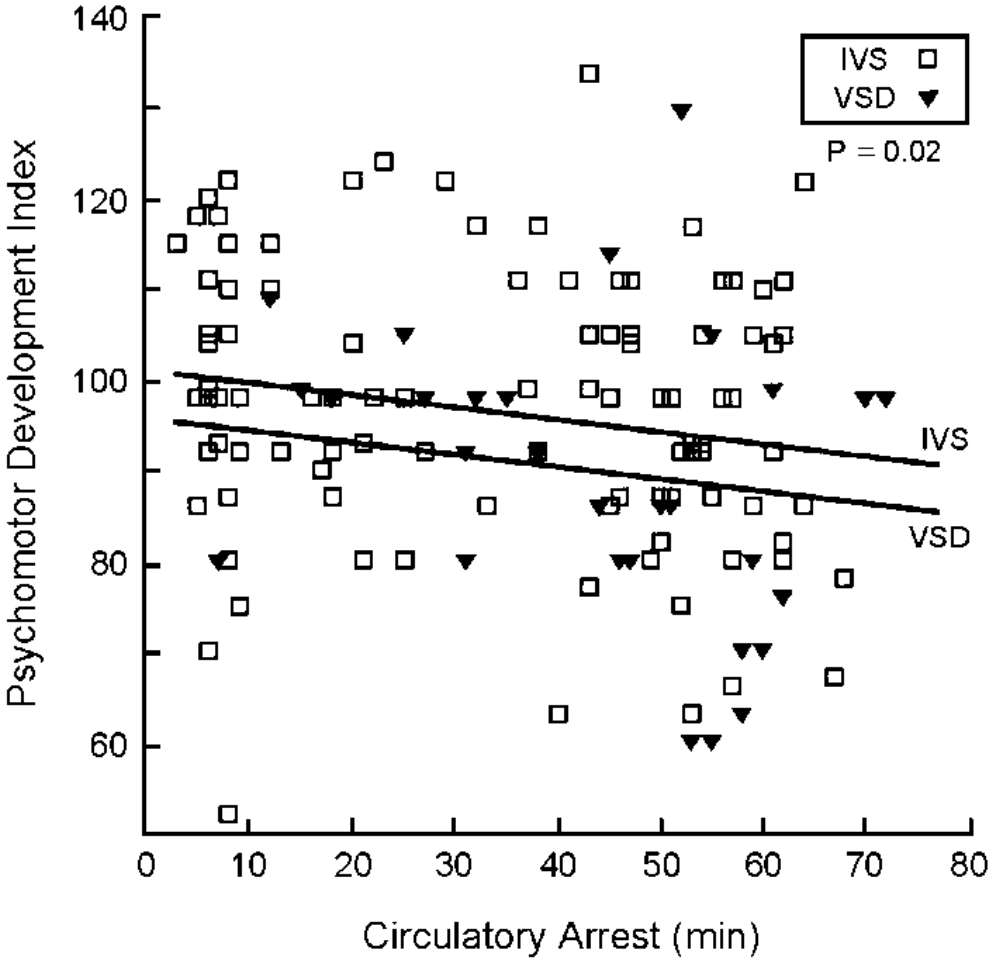


Figure 1. Score on the Psychomotor Development Index at One Year as a Function of the Duration of Total Circulatory Arrest.

Regression lines are shown for infants with an intact ventricular septum (IVS) and those with a ventricular septal defect (VSD). The P value shown was calculated by linear regression for the effect of the duration of total circulatory arrest on the score on the Psychomotor Development Index, with adjustment for diagnosis.

Outlying value?

Begin to develop general framework for approaching analyses

First -

- Learn the topic/study well, really well
- Collaborate to define motivating questions of interest, check PubMed
- What techniques might help to achieve answers? Which do the data warrant? (develop intuition, read literature)
- Possible Confounding or Effect Modification to account for?
- Keep an open mind, and the larger picture – there is no recipe

Begin to develop framework for analysis

Next -

- Diagnostics/Checking Assumptions:
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing (example: Lowess)
- Hypothesis testing/modeling:
  - t-test?
  - Correlation (r)?
  - ANOVA useful?
  - Nonparametric approach better?
  - Linear regression or extensions?
  - Generalizations

Begin to develop framework for analysis

Then -

- Assess Model Fit:
  - R^2
  - MSE
  - Confidence Intervals
  - Residual Analysis
- Interpretation and Inference, constant collaboration around data and meaning
- Backup and regroup as needed, delve deeper, use caution, stay organized