Statistics in Healthcare

Unit 8:

Overview/Teasers

Overview

- Regression I: Linear regression

Common statistics for various types of outcome data

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model (beyond the scope of this course)	Time-varying effects (beyond the scope of this course)

Teaser 1, Unit 8

Headline:

Brighten the twilight years: "Sunshine vitamin" boosts brain function in the elderly

- "Middle-aged and older men with high levels of vitamin D in their blood were mentally quicker than their peers, researchers report."
- "The findings are some of the strongest evidence yet of such a link because of the size of the study and because the researchers accounted for a number of lifestyle factors believed to affect mental ability when older, Lee said."

Teaser 2, Unit 8

My intriguing multivariate analysis: What predicts how much time Stanford students spend on homework?

Varsity Sports in High School increases homework time ($p=.02$)

Liberal politics increases homework time ($p<.0001$)

Liking President Clinton increases homework time ($p=.07$)

Liking President Regan increases homework time ($p=.002$)

Liking President Carter *decreases* homework time ($p=.004$)

Drinking more alcohol decreases homework time ($p=.149$)

Statistics in Medicine

Module 1:

Covariance and correlation

Assumptions of linear models

Assumptions for linear models (ttest, ANOVA, linear correlation, linear regression):

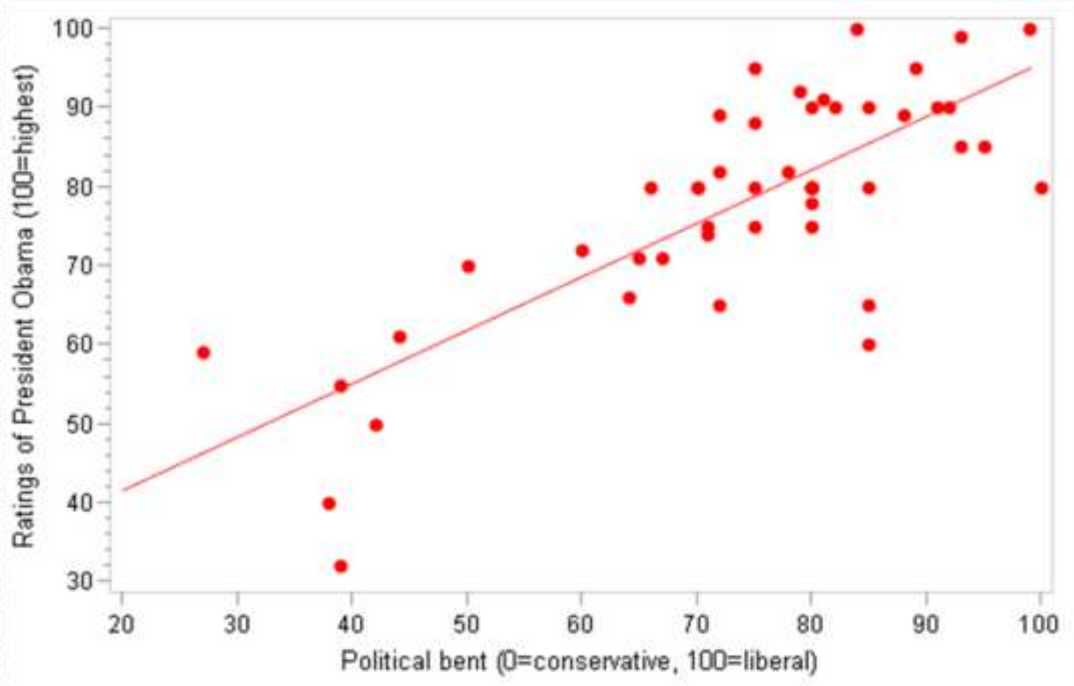
1. Normally distributed outcome variable
   - This assumption is most important for small samples; large samples are quite robust against this assumption because of the central limit theorem (averages are normally distributed even when the underlying trait is not!).
2. Homogeneity of variances
   - For linear regression, the assumption is that variances are equal at all levels of the predictor variable (which may be continuous).
   - Models are robust against this assumption.

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest** (2 groups) **ANOVA** (2 or more groups) **Pearson's correlation coefficient** (1 continuous predictor) **Linear regression** (multivariate regression technique)	**Paired ttest** (2 groups or time-points) **Repeated-measures ANOVA** (2 or more groups or time-points) **Mixed models/GEE modeling**: (multivariate regression techniques)	<u>Non-parametric statistics</u> **Wilcoxon sign-rank test** (alternative to the paired ttest) **Wilcoxon rank-sum test** (alternative to the ttest) **Kruskal-Wallis test** (alternative to ANOVA) **Spearman rank correlation coefficient** (alternative to Pearson's correlation coefficient)

Example: Stanford class data

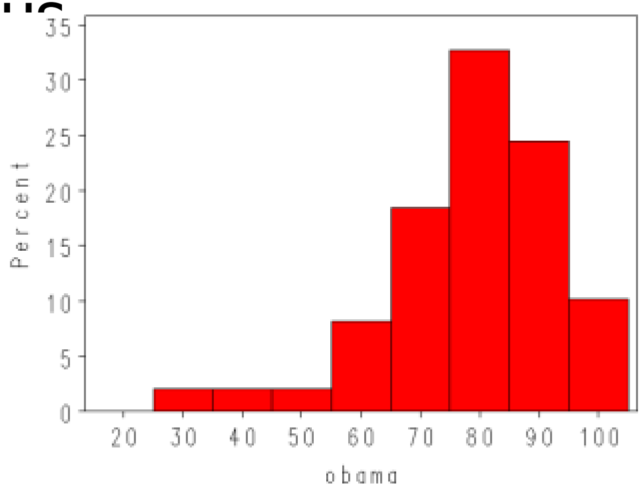
Political Leanings and Rating of Obama



Correlation coefficient

Statistical question: Is political bent related to ratings of President Obama?

- What is the outcome variable? Ratings of Obama
- What type of variable is it? Continuo
- Is it normally distributed? Close enou



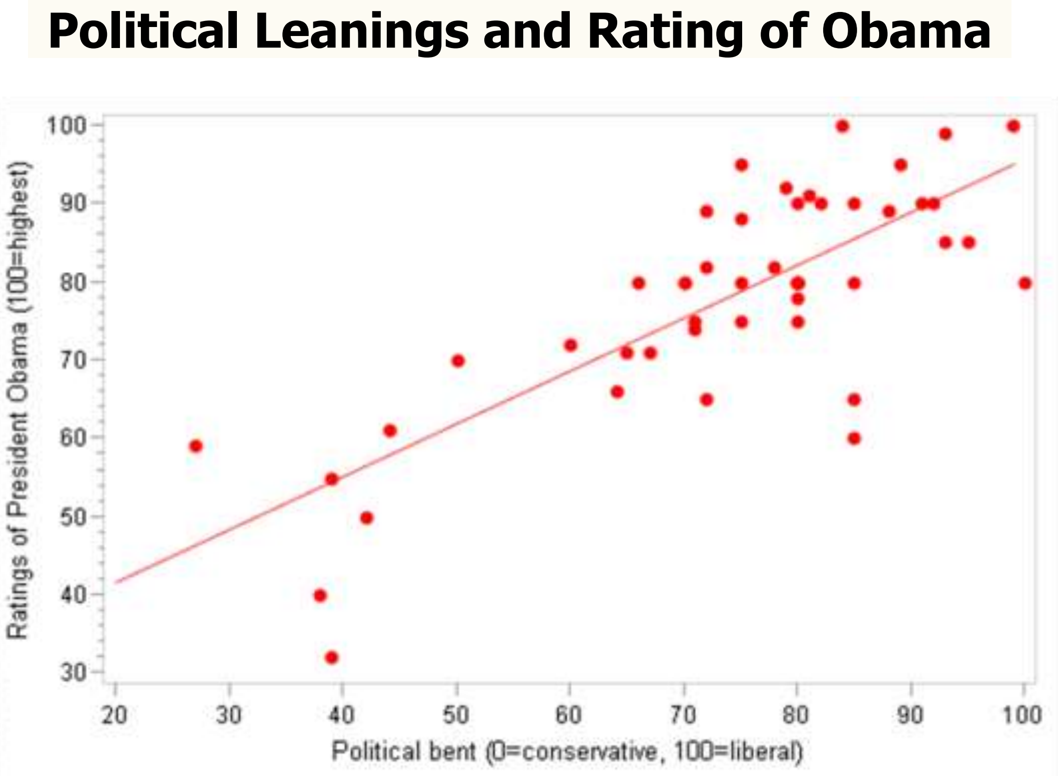
Correlation coefficient

Statistical question: Is political bent related to ratings of President Obama?

- What is the outcome variable? Ratings of Obama
- What type of variable is it? Continuous
- Is it normally distributed? Close enough!
- Are the observations correlated? No
- Are groups being compared? No—the independent variable is also continuous

→ Pearson's correlation coefficient

Example: Stanford class data



r=0.78

p<.0001

New concept: Covariance

$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Var(x) = Cov(x,x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

Interpreting Covariance

cov(X,Y) = 0	X and Y are independent (null value)
cov(X,Y) > 0	X and Y are positively correlated
cov(X,Y) < 0	X and Y are inversely correlated

$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

BUT the magnitude of the covariance depends on Units…

- E.g., kg*m has a different magnitude than lbs*in
- Thus, covariance is hard to interpret
- So…

Correlation coefficient= standardized covariance!

- Divide by the standard deviation of X and the standard deviation of Y to get rid of units and "standardized" covariance:

$$\boxed{r = \frac{\operatorname{cov}(x, y)}{s_x * s_y}}$$

Correlation coefficient

- Pearson's Correlation Coefficient is standardized covariance (unitless):

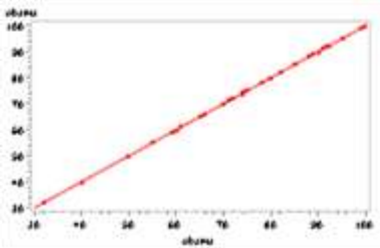
$$\boxed{r = \frac{\mathrm{cov}(x, y)}{s_x * s_y}}$$

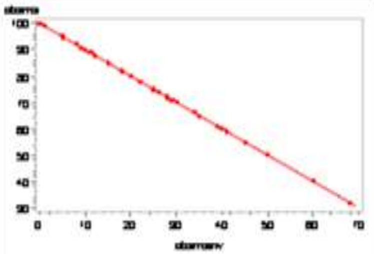
Calculating by hand…

$$\hat{r} = \frac{\mathrm{cov}\,ariance(x,y)}{s_x * s_y} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Correlation

- Measures the strength of the *linear* relationship between two variables
- Ranges between –1 and 1
  - 0: no correlation (independent)
  - -1: perfect inverse correlation
  - +1: perfect positive correlation



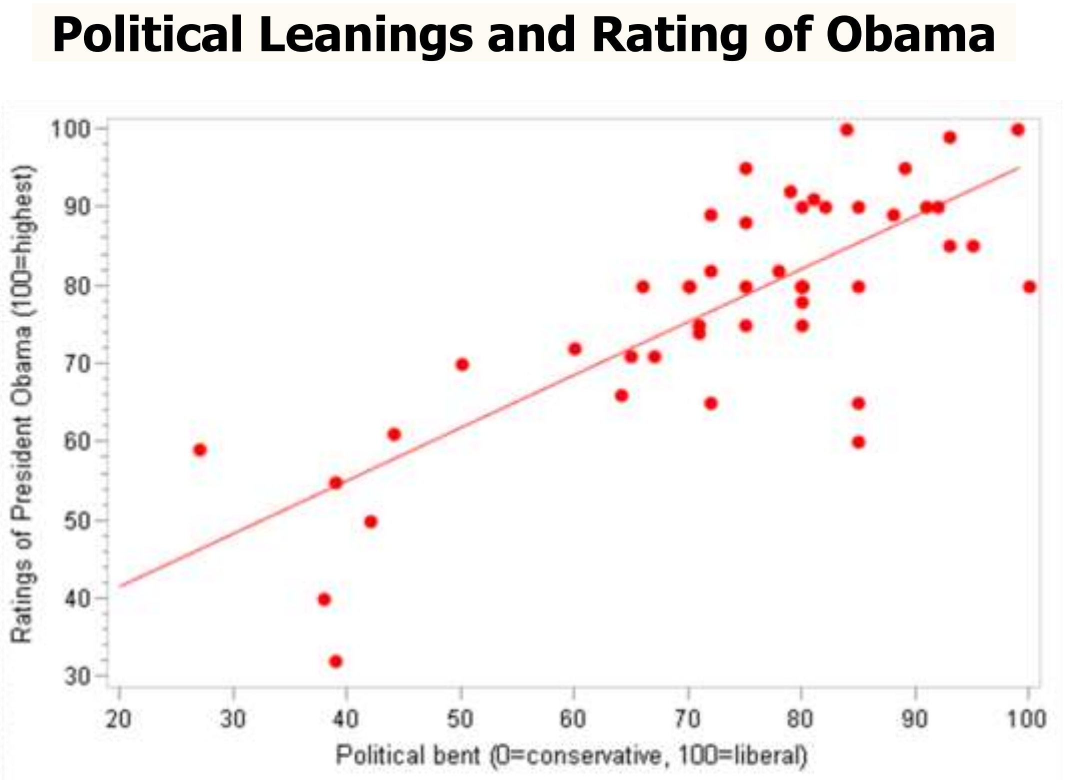


Greater scatter around the line means weaker correlations…

No correlation…

The correlation coefficient measures *linear* correlation!

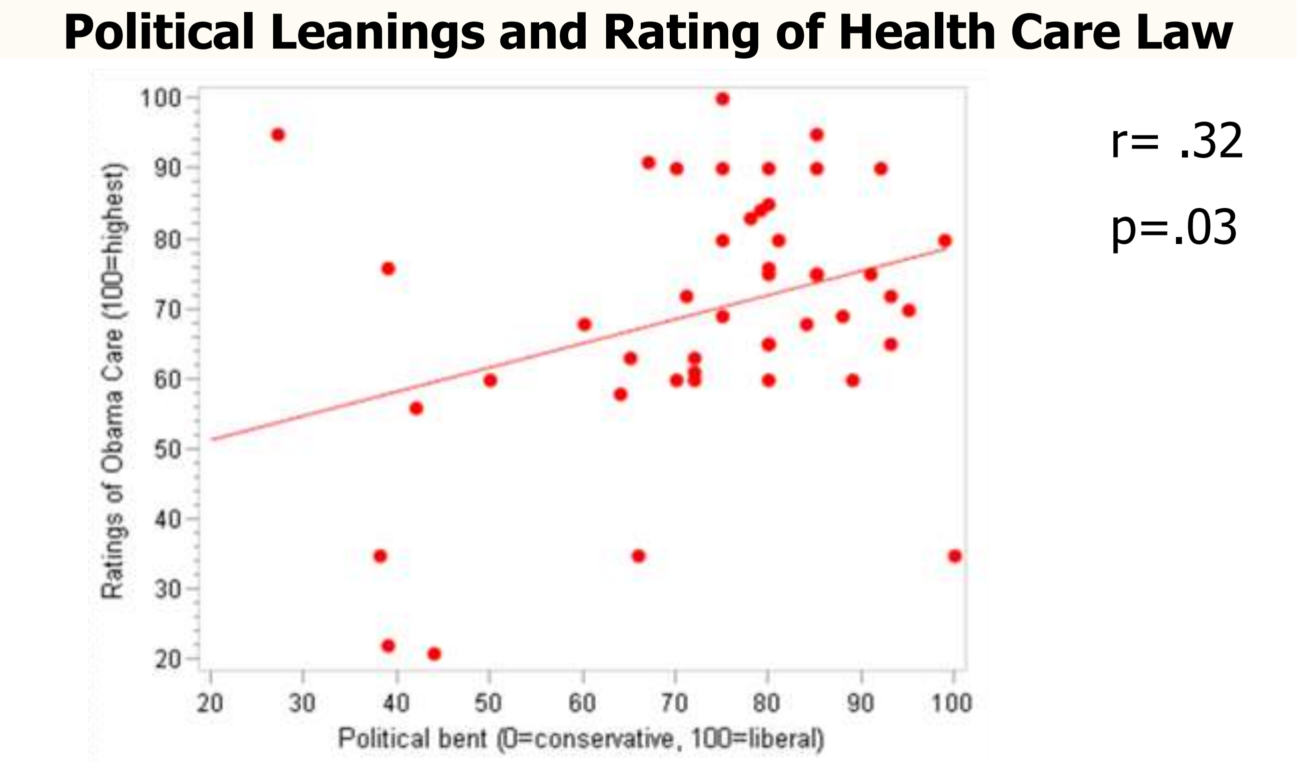
Example: Stanford class data



r=0.78

p<.0001

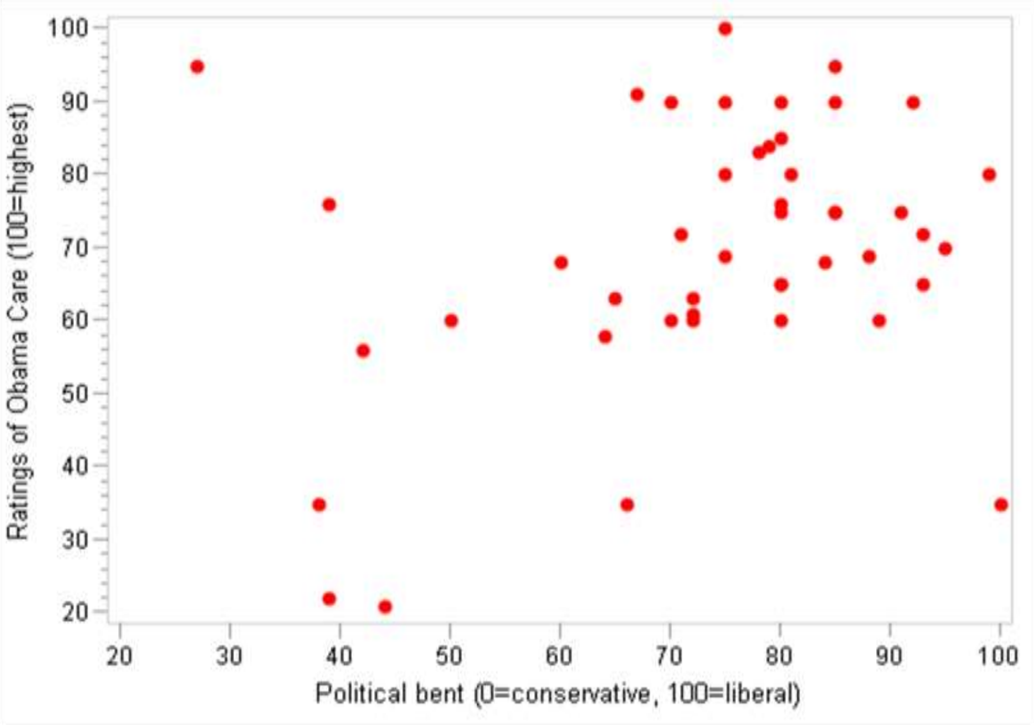
Example: class data



r= .32

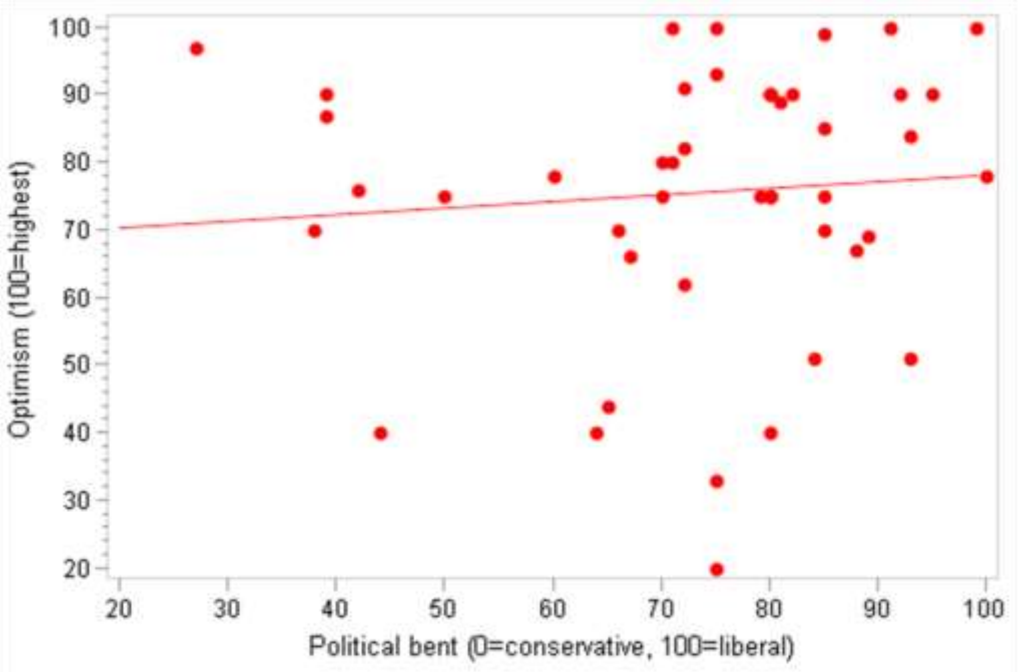
p=.03

With no line superimposed!



Example: class data

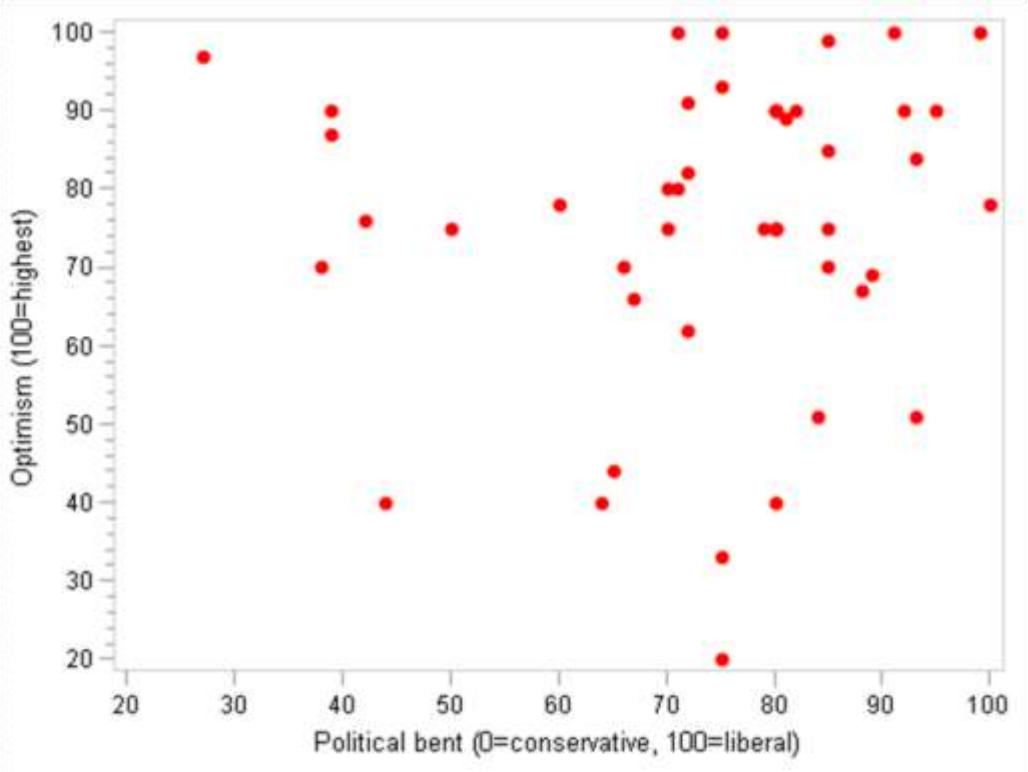
Political Leanings and Optimism



r= .09

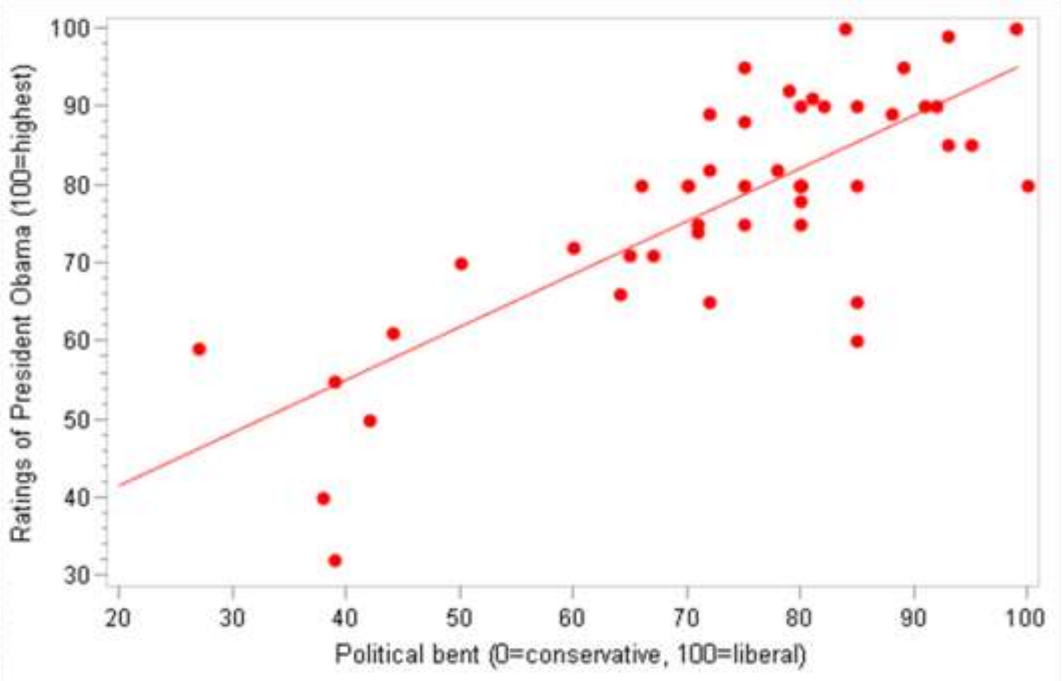
p=.57

With no line superimposed!



New concept: R-squared

Political Leanings and Rating of Obama

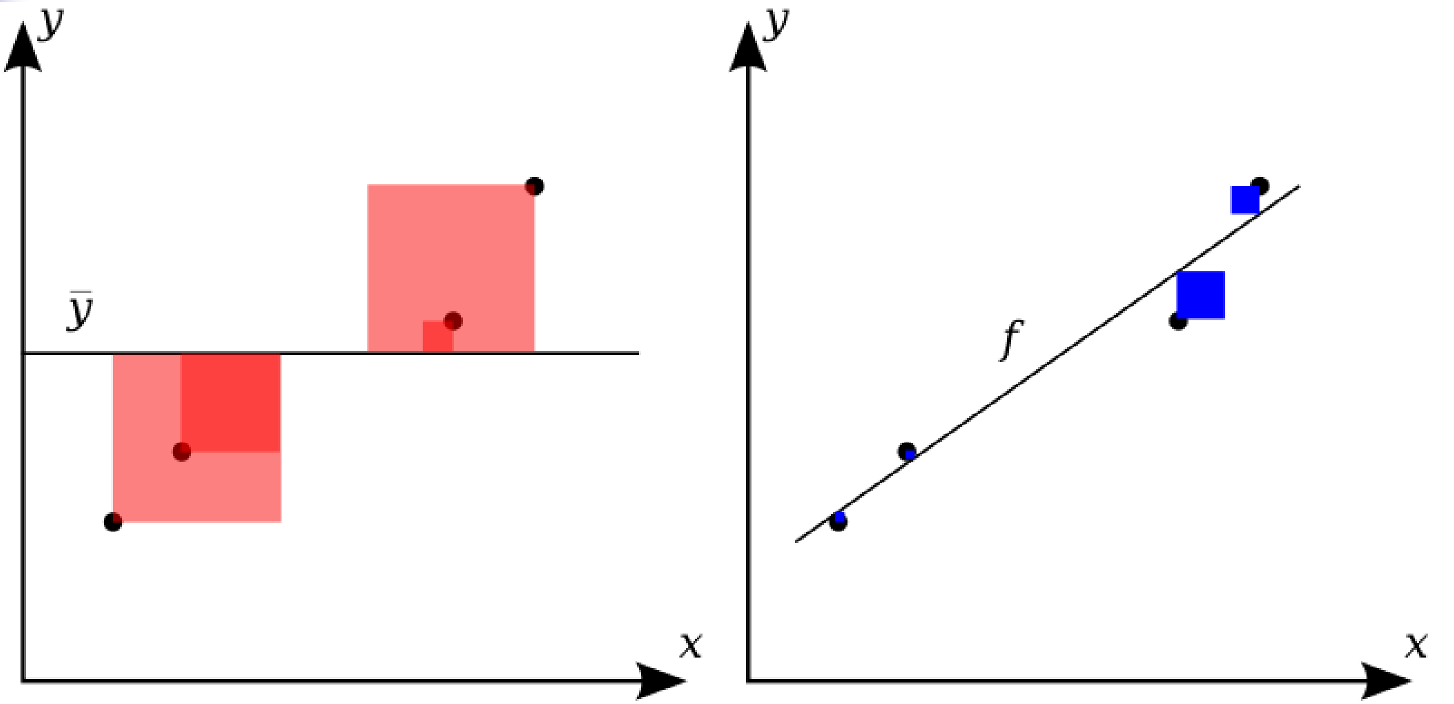


r=.78

R^2=.61

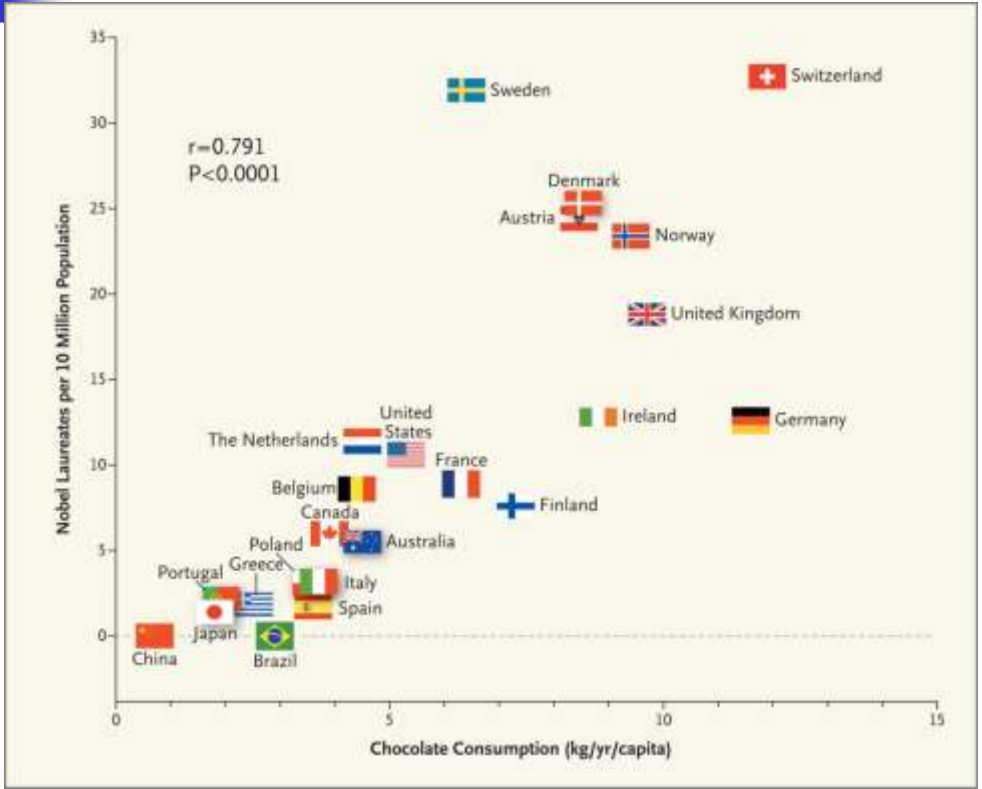
R-squared gives the proportion of variability in the outcome that is "explained by" the predictors. It is also a measure of model fit.

Good illustration: R-squared



Author: Orzetto/Wikipedia Commons

Recall: chocolate and Nobel prize winners:



r= .791

P<.0001

R^2=63%

Reproduced with permission from: Messerli FH. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population. *N Engl J Med* 2012;367:1562-1564.

Weak, moderate, strong?

- There are various rules of thumb with little consensus. A general guide:
  - >.70 is strong correlation
  - >.30 to <.70 is moderate correlation
  - >.10 to <.30 is weak correlation

Depends a lot on the problem at hand!

Inferences about r…

- Null hypothesis:
  - $r = 0$ (no linear relationship)
- Alternative hypothesis:
  - $r \neq 0$ (linear relationship does exist)

Recall: distribution of a correlation coefficient

- 1. Shape of the distribution
  - Normal distribution for larger n!
  - T-distribution for smaller n (<100).
- 2. Mean = true correlation coefficient (r)
- 3. Standard error $\approx \sqrt{\dfrac{1-r^2}{n}}$

To be precise: $\sqrt{\dfrac{1-r^2}{n-2}}$

Thus, for large n (>100):

- Hypothesis test:

$$Z = \frac{\text{r - 0}}{\sqrt{\frac{1-r^2}{n}}}$$

- Confidence Interval

$$\text{confidence interval} = \text{observed r} \pm Z_{\alpha/2} * (\sqrt{\frac{1-r^2}{n}})$$

And smaller n (<100):

- Hypothesis test:

$$T_{n-2} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

- Confidence Interval

$$\text{confidence interval} = \text{observed r} \pm \mathrm{T}_{n-2,\alpha/2} * \left(\sqrt{\frac{1 - r^2}{n - 2}}\right)$$

Sample size and statistical significance, correlation coefficient

The minimum correlation coefficient that will be statistically significant for various sample sizes. Calculated using the approximation, $r = \frac{2}{\sqrt{n}}$

Sample Size	Minimum correlation coefficient that will be statistically significant, p<.05
10	0.63
100	0.20
1000	0.06
10,000	0.02
100,000	0.006
1,000,000	0.002

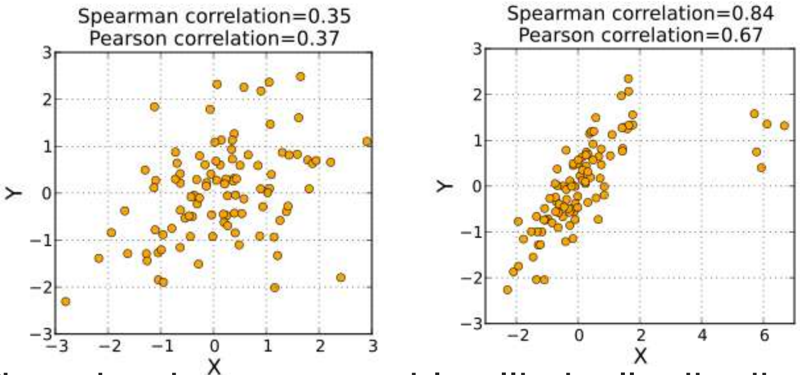
Sainani KL. Clinical versus statistical significance. *PM&R*. 2012;4:442-5.

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest** (2 groups) **ANOVA** (2 or more groups) **Pearson's correlation coefficient** (1 continuous predictor) **Linear regression** (multivariate regression technique)	**Paired ttest** (2 groups or time-points) **Repeated-measures ANOVA** (2 or more groups or time-points) **Mixed models/GEE modeling**: (multivariate regression techniques)	<u>Non-parametric statistics</u> **Wilcoxon sign-rank test** (alternative to the paired ttest) **Wilcoxon rank-sum test** (alternative to the ttest) **Kruskal-Wallis test** (alternative to ANOVA) **Spearman rank correlation coefficient** (alternative to Pearson's correlation coefficient)

Spearman rank correlation coefficient example

Data Element	X	Y	Rank(X)	Rank(Y)
1	0	0	11	11
2	0.25	1	10	10
3	0.5	2	9	9
4	1	3	8	8
5	2	3.5	7	7
6	3	3.75	6	6
7	4	4	5	5
8	5	4.5	4	4
9	5.5	6	3	3
10	6.5	6.5	2	2
11	7	7	1	1
Correlation	0.957		1.000	



- When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.
- The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's correlation limits the outlier to the value of its rank

Statistics in Medicine

Module 2:

Simple linear regression

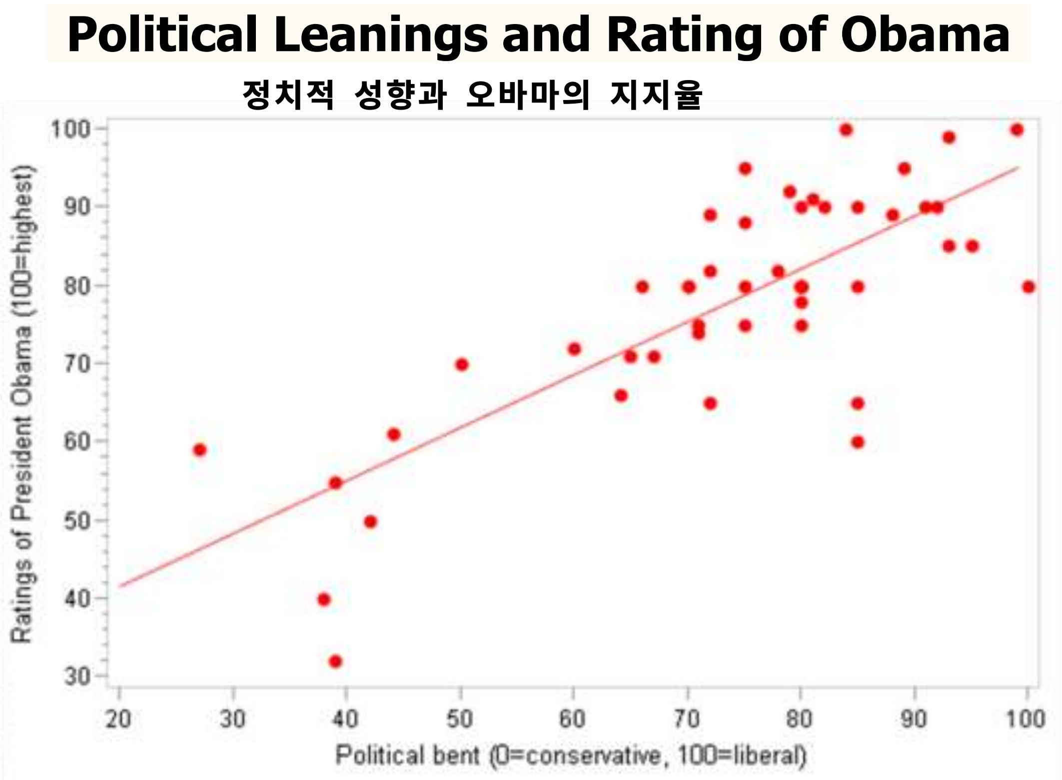
Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated <u>and</u> small sample size:
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest** (2 groups) **ANOVA** (2 or more groups) **Pearson's correlation coefficient** (1 continuous predictor) **Linear regression** (multivariate regression technique)	**Paired ttest** (2 groups or time-points) **Repeated-measures ANOVA** (2 or more groups or time-points) **Mixed models/GEE modeling**: (multivariate regression techniques)	<u>Non-parametric statistics</u> **Wilcoxon sign-rank test** (alternative to the paired ttest) **Wilcoxon rank-sum test** (alternative to the ttest) **Kruskal-Wallis test** (alternative to ANOVA) **Spearman rank correlation coefficient** (alternative to Pearson's correlation coefficient)

Linear regression vs. correlation:

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (*X*) and the other the dependent (=outcome) variable *Y*.

Example: class data



Simple linear regression

Statistical question: Does political leaning "predict" ratings of Obama?

결과는 오바마의 지지율

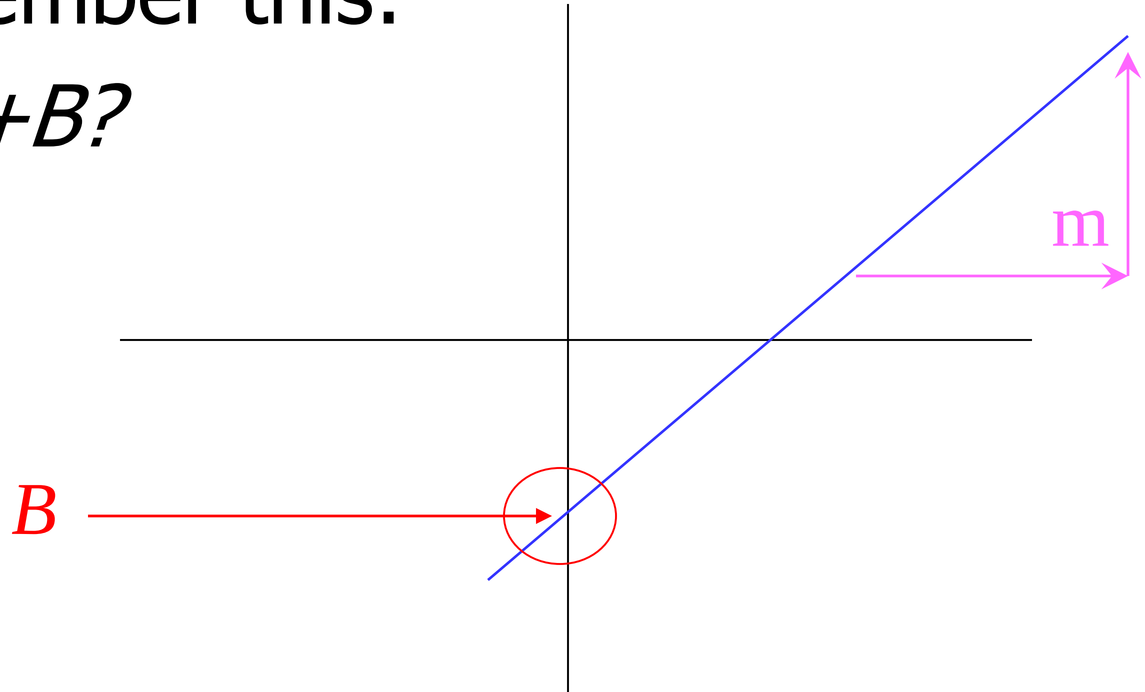
- What is the outcome variable? Obama ratings
- What type of variable is it? Continuous 변수의 종류는 연속적
- Is it normally distributed? Close enough! 정규 분포를 따라야 하고
- Are the observations correlated? No 관측피 간에 상관관계가 있는가?
- Are groups being compared? No—the independent variable is also continuous 두 변수가 연속적이고 상호 독립적이어야함

→ simple linear regression

What is “Linear”?

- Remember this:

Y=mX+B?



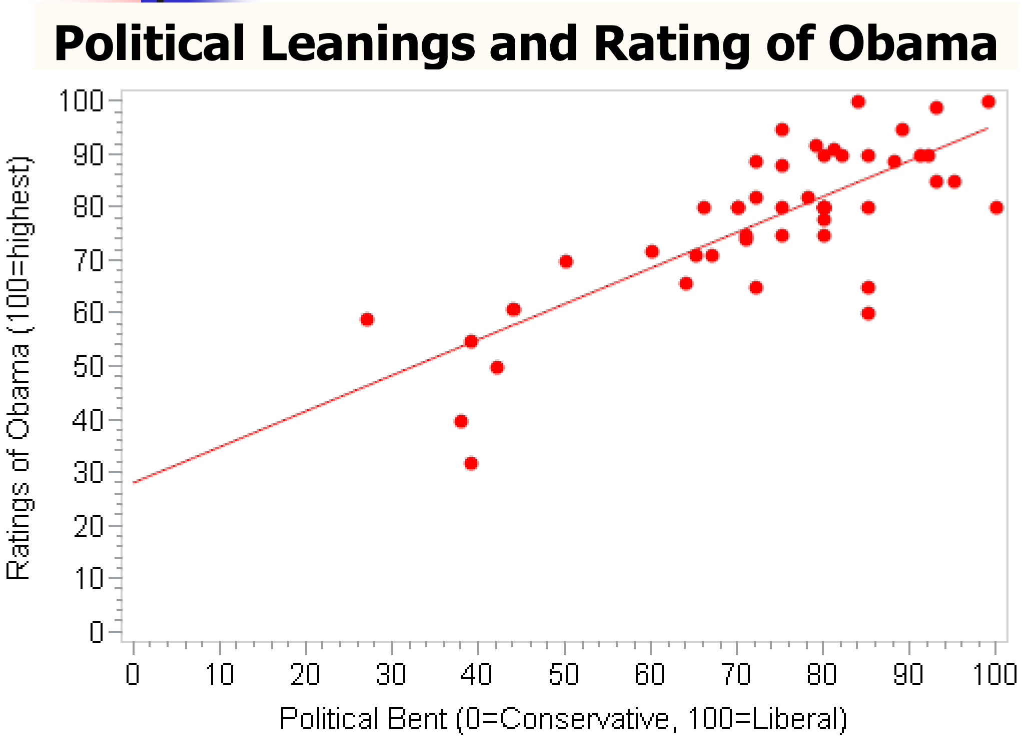
What’s Slope (or gradient)?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

What's Intercept?

The intercept is just the value of Y when X=0.

Example data:



What's the equation of this line (="The best fit line")?

Prediction

If you know something about X, this knowledge helps you predict something about Y. (Sound familiar?…sound like conditional probabilities?)

Regression equation…

Expected value of y at a given level of *x*=

$$\boxed{E(y_i / x_i) = \alpha + \beta x_i}$$

Predicted value for an individual…

$$\underbrace{\hat{y}_i = \alpha + \beta * x_i}_{\text{Fixed – exactly on the line}} + \boxed{\text{random error}_i}$$

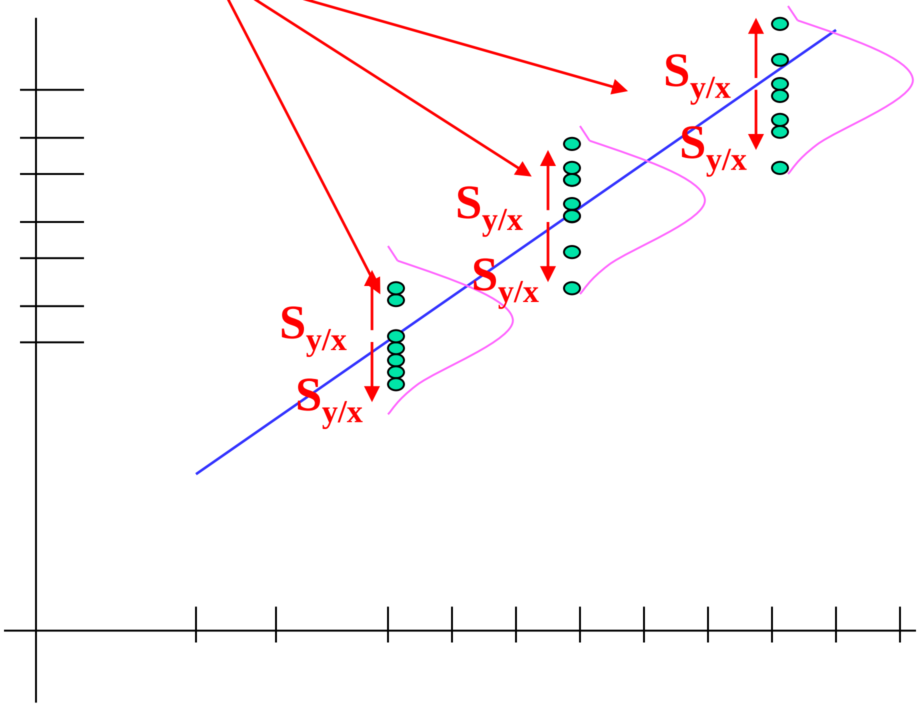
Fixed – exactly on the line

Follows a normal distribution

Assumptions (or the fine print)

- Linear regression assumes that…
  - **1. The relationship between X and Y is linear**
    x와 y 사이의 관계는 선형이다
  - 2. Y is distributed normally at each value of X
    Y는 모든 X에 대해서 정규분포를 따라야 한다
  - 3. The variance of Y at every value of X is the same (homogeneity of variances)
    모든 X에 대한 Y의 분산은 같아야 한다
  - 4. The observations are independent
    측정값은 독립적이어야한다

The standard error of Y given X is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.

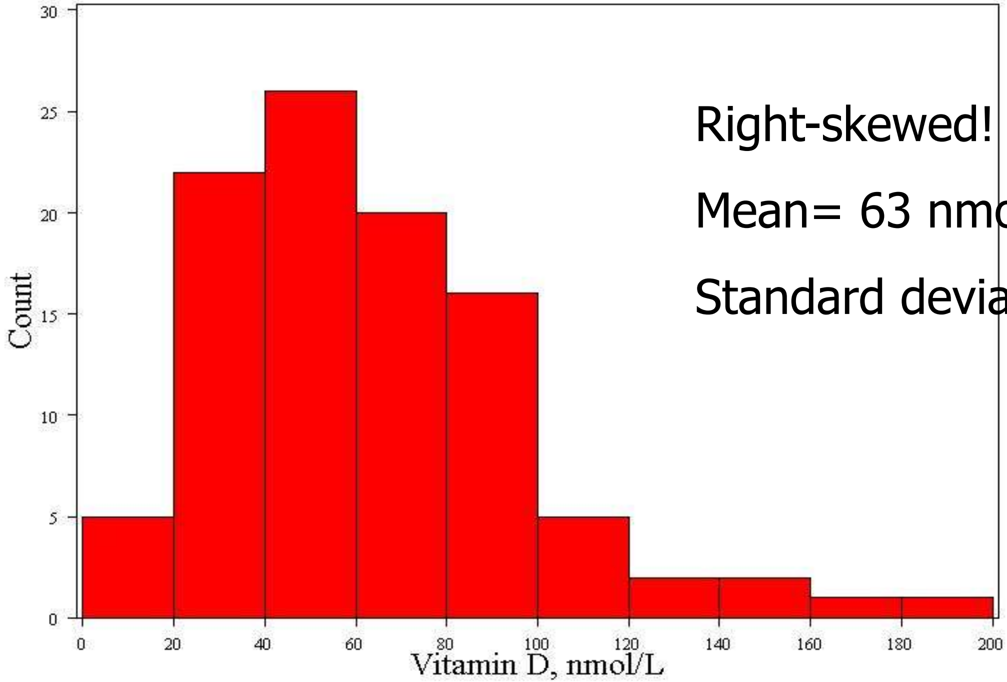


Recall example: cognitive function and vitamin D

- Hypothetical data loosely based on [1]; cross-sectional study of 100 middle-aged and older European men.
  - Cognitive function is measured by the Digit Symbol Substitution Test (DSST).

1. Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.

Sample data: vitamin D (n=100)

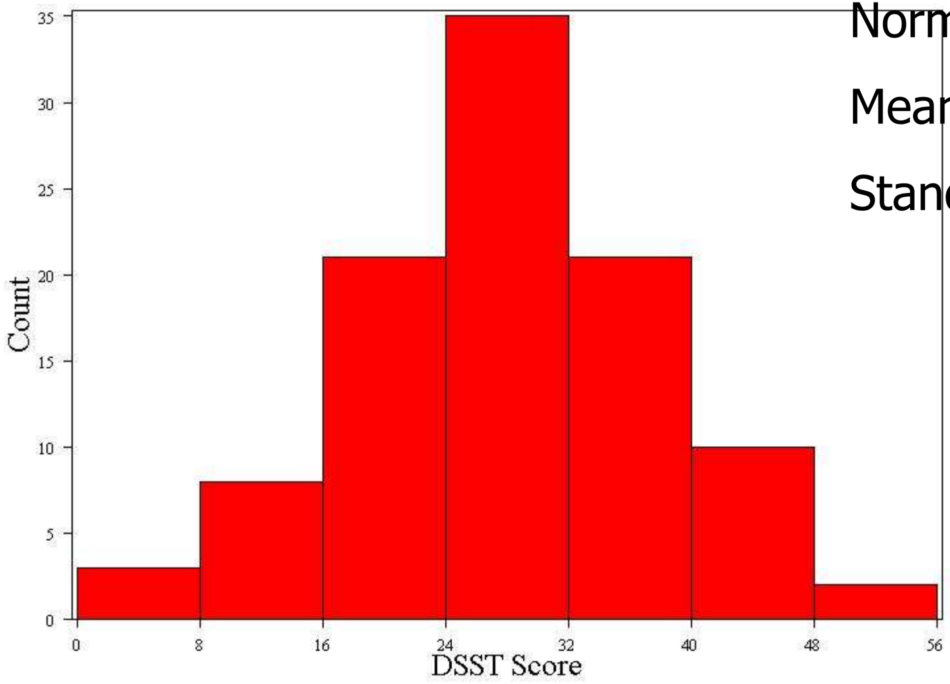


Right-skewed!

Mean= 63 nmol/L

Standard deviation = 33 nmol/L

Distribution of DSST



Normally distributed

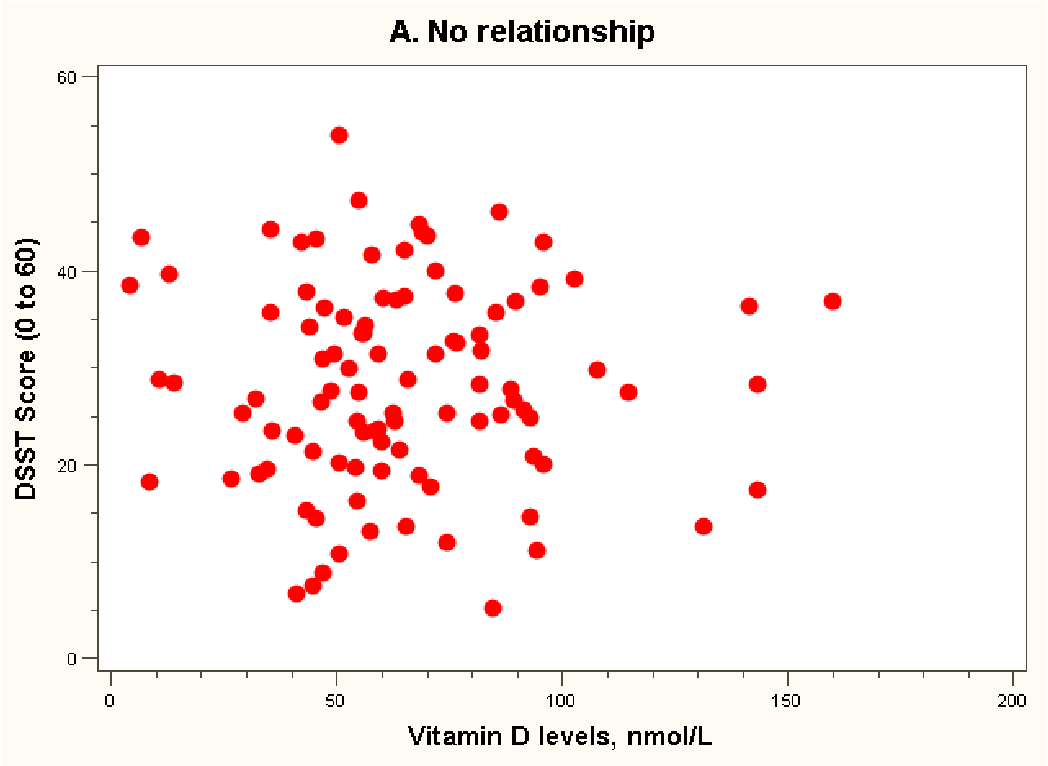
Mean = 28 points

Standard deviation = 10 points

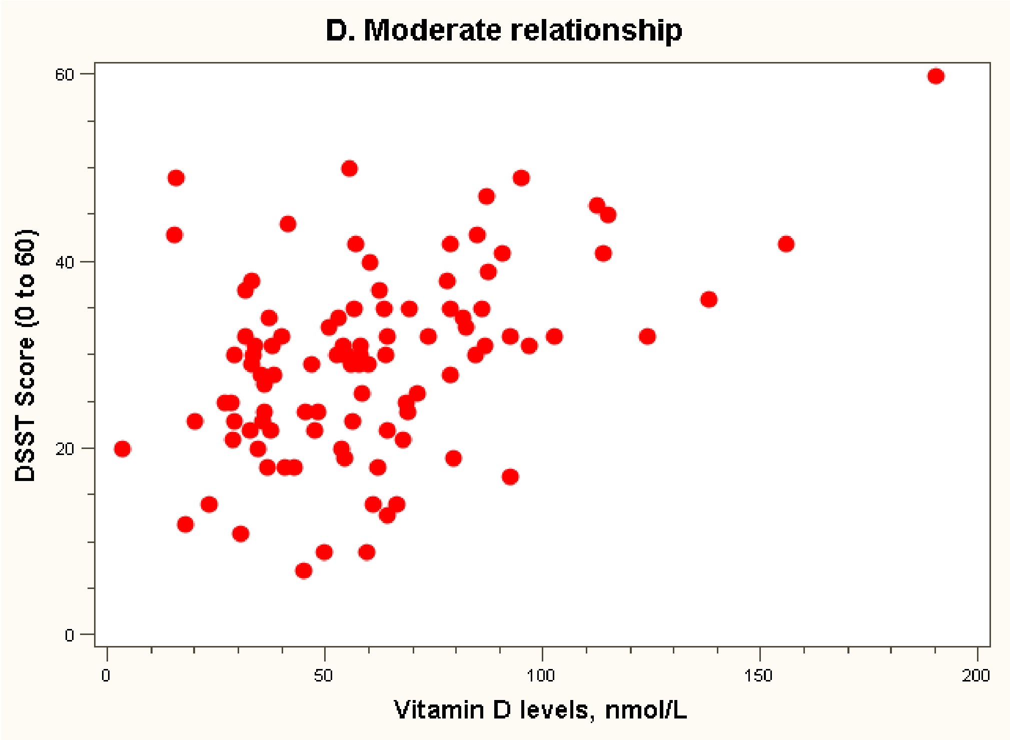
Four hypothetical datasets

- Four hypothetical datasets, with increasing TRUE slopes (between vit D and DSST) generated:
  - 0
  - 0.5 points per 10 nmol/L
  - 1.0 points per 10 nmol/L
  - 1.5 points per 10 nmol/L

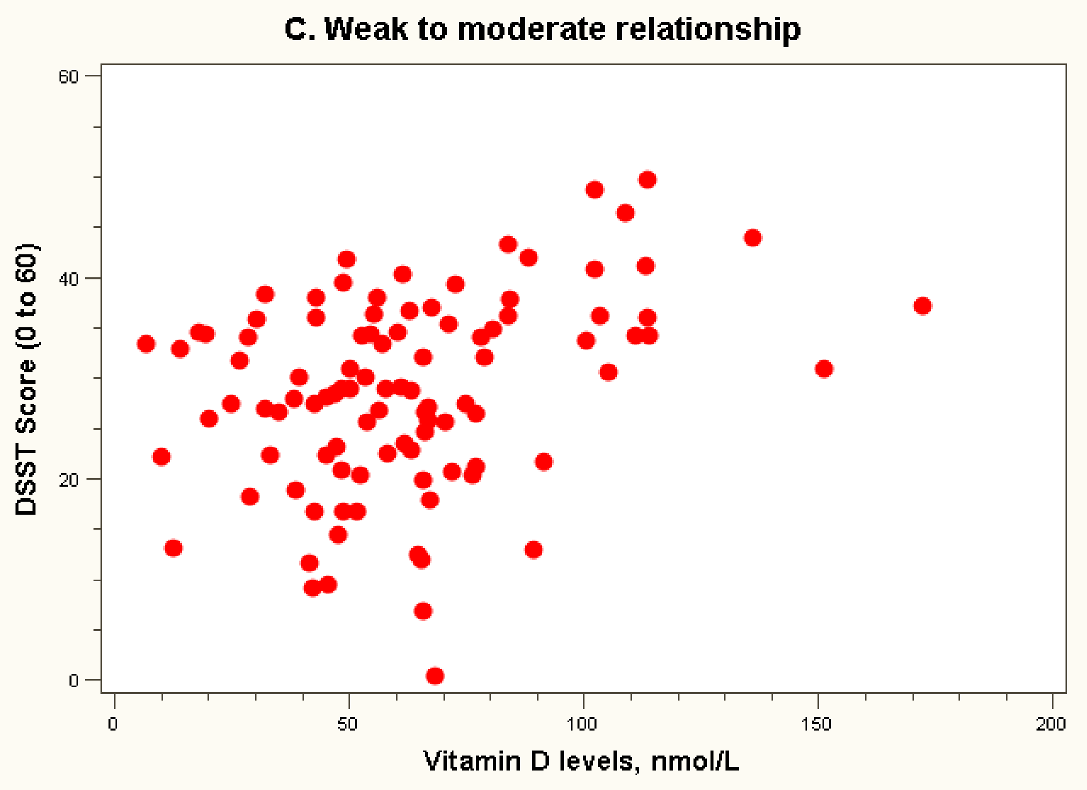
Dataset 1: no relationship



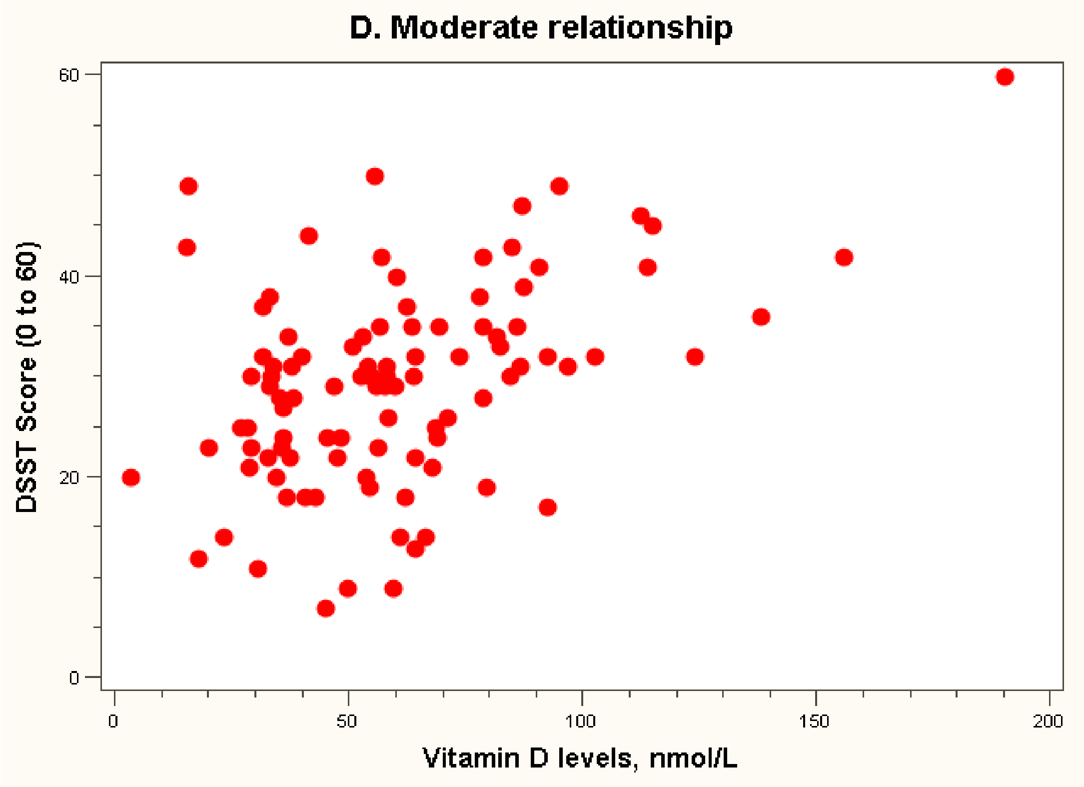
Dataset 2: weak relationship



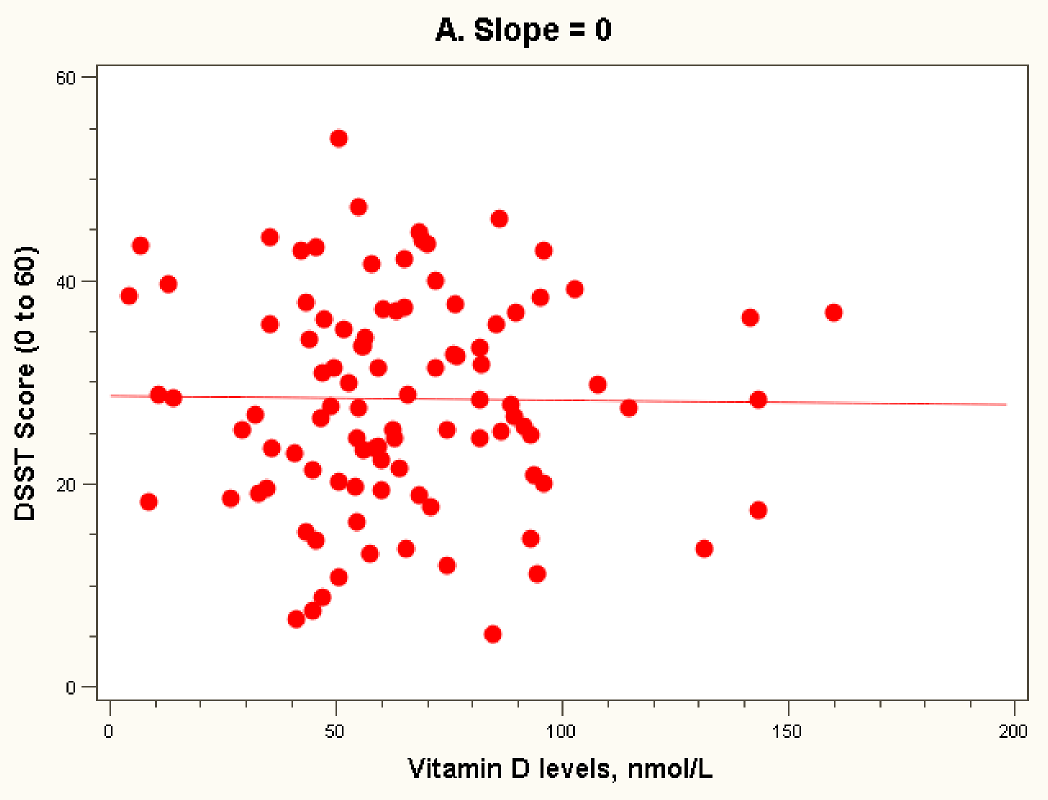
Dataset 3: weak to moderate relationship



Dataset 4: moderate relationship



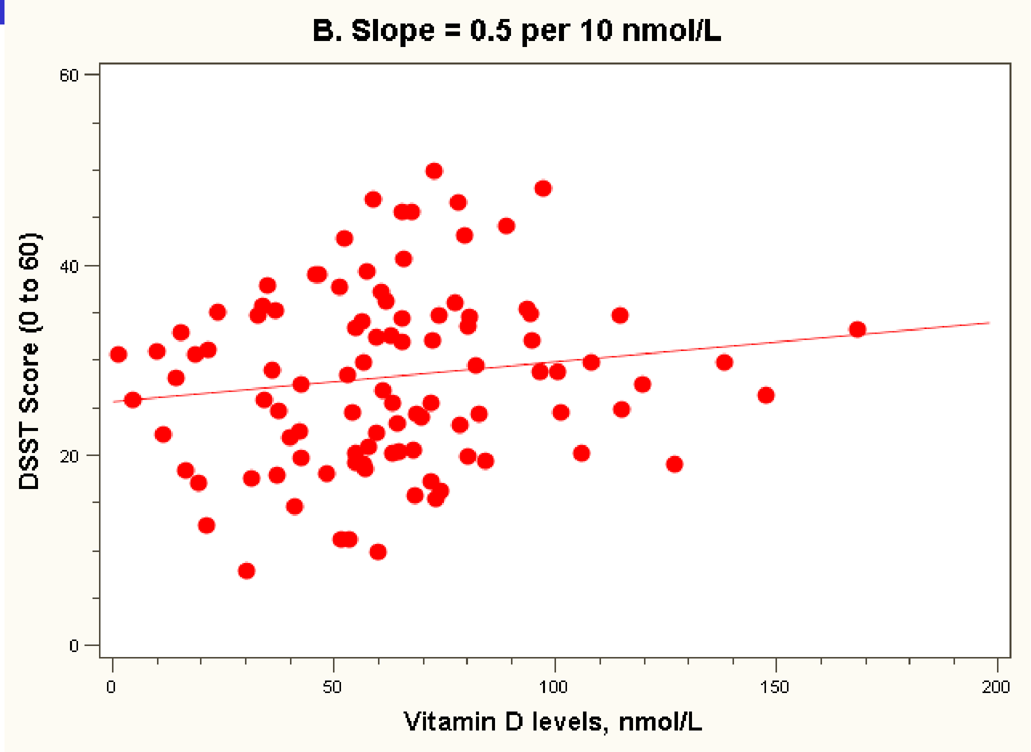
The "Best fit" line



Regression equation:

$E(Y_i) = 28 + 0*vit\ D_i$ (in 10 nmol/L)

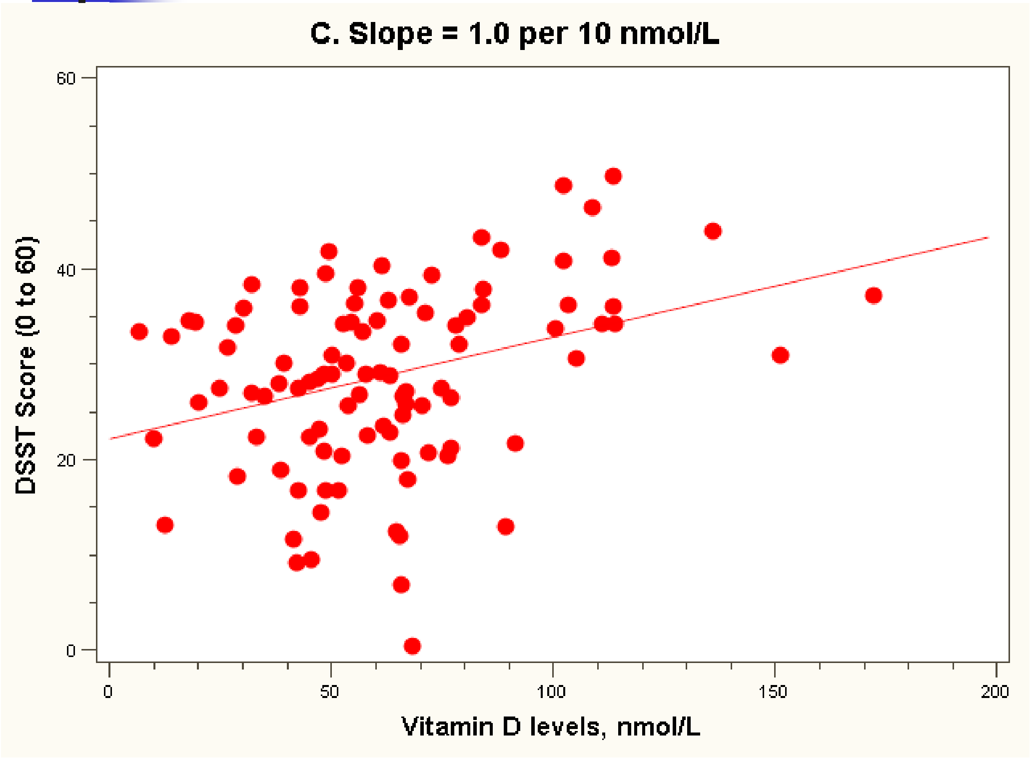
The "Best fit" line



Regression equation:

$E(Y_i) = 26 + 0.5*\text{vit } D_i$ (in 10 nmol/L)

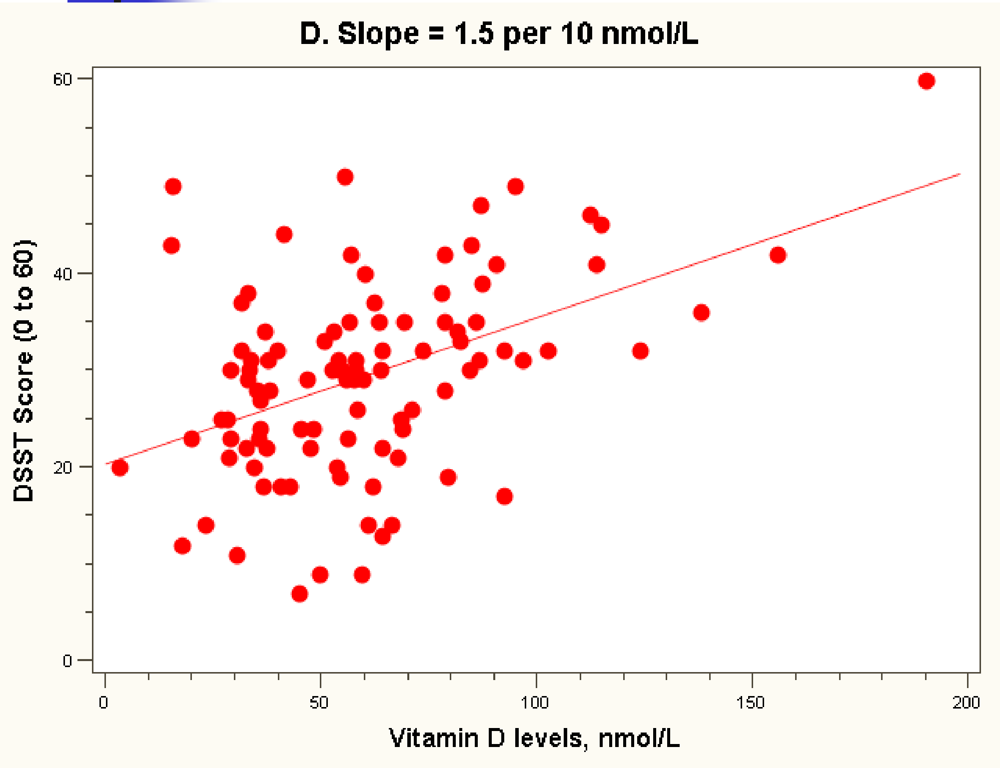
The “Best fit” line



Regression equation:

$E(Y_i) = 22 + 1.0*vitD_i$ (in 10 nmol/L)

The "Best fit" line



Regression equation:

$E(Y_i) = 20 + 1.5*\text{vit } D_i$ (in 10 nmol/L)

Note: all the lines go through the point (63, 28)!

How is the “best fit” line estimated?

- Least squares estimation!

Estimating the intercept and slope: least squares estimation

A little calculus….

What are we trying to estimate? **β, the slope,** from

What's the constraint? We are trying to minimize the squared distance (hence the "least squares") between the observations themselves and the predicted values , or (<u>also called the "residuals", or left-over unexplained variability</u>)

Difference$_i$ = $y_i - (\beta x + \alpha)$ Difference$_i^2$ = $(y_i - (\beta x + \alpha))^2$

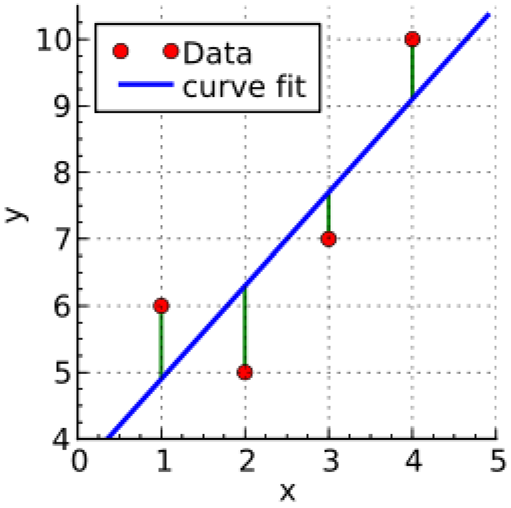
Find the β that gives the minimum sum of the squared differences. How do you minimize/maximize a function? Take the derivative; set it equal to zero; and solve. Typical max/min problem from calculus….

$$\frac{d}{d\beta}\sum_{i=1}^{n}(y_i - (\beta x_i + \alpha))^2 = 2(\sum_{i=1}^{n}(y_i - \beta x_i - \alpha)(-x_i))$$

$$2(\sum_{i=1}^{n}(-y_i x_i + \beta x_i^2 + \alpha x_i)) = 0...$$

From here takes a little math trickery to solve for β…

Linear regression example



We hope to find a line $y = \beta_1 + \beta_2 x$ that best fits these four points.

Data:
$$\begin{aligned} \beta_1 + 1\beta_2 &= 6 \\ \beta_1 + 2\beta_2 &= 5 \\ \beta_1 + 3\beta_2 &= 7 \\ \beta_1 + 4\beta_2 &= 10 \end{aligned}$$

Sum of squares:
$$\begin{aligned} S(\beta_1, \beta_2) &= [6 - (\beta_1 + 1\beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 \\ &+ [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2 \\ &= 4\beta_1^2 + 30\beta_2^2 + 20\beta_1\beta_2 - 56\beta_1 - 154\beta_2 + 210. \end{aligned}$$

Minimize: $\frac{\partial S}{\partial \beta_1} = 0 = 8\beta_1 + 20\beta_2 - 56$ $\quad \frac{\partial S}{\partial \beta_2} = 0 = 20\beta_1 + 60\beta_2 - 154.$

Results: $\beta_1 = 3.5$ $\quad \beta_2 = 1.4$ $\quad y = 3.5 + 1.4x$

Also works for other models (quadratic, etc)

Resulting formulas…

Slope (beta coefficient) = $$\hat{\beta} = \frac{Cov(x,y)}{Var(x)}$$

Intercept= $$\text{Calculate}: \hat{\alpha} = \overline{\text{y}} - \hat{\beta}\overline{\text{x}}$$

Regression line always goes through the point: $(\bar{x}, \bar{y})$

Relationship with correlation

$$\hat{r} = \hat{\beta}\frac{SD_x}{SD_y}$$

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (*X*) and the other the dependent (=outcome) variable *Y*.

Inferences about beta coefficients:

- Null hypothesis:
  - $\beta_1 = 0$ (no linear relationship)
- Alternative hypothesis:
  - $\beta_1 \neq 0$ (linear relationship does exist)

What's the distribution of a beta coefficient?

- Shape: T-distribution
- Mean: True slope
- Standard error:

$$s_{\hat{\beta}} = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{s_{y/x}^{2}}{\sum (x_i - \bar{x})^2}}$$

You don't need to calculate this by hand!!

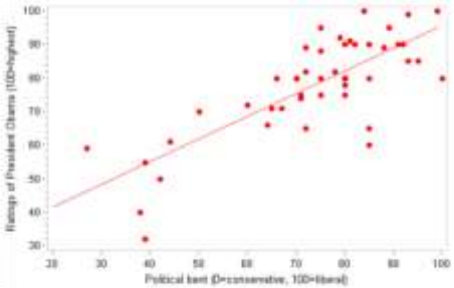
Example: hypothetical dataset 4

- Standard error (beta) = 0.3
- $T_{98} = 1.\ 5/0.3 = 5$, $p<.0001$

- 95% Confidence interval = 0.9 to 2.1 (per 10 nmol/L vit D)

Example: Obama and politics

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	27.97263	6.07079	4.61	<.0001
politics	politics	1	0.67550	0.08020	8.42	<.0001

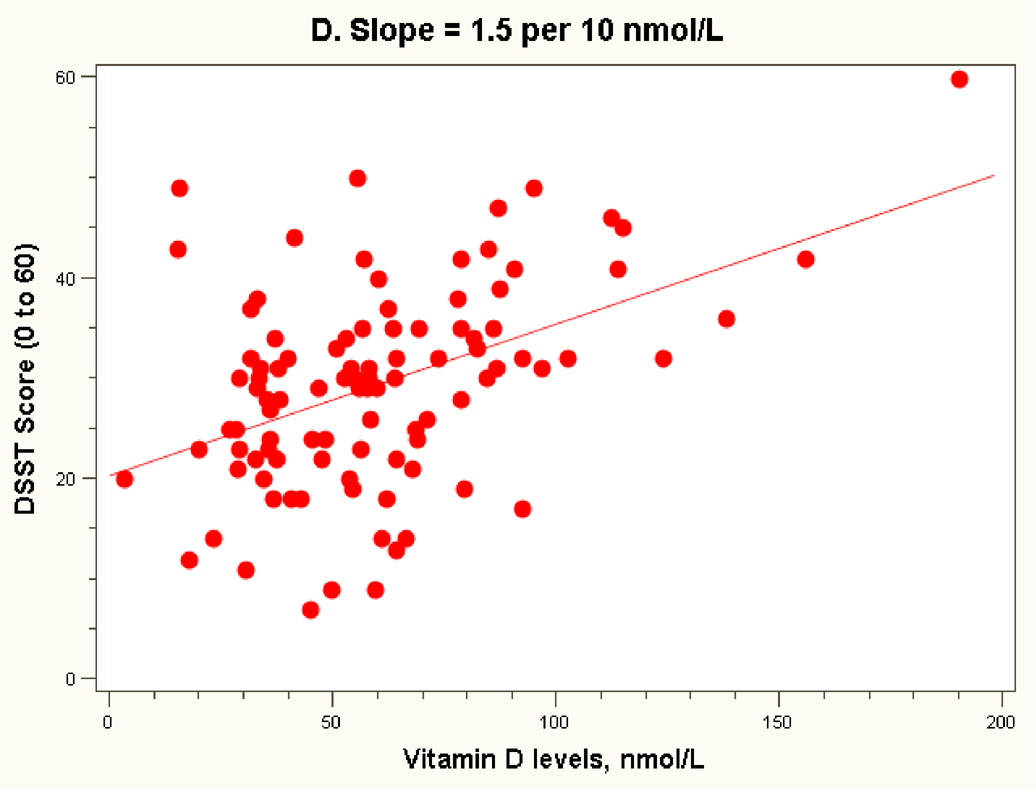


Statistics in Medicine

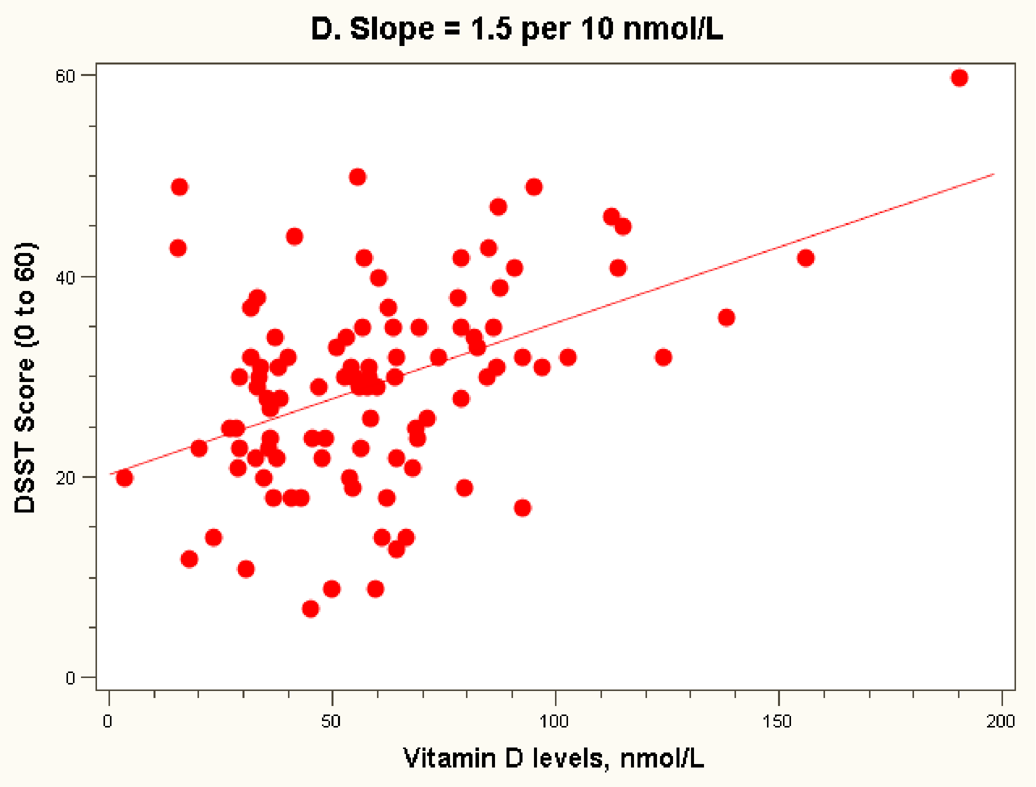
Module 3:

Residual analysis

"Predicted" values

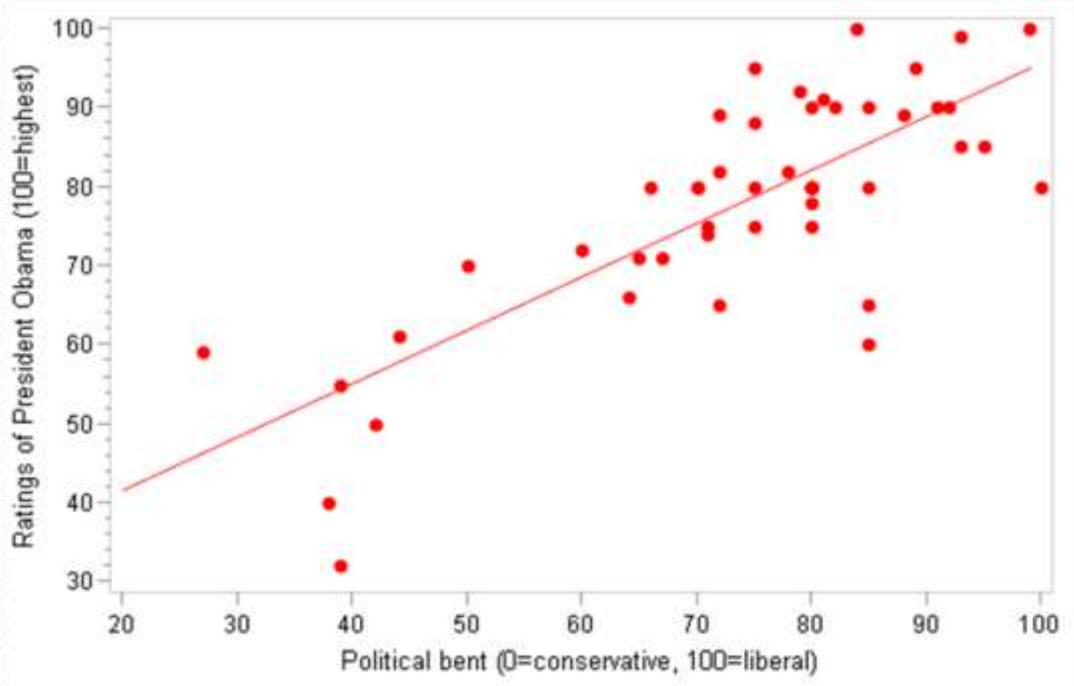


Residual = observed - predicted

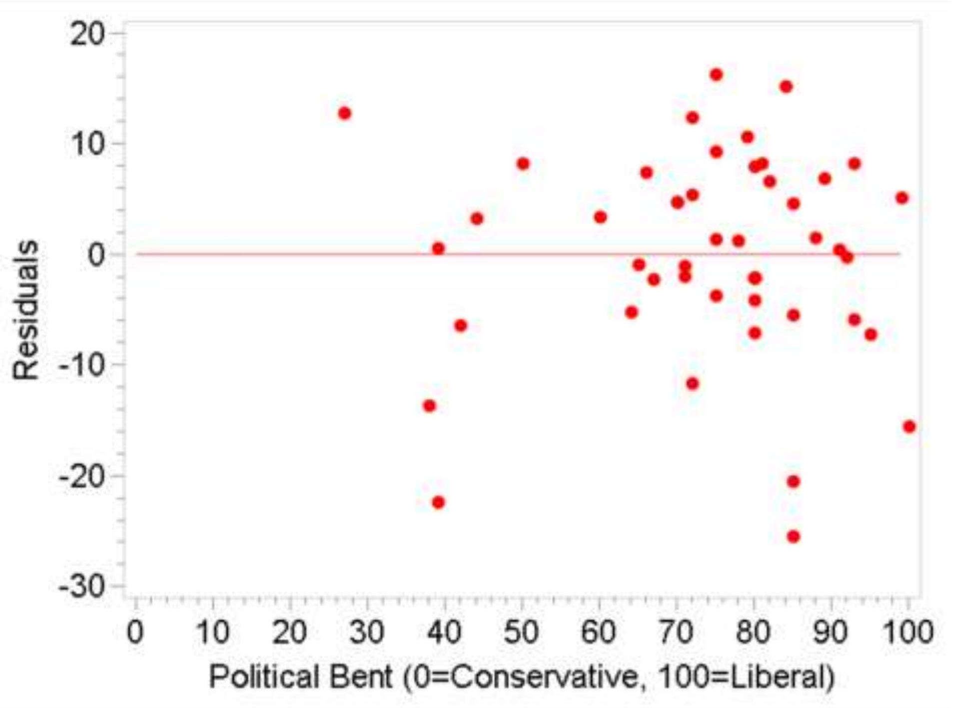


Example: Stanford class data

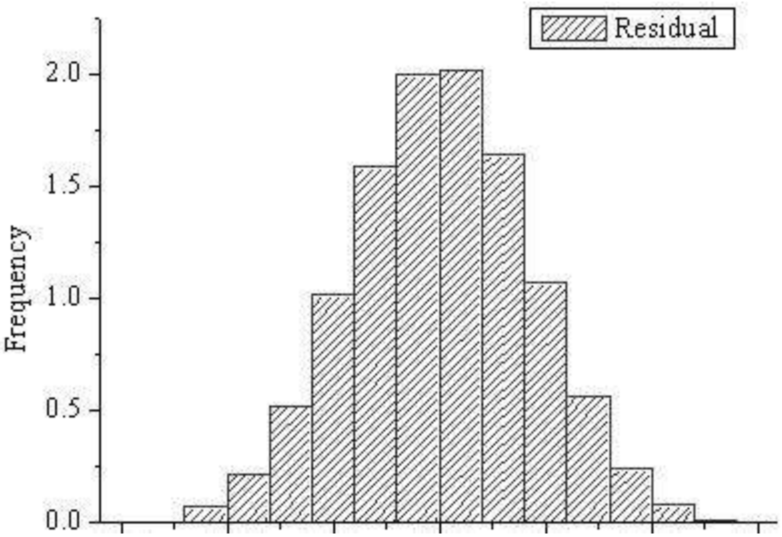
Political Leanings and Rating of Obama



Residual plot: Obama and politics



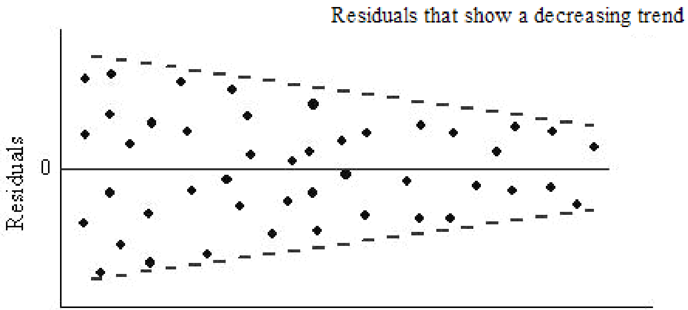
Residual analysis for normality



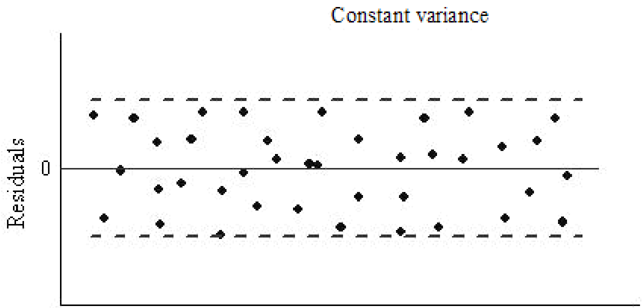
Residual values

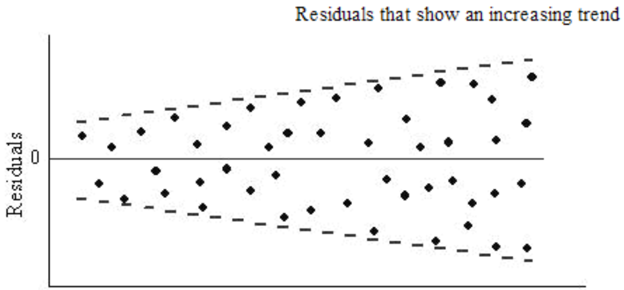
http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

Residual analysis for homogeneity of variances

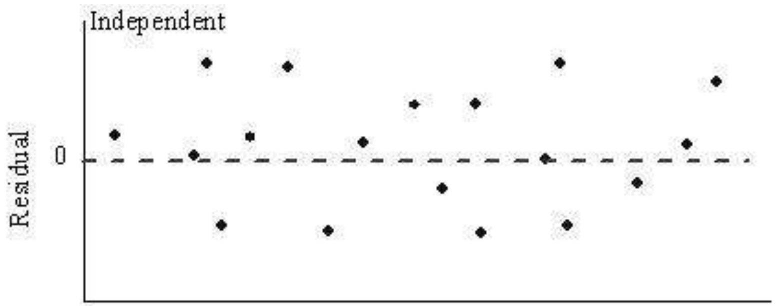


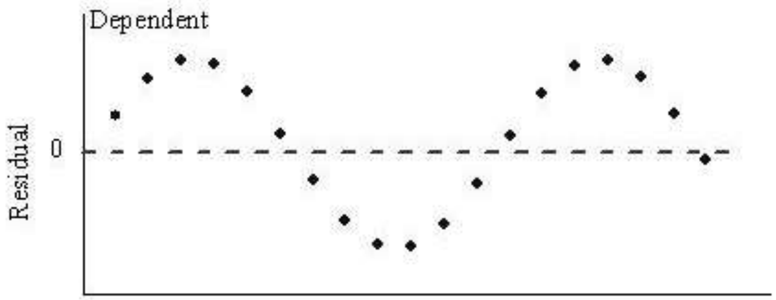
Predictor





Residual analysis for independence





Statistics in Medicine

Module 4:

Multiple linear regression and statistical adjustment

Multiple Linear Regression

- More than one predictor…

$E(y) = \alpha + \beta_1 * X + \beta_2 * W + \beta_3 * Z\ldots$

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

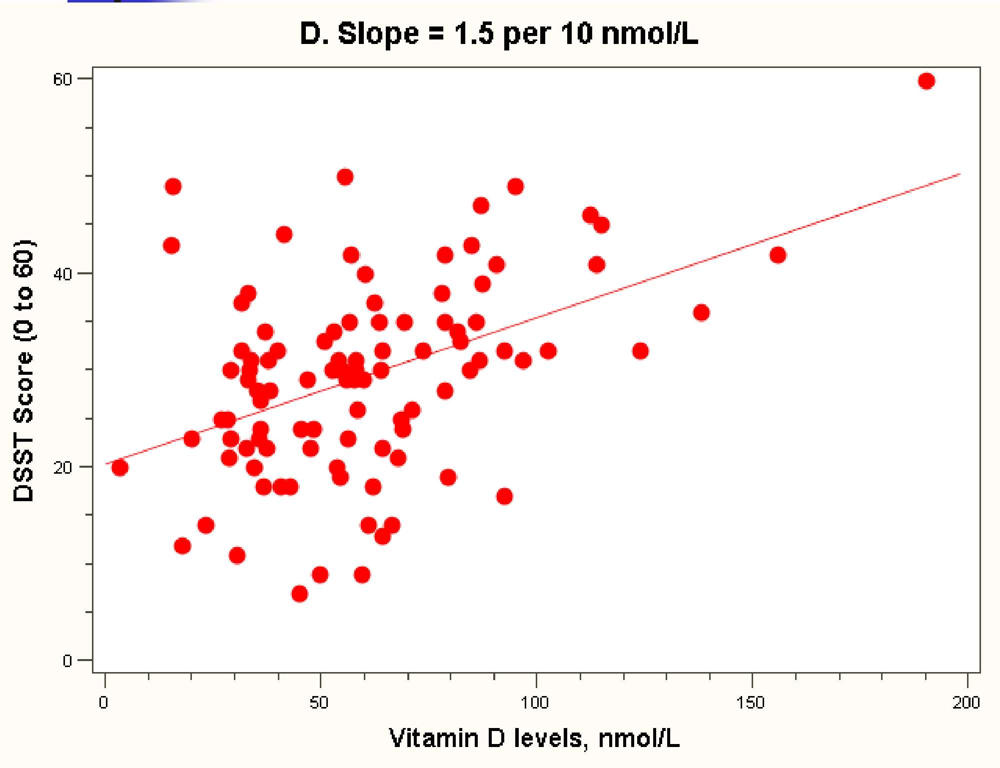
Functions of multivariate analysis:

- Control for confounders
- Improve predictions
- Test for interactions between predictors (effect modification)

Functions of multivariate analysis:

- **Control for confounders**
- Improve predictions
- Test for interactions between predictors (effect modification)

The "Best fit" line



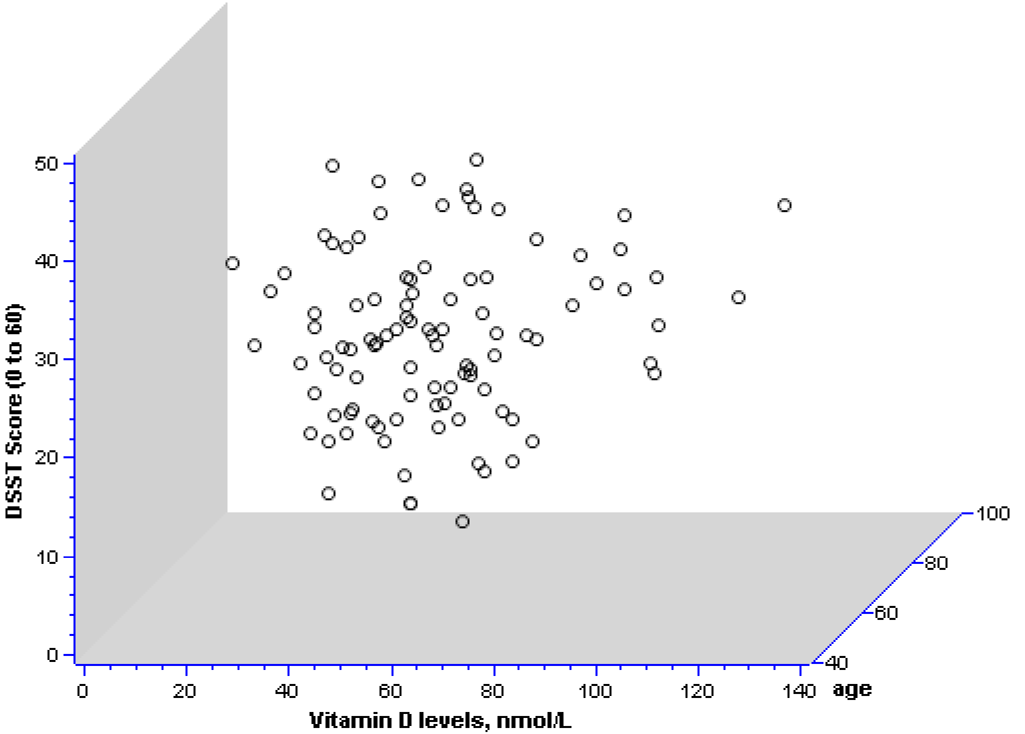
Regression equation:

$E(Y_i) = 20 + 1.5*\text{vit } D_i$ (in 10 nmol/L)

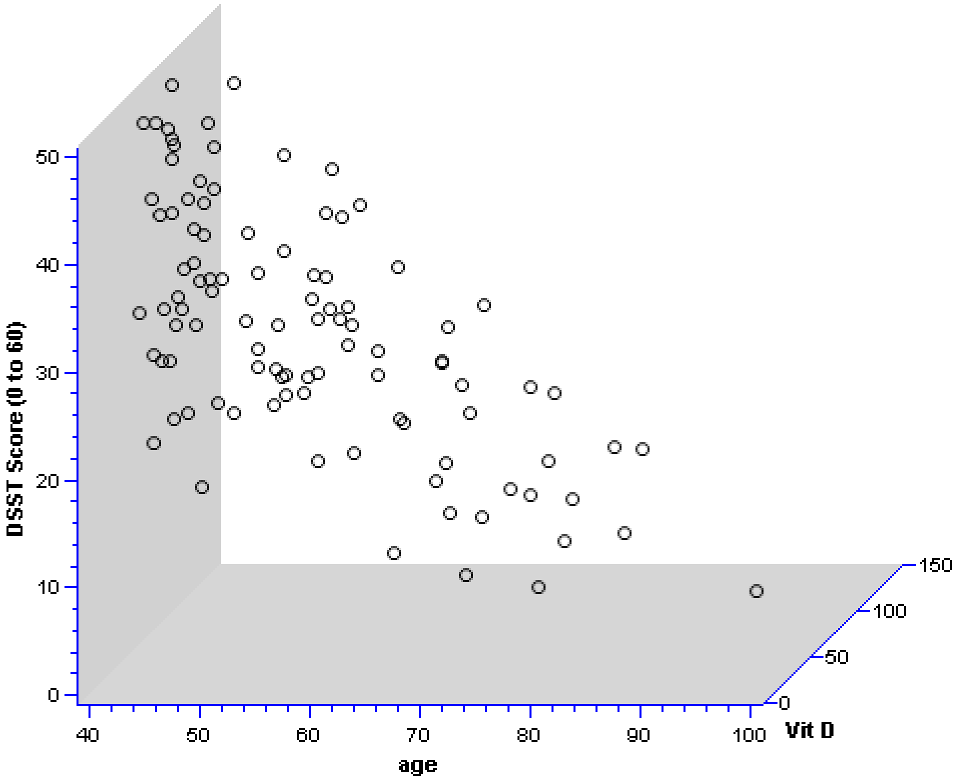
Example: adjustment for confounding:

- What if age is a confounder here?
  - Older men have lower vitamin D
  - Older men have poorer cognition
- "Adjust" for age by putting age in the model:
  - DSST score = intercept + slope_1 *x* vitamin D + slope_2 *x* age

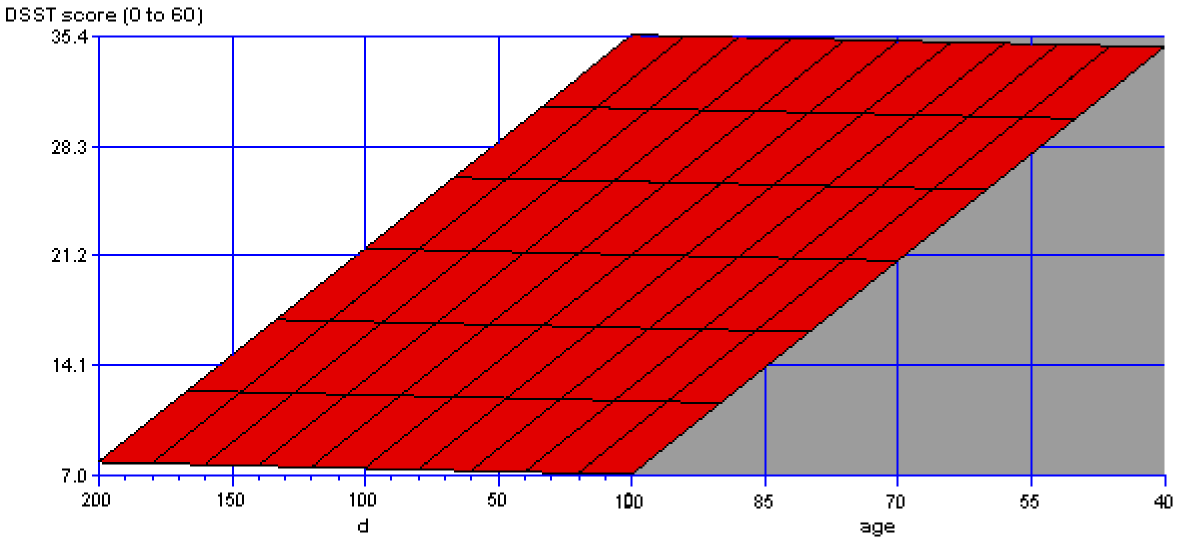
2 predictors: age and vit D…



Different 3D view…



Fit a plane rather than a line…



On the plane, the slope for vitamin D is the same at every age; thus, the slope for vitamin D represents the effect of vitamin D when age is held constant.

Multivariate regression results:

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	53.29472	6.31183	8.44	<.0001
age		1	-0.45689	0.07647	-5.97	<.0001
d	Vitamin D, per 10 nmol/L	1	0.03944	0.04398	0.09	0.9289

DSST score = 53 + 0.039 *x* vitamin D (in 10 nmol/L) - 0.46 *x* age (in years)

Equation of the "Best fit" plane…

- DSST score = 53 + 0.039 x vitamin D (in 10 nmol/L) - 0.46 x age (in years)

Multivariate regression results:

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	53.29472	6.31183	8.44	<.0001
age		1	-0.45689	0.07647	-5.97	<.0001
d	**Vitamin D levels, per 10 nmol/L**	**1**	**0.03944**	**0.04398**	**0.09**	**0.9289**

DSST score = 53 + 0.039*x*vitamin D (in 10 nmol/L) - 0.46 *x*age (in years)

There is no independent effect of Vitamin D after accounting for age.

Evidence of confounding

- Unadjusted beta: 1.5 per 10 nmol/L
- Age-adjusted beta: .039 per 10 nmol/L

- The relationship between DSST and vitamin D disappears after adjusting for age. Thus, this apparent relationship was due to confounding by age!
- The beta has changed by (1.5-.039)/1.5= 97.4%!

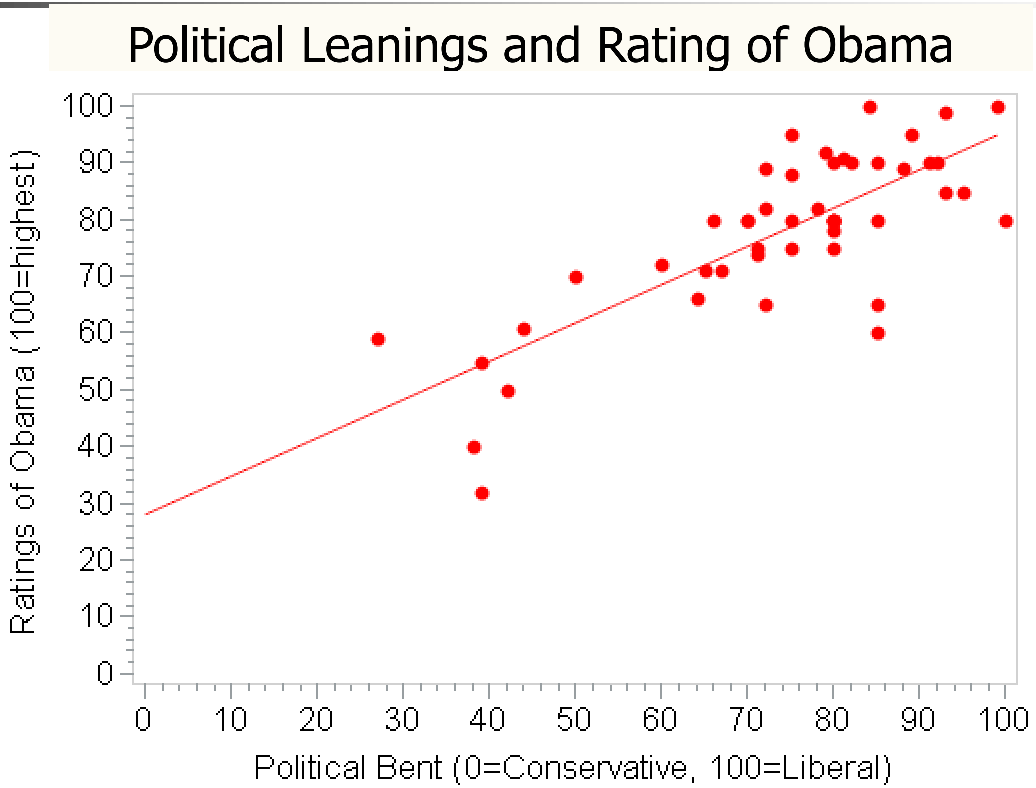
Confounding rule of thumb

- If a potential confounder changes the beta coefficient between the predictor of interest and the outcome variable by more than 10%, then it is considered a confounder.
- Do *not* judge confounders by their effect on p-values!

Functions of multivariate analysis:

- Control for confounders
- **Improve predictions**
- Test for interactions between predictors (effect modification)

Recall: political leaning and Obama (single predictor)



R^2=.61

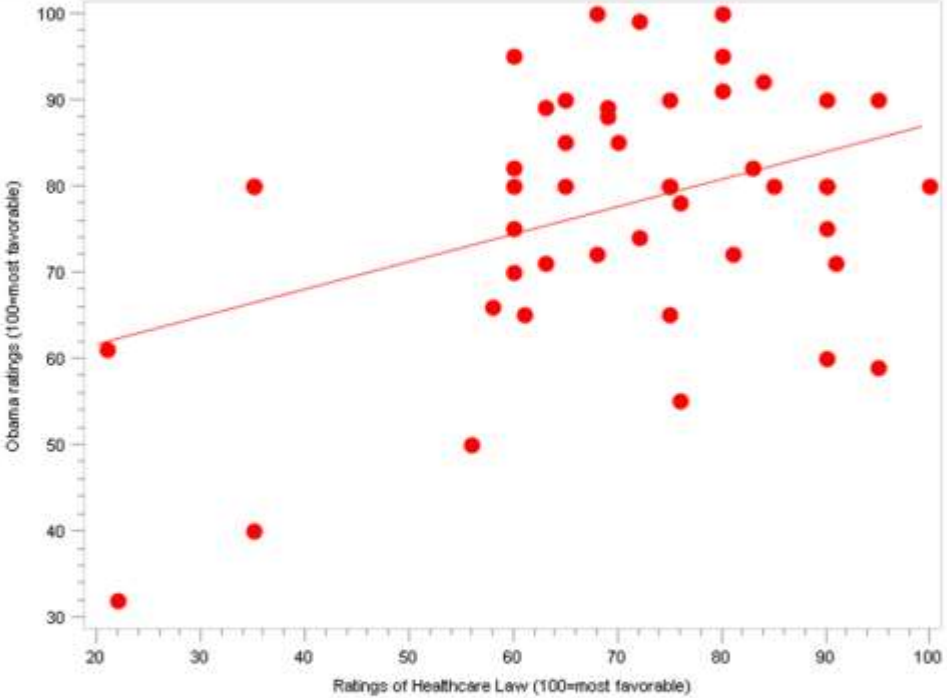
What else predicts Obama ratings from our dataset?

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	46.36516	10.70167	4.33	<.0001
clinton	clinton	1	0.40502	0.13667	2.96	0.0049

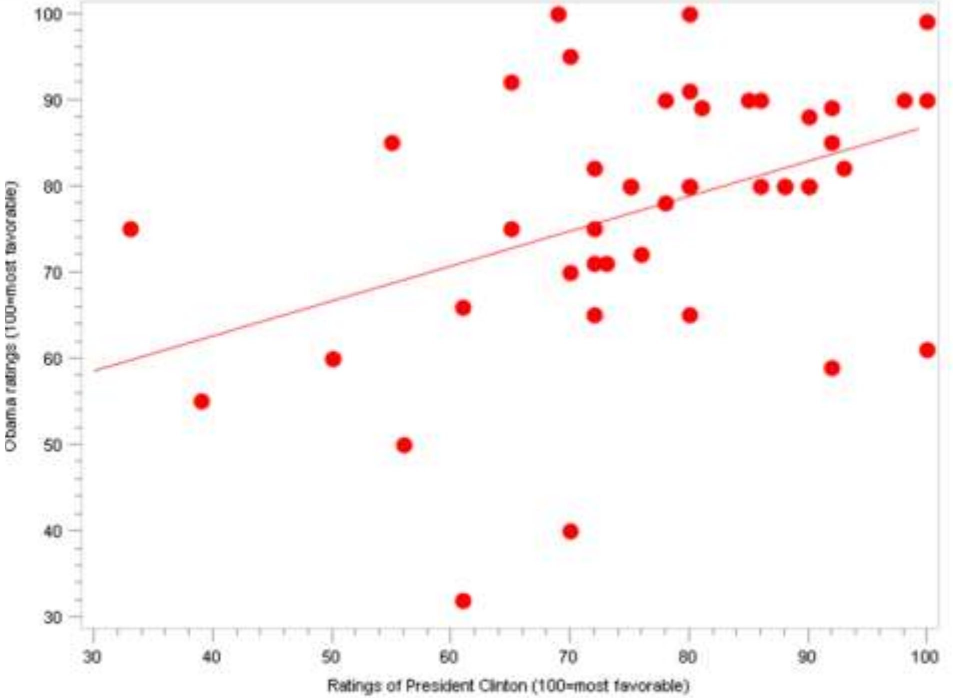
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	55.24831	8.02624	6.88	<.0001
healthcare	healthcare	1	0.31823	0.11110	2.86	0.0063

Don't forget to graph!

Ratings of healthcare law and Obama



Ratings of Clinton and Obama



Can we improve our prediction?

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.90116	9.20794	0.64	0.5254
politics	politics	1	0.56069	0.08358	6.71	<.0001
healthcare	healthcare	1	0.14235	0.07801	1.82	0.0757
clinton	clinton	1	0.26080	0.10129	2.57	0.0139

The Model: Predicted Obama Rating = 5.9 + .56*politics + .14*healthcare+.26*Clinton

R^2=.69; adjusted R^2 =.67

Adjusted R^2 corrects for the number of predictors in the model (since more predictors always increases R^2)

Prediction

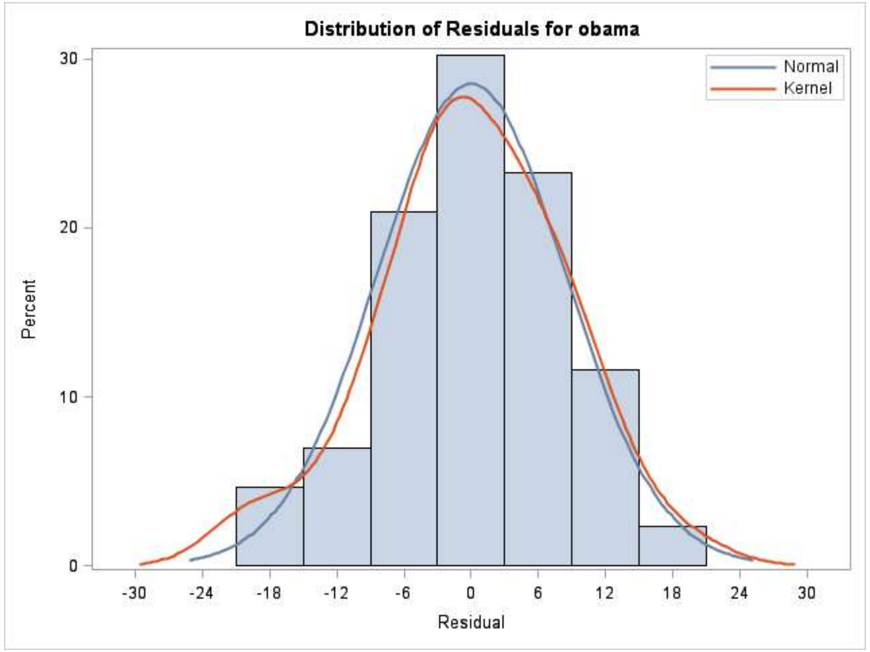
- What is the predicted Obama rating for a person with the following characteristics?:
  - Political leaning: 60
  - Rating of Obama care (healthcare reform in the U.S.): 90
  - Rating of Clinton: 60

The Model: Predicted Obama Rating = 5.9 + .56*politics + .14*healthcare+.26*Clinton

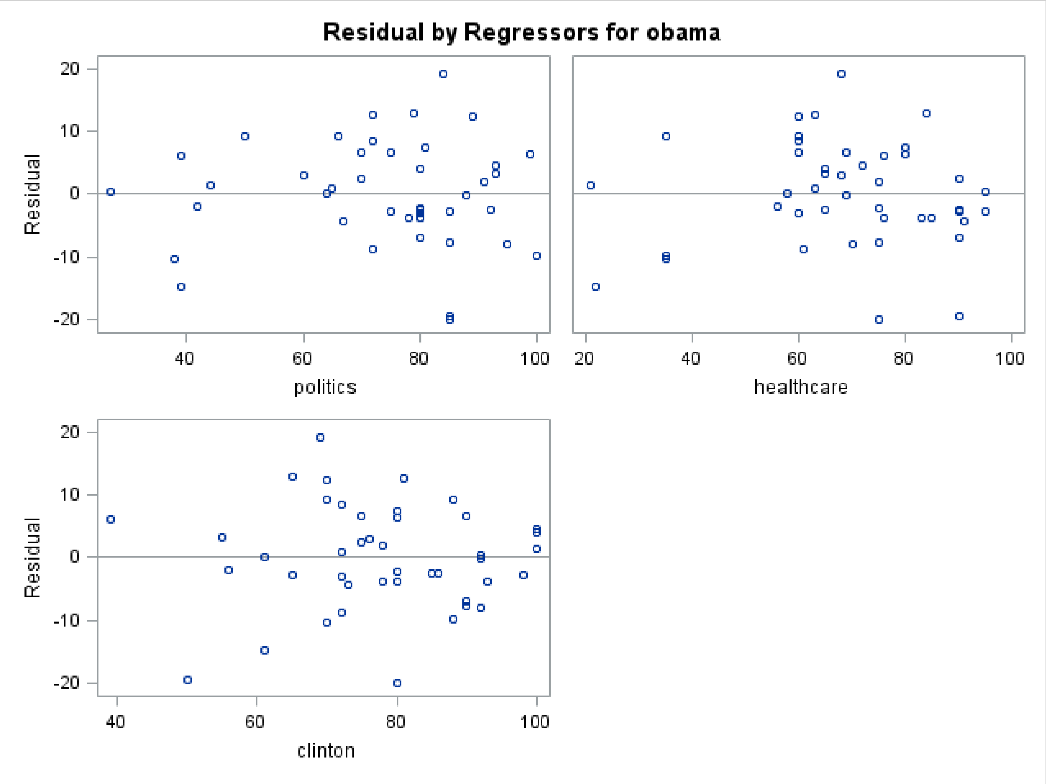
Answer: Predicted Obama Rating = 5.9 + .56*(60) + .14*(90)+.26*(60) = 67.7

Residual analysis: Obama model

Normality assumption: histogram of residuals

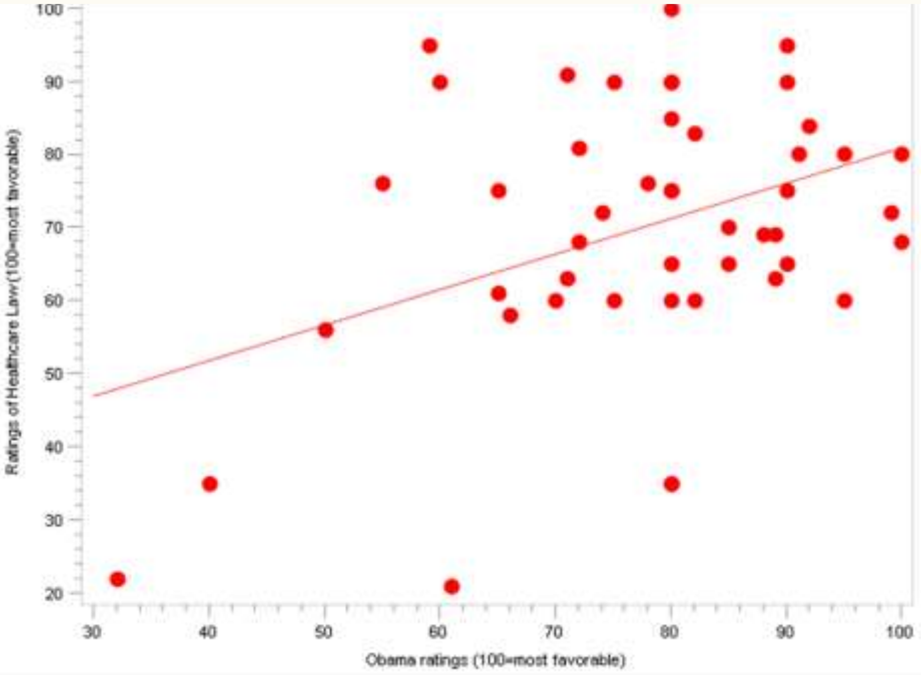


Residual analysis: Obama model



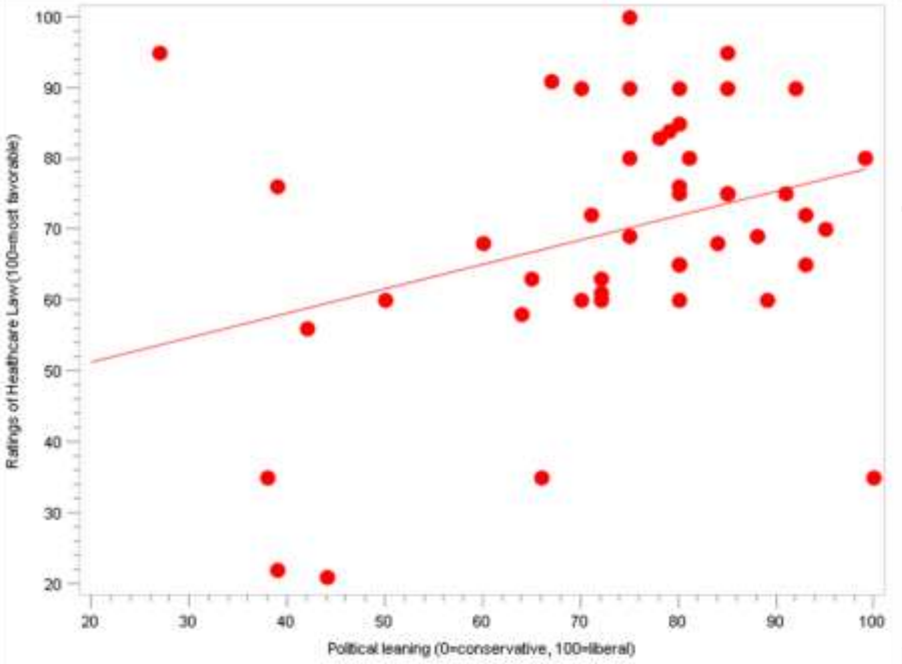
What predicts ratings of healthcare reform (“Obama care”)?

Ratings of Obama and Ratings of Healthcare



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	32.39931	13.34347	2.43	0.0192
obama	obama	1	0.48455	0.16917	2.86	0.0063

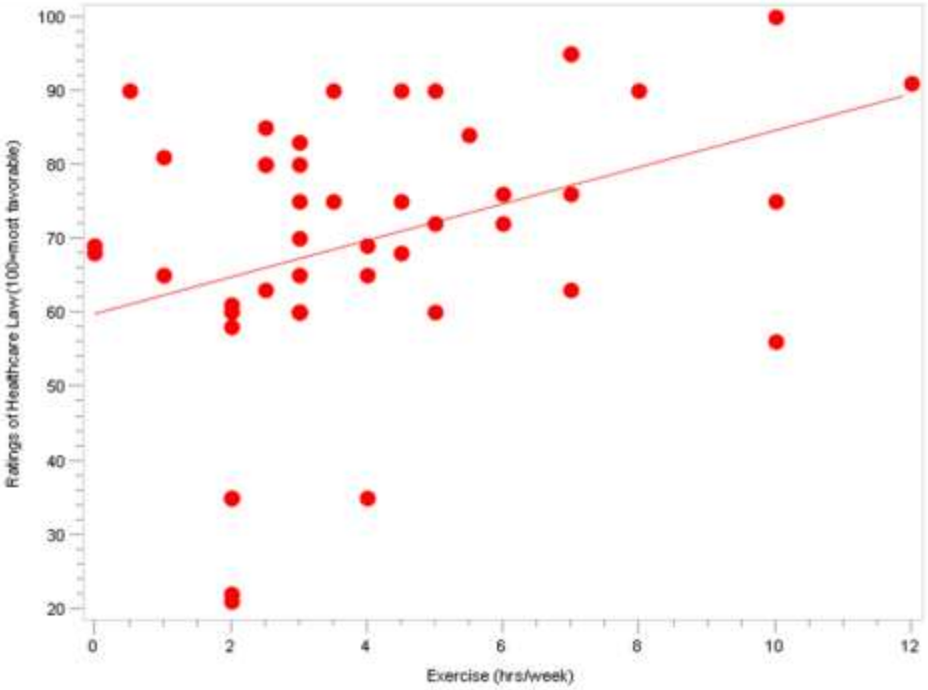
Political Bent and Ratings of Healthcare



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	44.37726	11.46081	3.87	0.0004
politics	politics	1	0.34385	0.15150	2.27	0.0282

Exercise (hrs/wk) and Ratings of Healthcare



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	59.76668	4.53566	13.18	<.0001
exercise	exercise	1	2.47528	0.92025	2.69	0.0100

Multivariate model (all 3 predictors):

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.52389	12.98783	1.20	0.2387
obama	obama	1	0.50387	0.24453	2.06	0.0456
politics	politics	1	0.03916	0.21076	0.19	0.8535
exercise	exercise	1	2.91337	0.84147	3.46	0.0012

R^2=.35; adjusted R^2 =.30

What predicts ratings of healthcare reform ("Obama care")?

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	15.52389	12.98783	1.20	0.2387
obama	obama	1	0.50387	0.24453	2.06	0.0456
politics	**politics**	**1**	**0.03916**	**0.21076**	**0.19**	**0.8535**
exercise	exercise	1	2.91337	0.84147	3.46	0.0012

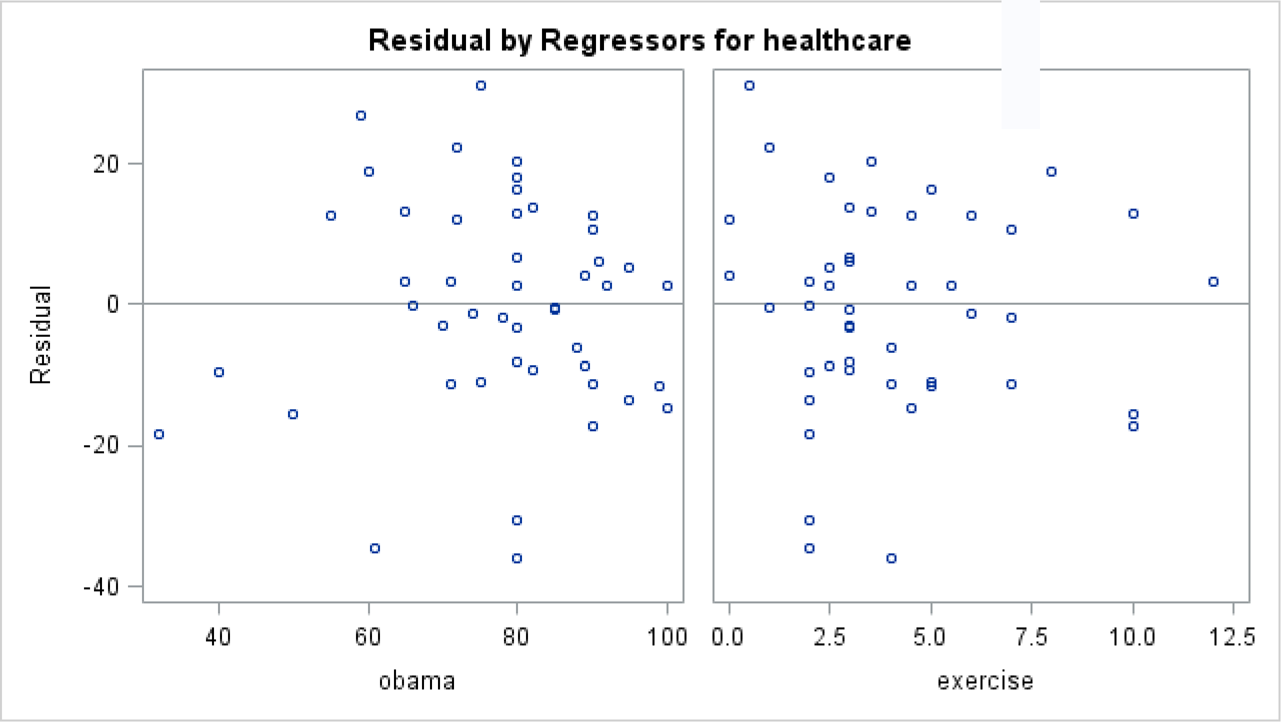
So, after accounting for Ratings of Obama, political bent doesn't help improve our prediction (the adjusted beta is close to zero and p=.85). We say that there is no "independent" effect of political bent after accounting for Ratings of Obama and exercise.

Final model

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	18.33123	12.87929	1.42	0.1617
obama	obama	1	0.52304	0.15419	3.39	0.0015
exercise	exercise	1	2.69247	0.83105	3.24	0.0023

R^2=.32; adjusted R^2 =.29

Residual plots:



Statistics in Medicine

Module 5:

Categorical predictors in regression

How does linear regression handle categorical predictors?

- Binary
  - Treats the "0" and "1" as quantitative (numbers)!
- Categorical
  - Dummy coding!
    - Re-code the categorical predictor as a series of binary predictors!

Common statistics for various types of outcome data

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model (beyond the scope of this course)	Time-varying effects (beyond the scope of this course)

Binary variables

- Imagine that vitamin D was a binary predictor:
  - 1=Insufficient (<50 nmol/L)
  - 0=Sufficient (>=50 nmol/L)

Binary predictor!

Two points always make a straight line so we don't have to worry about the linearity assumption!

Binary predictor:

Intercept represents the mean value in the sufficient group.

Slope represents the difference in means between the groups. Difference is significant.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	40.07407	1.47511	27.17	<.0001
Insufficient	Insufficient	1	-7.53060	2.17493	-3.46	0.0008

A ttest is linear regression!

We can evaluate these data with a ttest also; the effect size (difference in means) and p-value are identical!

$$T_{98} = \frac{40 - 32.5 = 7.5}{\sqrt{\frac{10.8^2}{54} + \frac{10.8^2}{46}}} = 3.46; p = .0008$$

Dummy coding

- Imagine that vitamin D was a categorical predictor:
  - Deficient (<25 nmol/L)
  - Insufficient (>=25 and <50 nmol/L)
  - Sufficient (>=50 nmol/L), reference group

Dummy coding

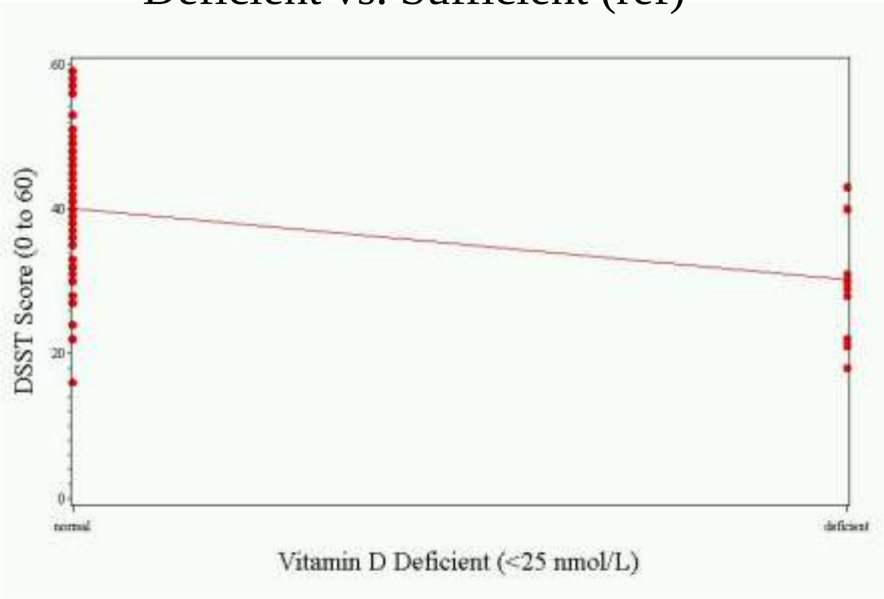
- A 3-level categorical variable can be represented with 2 binary variables:

Vitamin D category	Deficient (Yes/No)	Insufficient (Yes/No)
Deficient	1	0
Insufficient	0	1
Sufficient (ref)	0	0

The picture…

Insufficient vs. Sufficient (ref)

Deficient vs. Sufficient (ref)



Linear regression output:

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	40.07407	1.47511	27.17	<.0001
Insufficient	<50 nmol/L	1	**-6.87963**	2.33719	-2.94	0.0041
Deficient	>=50 nmol/L	1	**-9.87407**	3.73950	-2.64	0.0096

The linear regression equation

DSST= 40 -6.87*(1 if insufficient) -9.87 *(1 if deficient)

40= mean DSST in the sufficient (reference) group

-6.87= the difference in means between insufficient and sufficient groups

-9.87= difference in means between deficient and sufficient groups

Applying the equation

DSST= 40 -6.87*(1 if insufficient) -9.87 *(1 if deficient)

Thus, a person who is in the deficient group has:

expected DSST = 40 -6.87*0 -9.87*1 = 30.13

Thus, a person in the insufficient group has:

expected DSST = 40 -6.87*1 -9.87*0 = 33.13

Thus, a person in the sufficient group has:

expected DSST = 40 -6.87*0 -9.87*0 = 40

ANOVA is linear regression!

- P-value for overall model fit from linear regression with a categorical predictor = p-value from ANOVA.

Why is this useful?

- Because you can adjust the model for covariates (confounders, other predictors)!
- For example, you can get the age-adjusted means in each Vitamin D group.

To adjust for age:

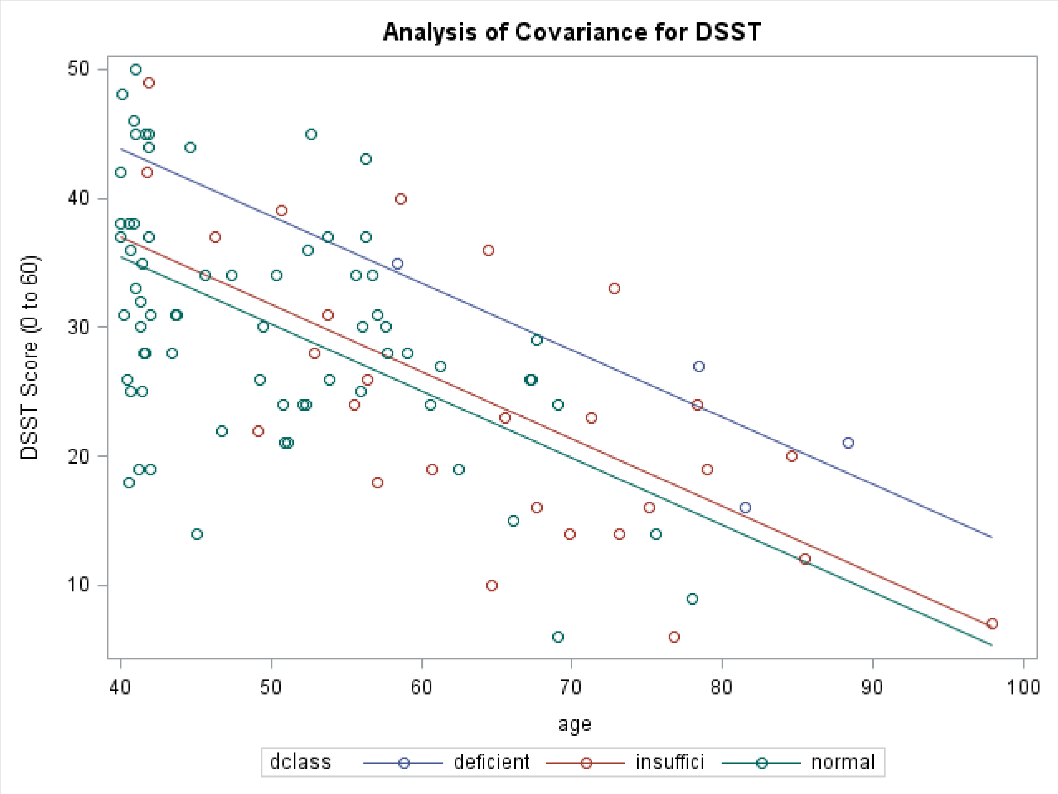
DSST= α + β_1*(1 if insufficient) + β_2 *(1 if deficient) + β_3*age (in years)

Model Output (Hypothetical data)

Parameter	Estimate	Standard Error	t Value	Pr > $\lvert t \rvert$
Intercept	56.32002262	3.71613448	15.16	<.0001
dclass deficient	8.34866787	4.54318617	1.84	0.0692
dclass insuffici	1.45706752	2.11209641	0.69	0.4919
dclass normal	reference			
age	-0.52075353	0.07196455	-7.24	<.0001

Expected DSST = 56.3 + 8.34*(1 if deficient) + 1.45*(1 if insufficient) + -.52*age

The picture!



Analysis of Covariance for DSST
DSST Score (0 to 60)
50
40
30
20
10
40
50
60
70
80
90
100
age
dclass
deficient
insuffici
normal

Calculating "age-adjusted means" for Vitamin D groups

Expected DSST = 56.3 + 8.34*(1 if deficient) + 1.45*(1 if insufficient) + -.52*(age)

Plug in the mean age (55 years):

The expected DSST for a vitamin D deficient person who is 55 years old:

expected DSST = 56.3 +8.34*1+1.45*0-.52*55 =36.0

The expected DSST for a vitamin D insufficient person who is 55 years old:

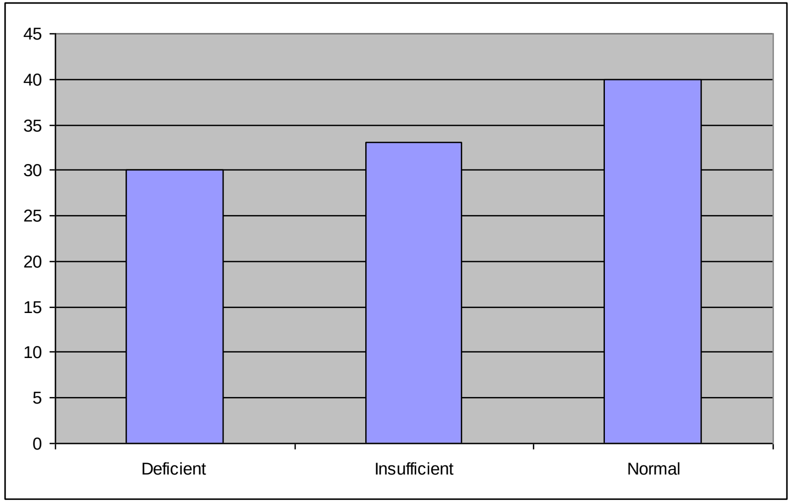
expected DSST = 56.3 +8.34*0+1.45*1-.52*55 =29.1

The expected DSST for a vitamin D sufficient person who is 55 years old:

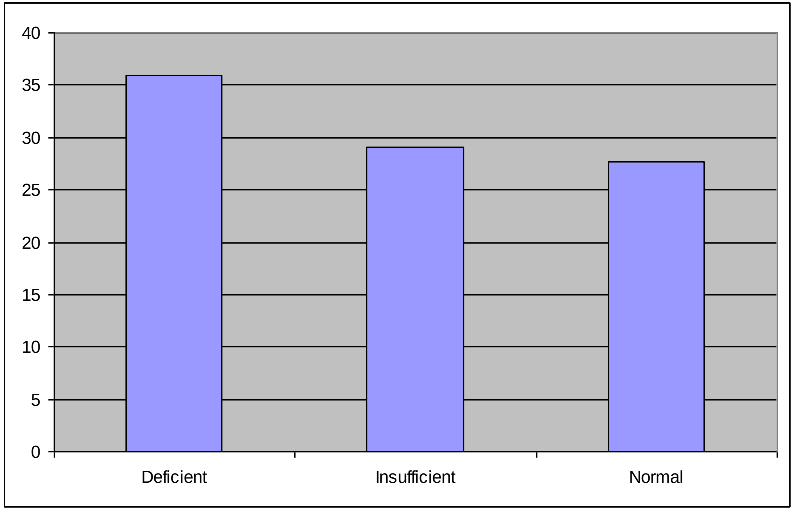
expected DSST = 56.3 +8.34*0+1.45*0-.52*55 =27.7

Unadjusted vs. Adjusted means:

Unadjusted group means



Age-adjusted group means



Statistics in Medicine

Module 6:

Practice interpreting linear regression results

Recall:

Headline:

Brighten the twilight years: "Sunshine vitamin" boosts brain function in the elderly

- "Middle-aged and older men with high levels of vitamin D in their blood were mentally quicker than their peers, researchers report."
- "The findings are some of the strongest evidence yet of such a link because of the size of the study and because the researchers accounted for a number of lifestyle factors believed to affect mental ability when older, Lee said."

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
Age (years)	-0.127 (-0.142, -0.113)*	-0.228 (-0.247, -0.208)*	-0.163 (-0.177, -0.149)*	-0.415 (-0.439, -0.391)*	-0.0135 (-0.114, 0.087)
Age left education (years)	0.097 (0.077, 0.116)*	0.129 (0.101, 0.157)*	0.060 (0.039, 0.080)*	0.199 (0.165, 0.233)*	-0.127 (-0.272, 0.017)
BDI score	-0.044 (-0.068, -0.021)*	-0.088 (-0.122, -0.055)*	-0.071 (-0.095, -0.047)*	-0.177 (-0.217, -0.136)*	-0.815 (-0.985, -0.644)*
BDI category:					
Normal (0 – 10)	Reference	Reference	Reference	Reference	Reference
Mild - Borderline (11 – 20)	-0.218 (-0.610, 0.173)	-0.732 (-1.293, -0.172)*	-0.849 (-1.252, -0.446)*	-1.743 (-2.427, -1.058)*	-9.641 (-12.51, -6.774)*
Moderate - Extreme (21+)	-1.585 (-2.361, -0.809)*	-2.264 (-3.375, -1.154)*	-1.433 (-2.231, -0.636)*	-4.050 (-5.405, -2.695)*	-13.93 (-19.61, -8.249)*
Body Mass Index (kg/m^2)	-0.059 (-0.096, -0.022)*	-0.041 (-0.093, 0.012)	-0.037 (-0.075, 0.001)	-0.115 (-0.179, -0.050)*	-0.811 (-1.081, -0.541)*
PASE score tertiles:					
Lower	Reference	Reference	Reference	Reference	Reference
Mid	0.809 (0.419, 1.200)*	1.358 (0.796, 1.919)*	1.283 (0.882, 1.685)*	1.836 (1.148, 2.523)*	5.148 (2.244, 8.052)*
Upper	0.592 (0.176, 1.008)*	1.166 (0.569, 1.764)*	1.096 (0.669, 1.524)*	1.381 (0.649, 2.113)*	7.072 (3.972, 10.17)*
PPT total tertiles:					
Lower	Reference	Reference	Reference	Reference	Reference
Mid	1.155 (0.783, 1.526)*	0.899 (0.363, 1.434)*	1.243 (0.860, 1.625)*	3.035 (2.396, 3.673)*	6.433 (3.677, 9.188)*
Upper	1.285 (0.909, 1.661)*	1.089 (0.547, 1.632)*	1.409 (1.021, 1.796)*	4.553 (3.907, 5.199)*	7.302 (4.514, 10.09)*
Current smoker					
No	Reference	Reference	Reference	Reference	Reference
Yes	-0.735 (-1.112, -0.359)*	-1.151 (-1.687, -0.615)*	-1.190 (-1.575, -0.805)*	-2.502 (-3.158, -1.847)*	-10.95 (-13.69, -8.207)*
Alcohol (≥ 1day/week)					
No	Reference	Reference	Reference	Reference	Reference
Yes	0.257 (-0.047, 0.562)	0.180 (-0.255, 0.615)	1.014 (0.704, 1.324)*	2.159 (1.630, 2.687)*	8.521 (6.307, 10.74)*

[†]Adjusted for age where applicable *$P < 0.05$

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
Age (years)	-0.127 (-0.142, -0.113)*	-0.228 (-0.247, -0.208)*	-0.163 (-0.177, -0.149)*	-0.415 (-0.439, -0.391)*	-0.0135 (-0.114, 0.087)

[†]Adjusted for age where applicable *$P < 0.05$

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
BDI category:					
Normal (0 – 10)	Reference	Reference	Reference	Reference	Reference
Mild - Borderline (11 – 20)	-0.218 (-0.610, 0.173)	-0.732 (-1.293, -0.172)*	-0.849 (-1.252, -0.446)*	-1.743 (-2.427, -1.058)*	-9.641 (-12.51, -6.774)*
Moderate - Extreme (21+)	-1.585 (-2.361, -0.809)*	-2.264 (-3.375, -1.154)*	-1.433 (-2.231, -0.636)*	-4.050 (-5.405, -2.695)*	-13.93 (-19.61, -8.249)*

[†]Adjusted for age where applicable *$P < 0.05$

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
Body Mass Index (kg/m^2)	-0.059 (-0.096, -0.022)*	-0.041 (-0.093, 0.012)	-0.037 (-0.075, 0.001)	-0.115 (-0.179, -0.050)*	-0.811 (-1.081, -0.541)*

[†]Adjusted for age where applicable   *P < 0.05

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
PASE score tertiles:					
Lower	Reference	Reference	Reference	Reference	Reference
Mid	0.809 (0.419, 1.200)*	1.358 (0.796, 1.919)*	1.283 (0.882, 1.685)*	1.836 (1.148, 2.523)*	5.148 (2.244, 8.052)*
Upper	0.592 (0.176, 1.008)*	1.166 (0.569, 1.764)*	1.096 (0.669, 1.524)*	1.381 (0.649, 2.113)*	7.072 (3.972, 10.17)*

[†]Adjusted for age where applicable *$P < 0.05$

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
Current smoker					
No	Reference	Reference	Reference	Reference	Reference
Yes	-0.735 (-1.112, -0.359)*	-1.151 (-1.687, -0.615)*	-1.190 (-1.575, -0.805)*	-2.502 (-3.158, -1.847)*	-10.95 (-13.69, -8.207)*

[†]Adjusted for age where applicable *$P < 0.05$

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Table 2. Determinants of cognitive test scores and 25(OH)D levels: linear regression analyses

	ROCF copy score	ROCF recall score	CTRM score	DSST score	25(OH)D (nmol/L)
	β coefficient (95% CI)[†]				
Alcohol (≥ 1day/week)					
No	Reference	Reference	Reference	Reference	Reference
Yes	0.257 (-0.047, 0.562)	0.180 (-0.255, 0.615)	1.014 (0.704, 1.324)*	2.159 (1.630, 2.687)*	8.521 (6.307, 10.74)*

[†]Adjusted for age where applicable *$P < 0.05$

ROCF = Rey-Osterrieth Complex Figure, CTRM = Camden Topographical Recognition Memory, DSST = Digit Symbol Substitution Test, BDI = Beck Depression Inventory, PASE = Physical Activity Scale for the Elderly, PPT = Reuben's Physical Performance Test.

Final study results, Vitamin D and Cognitive Function:

Association between cognitive test scores and serum 25(OH)D levels: linear regression analyses ; β coefficient (95% CI)

	CTRM test score†	DSST score†
25(OH)D level (per 10 nmol/l)‡	**0.075 (0.026 to 0.124)****	**0.318 (0.235 to 0.401)****
25(OH)D level (per 10 nmol/l)§	**−0.001 (−0.146 to 0.144)**	**0.152 (0.051 to 0.253)****
25(OH)D categories§		
Sufficient (≥75.0 nmol/l)	**Reference**	**Reference**
Suboptimum (50.0–74.9 nmol/l)	**−0.143 (−0.752 to 0.465)**	**−0.759 (−1.313 to −0.204)***
Insufficient (25.0–49.9 nmol/l)	**0.084 (−0.874 to 1.043)**	**−0.768 (−1.822 to 0.287)**
Deficient (<25.0 nmol/l)	**−0.125 (−1.304 to 1.054)**	**−1.404 (−2.681 to −0.127)***

Adjusted for age only

Adjusted for other factors

*p<0.05; **p<0.01; †dependent variables; ‡adjusted for age; §adjusted for age, education, depression, body mass index, physical activity, physical performance, smoking, alcohol consumption, centre and season.

25(OH)D, 25-hydroxyvitamin D; CTRM, Camden Topographical Recognition Memory; DSST, Digit Symbol Substitution Test.

Statistics in Medicine

Module 7:

Regression worries: Overfitting and missing data

Example

- What predicts homework time in Stanford students?
- Statistical strategy (be careful, though):
  - Homework time is continuous, so use linear regression (not a normal distribution, but n=50).
  - Find the best set of predictors using an automatic selection procedure, such as stepwise selection (other common procedures include forward and backward selection).

Violà!

- SAS can automatically find predictors of homework in the example class dataset, using stepwise selection (no graphing, no thinking!). Here's the resulting linear regression model:

Variable	Parameter Estimate	Standard Error	F Value	Pr > F
Intercept	-31.05684	13.76487	5.09	0.0334
Varsity	7.04380	3.02870	5.41	0.0288
politics	0.37203	0.07768	22.94	<.0001
clinton	0.22184	0.11741	3.57	0.0710
regan	0.31167	0.09157	11.58	0.0023
carter	-0.26015	0.08259	9.92	0.0043
alcohol	-0.89198	0.59871	2.22	0.1493

But if something seems to good to be true…

Varsity Sports in High School, Univariate predictor…						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	12.35000	2.37537	5.20	<.0001
Varsity	Varsity	1	-1.59138	3.08767	-0.52	0.6087

Politics, Univariate model…						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-7.89181	6.35044	-1.24	0.2204
politics	politics	1	0.26561	0.08381	3.17	0.0027

Univariate models:

Carter ratings, Univariate model:						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.76205	6.20054	2.54	0.0156
carter	carter	1	-0.05379	0.10779	-0.50	0.6209

Regan ratings, Univariate model:						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.85189	3.50699	3.66	0.0007
regan	regan	1	-0.01403	0.07936	-0.18	0.8606

Univariate models:

Clinton ratings, Univariate model:						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.37183	8.30210	0.65	0.5211
clinton	clinton	1	0.09169	0.10611	0.86	0.3925

Alcohol, Univariate model:						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.66313	2.23184	5.67	<.0001
alcohol	alcohol	1	-0.55399	0.72449	-0.76	0.4483

What’s going on?

- Over-fitting!

Overfitting

- In multivariate modeling, you can get highly significant but meaningless results if you put too many predictors in the model (small dataset).
- The model is fit perfectly to outliers of your particular sample, but has no predictive ability in a new sample.

Overfitting

Rule of thumb: You need at least 10 subjects for each predictor variable in the multivariate regression model (and the intercept).

Always check your N's!!

- What was the N in my multivariate model here?
  - N=50??

Computer Output	
Number of Observations Read	50
Number of Observations Used	31
Number of Observations with Missing Values	19

Where did everybody go?

- Most regression analyses automatically throw out incomplete observations, so if a subject is missing the value for just one of the variables in the model, that subject will be excluded.
- This can add up to lots of omissions!
- Always check your N's!