Simple Linear Regression

Correlation indicates the magnitude and direction of the linear relationship between two variables.

Linear Regression: variable Y (criterion) is predicted by variable X (predictor) using a linear equation.

Advantages:

Scores on X allow prediction of scores on Y.

Allows for multiple predictors (continuous and categorical) so you can control for variables.

Linear Regression Equation

Geometry equation for a line:

$$y = mx + b$$

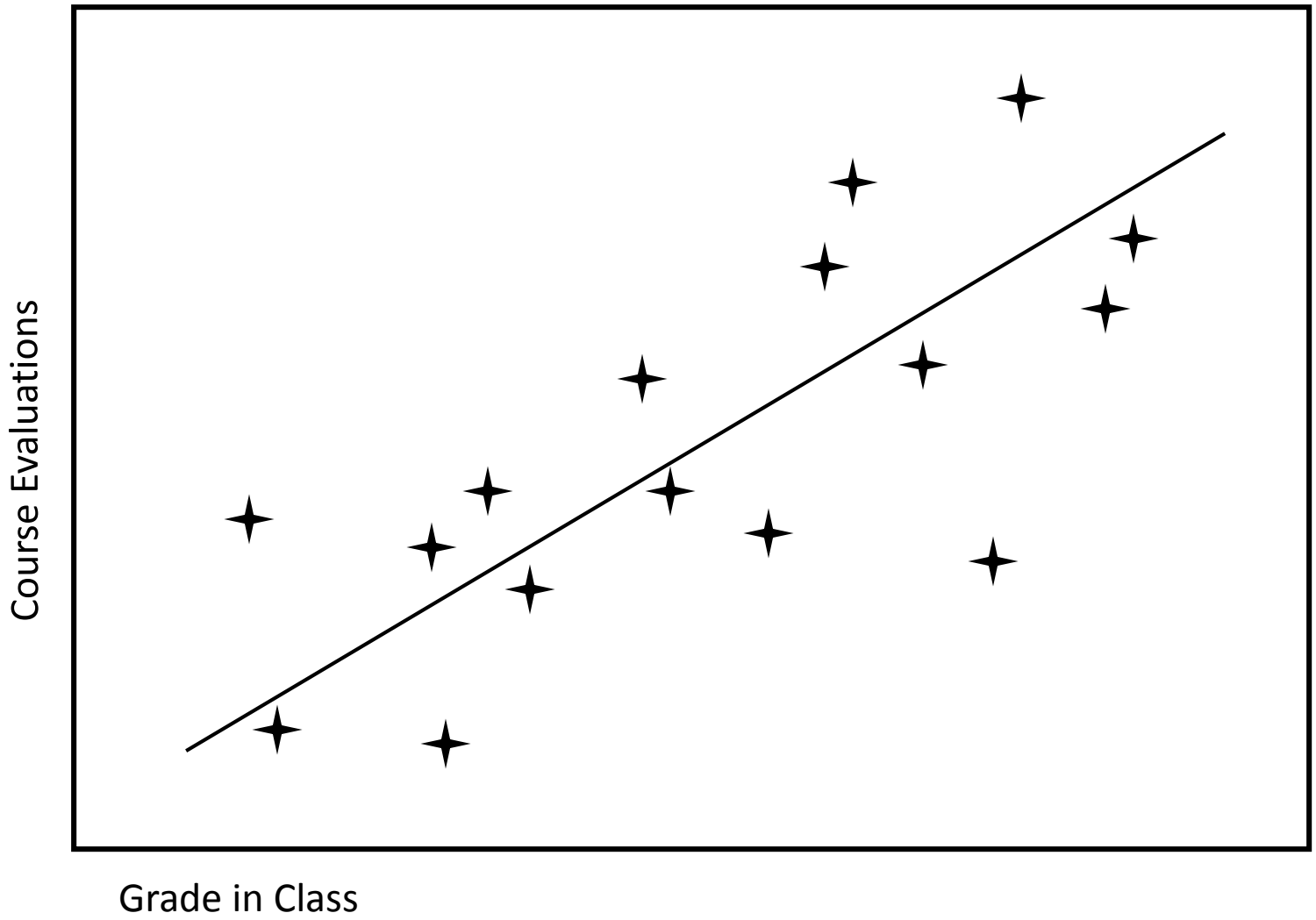
Regression equation for a line (population):

$$y = \beta_0 + \beta_1 x$$

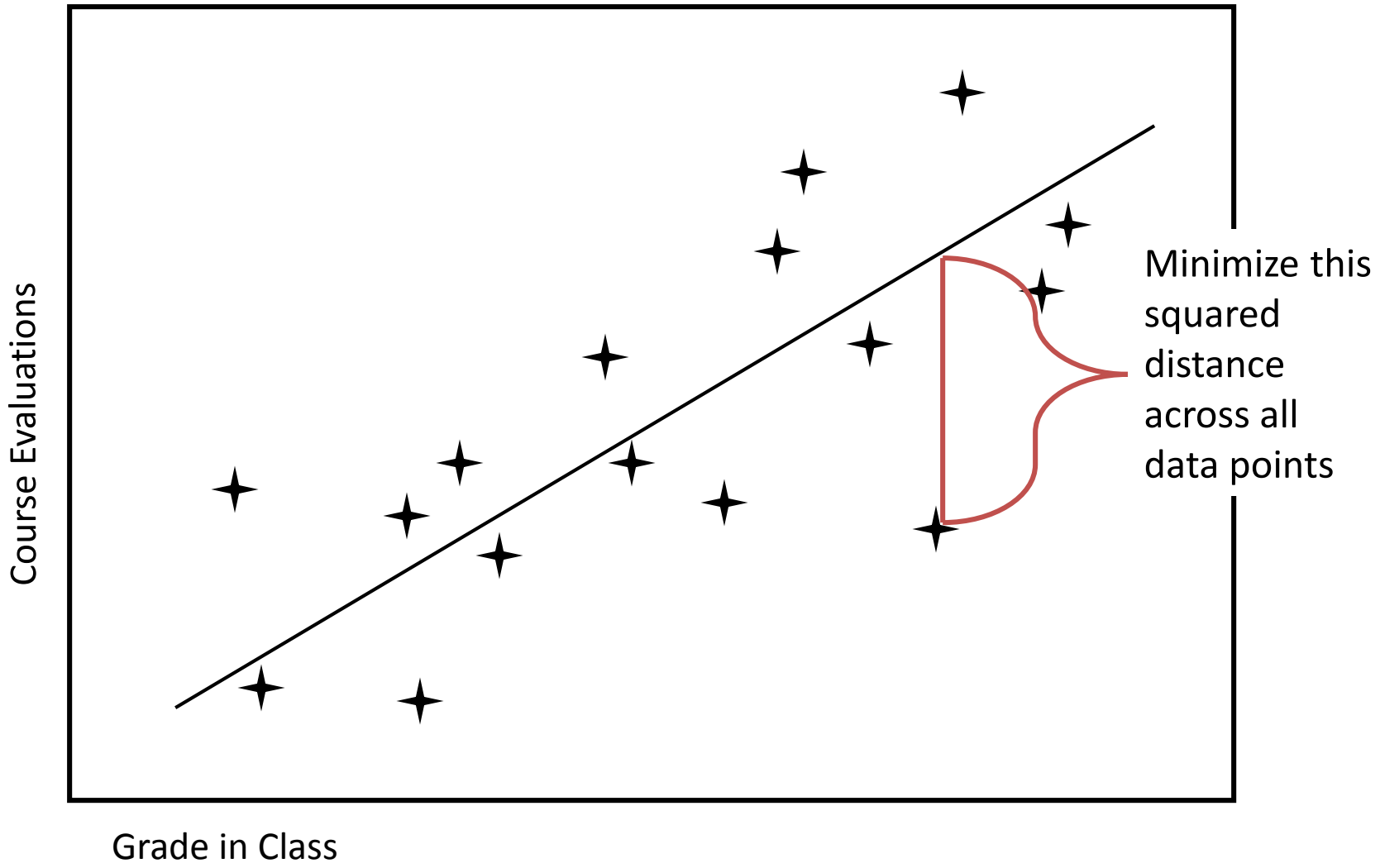
β_0 : point where the line intercepts y-axis

β_1 : slope of the line

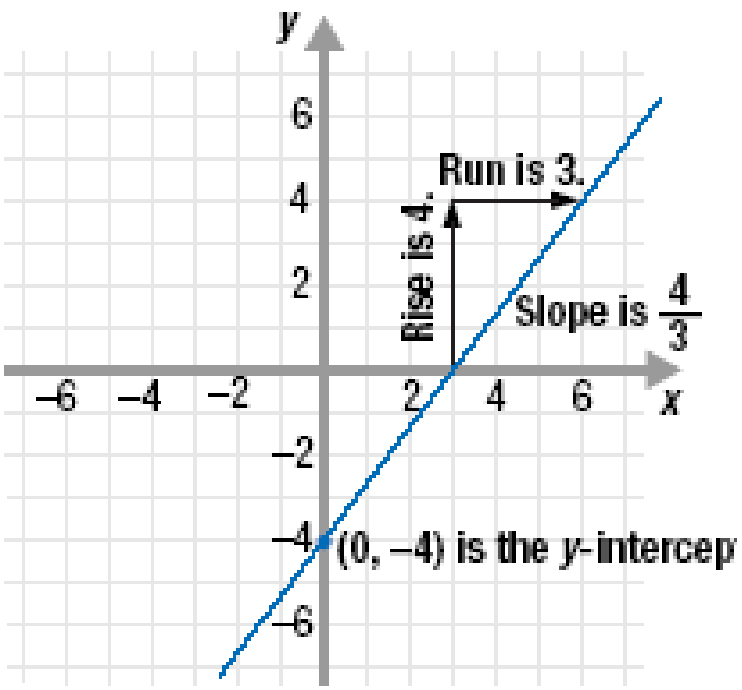
Regression: Finding the Best-Fitting Line

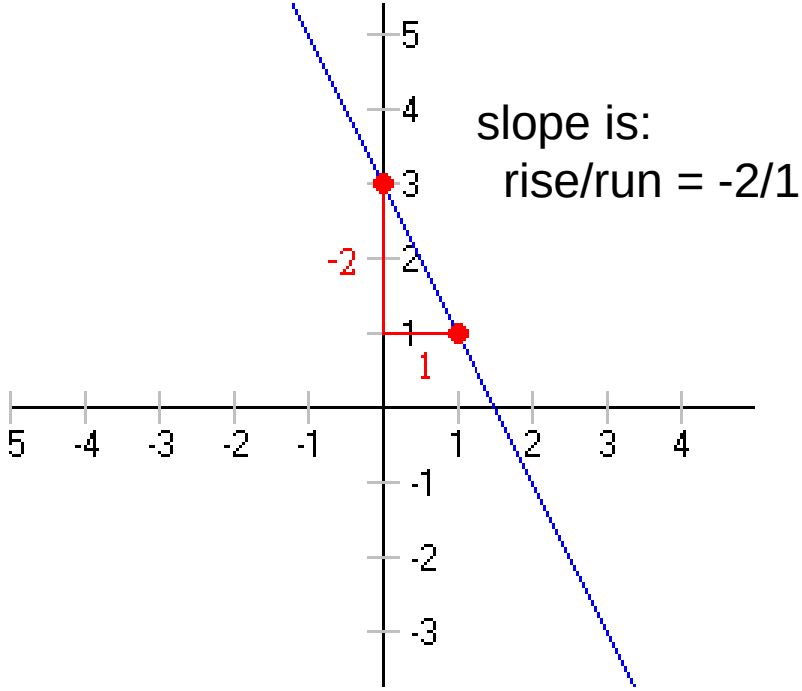


Best-Fitting Line



Slope and Intercept in Scatterplots





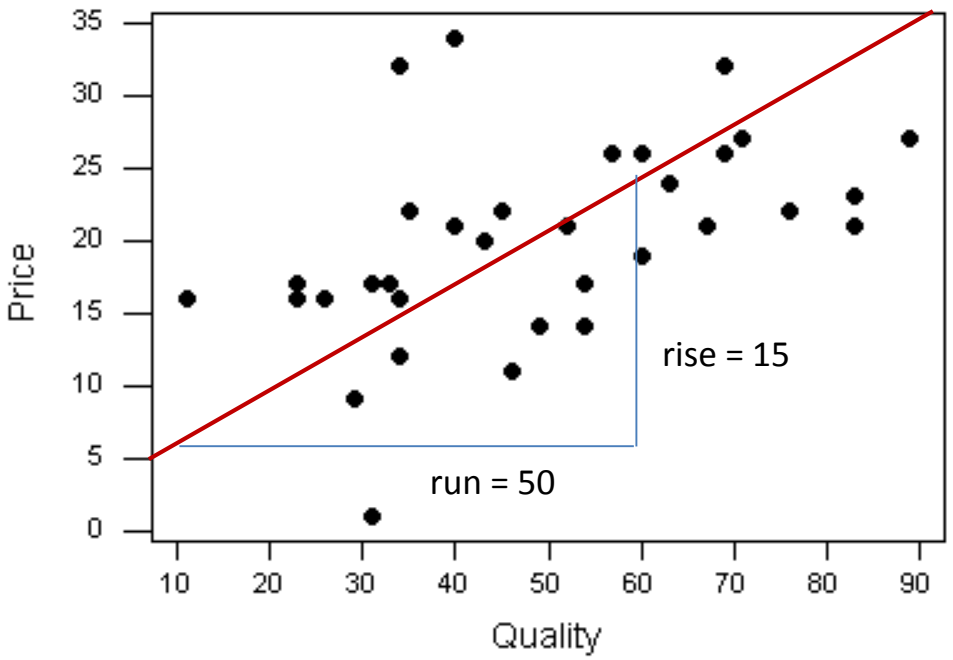
$y = b_0 + b_1x + e$

$y = -4 + 1.33x + e$

$y = b_0 + b_1x + e$

$y = 3 - 2x + e$

Estimating Equation from Scatterplot



$y = b_0 + b_1x + e$

slope = 15/50 = .3

y = 5 + .3x + e

Predict price at quality = 90

y = 5 + .3x + e

y = 5 + .3*90 = 35

Example Van Camp, Barden & Sloan (2010)

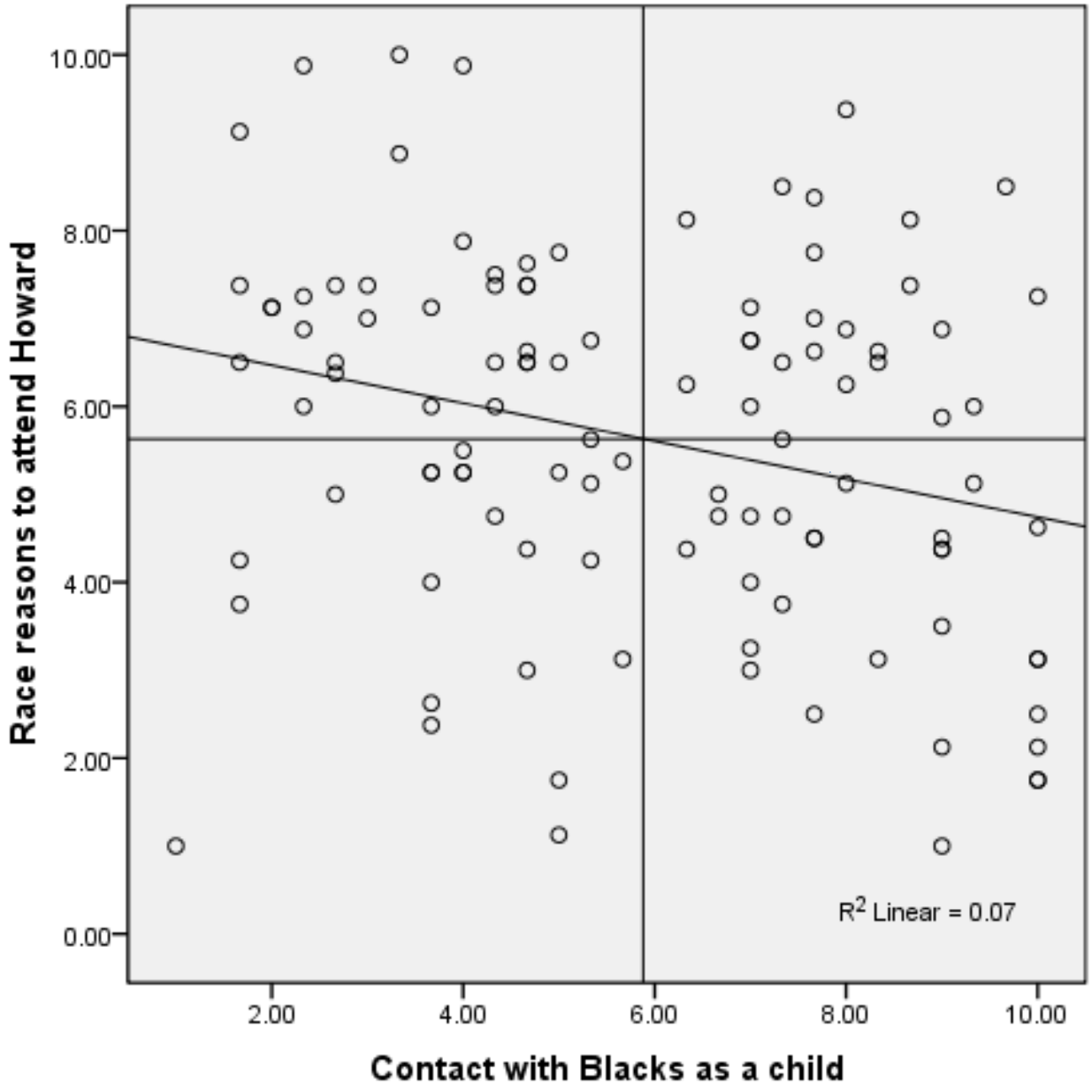
Contact with Blacks Scale:

Ex: “What percentage of your neighborhood growing up was Black?” 0%-100%

Race Related Reasons for College Choice:

Ex: “To what extent did you come to Howard specifically because the student body is predominantly Black?”
1(not very much) – 10 (very much)

Your predictions, how would prior contact predict race related reasons?


Race reasons to attend Howard
10.00
8.00
6.00
4.00
2.00
0.00
2.00
4.00
6.00
8.00
10.00
Contact with Blacks as a child
R2 Linear = 0.07

Results Van Camp, Barden & Sloan (2010)

Regression equation (sample):

$y = b_0 + b_1x + e$

Contact(x) predict Reasons:

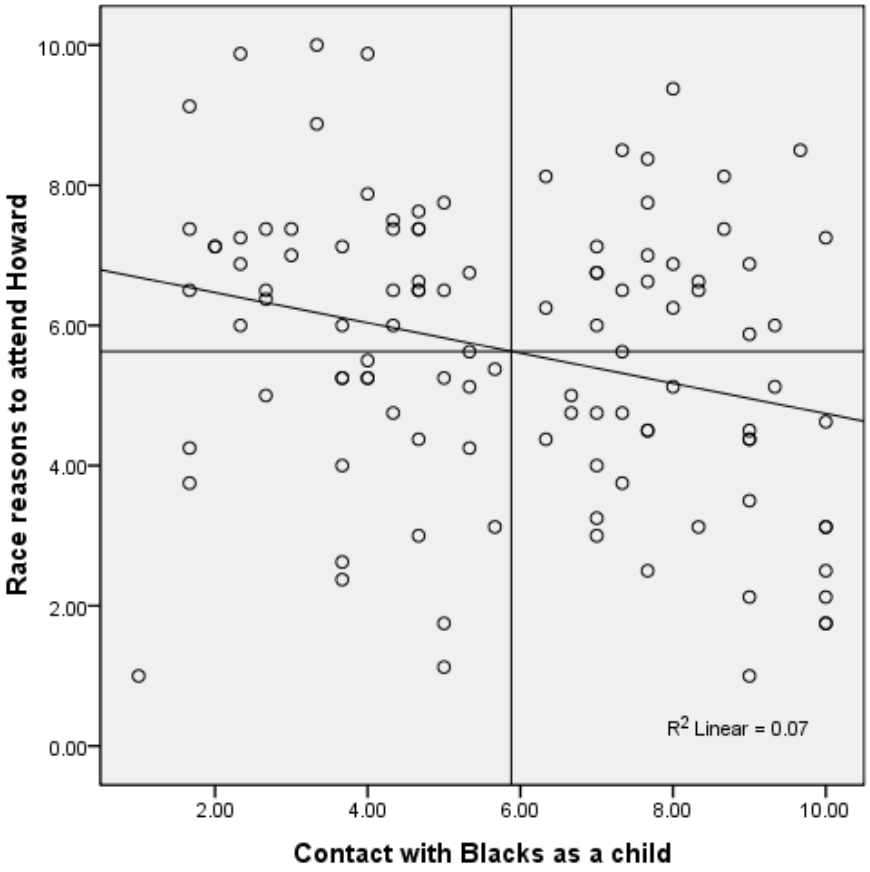
$y = 6.926 - .223x + e$

b_0: $t(107) = 14.17$, $p < .01$

b_1: $t(107) = -2.93$, $p < .01$

$df = N - k - 1 = 109 - 1 - 1$

k: predictors entered



Unstandardized and Standardized *b*

unstandardized *b*: in the original units of X and Y

- tells us how much a change in X will produce a change in Y in the original units (meters, scale points…)
- not possible to compare relative impact of multiple predictors

standardized *b*: scores 1st standardized to SD units

- +1 *SD* change in X produces *b*SD* change in Y
- indicates relative importance of multiple predictors of Y

Results Van Camp, Barden & Sloan (2010)

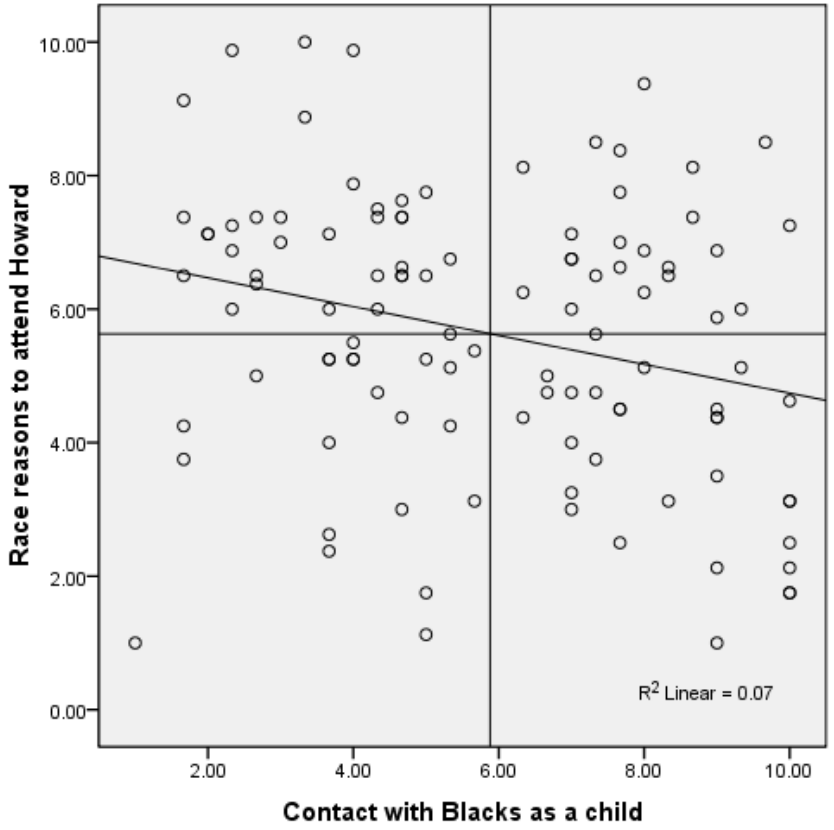
Contact predicts Reasons:

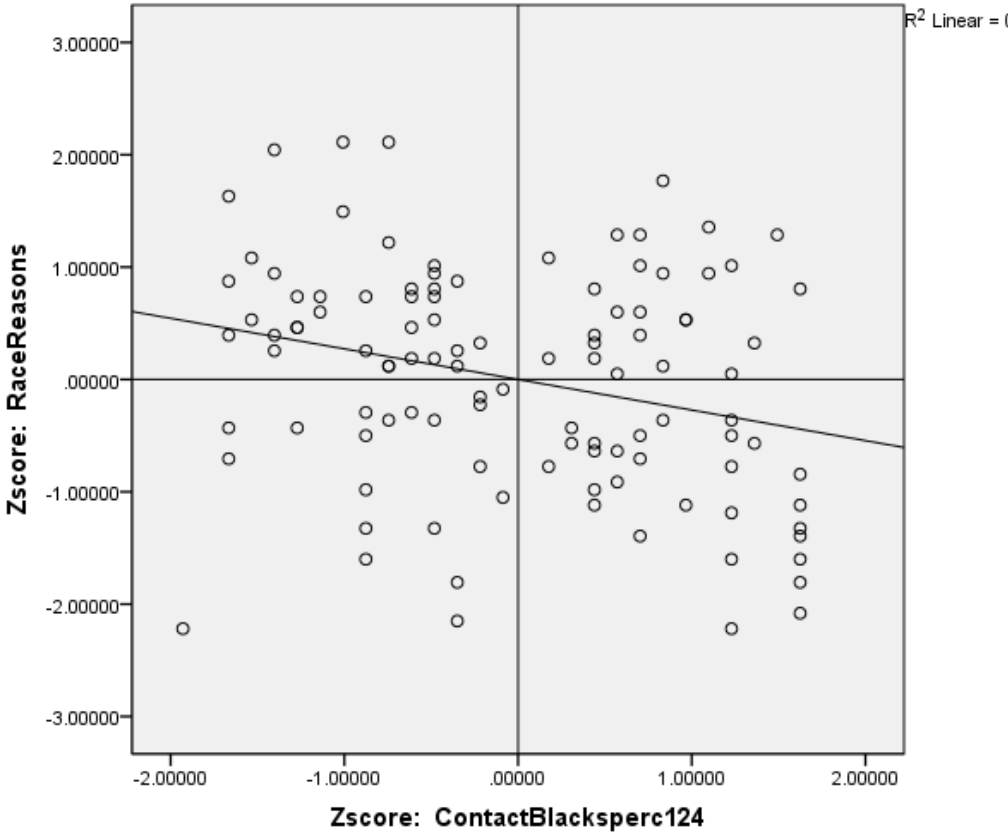
Unstandardized: $y = 6.926 - .223x + e$

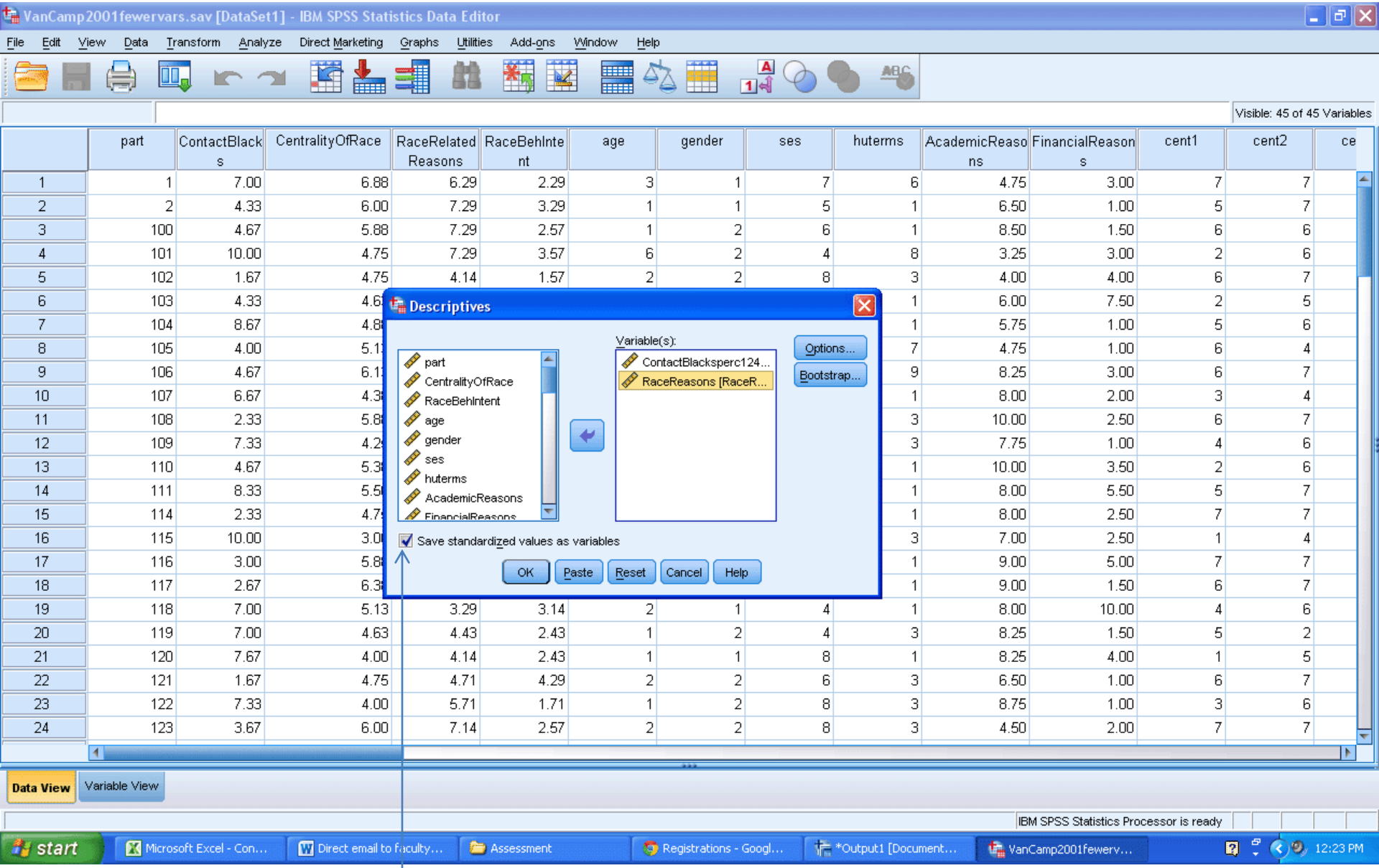
($M_x = 5.89$, $SD_x = 2.53$; $M_y = 5.61$, $SD_y = 2.08$)

Standardized: $y = 0 - .272x + e$

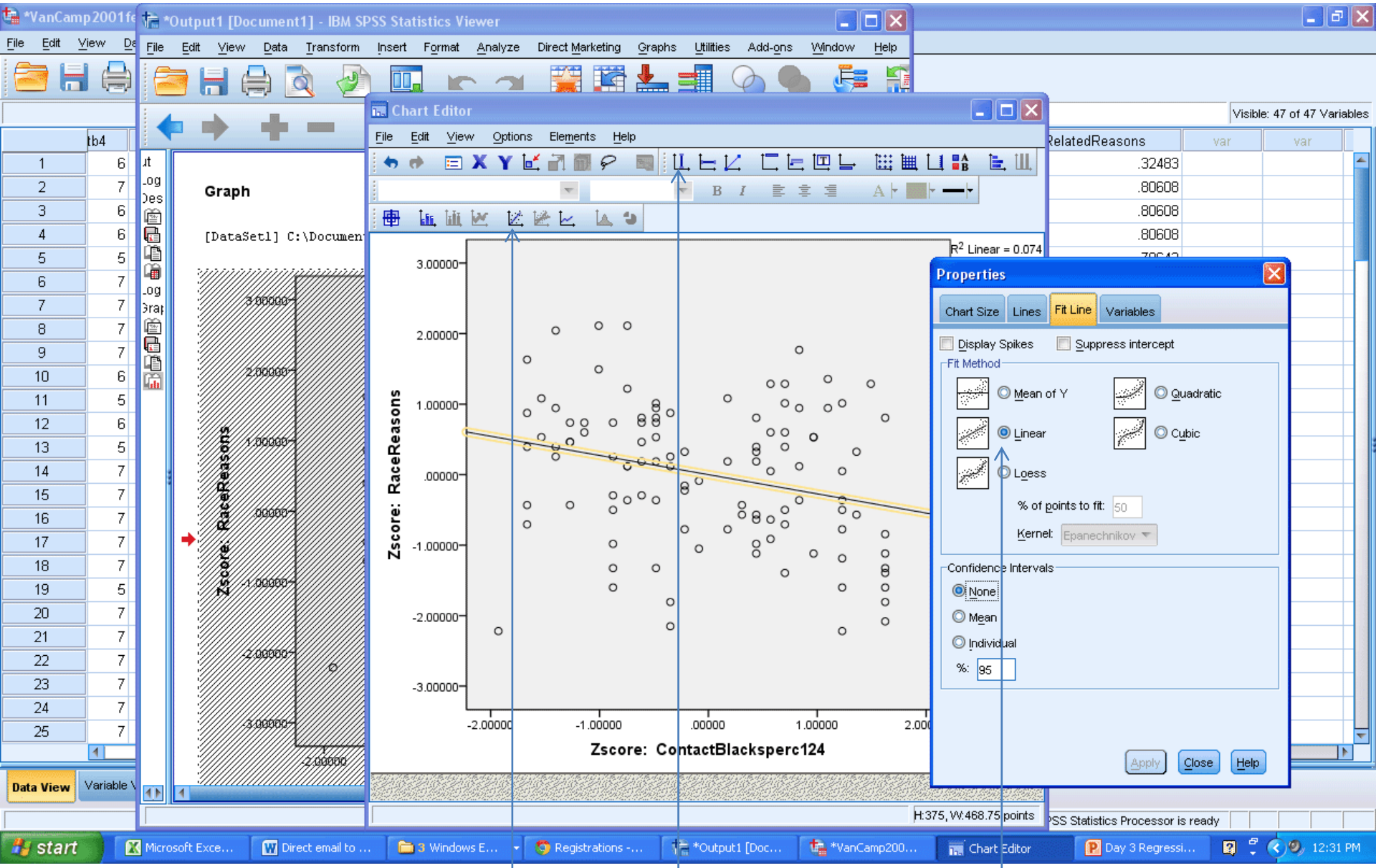
($M_x = 0$, $SD_x = 1.00$; $M_y = 0$, $SD_y = 1.00$)







save new variables that are standardized versions of current variables



add fit lines

add reference lines
(may need to adjust to mean)

select fit line

Predicting Y from X

Once we have a straight line we can know what the change in Y is with each change in X

Y prime (Y') is the prediction of Y at a given X, and it is the average Y score at that X score.

Warning: Predictions can only be made:

(1) within the range of the sample

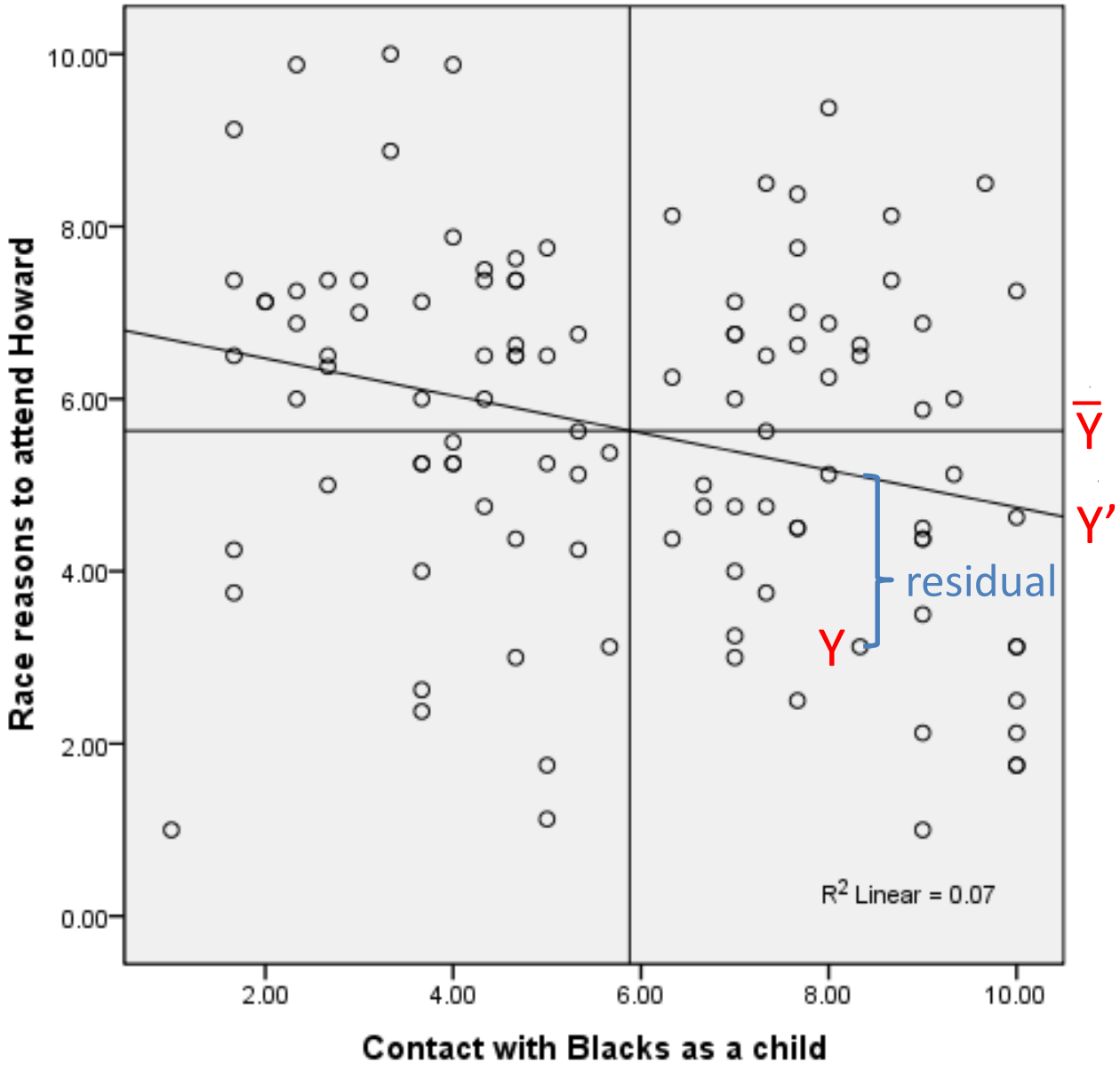
(2) for individuals taken from a similar population under similar circumstances.

Errors around the regression line

Regression equation give us the straight line that minimizes the error involved in making predictions (least squares regression line).

Residual: difference between an actual Y value and predicted (Y') value: Y – Y'

- It is the amount of the original value that is left over after the prediction is subtracted out
- The amount of error above and below the line is the same


Race reasons to attend Howard
10.00
8.00
6.00
4.00
2.00
0.00
2.00
4.00
6.00
8.00
10.00
Contact with Blacks as a child
$\bar{Y}$
Y'
residual
Y
R^2 Linear = 0.07

Dividing up Variance

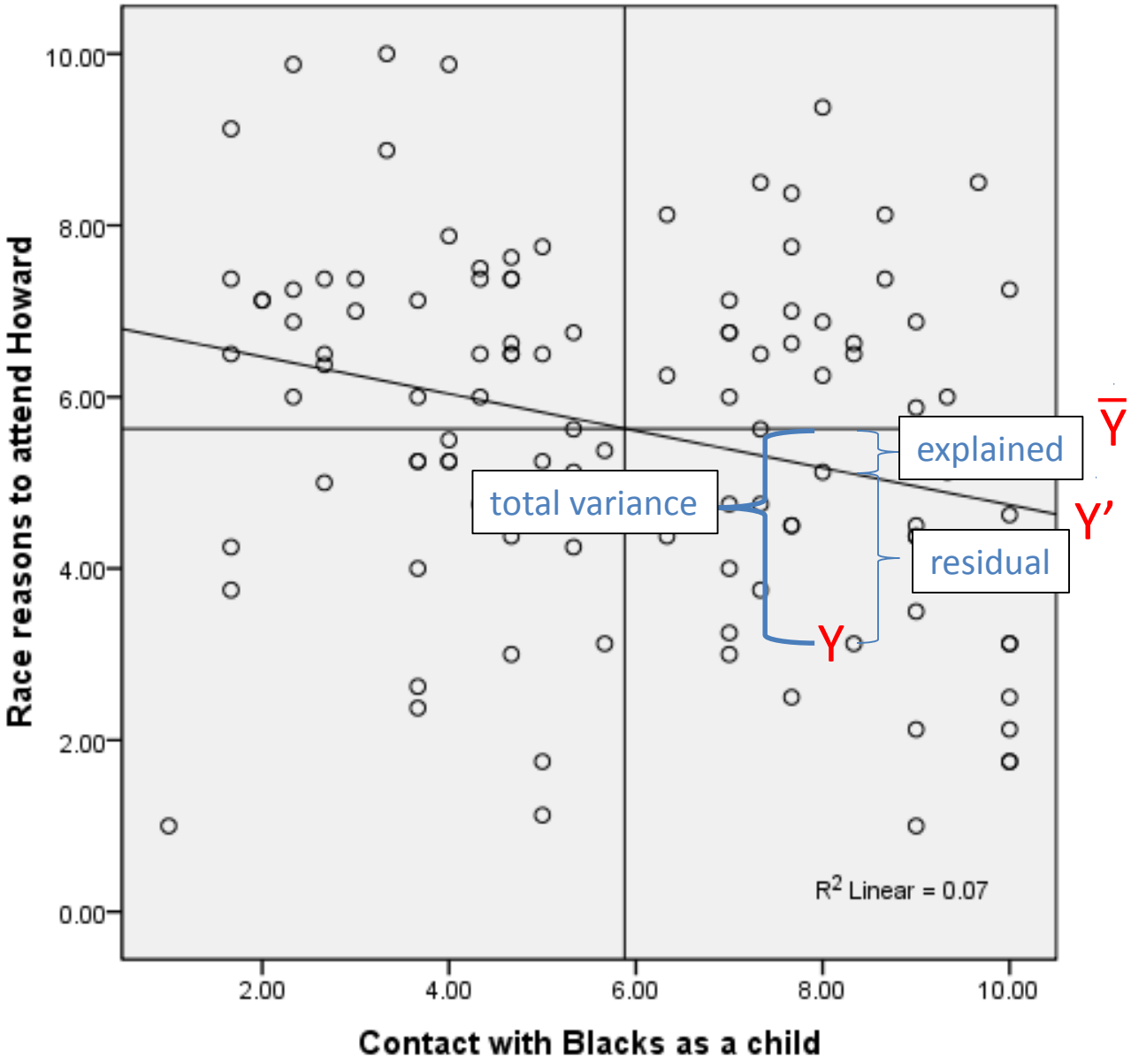
Total: deviation of individual data points from the sample mean

Explained: deviation of the regression line from the mean

Unexplained: deviation of individual data points from the regression line (error in prediction)

$$\left(Y - Y'\right) + \left(Y' - \overline{Y}\right) = \left(Y - \overline{Y}\right)$$

unexplained variance (residual)	explained variance	total variance



$$(Y - Y') + (Y' - \bar{Y}) = (Y - \bar{Y})$$

unexplained variance (residual) | explained variance | total variance

Coefficient of determination: proportion of the total variance that is explained by the predictor variable

$$R^2 = \frac{\text{explained variance}}{\text{total variance}}$$

SPSS - regression

Analyze → regression → linear

Select criterion variable (Y) [Racereas]

[SPSS calls DV]

Select predictor variable (X) [ContactBlacks]

[SPSS calls IV]

OK

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.272[a]	.074	.066	2.00872

a. Predictors: (Constant), ContactBlacksperc124

b. Dependent Variable: RaceReasons

coefficient of determination

SS_{error}: minimized in OLS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.582	1	34.582	8.571	.004[a]
	Residual	431.739	107	4.035		
	Total	466.321	108			

a. Predictors: (Constant), ContactBlacksperc124

b. Dependent Variable: RaceReasons

Reporting in Results:

$b = -.27$, $t(107) = -2.93$, $p < .01$.

(pp. 240 in Van Camp 2010)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6.926	.489		14.172	.000
	ContactBlacksperc124	-.223	.076	-.272	-2.928	.004

a. Dependent Variable: RaceReasons

Unstandardized: $y = 6.926 - .223x + e$

Standardized: $y = 0 - .272x + e$

Assumptions Underlying Linear Regression

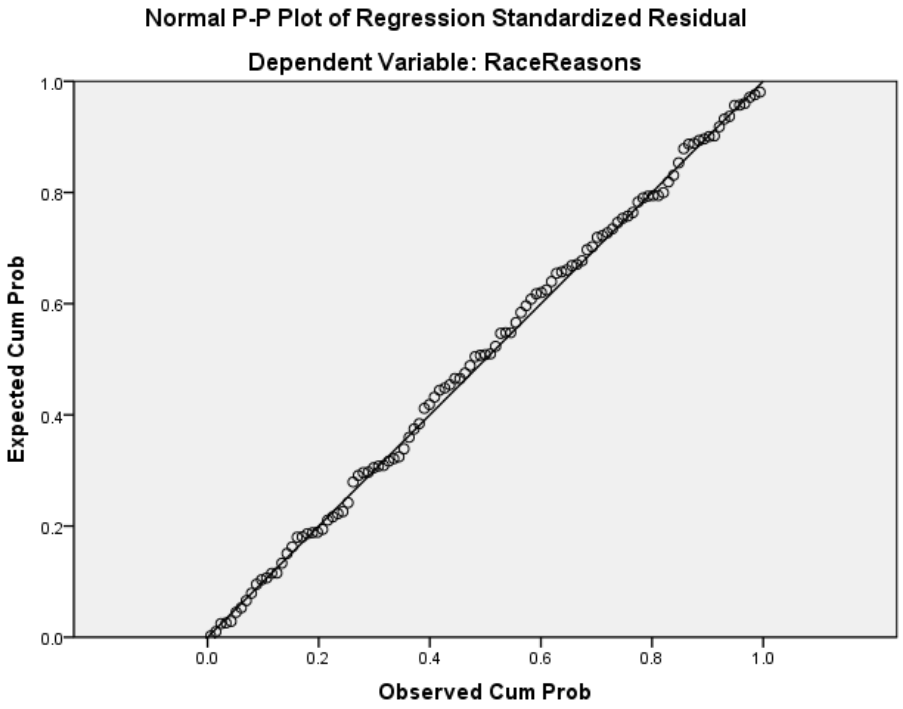
1. Independent random sampling
2. Normal distribution
3. Linear relationships (not curvilinear)
4. Homoscedasticity of errors (homogeneity)

Best way to check 2-4? Diagnostic Plots.

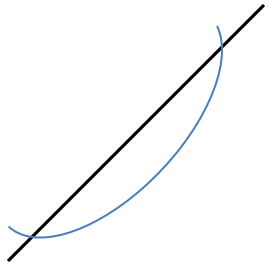
Test for Normality

Solution?

Normal Distribution:

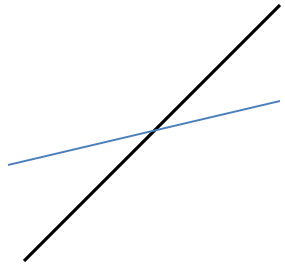


Right (positive) Skew



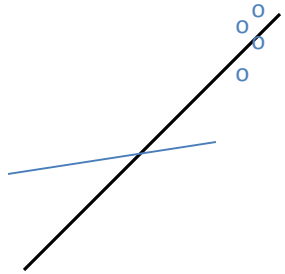
transform data

Narrow Distribution



not serious

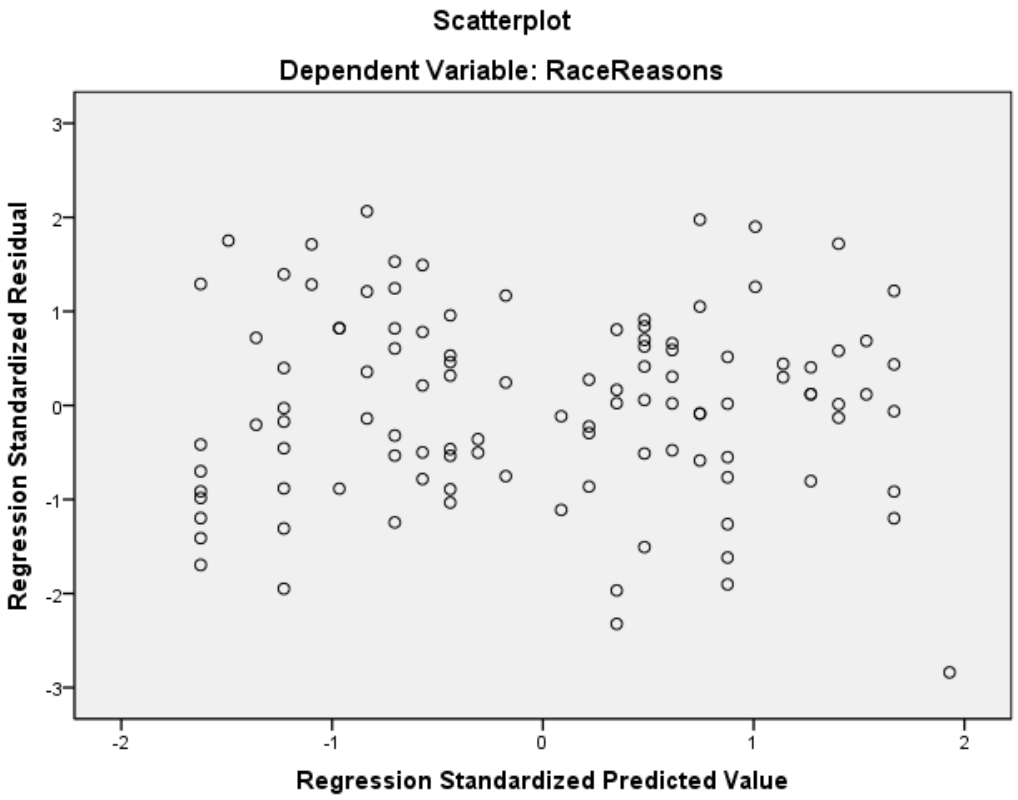
Positive Outliers



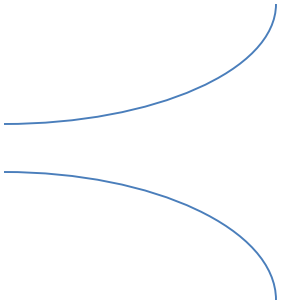
investigate further

Homoscedastic? Linear Appropriate?

Homoscedastic residual errors
& Linear relationship

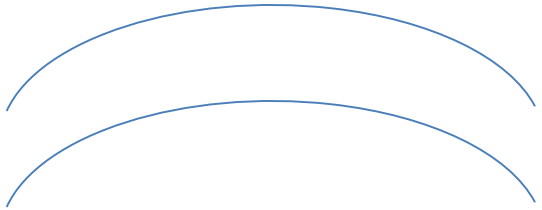


Heteroscedasticity (of residual errors)



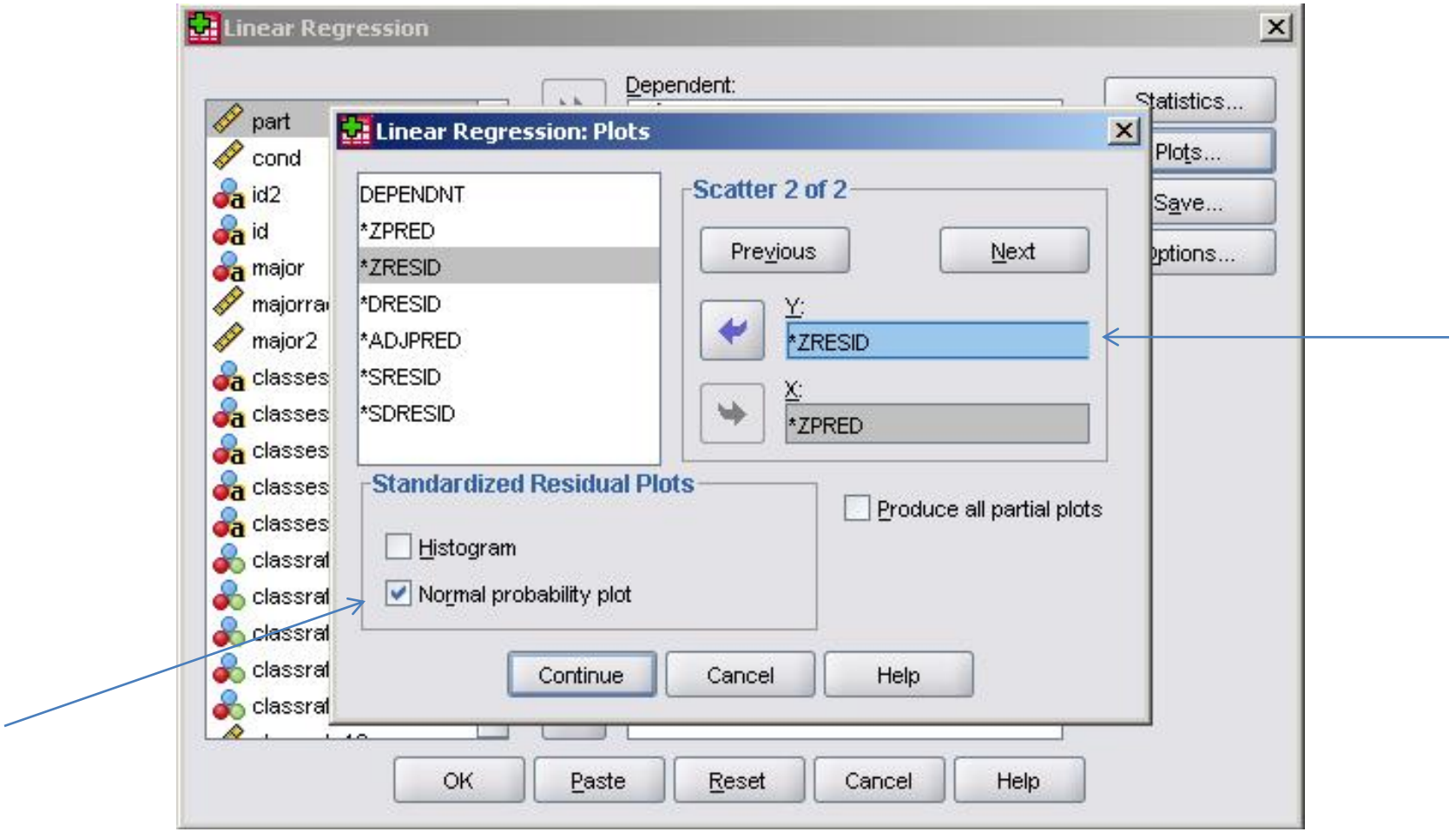
Solution: transform data or
weighted least squares (WLS)

Curvilinear relationship



Solution: add x^2 as predictor
(linear regression not appropriate)

SPSS—Diagnostic Graphs



END