Linear correlation and linear regression

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated? independent	correlated	Alternatives if the normality assumption is violated (and small sample size):
Continuous (e.g. pain scale, cognitive function)	**Ttest:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling:** multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

Recall: Covariance

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

Interpreting Covariance

$\mathrm{cov}(X,Y) > 0$ → X and Y are positively correlated

$\mathrm{cov}(X,Y) < 0$ → X and Y are inversely correlated

$\mathrm{cov}(X,Y) = 0$ → X and Y are independent

Correlation coefficient

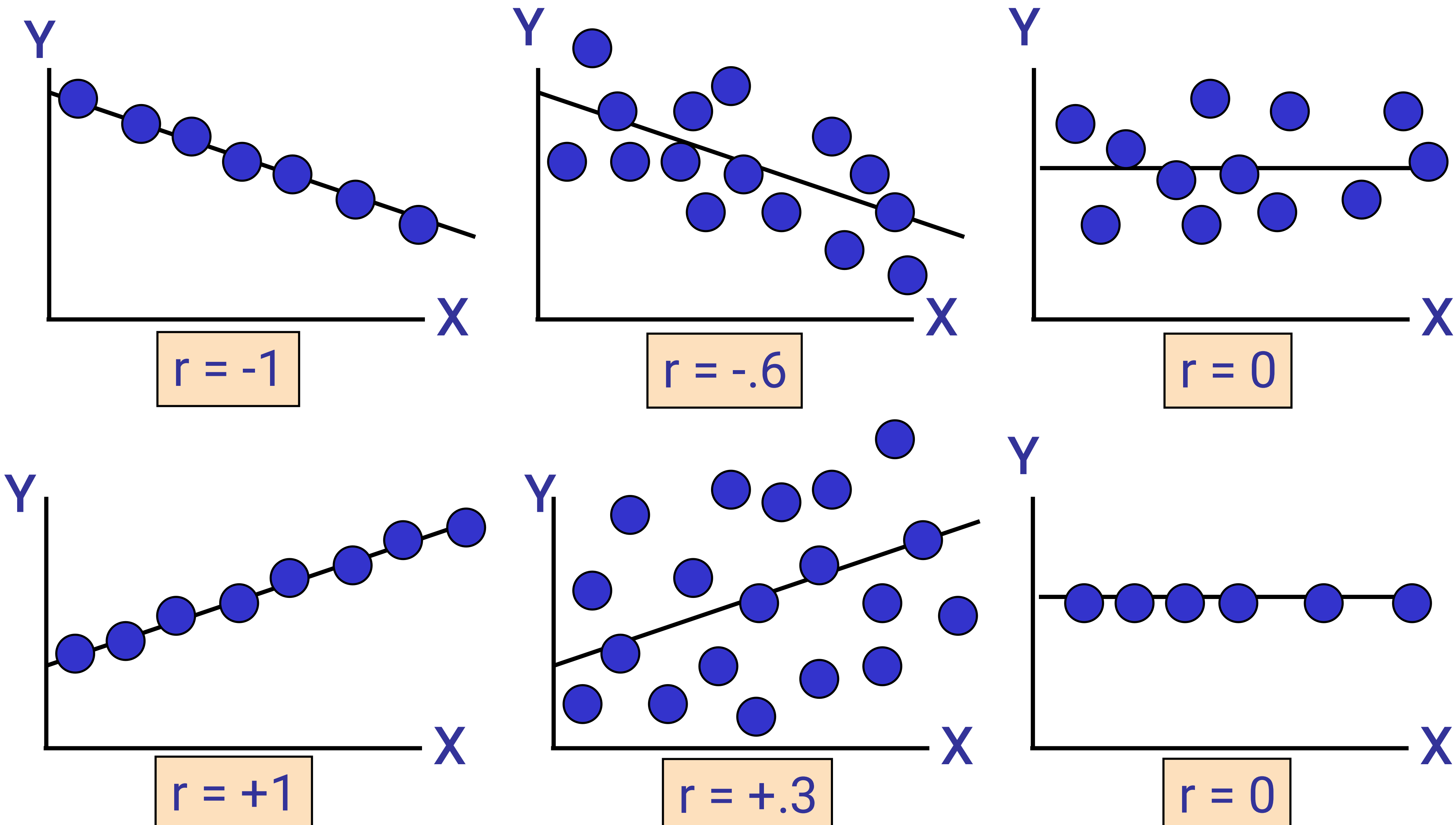
- Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{\operatorname{cov}\mathit{ariance}\,(x, y)}{\sqrt{\operatorname{var} x}\sqrt{\operatorname{var} y}}$$

Correlation

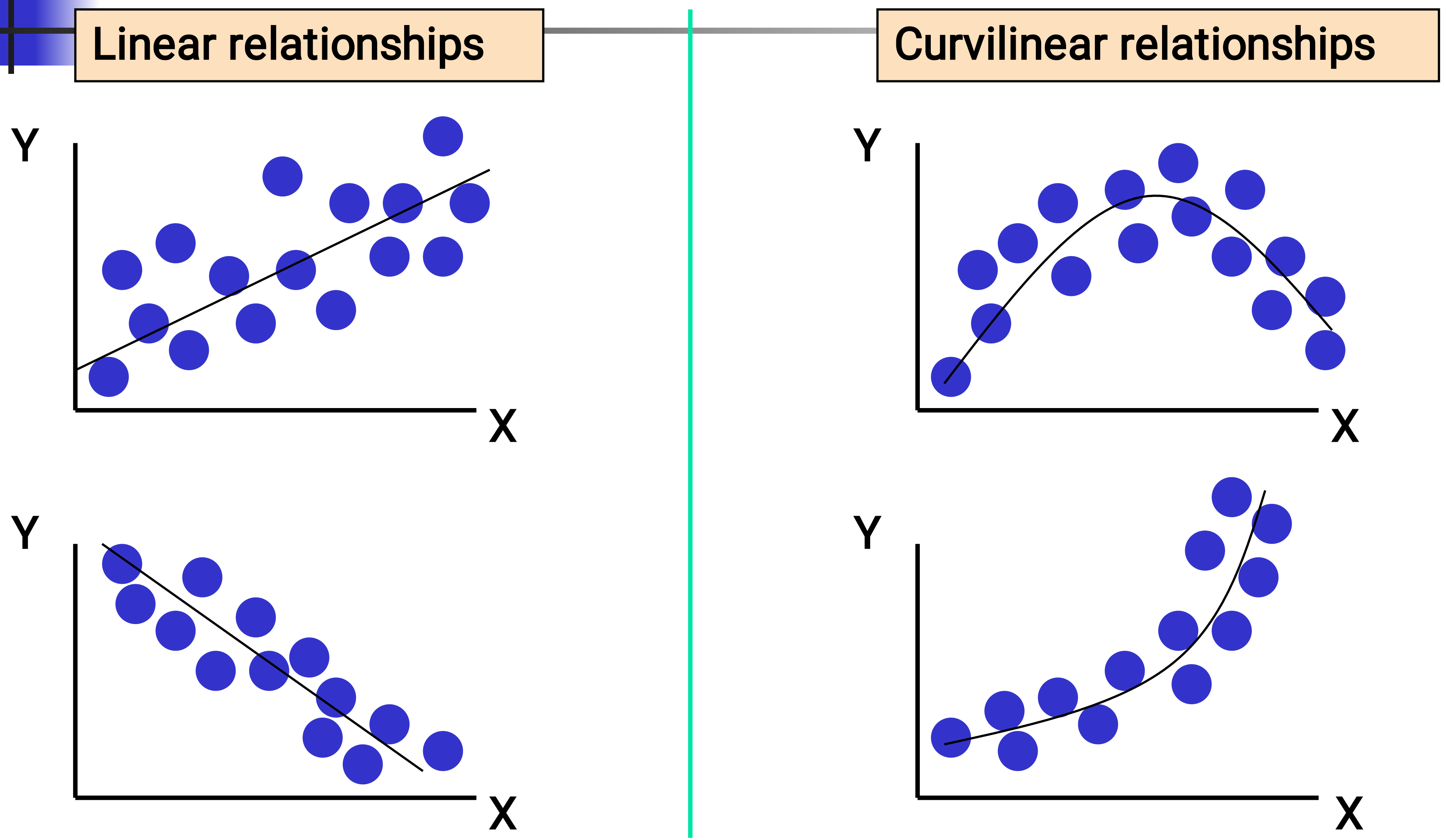
- Measures the relative strength of the *linear* relationship between two variables
- Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients



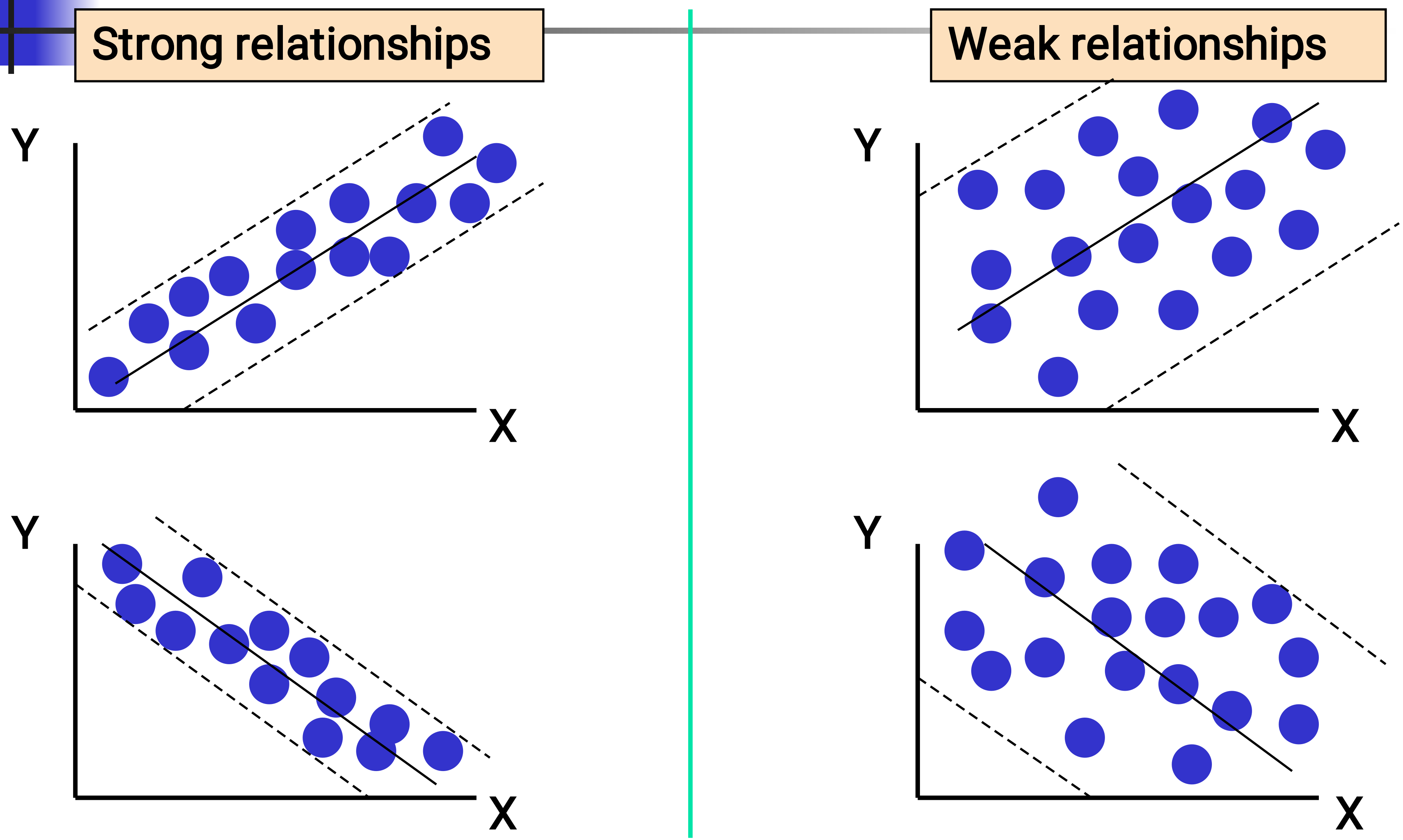
Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

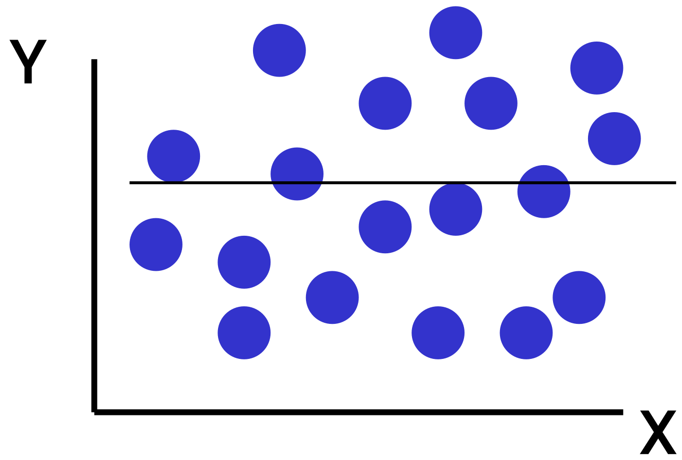
Linear Correlation

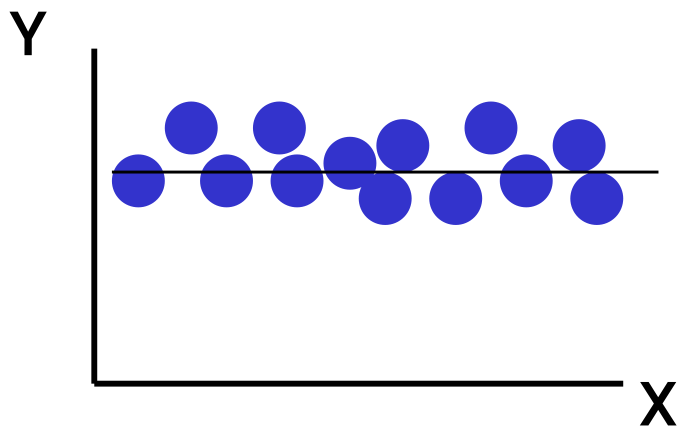


Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation

No relationship





Calculating by hand…

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

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\cancel{n-1}}}} =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

$$\longrightarrow \hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

Numerators of variance

Distribution of the correlation coefficient:

$$SE(\hat{r}) = \sqrt{\frac{1 - r^2}{n - 2}}$$

The sample correlation coefficient follows a T-distribution with n-2 degrees of freedom (since you have to estimate the standard error).

*note, like a proportion, the variance of the correlation coefficient depends on the correlation coefficient itself→substitute in estimated r

Continuous outcome (means)

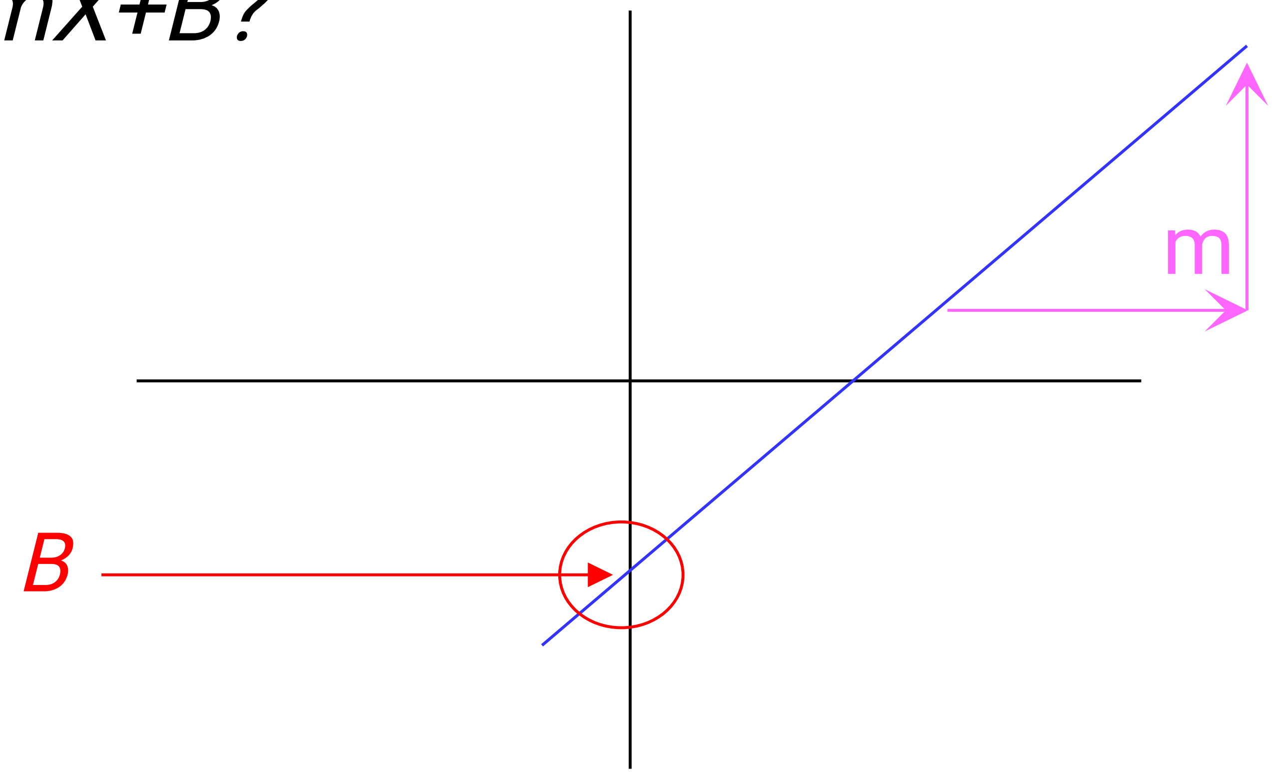
Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated (and small sample size):
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling:** multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	<u>Non-parametric statistics</u> **Wilcoxon sign-rank test**: non-parametric alternative to the paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

Linear regression

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.

What is "Linear"?

- Remember this:
- *Y=mX+B?*



What’s Slope?

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

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*=

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

Predicted value for an individual…

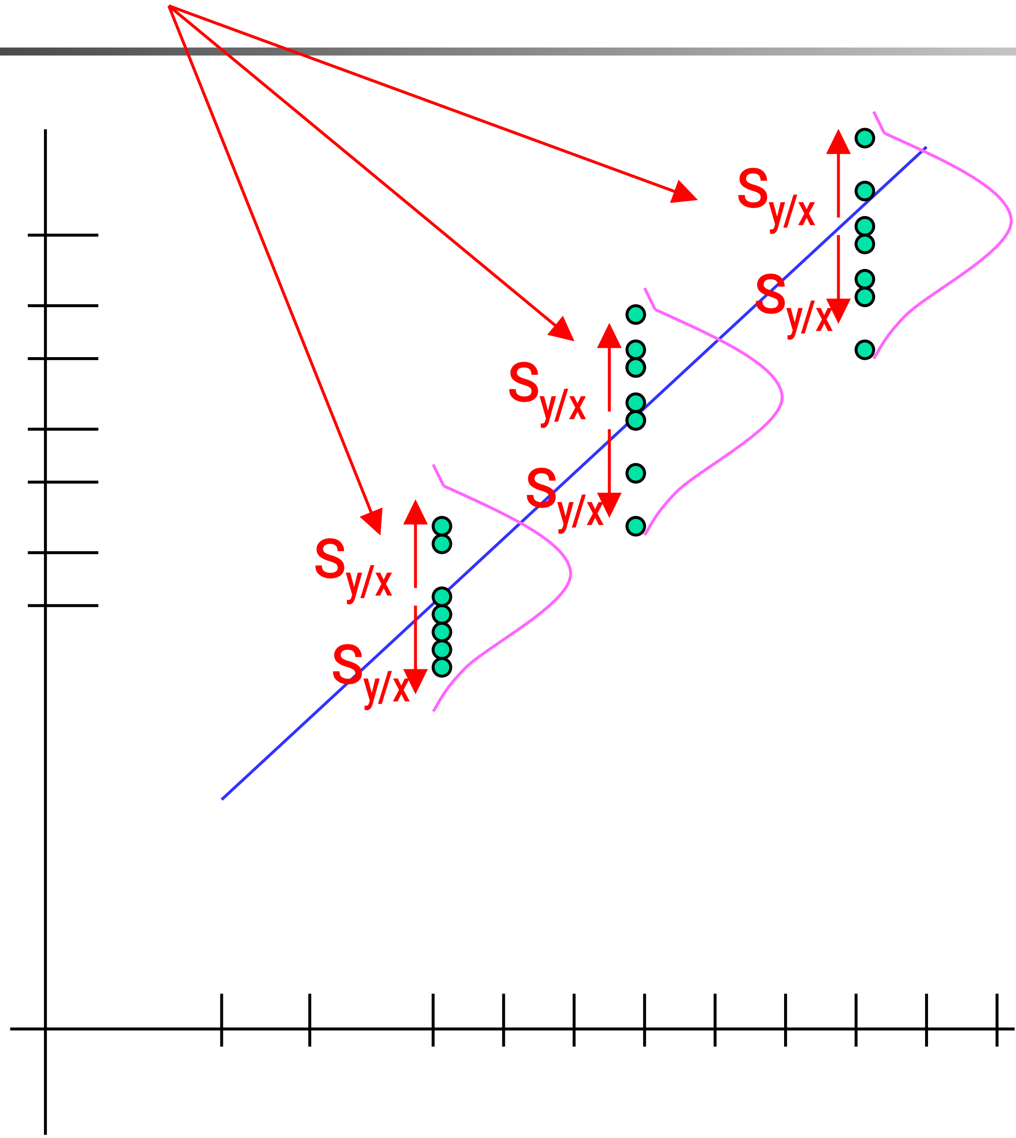
$$\hat{y}_i = \underbrace{\alpha + \beta * x_i}_{\text{Fixed – exactly on the line}} + \boxed{\text{random error}_i}$$

Follows a normal distribution

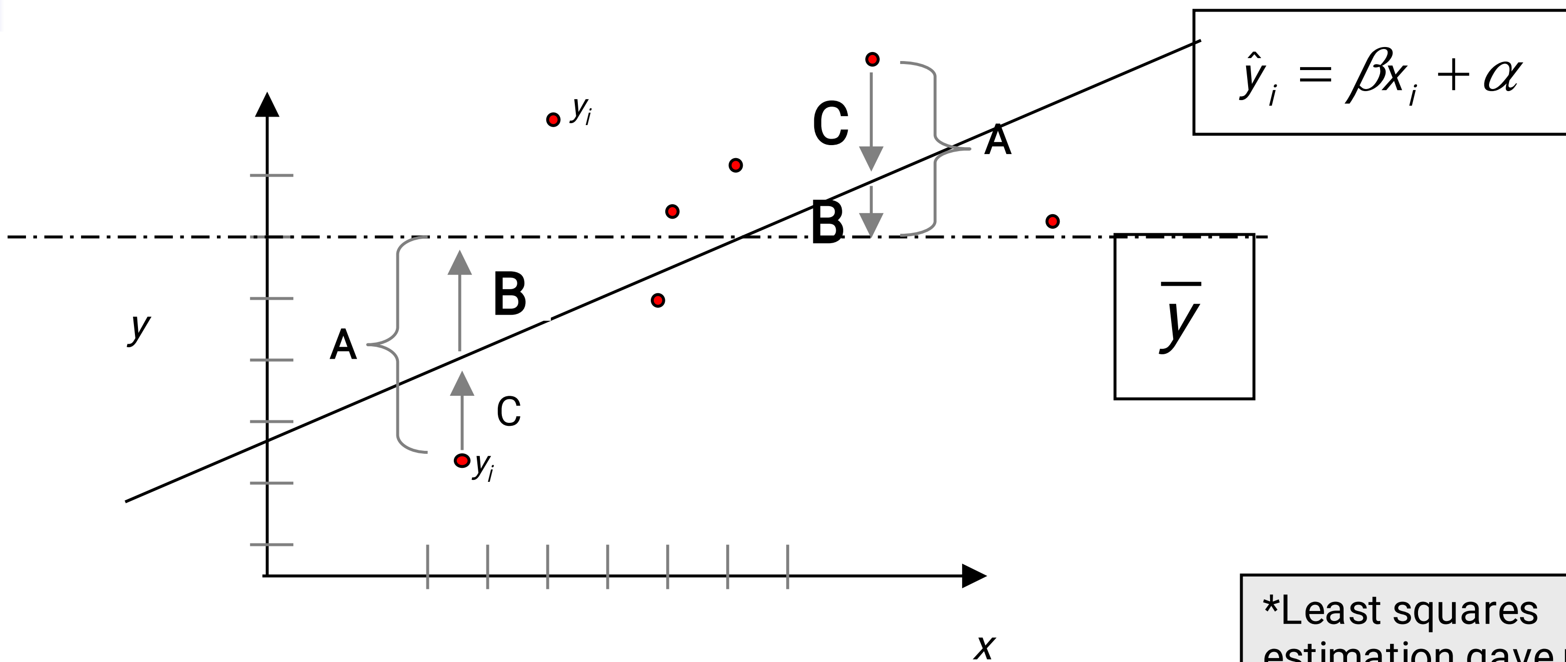
Assumptions (or the fine print)

- Linear regression assumes that…
  - 1. The relationship between X and Y is linear
  - 2. Y is distributed normally at each value of X
  - 3. The variance of Y at every value of X is the same (homogeneity of variances)
  - 4. The observations are independent

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.



Regression Picture



*Least squares estimation gave us the line (β) that minimized C^2

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$R^2 = SS_{reg}/SS_{total}$

A^2
SS_{total}
Total squared distance of observations from naïve mean of y
Total variation

B^2
SS_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

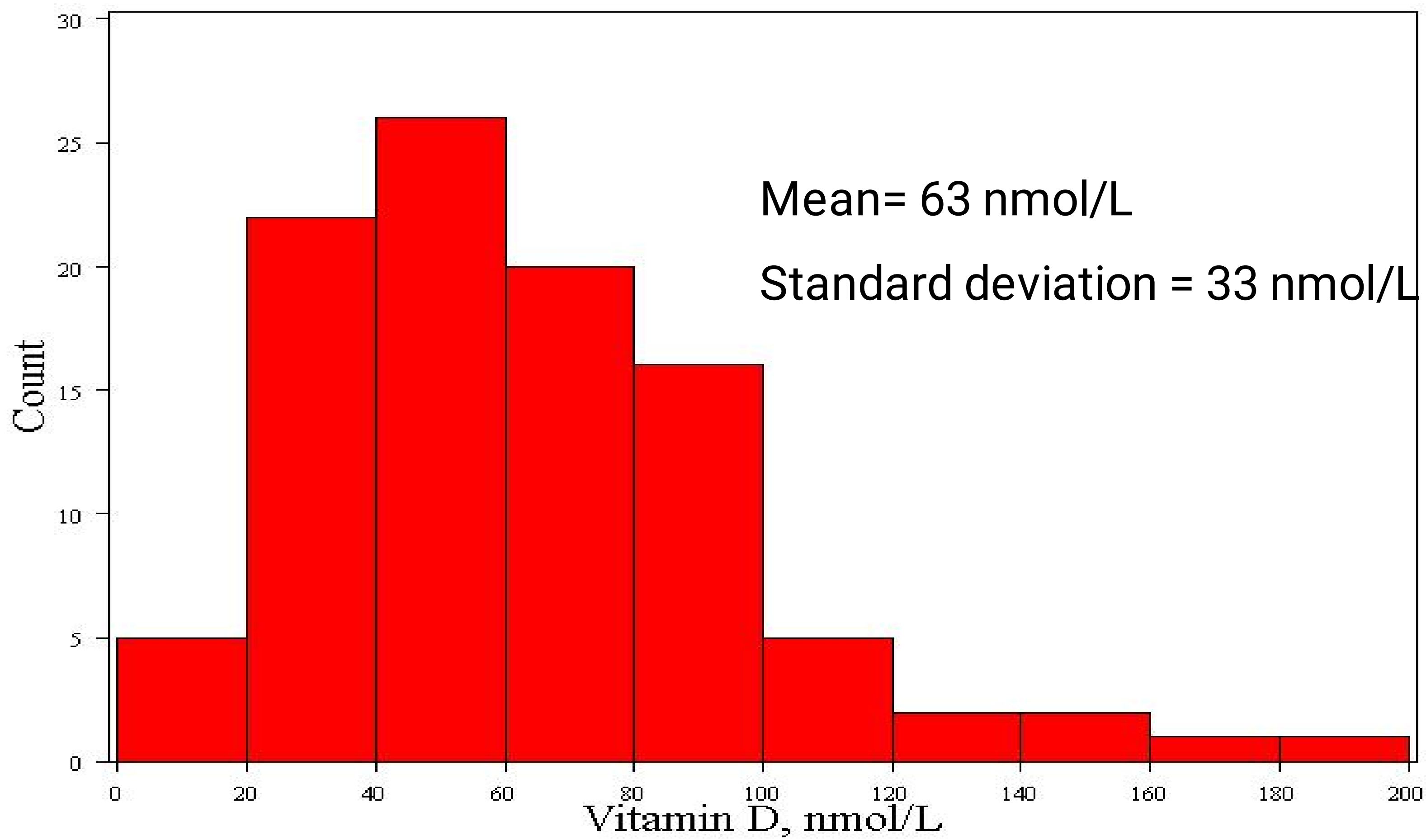
C^2
$SS_{residual}$
Variance around the regression line
Additional variability not explained by x—what least squares method aims to minimize

Recall example: cognitive function and vitamin D

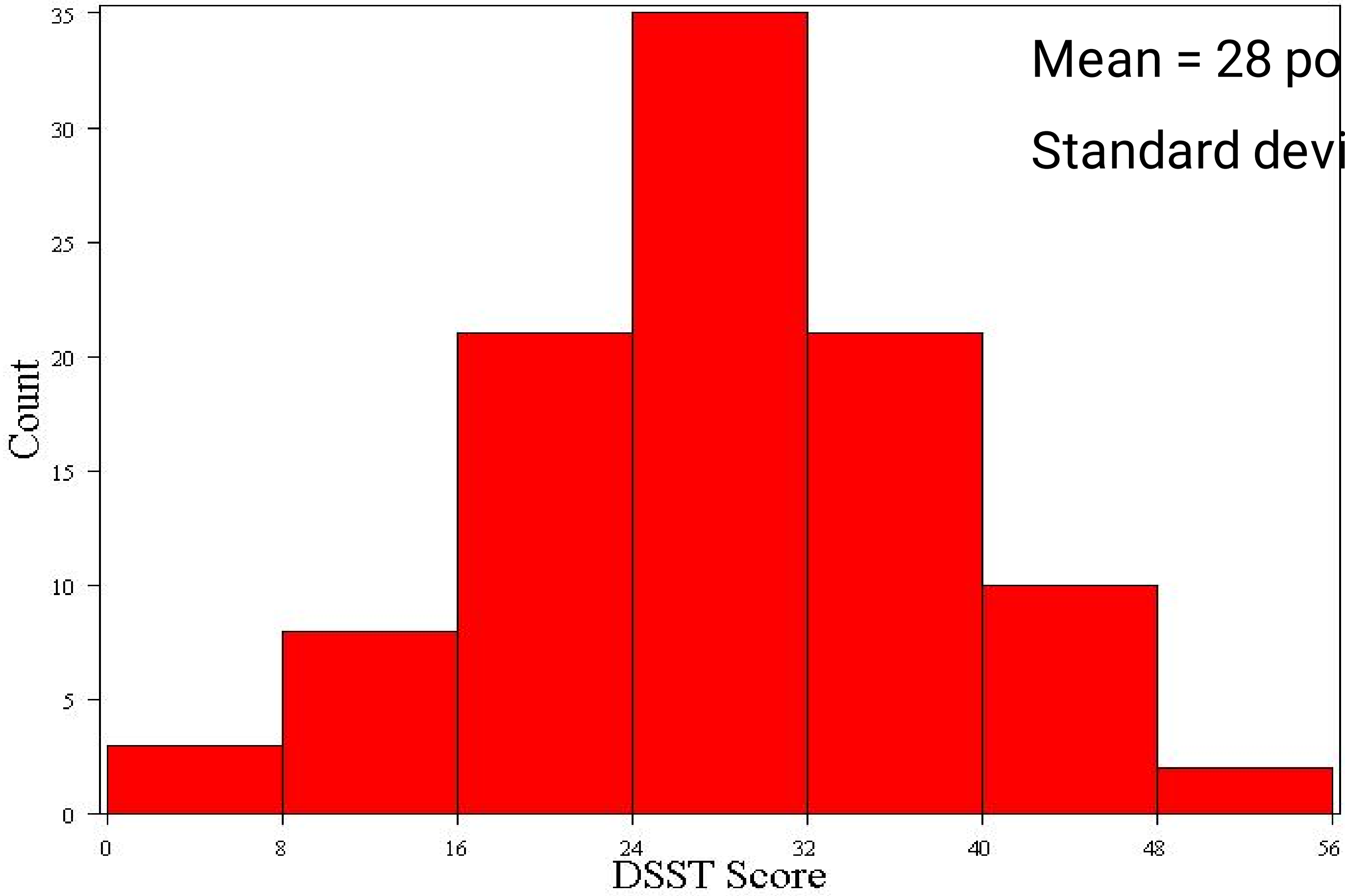
- Hypothetical data loosely based on [1]; cross-sectional study of 100 middle-aged and older European men.
  - Cognitive function is measured by the Digit Symbol Substitution Test (DSST).

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.

Distribution of vitamin D



Distribution of DSST



Normally distributed

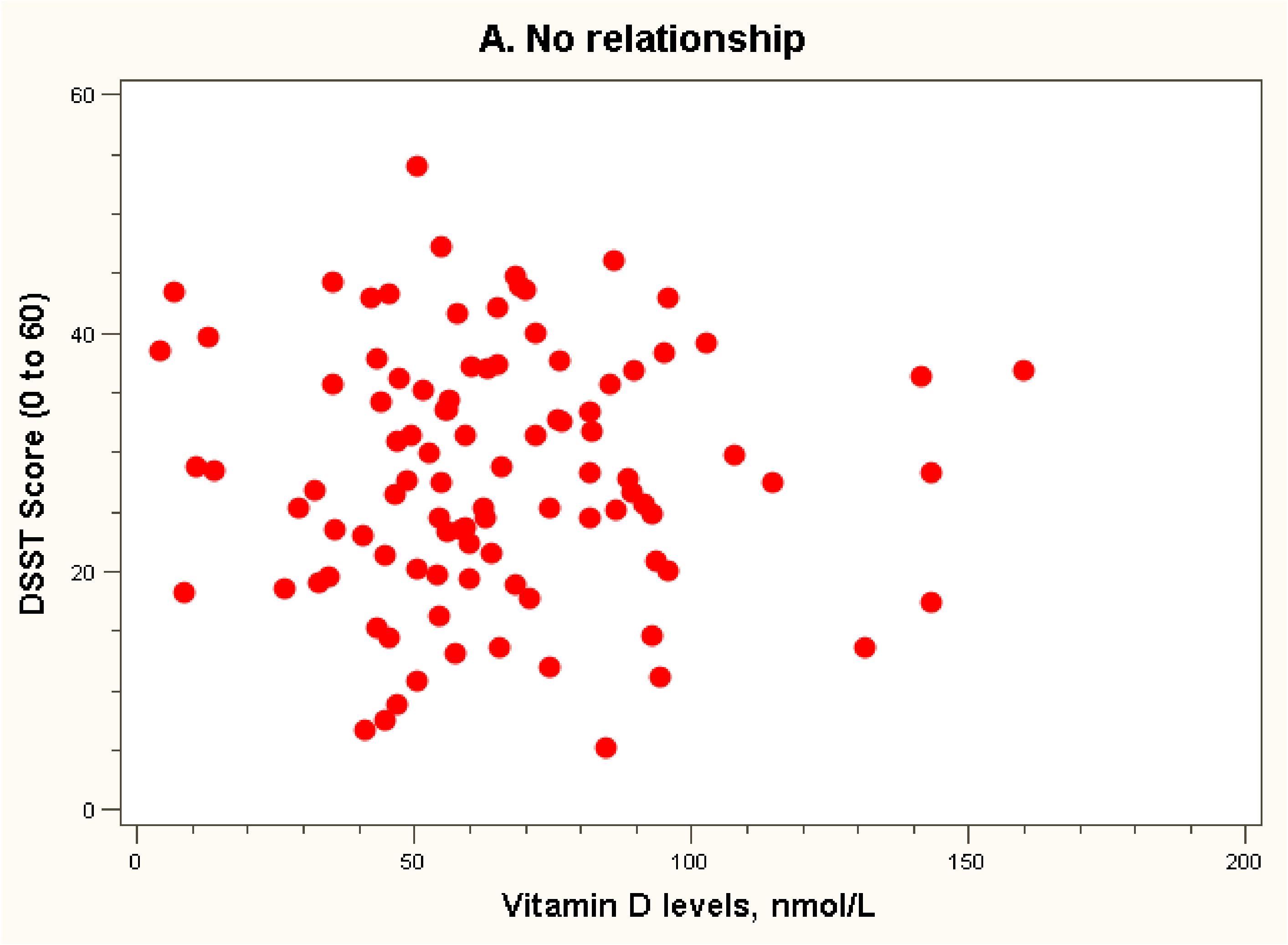
Mean = 28 points

Standard deviation = 10 points

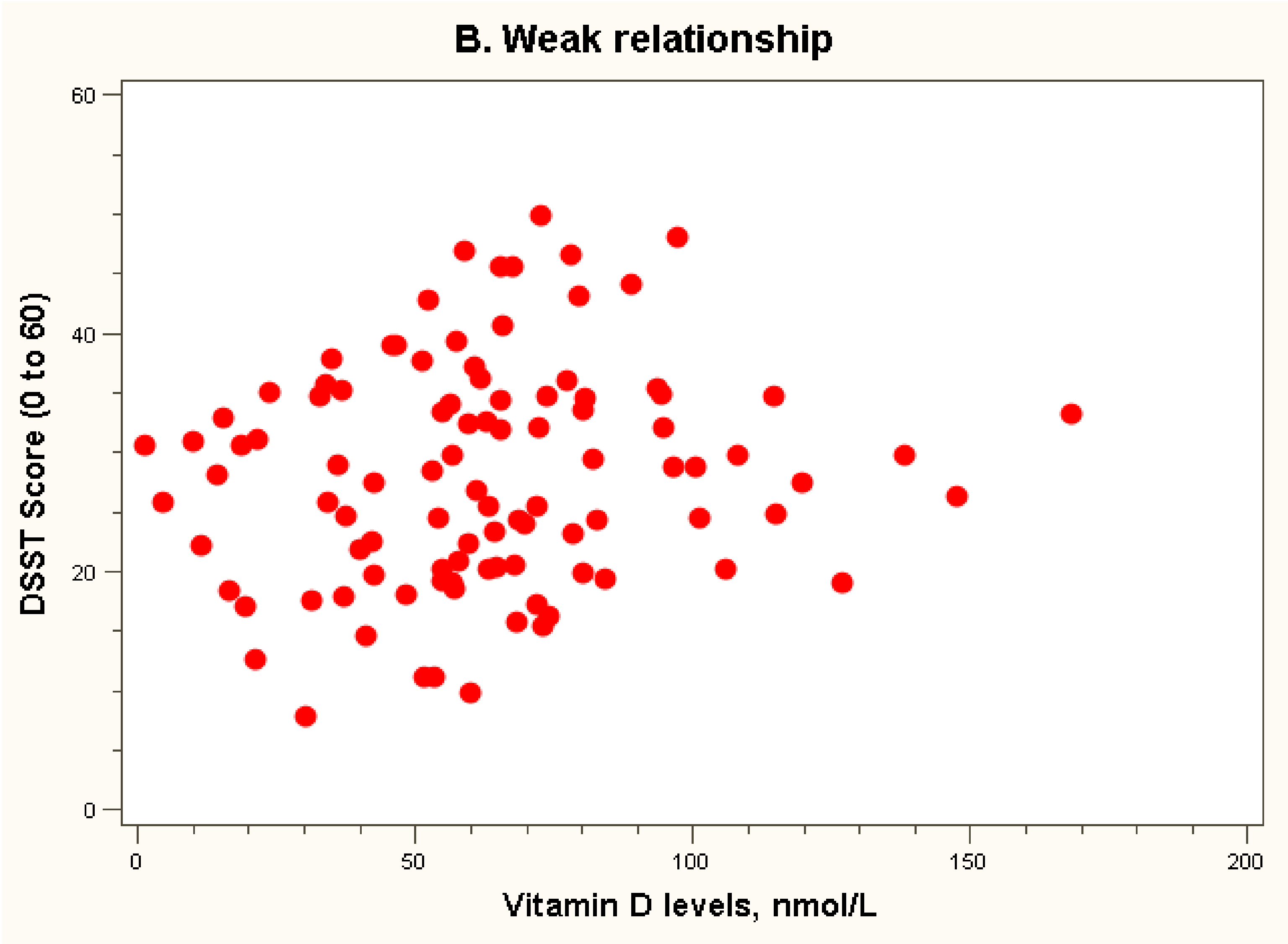
Four hypothetical datasets

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

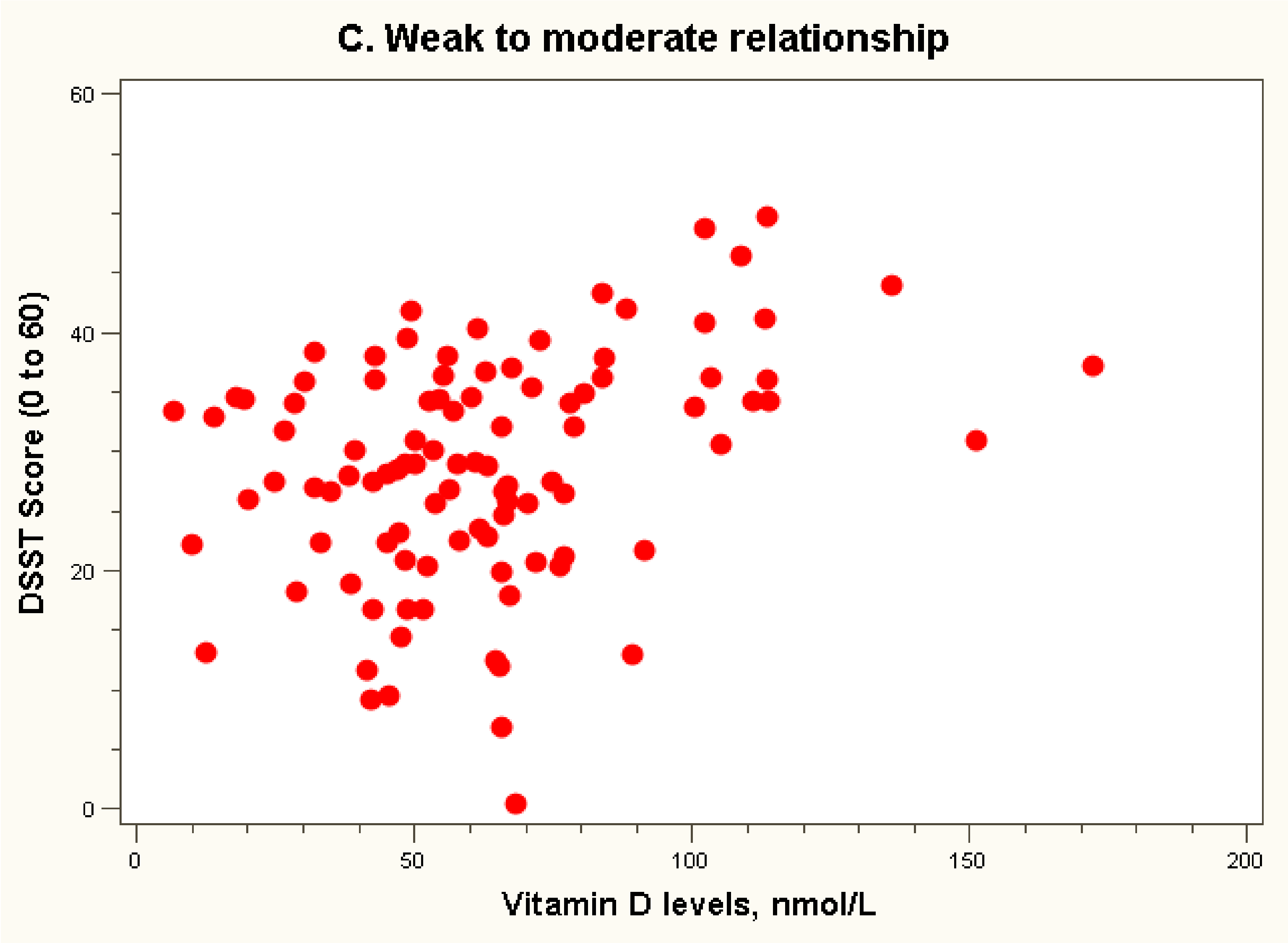
Dataset 1: no relationship



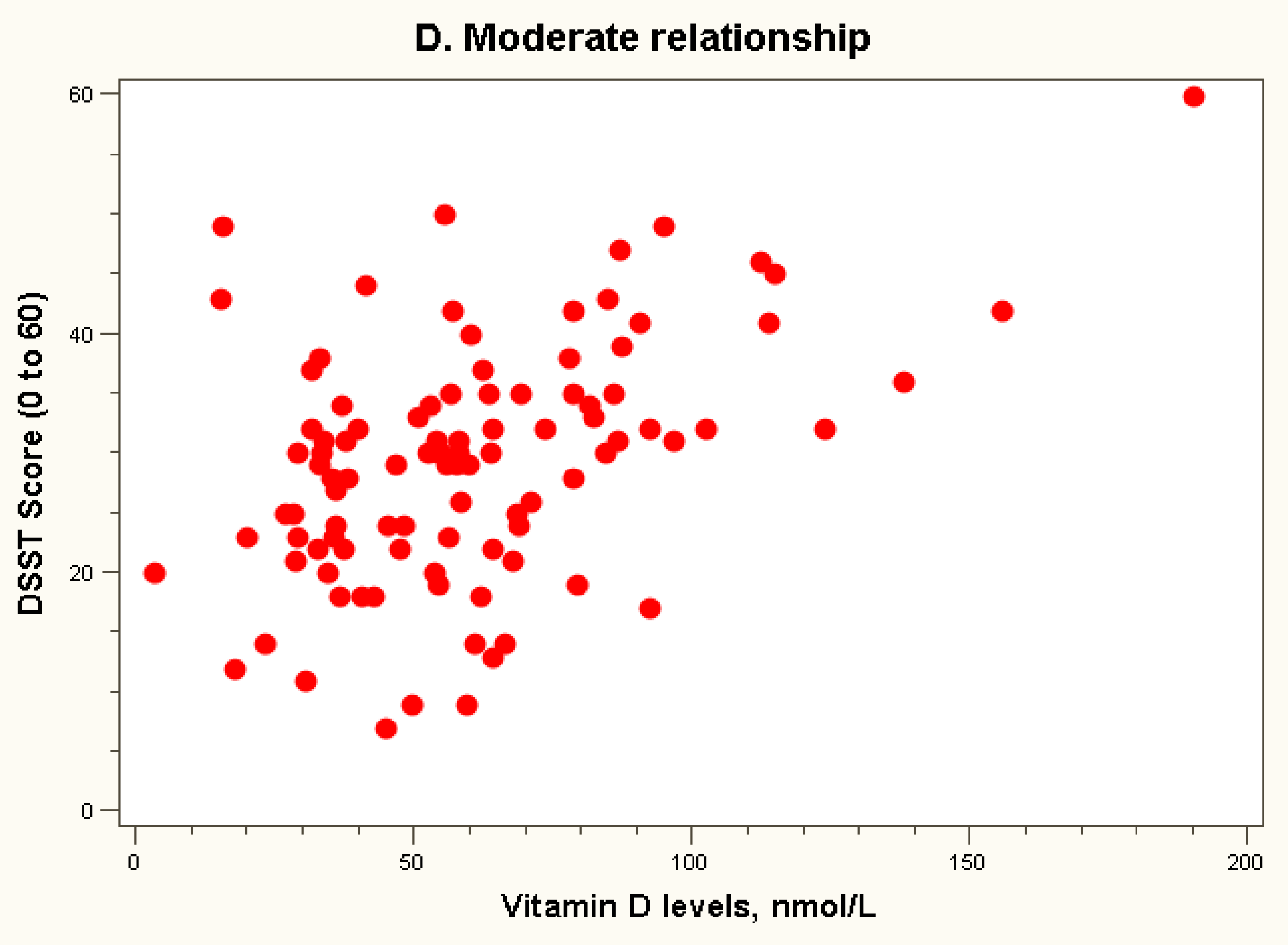
Dataset 2: weak relationship



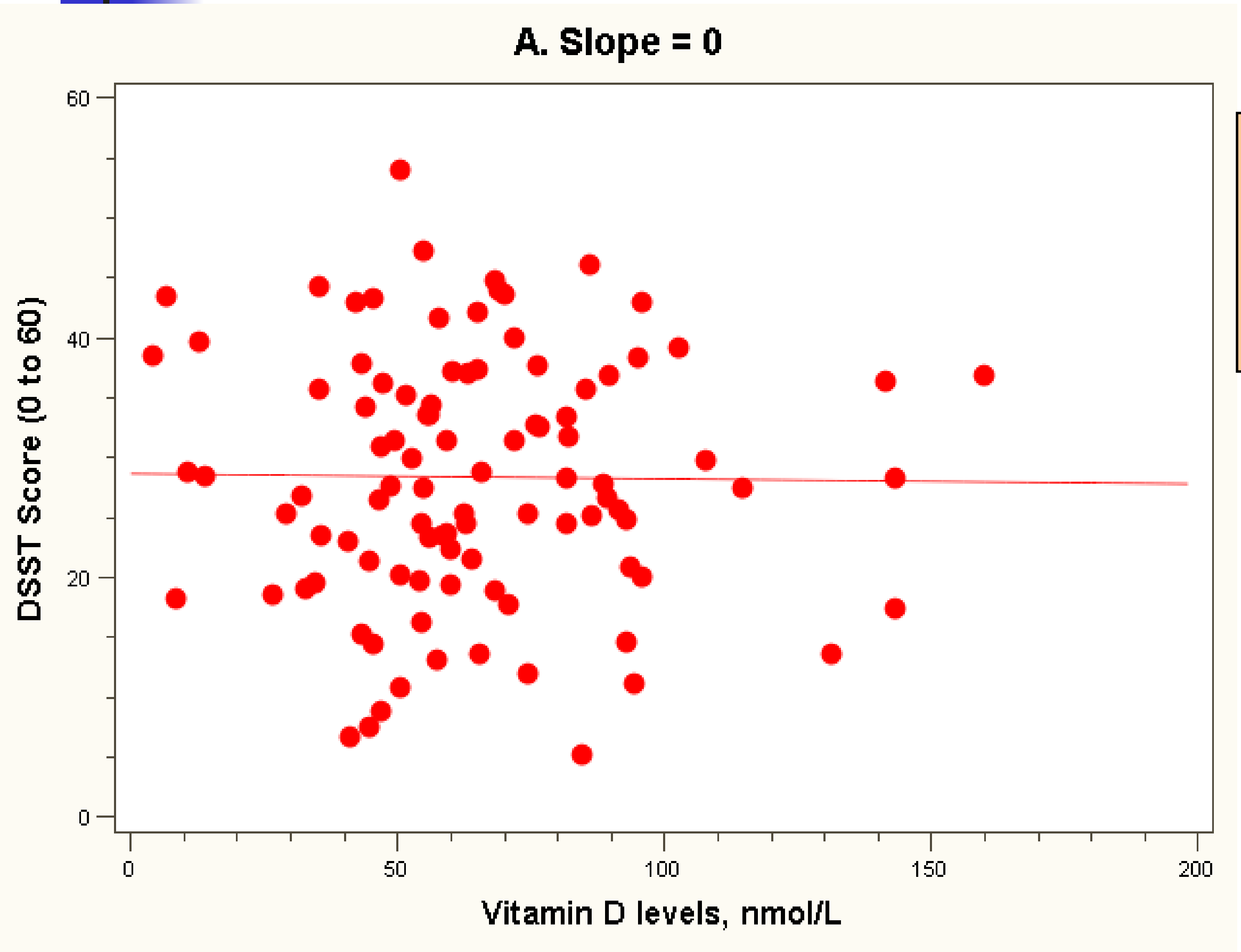
Dataset 3: weak to moderate relationship



Dataset 4: moderate relationship



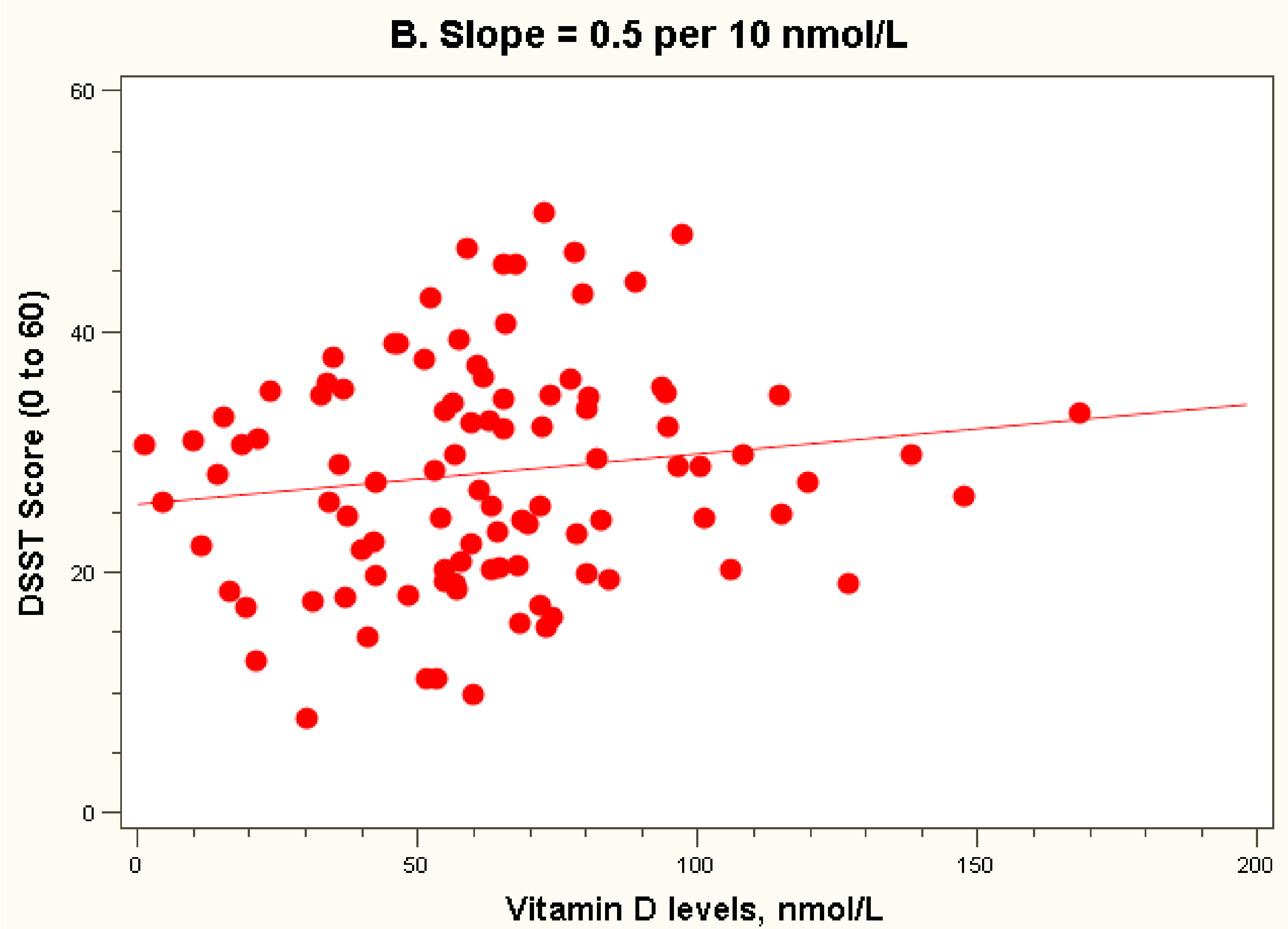
The "Best fit" line



Regression equation:

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

The "Best fit" line

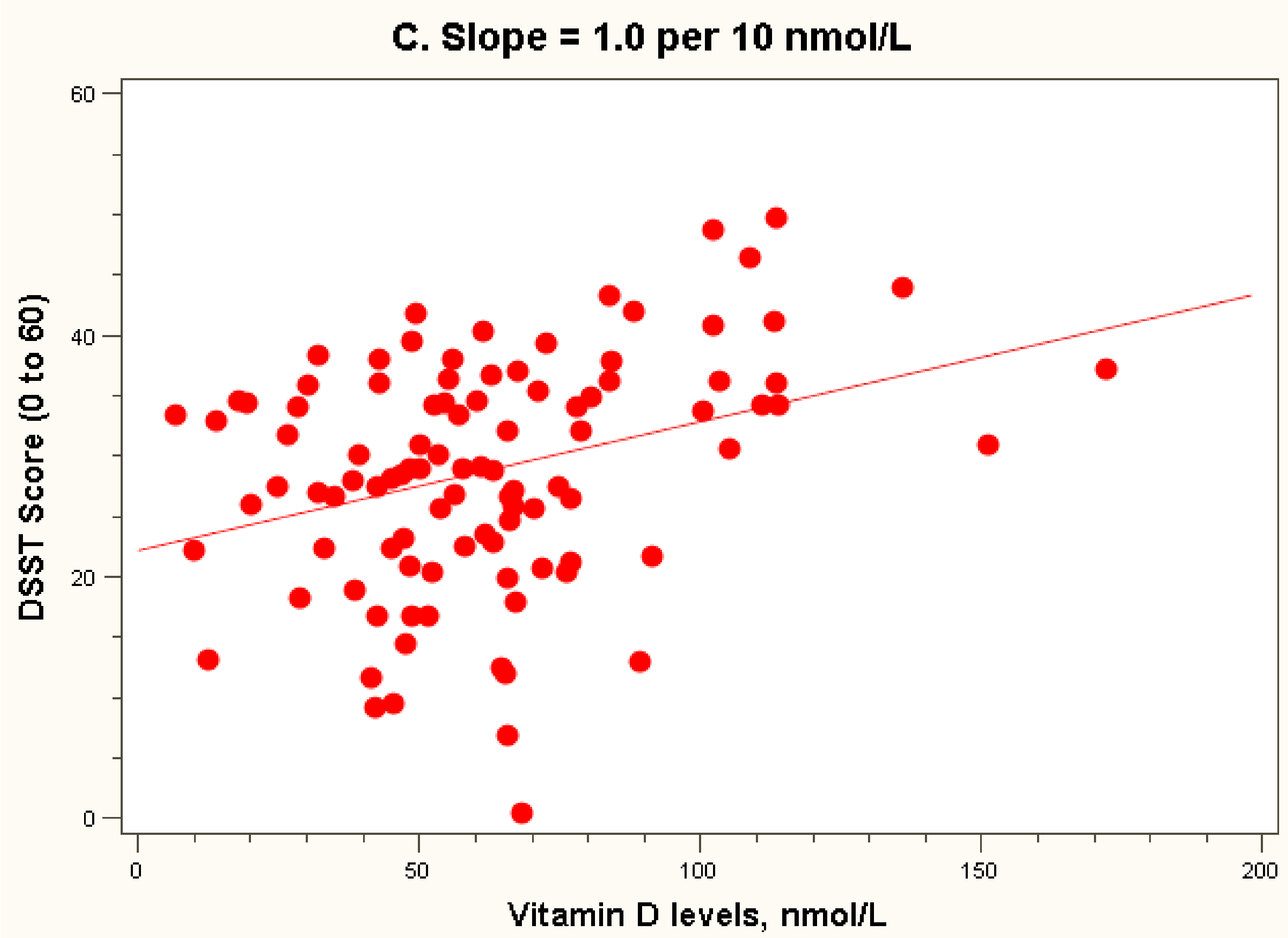


Note how the line is a little deceptive; it draws your eye, making the relationship appear stronger than it really is!

Regression equation:

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

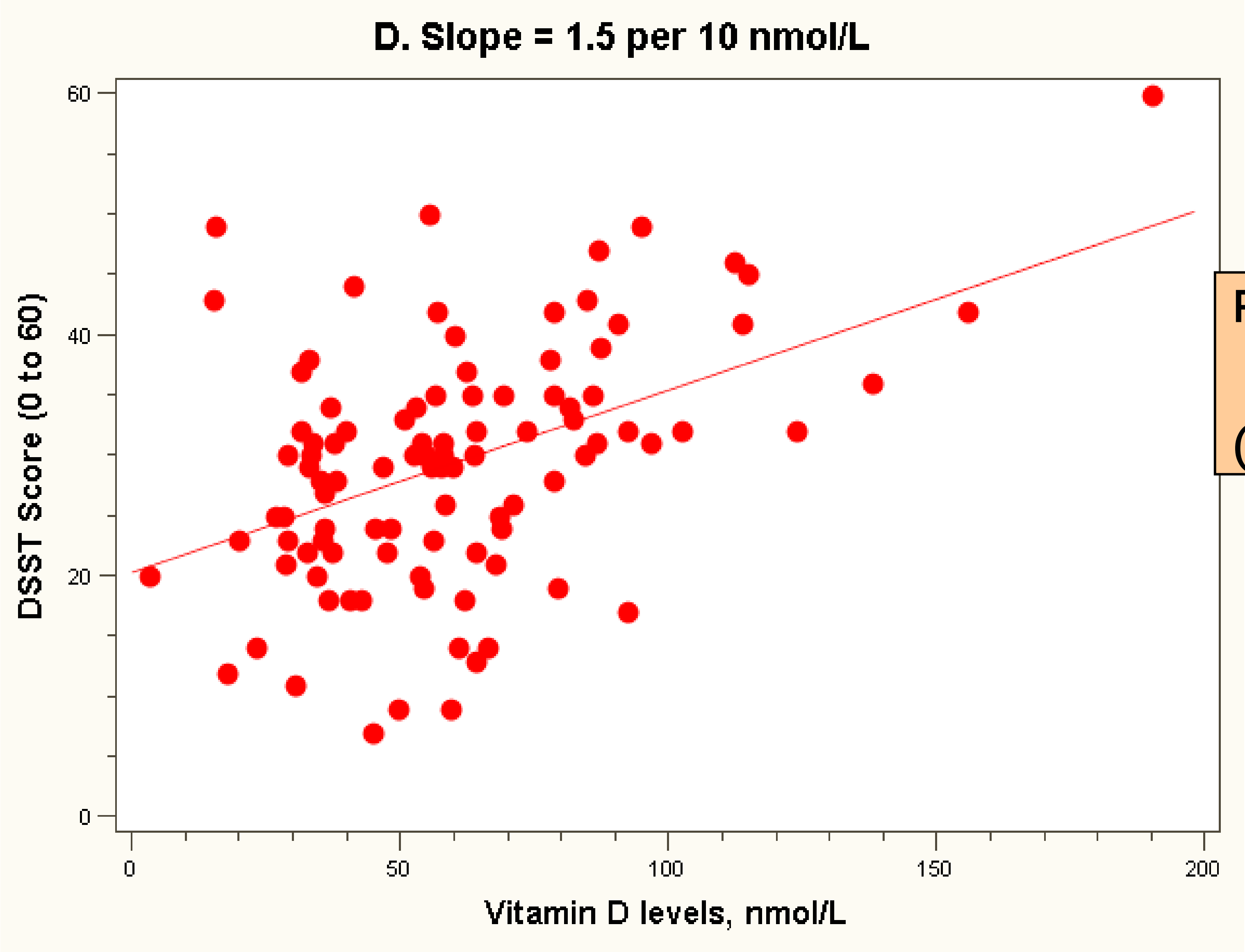
The "Best fit" line



Regression equation:

$E(Y_i) = 22 + 1.0 * vit$ D_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)!

Estimating the intercept and slope: least squares estimation

** 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 maximize a function? Take the derivative; se

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

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

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

Resulting formulas…

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

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

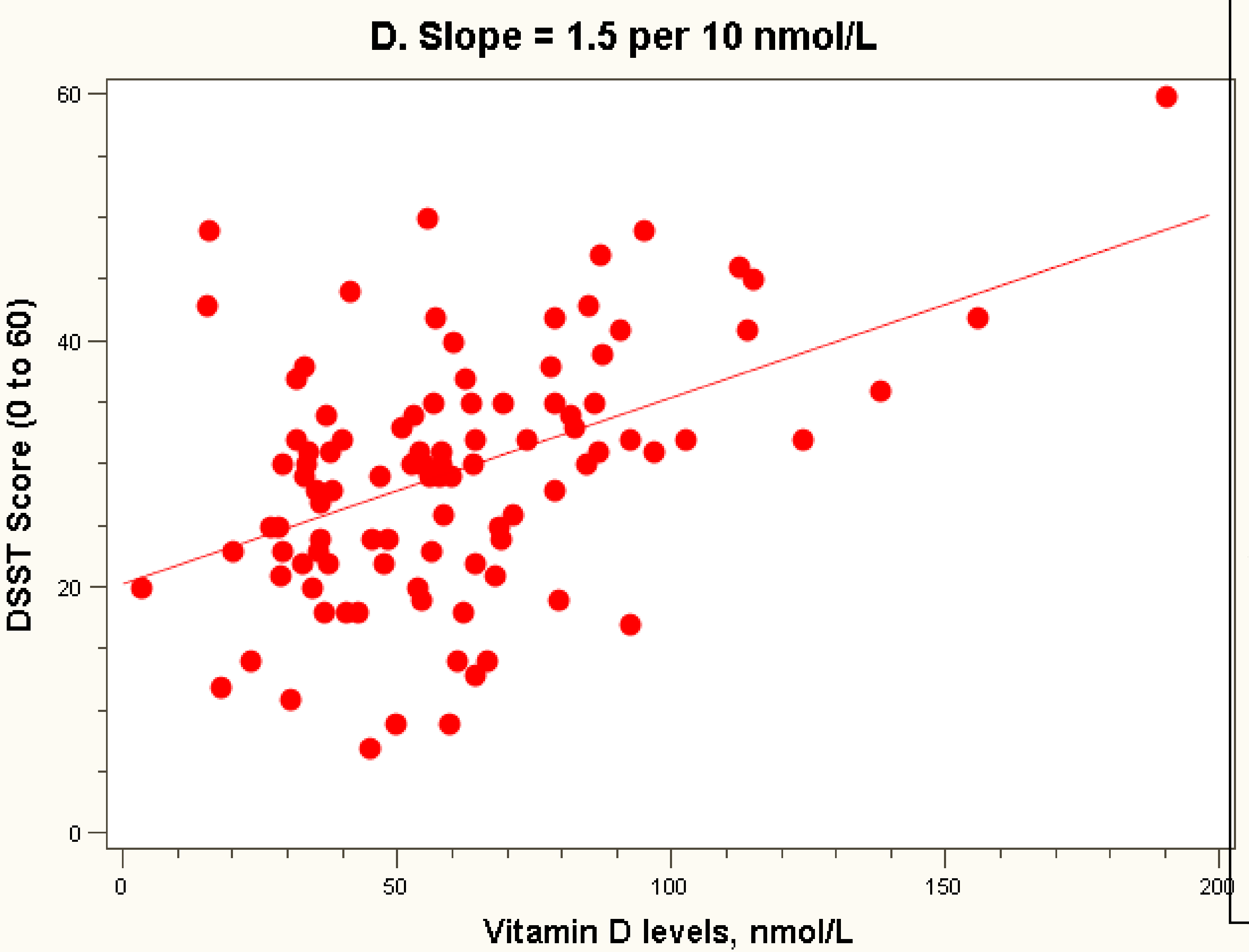
Regression line always goes through the point: $(\bar{x}, \bar{y})$

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*.

Example: dataset 4



SDx = 33 nmol/L

SDy= 10 points

Cov(X,Y) = 163 points*nmol/L

Beta = $163/33^2$ = 0.15 points per nmol/L

= 1.5 points per 10 nmol/L

r = 163/(10*33) = 0.49

Or

r = 0.15 * (33/10) = 0.49

Significance testing…

<u>Slope</u>

Distribution of slope ~ $T_{n-2}(\beta, s.e.(\hat{\beta}))$

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship does exist)

$$T_{n-2} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

Formula for the standard error of beta (you will not have to calculate by hand!):

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

where $SS_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$

and $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Example: dataset 4

- Standard error (beta) = 0.03
- T_{98} = 0.15/0.03 = 5, p<.0001

- 95% Confidence interval = 0.09 to 0.21

Residual Analysis: check assumptions

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

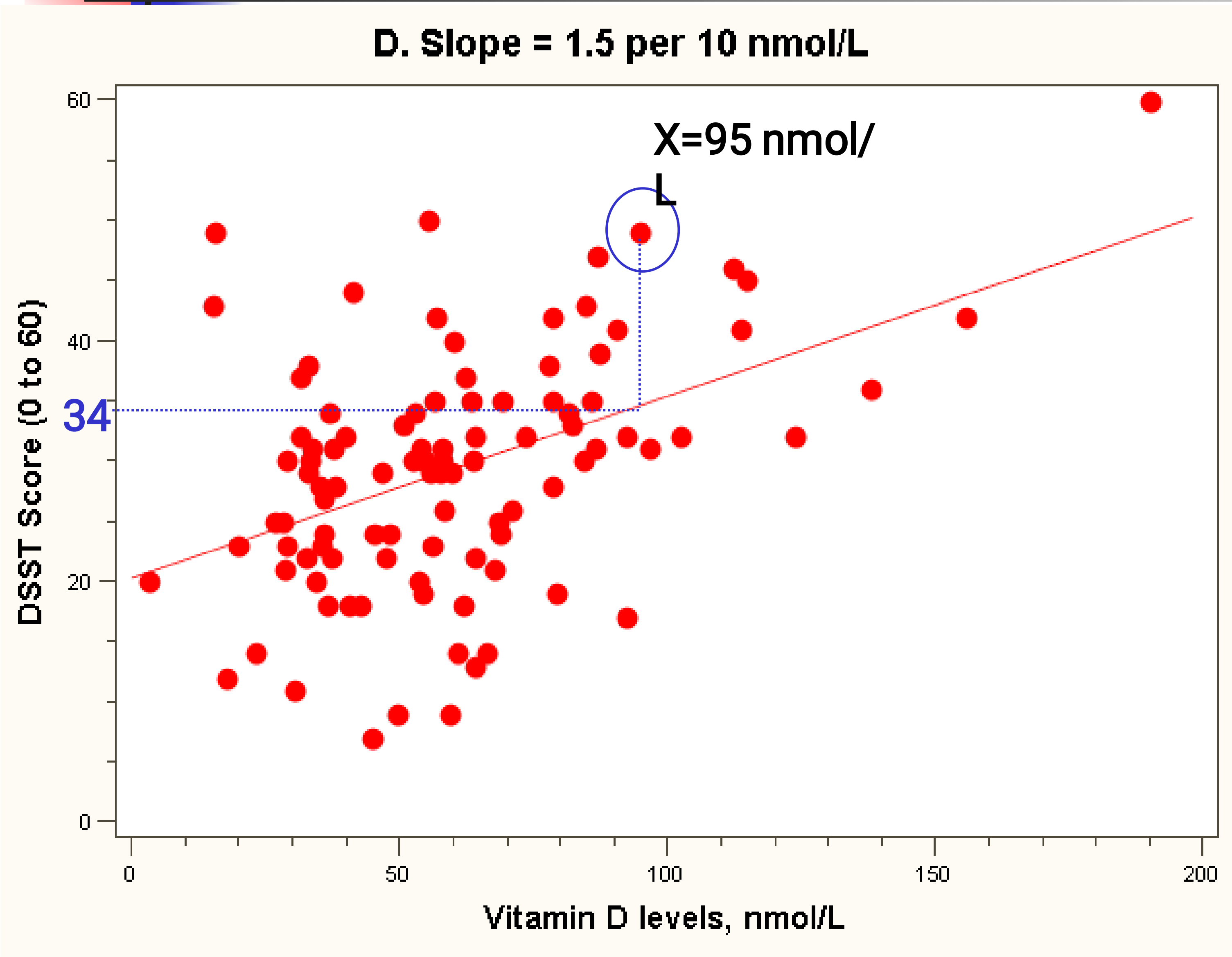
Predicted values…

$$\hat{y}_i = 20 + 1.5 x_i$$

For Vitamin D = 95 nmol/L (or 9.5 in 10 nmol/L):

$$\hat{y}_i = 20 + 1.5(9.5) = 34$$

Residual = observed - predicted

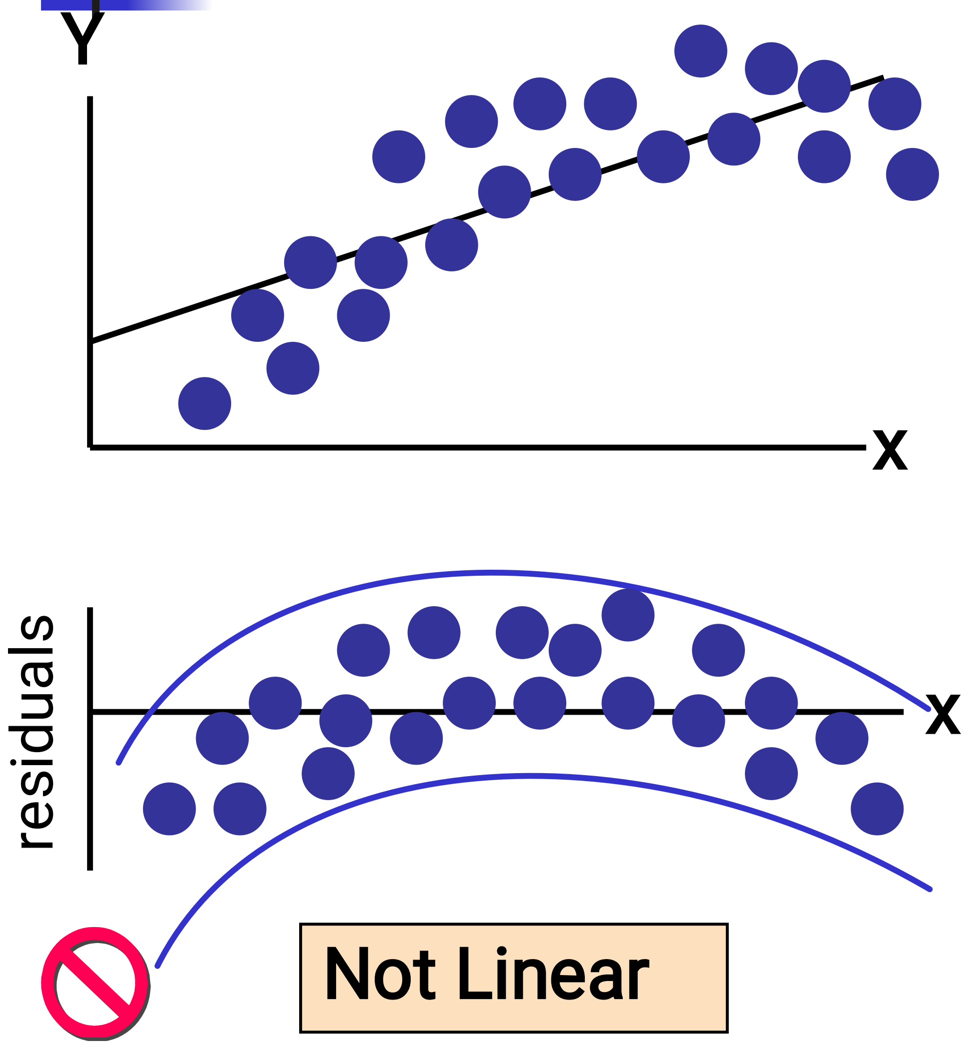


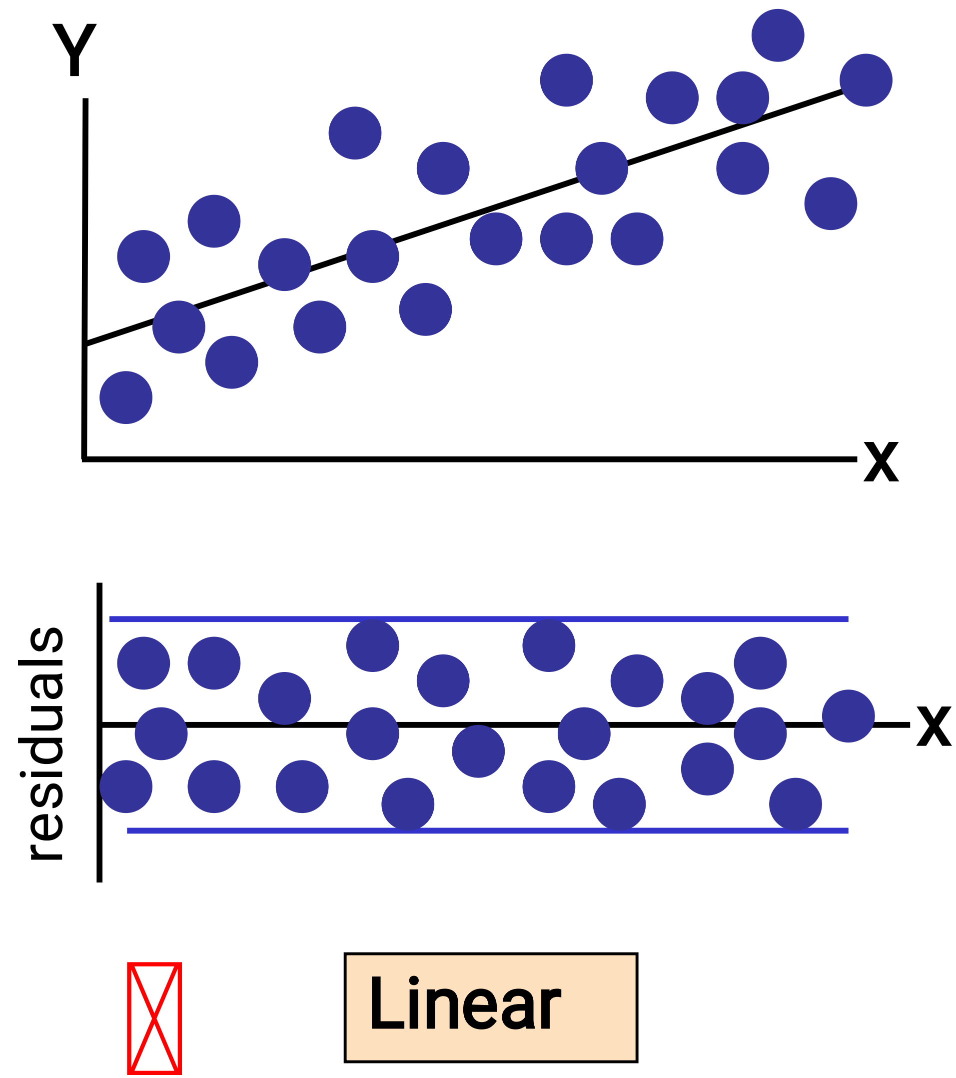
$$y_i = 48$$

$$\hat{y}_i = 34$$

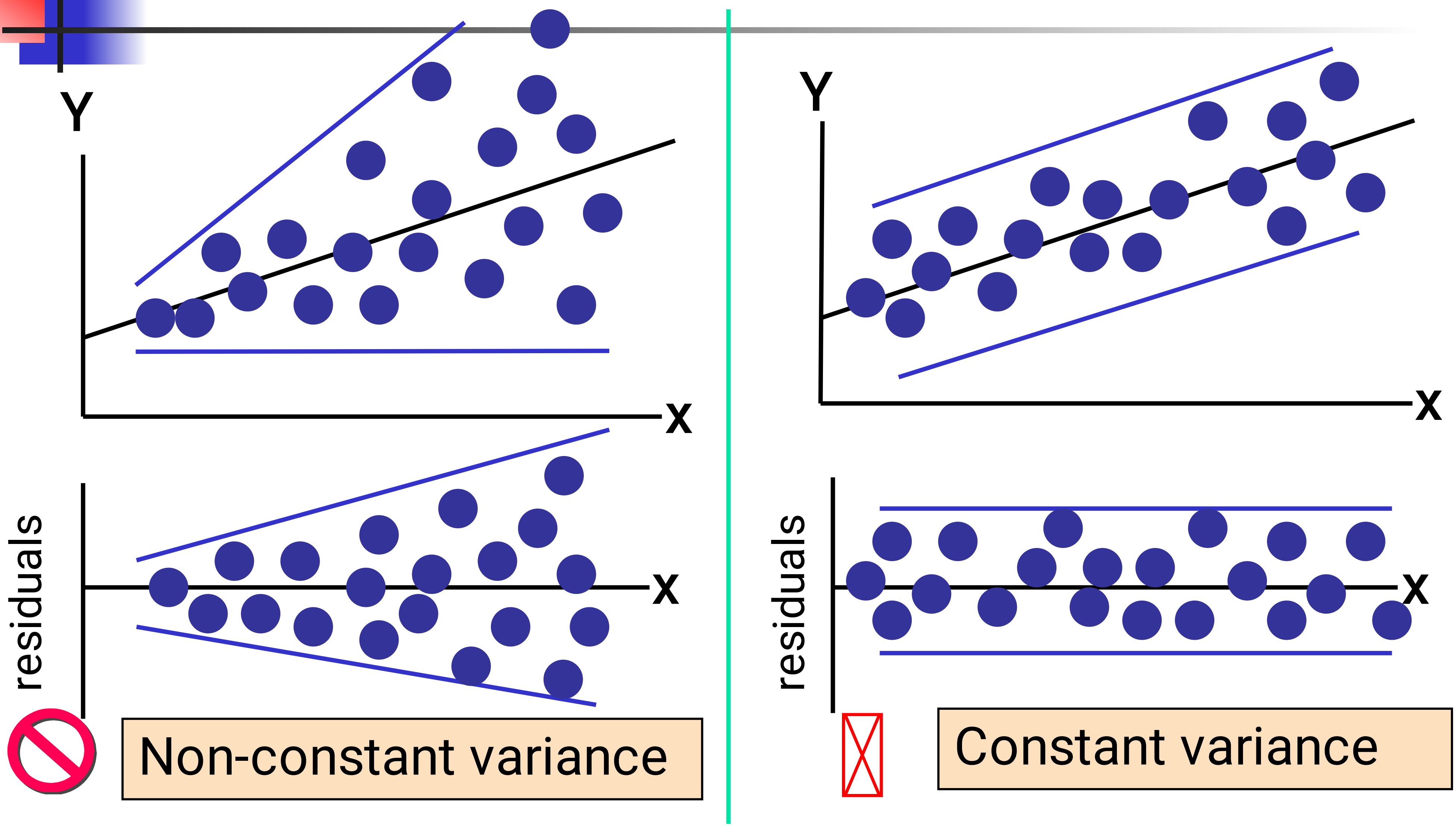
$$y_i - \hat{y}_i = 14$$

Residual Analysis for Linearity



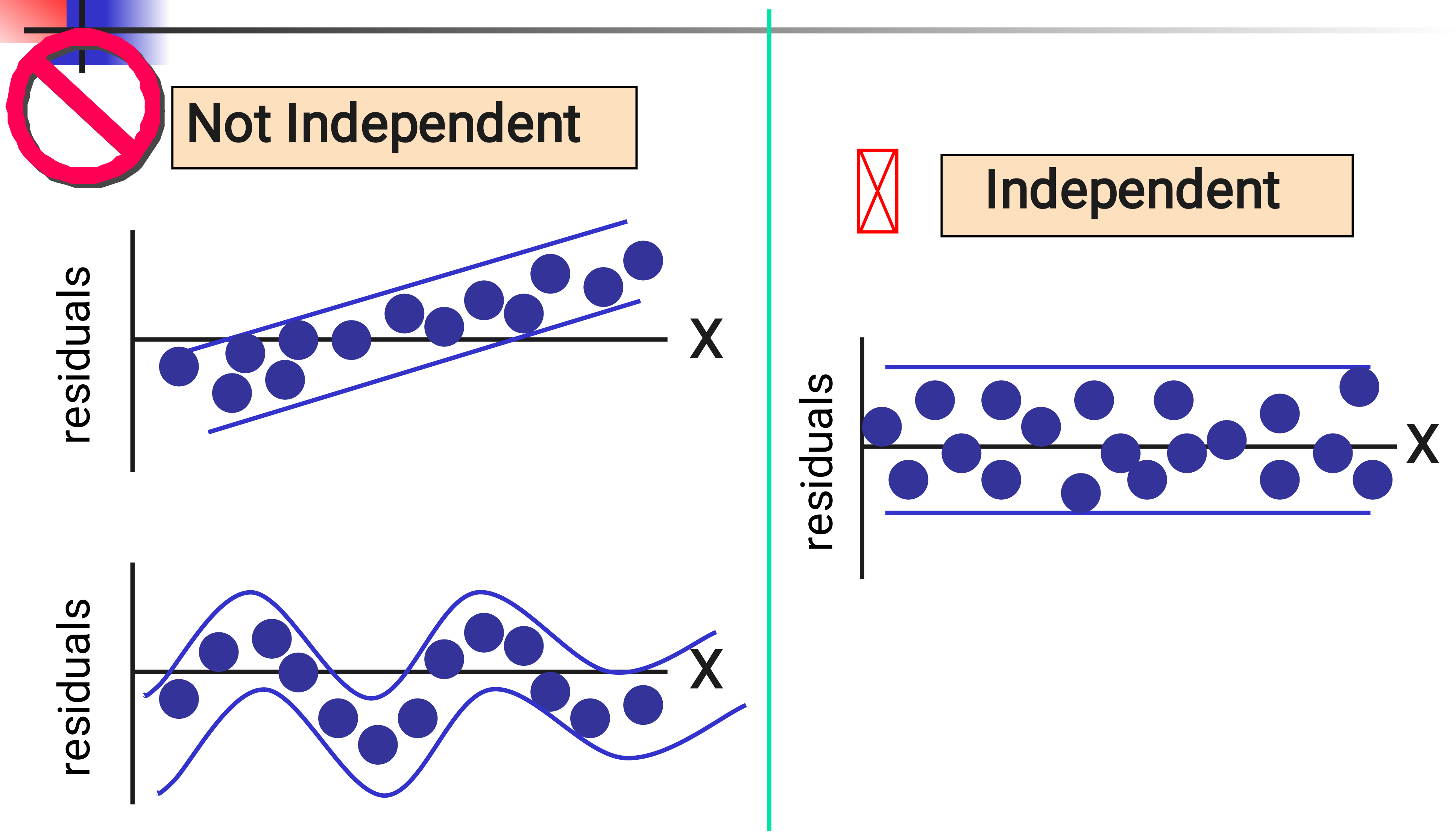


Residual Analysis for Homoscedasticity



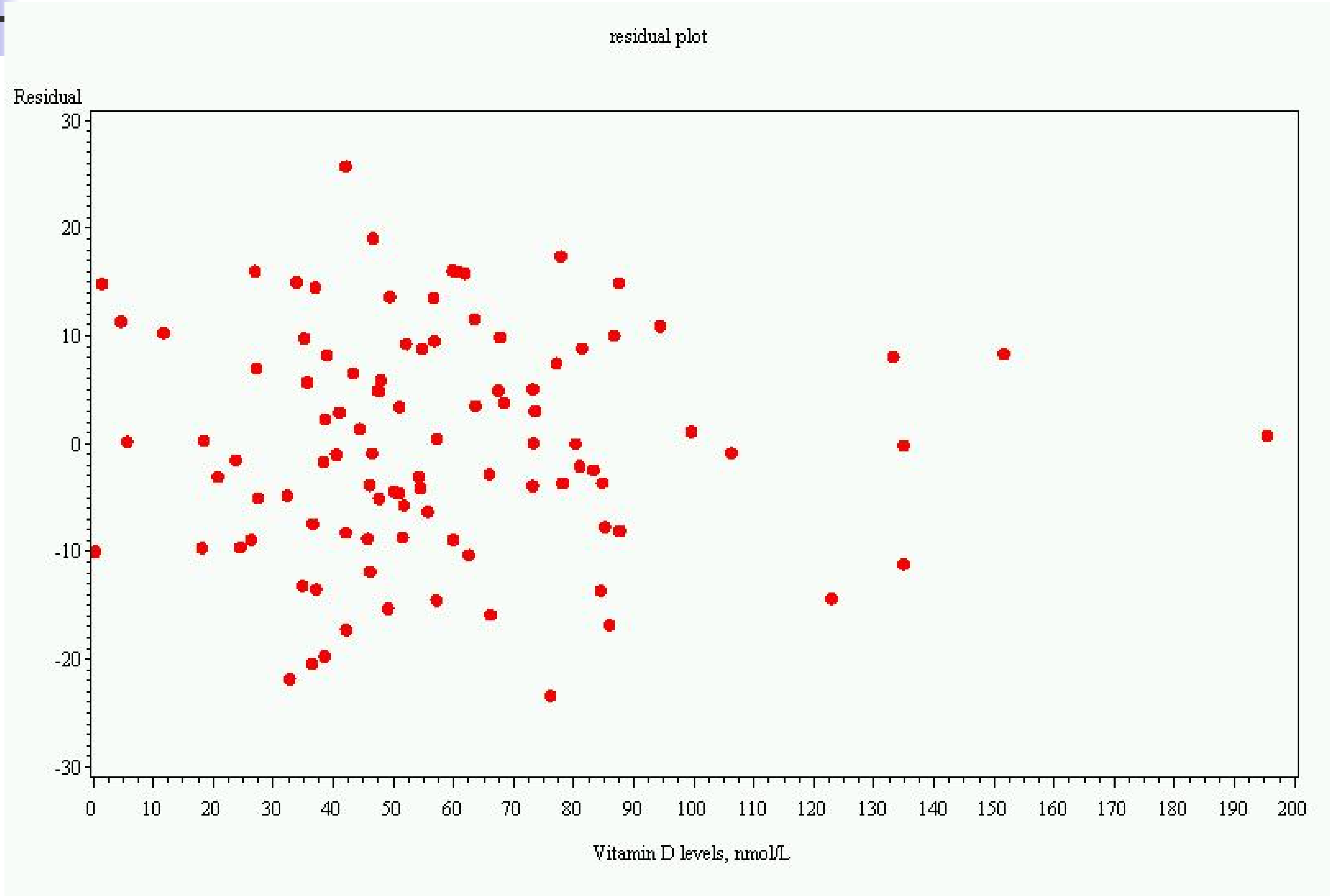
Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Residual Analysis for Independence



■Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

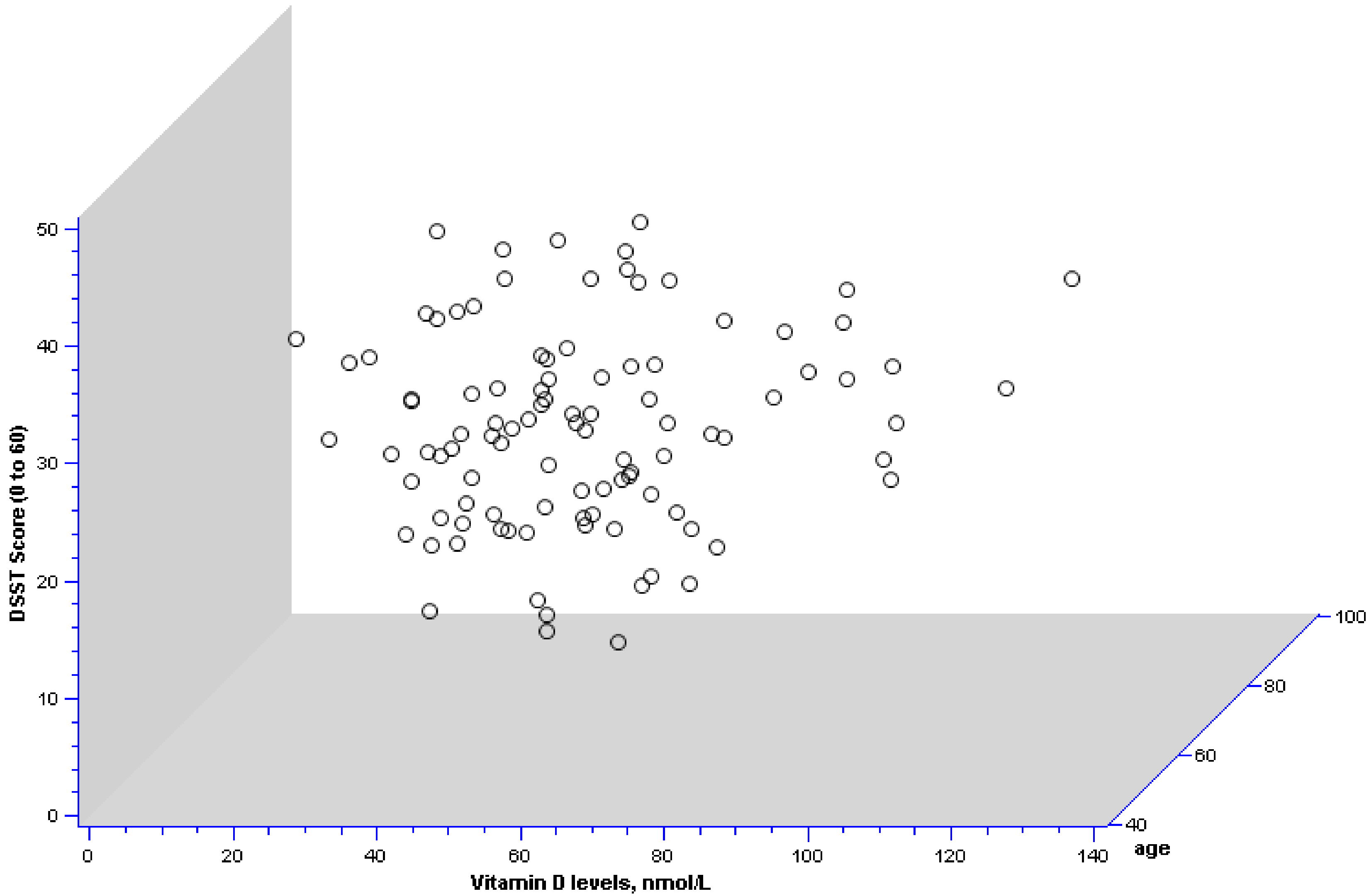
Residual plot, dataset 4



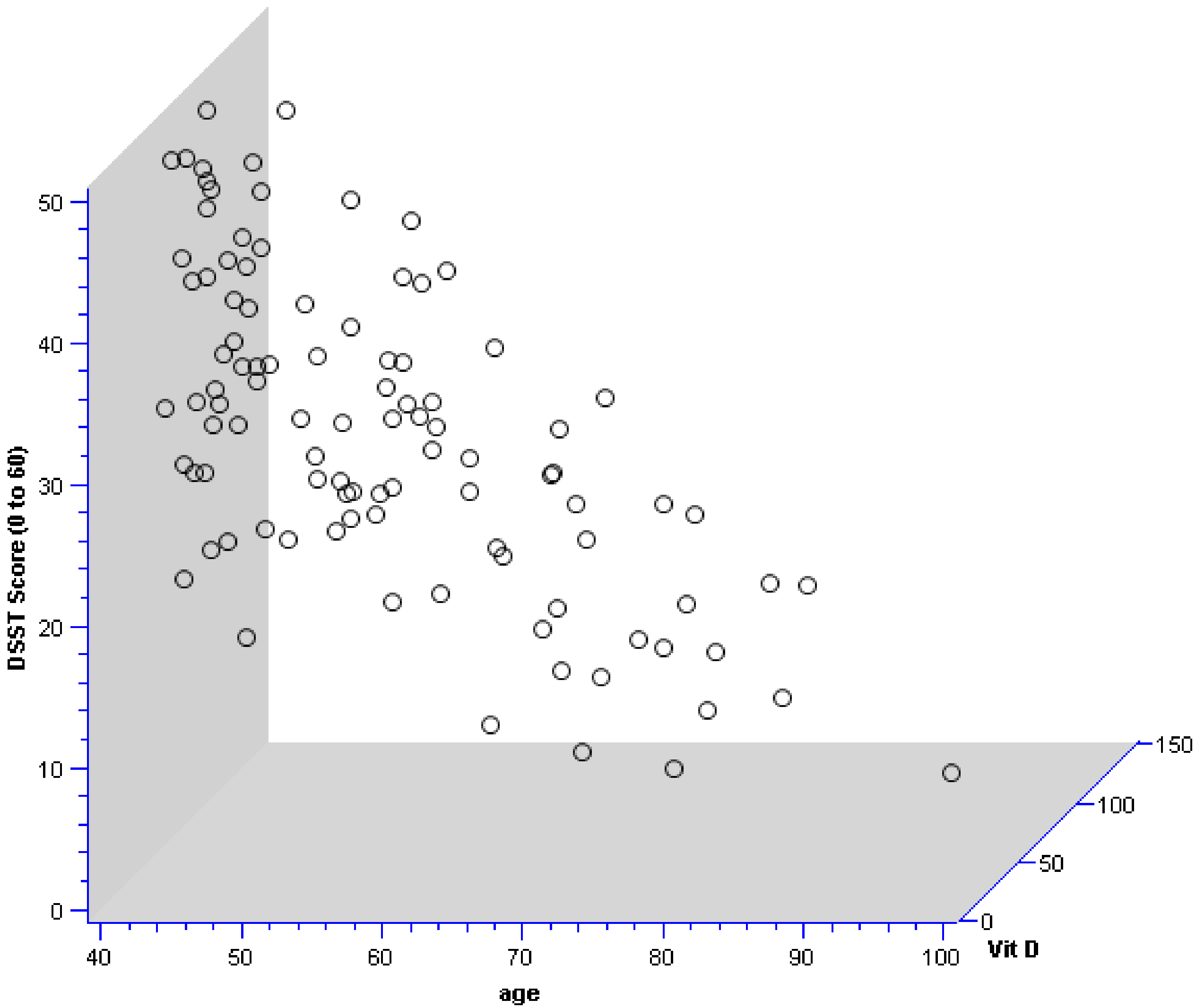
Multiple linear regression…

- 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 + $\text{slope}_1 x$vitamin D + $\text{slope}_2\ x$age

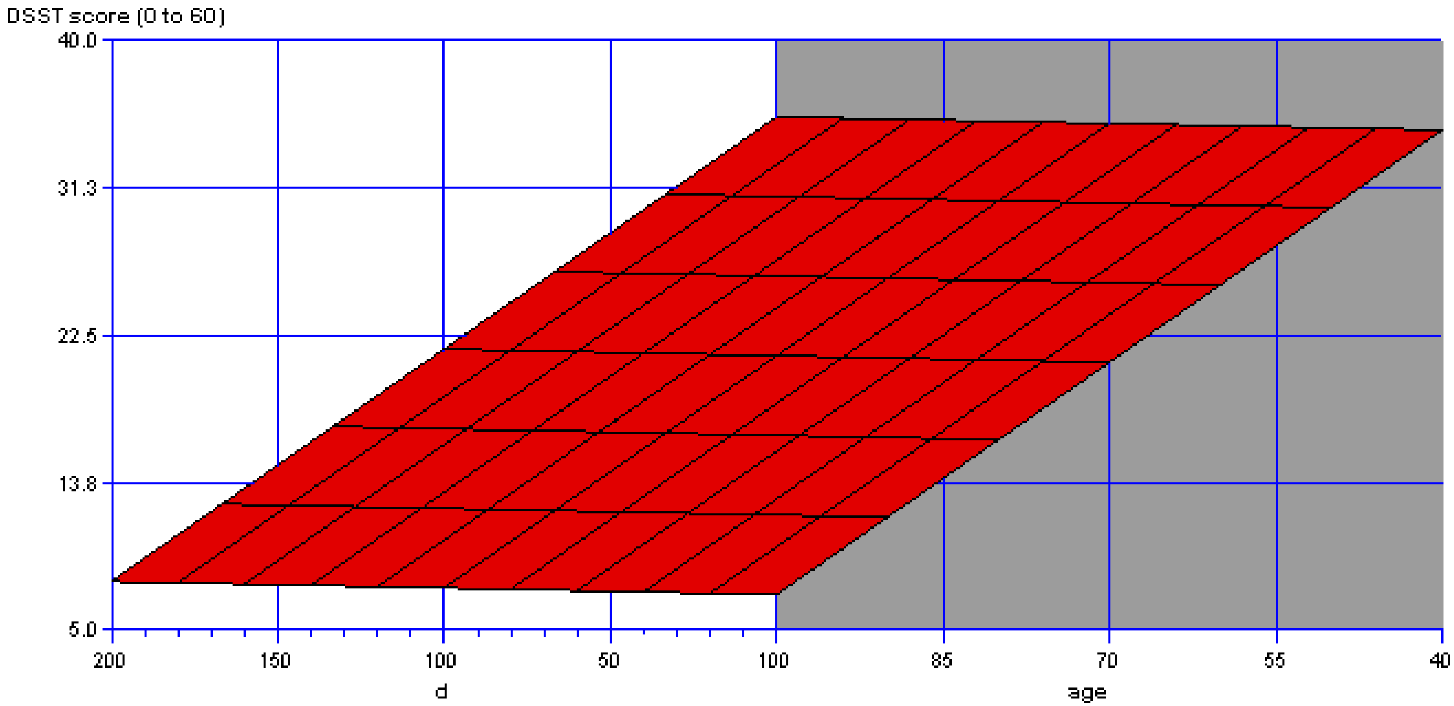
2 predictors: age and vit D…



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.

Equation of the "Best fit" plane…

- DSST score = 53 + 0.0039 *x* vitamin D (in 10 nmol/L) - 0.46 *x* age (in years)

- P-value for vitamin D >>.05
- P-value for age <.0001

- Thus, relationship with vitamin D was due to confounding by age!

Multiple Linear Regression

- More than one predictor…

E(y)= $\alpha + \beta_1$*X + β_2*W + β_3 *Z…

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

A ttest is linear regression!

- Divide vitamin D into two groups:
  - Insufficient vitamin D (<50 nmol/L)
  - Sufficient vitamin D (>=50 nmol/L), reference group
- We can evaluate these data with a ttest or a linear regression…

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

As a linear regression…

Intercept represents the mean value in the sufficient group.

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

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	40.07407	1.47511	27.17	<.0001
insuff	-7.53060	2.17493	-3.46	0.0008

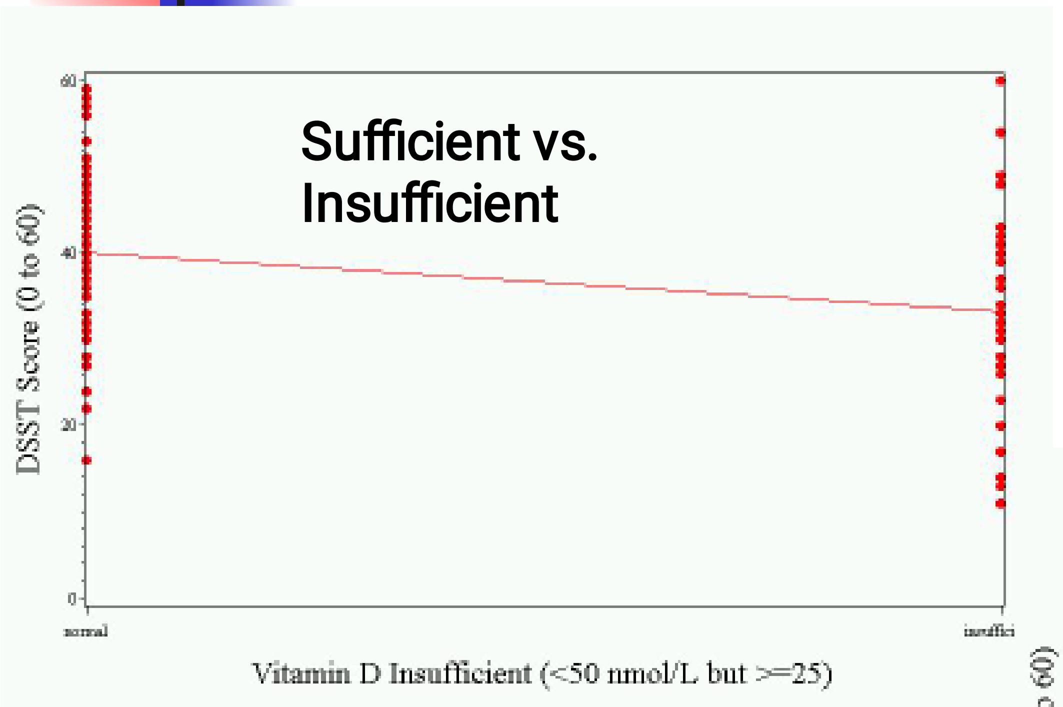
ANOVA is linear regression!

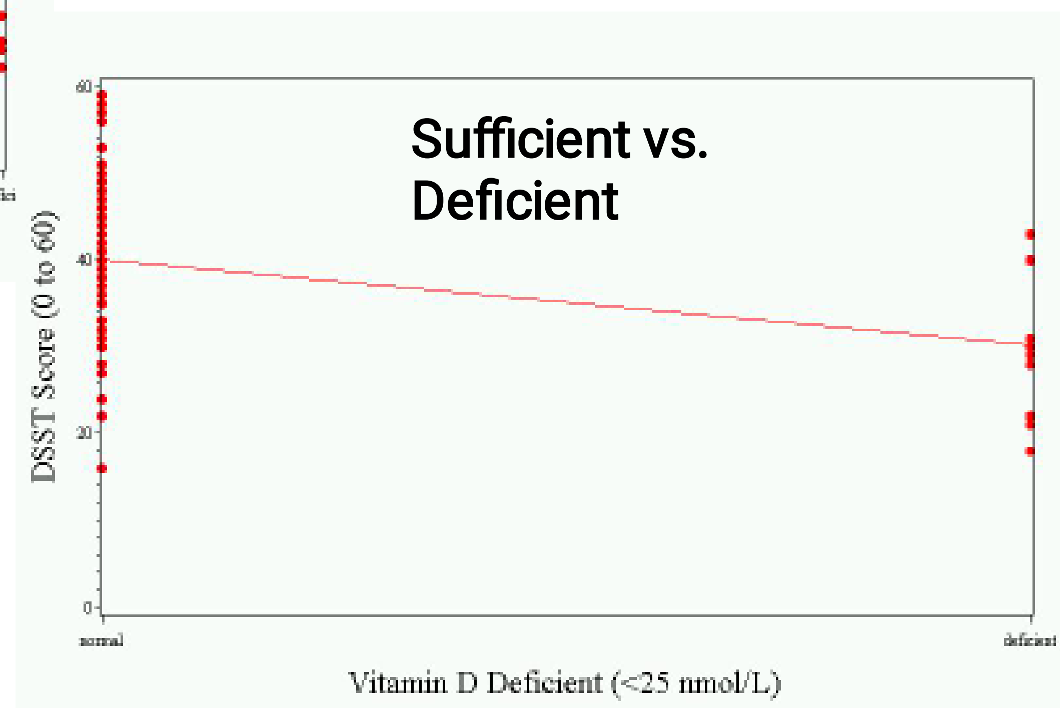
- Divide vitamin D into three groups:
  - Deficient (<25 nmol/L)
  - Insufficient (>=25 and <50 nmol/L)
  - Sufficient (>=50 nmol/L), reference group

DSST$= \alpha$ (=value for sufficient) + $\beta_{insufficient}$*(1 if insufficient) + β_2*(1 if deficient)

This is called "dummy coding"—where multiple binary variables are created to represent being in each category (or not) of a categorical variable

The picture…


Sufficient vs.
Insufficient
DSST Score (0 to 60)
60
40
20
0
normal
insuffici
Vitamin D Insufficient (<50 nmol/L but >=25)


Sufficient vs.
Deficient
DSST Score (0 to 60)
60
40
20
0
normal
deficient
Vitamin D Deficient (<25 nmol/L)

Results…

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	40.07407	1.47817	27.11	<.0001
deficient	**1**	**-9.87407**	**3.73950**	**-2.64**	**0.0096**
insufficient	**1**	**-6.87963**	**2.33719**	**-2.94**	**0.0041**

- Interpretation:
  - The deficient group has a mean DSST 9.87 points lower than the reference (sufficient) group.
  - The insufficient group has a mean DSST 6.87 points lower than the reference (sufficient) group.

Other types of multivariate regression

- Multiple linear regression is for normally distributed outcomes
- Logistic regression is for binary outcomes
- Cox proportional hazards regression is used when time-to-event is the outcome

Common multivariate regression models.

Outcome (dependent variable)	Example outcome variable	Appropriate multivariate regression model	Example equation	What do the coefficients give you?
Continuous	Blood pressure	**Linear regression**	blood pressure (mmHg) = $\alpha + \beta$salt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	slopes—tells you how much the outcome variable increases for every 1-unit increase in each predictor.
Binary	High blood pressure (yes/no)	**Logistic regression**	ln (odds of high blood pressure) = $\alpha + \beta$salt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	odds ratios—tells you how much the odds of the outcome increase for every 1-unit increase in each predictor.
Time-to-event	Time-to-death	**Cox regression**	ln (rate of death) = $\alpha + \beta$salt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	hazard ratios—tells you how much the rate of the outcome increases for every 1-unit increase in each predictor.

Multivariate regression pitfalls

- Multi-collinearity
- Residual confounding
- Overfitting

Multicollinearity

- **Multicollinearity** arises when two variables that measure the same thing or similar things (e.g., weight and BMI) are both included in a multiple regression model; they will, in effect, cancel each other out and generally destroy your model.

- Model building and diagnostics are tricky business!

Residual confounding

- You cannot completely wipe out confounding simply by adjusting for variables in multiple regression unless variables are measured with zero error (which is usually impossible).
- Example: meat eating and mortality

Men who eat a lot of meat are unhealthier for many reasons!

Table 1. Selected Age-Adjusted Characteristics of the National Institutes of Health–AARP Cohort by Red Meat Quintile Category[a]

Characteristic	Red Meat Intake Quintile, g/1000 kcal Q1	Q2	Q3	Q4	Q5
Men (n=322 263)					
Meat intake					
Red meat, g/1000 kcal	9.3	21.4	31.5	43.1	68.1
White meat, g/1000 kcal	36.6	32.2	30.7	30.4	30.9
Processed meat, g/1000 kcal	5.1	7.8	10.3	13.3	19.4
Age, y	62.8	62.8	62.5	62.3	61.7
Race, %					
Non-Hispanic white	88.6	91.8	93.1	94.0	94.1
Non-Hispanic black	4.2	3.2	2.7	2.2	1.9
Hispanic/Asian/Pacific Islander/American Indian/Alaskan native/unknown	7.2	5.0	4.2	3.8	4.0
Positive family history of cancer,%	47.0	47.7	48.4	48.6	47.8
Currently married, %	80.8	84.4	86.1	86.7	85.6
BMI	25.9	26.7	27.1	27.6	28.3
Smoking history, %[b]					
Never smoker	34.4	30.5	28.8	27.6	25.4
Former smoker	56.5	58.1	57.5	57.1	55.8
Current smoker or having quit <1 y prior	4.9	7.6	9.9	11.4	14.8
Education, college graduate or postgraduate, %	53.0	47.3	45.1	42.3	39.1
Vigorous physical activity ≥5 times/wk, %	30.7	23.6	20.5	18.6	16.3
Dietary intake					
Energy, kcal/d	1899	1955	1998	2038	2116
Fruit, servings/1000 kcal	2.3	1.8	1.6	1.4	1.1
Vegetables, servings/1000 kcal	2.4	2.1	2.0	2.0	1.9
Alcohol, g/d	20.2	20.4	17.6	15.3	12.5
Total fat, g/1000 kcal	25.8	30.5	33.5	35.9	39.4
Saturated fat, g/1000 kcal	7.6	9.4	10.5	11.3	12.7
Fiber, g/1000 kcal	13.2	11.0	10.2	9.6	8.8
Vitamin supplement use ≥1/mo	67.3	62.1	59.1	55.8	52.0

Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71

Mortality risks…

Table 2. Multivariate Analysis for Red, White, and Processed Meat Intake and Total and Cause-Specific Mortality in Men in the National Institutes of Health–AARP Diet and Health Study[a]

Mortality in Men (n=322 263)	Quintile Q1	Q2	Q3	Q4	Q5	*P* Value for Trend
			Red Meat Intake[b]			
All mortality						
Deaths	6437	7835	9366	10 988	13 350	
Basic model[c]	1 [Reference]	1.07 (1.03-1.10)	1.17 (1.13-1.21)	1.27 (1.23-1.31)	1.48 (1.43-1.52)	<.001
Adjusted model[d]	1 [Reference]	1.06 (1.03-1.10)	1.14 (1.10-1.18)	1.21 (1.17-1.25)	1.31 (1.27-1.35)	<.001
Cancer mortality						
Deaths	2136	2701	3309	3839	4448	
Basic model[c]	1 [Reference]	1.10 (1.04-1.17)	1.23 (1.16-1.29)	1.31 (1.24-1.39)	1.44 (1.37-1.52)	<.001
Adjusted model[d]	1 [Reference]	1.05 (0.99-1.11)	1.13 (1.07-1.20)	1.18 (1.12-1.25)	1.22 (1.16-1.29)	<.001
CVD mortality						
Deaths	1997	2304	2703	3256	3961	
Basic model[c]	1 [Reference]	1.02 (0.96-1.08)	1.10 (1.04-1.17)	1.24 (1.17-1.31)	1.44 (1.37-1.52)	<.001
Adjusted model[d]	1 [Reference]	0.99 (0.96-1.09)	1.08 (1.02-1.15)	1.18 (1.12-1.26)	1.27 (1.20-1.35)	<.001
Mortality from injuries and sudden deaths						
Deaths	184	216	228	280	343	
Basic model[c]	1 [Reference]	1.02 (0.84-1.24)	0.97 (0.80-1.18)	1.09 (0.90-1.31)	1.24 (1.03-1.49)	.01
Adjusted model[d]	1 [Reference]	1.06 (0.86-1.29)	1.01 (0.83-1.24)	1.14 (0.94-1.39)	1.26 (1.04-1.54)	.008
All other deaths						
Deaths	1268	1636	1971	2239	2962	
Basic model[c]	1 [Reference]	1.13 (1.05-1.22)	1.25 (1.17-1.35)	1.33 (1.24-1.42)	1.68 (1.57-1.80)	<.001
Adjusted model[d]	1 [Reference]	1.17 (1.09-1.26)	1.28 (1.19-1.38)	1.34 (1.25-1.44)	1.58 (1.47-1.70)	<.001

Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71

Overfitting

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

Overfitting: class data example

- I asked SAS to automatically find predictors of optimism in our class dataset. Here's the resulting linear regression model:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	11.80175	2.98341	11.96067	15.65	0.0019
exercise	-0.29106	0.09798	6.74569	8.83	0.0117
sleep	-1.91592	0.39494	17.98818	23.53	0.0004
obama	1.73993	0.24352	39.01944	51.05	<.0001
Clinton	-0.83128	0.17066	18.13489	23.73	0.0004
mathLove	0.45653	0.10668	13.99925	18.32	0.0011

Exercise, sleep, and high ratings for Clinton are negatively related to optimism (*highly* significant!) and high ratings for Obama and high love of math are positively related to optimism (*highly* significant!).

If something seems to good to be true…

Clinton, univariate:

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.43688	2.13476	2.55	0.0188
Clinton	**Clinton**	**1**	**0.24973**	**0.27111**	**0.92**	**0.3675**

Sleep, Univariate:

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	8.30817	4.36984	1.90	0.0711
					.22	**0.8270**

Exercise, Univariate:

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6.65189	0.89153	7.46	<.0001
exercise	**exercise**	**1**	**0.19161**	**0.20709**	**0.93**	**0.3658**

More univariate models...

Obama, Univariate:

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.82107	2.43137	0.34	0.7389
obama	**obama**	**1**	**0.87276**	**0.31973**	**2.73**	**0.0126**

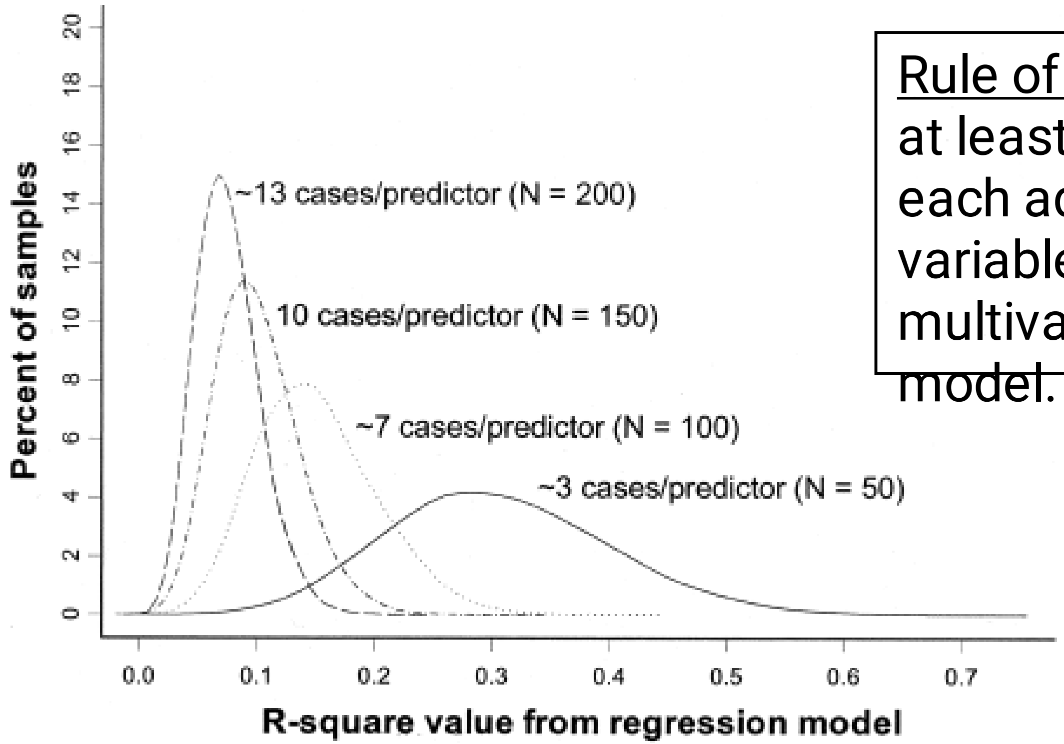
Compare with multivariate result; p<.0001

Love of Math, univariate:

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.70270	1.25302	2.96	0.0076
mathLove	**mathLove**	**1**	**0.59459**	**0.19225**	**3.09**	**0.0055**

Compare with multivariate result; p=.0011

Overfitting



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

Pure noise variables still produce good R^2 values if the model is overfitted. The distribution of R^2 values from a series of simulated regression models containing only noise variables.

(Figure 1 from: Babyak, MA. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine* 66:411-421 (2004).)

Review of statistical tests

The following table gives the appropriate choice of a statistical test or measure of association for various types of data (outcome variables and predictor variables) by study design.

e.g., blood pressure= pounds + age + treatment

Continuous outcome

Continuous predictors

Binary predictor

Types of variables to be analyzed: Predictor variable/s	Outcome variable	Statistical procedure or measure of association
Cross-sectional/case-control studies		
Binary (two groups)	Continuous	T-test
Binary	Ranks/ordinal	Wilcoxon rank-sum test
Categorical (>2 groups)	Continuous	ANOVA
Continuous	Continuous	Simple linear regression
Multivariate (categorical and continuous)	Continuous	Multiple linear regression
Categorical	Categorical	Chi-square test (or Fisher's exact)
Binary	Binary	Odds ratio, risk ratio
Multivariate	Binary	Logistic regression
Cohort Studies/Clinical Trials		
Binary	Binary	Risk ratio
Categorical	Time-to-event	Kaplan-Meier/ log-rank test
Multivariate	Time-to-event	Cox-proportional hazards regression, hazard ratio
Categorical	Continuous	Repeated measures ANOVA
Multivariate	Continuous	Mixed models; GEE modeling

Alternative summary: statistics for various types of outcome data

Outcome Variable	Are the observations independent or correlated? independent	correlated	Assumptions
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Outcome is normally distributed (important for small samples). Outcome and predictor have a linear relationship.
Binary or categorical (e.g. fracture yes/no)	Difference in proportions Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Chi-square test assumes sufficient numbers in each cell (>=5)
Time-to-event (e.g. time to fracture)	Kaplan-Meier statistics Cox regression	n/a	Cox regression assumes proportional hazards between groups

Continuous outcome (means); HRP 259/HRP 262

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated (and small sample size):
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling:** multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

Binary or categorical outcomes (proportions); HRP 259/HRP 261

Outcome Variable	Are the observations correlated?		Alternative to the chi-square test if sparse cells:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	**Chi-square test:** compares proportions between two or more groups **Relative risks:** odds ratios or risk ratios **Logistic regression:** multivariate technique used when outcome is binary; gives multivariate-adjusted odds ratios	**McNemar's chi-square test:** compares binary outcome between correlated groups (e.g., before and after) **Conditional logistic regression:** multivariate regression technique for a binary outcome when groups are correlated (e.g., matched data) **GEE modeling:** multivariate regression technique for a binary outcome when groups are correlated (e.g., repeated measures)	**Fisher's exact test:** compares proportions between independent groups when there are sparse data (some cells <5). **McNemar's exact test:** compares proportions between correlated groups when there are sparse data (some cells <5).

Time-to-event outcome (survival data); HRP 262

Outcome Variable	Are the observation groups independent or correlated?		Modifications to Cox regression if proportional-hazards is violated:
	independent	correlated	
Time-to-event (e.g., time to fracture)	**Kaplan-Meier statistics:** estimates survival functions for each group (usually displayed graphically); compares survival functions with log-rank test **Cox regression:** Multivariate technique for time-to-event data; gives multivariate-adjusted hazard ratios	n/a (already over time)	Time-dependent predictors or time-dependent hazard ratios (tricky!)