

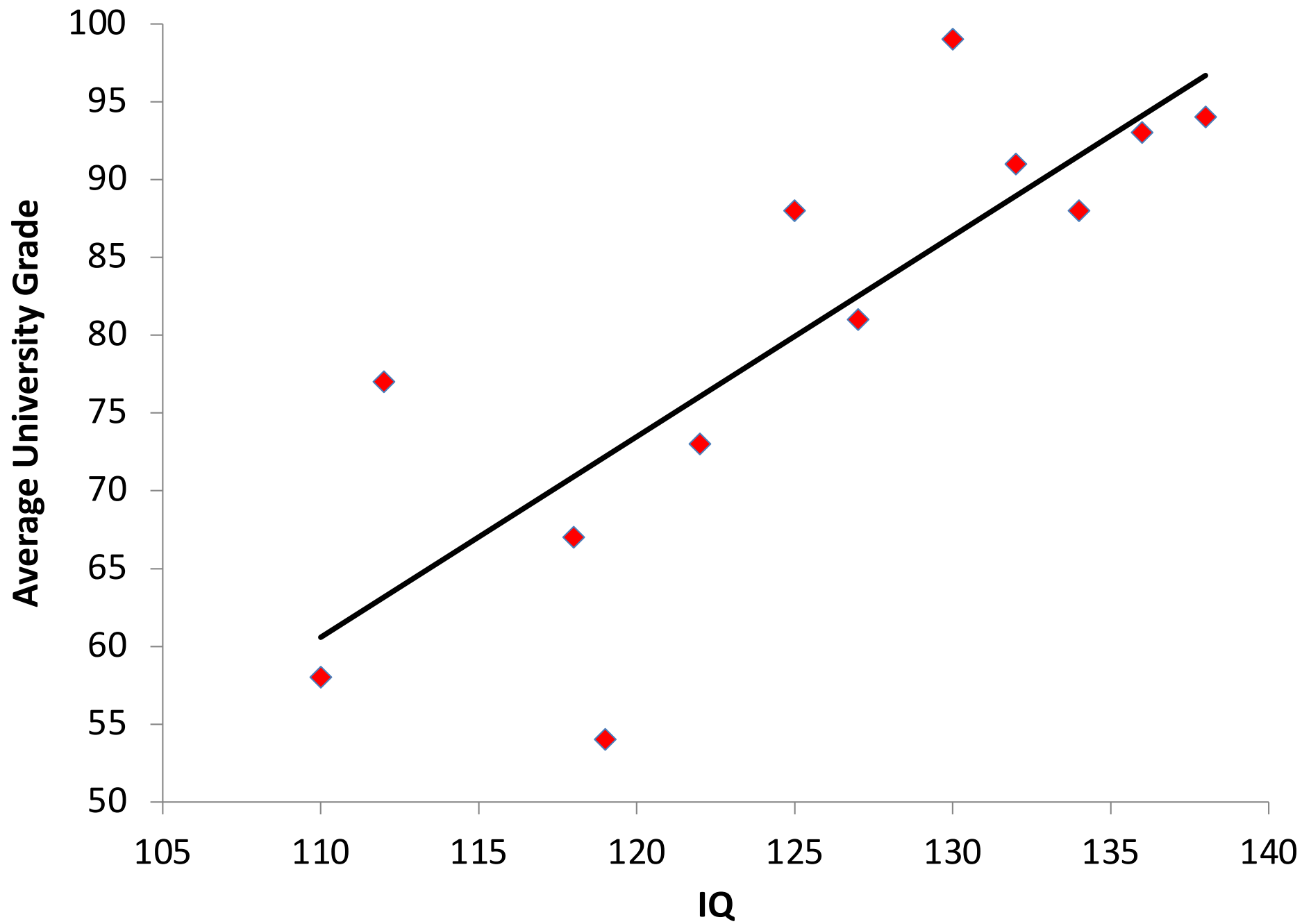
EPHE 357

Introduction to Research

Correlation and Regression

Olav Krigolson, PhD
krigolson@uvic.ca

Relationships



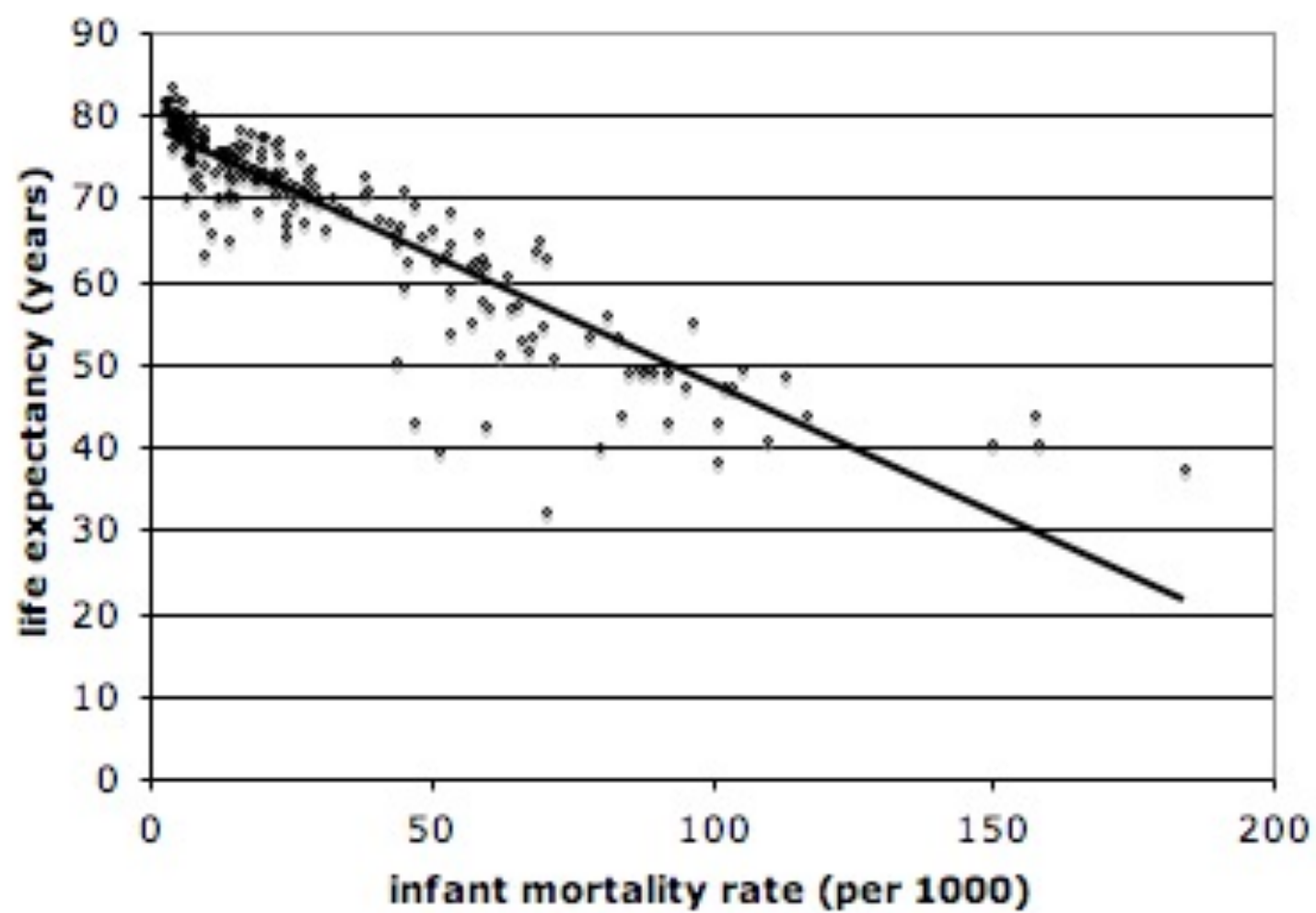
Finding the equation of the line

$$y = mx + b$$

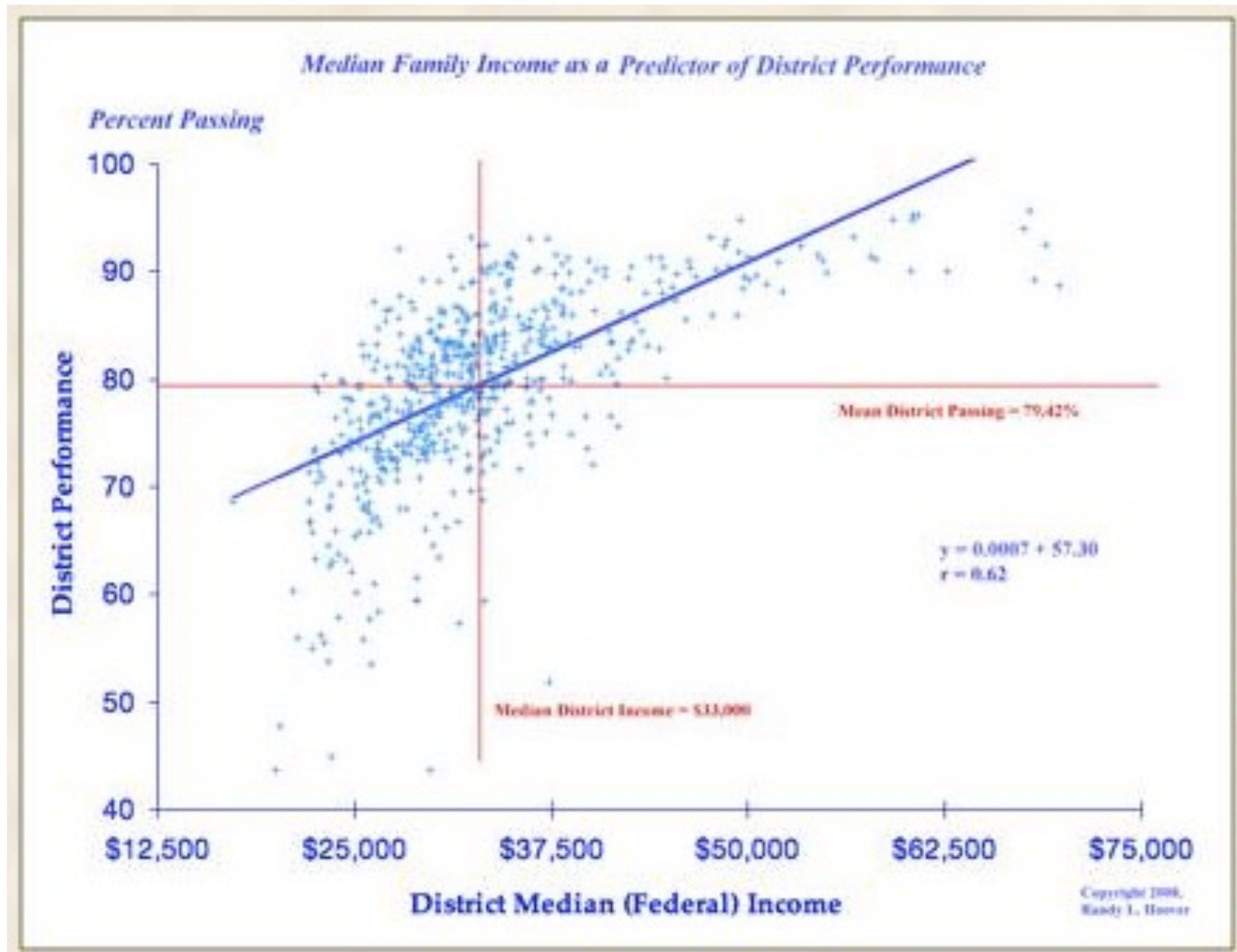
Where:

m = slope

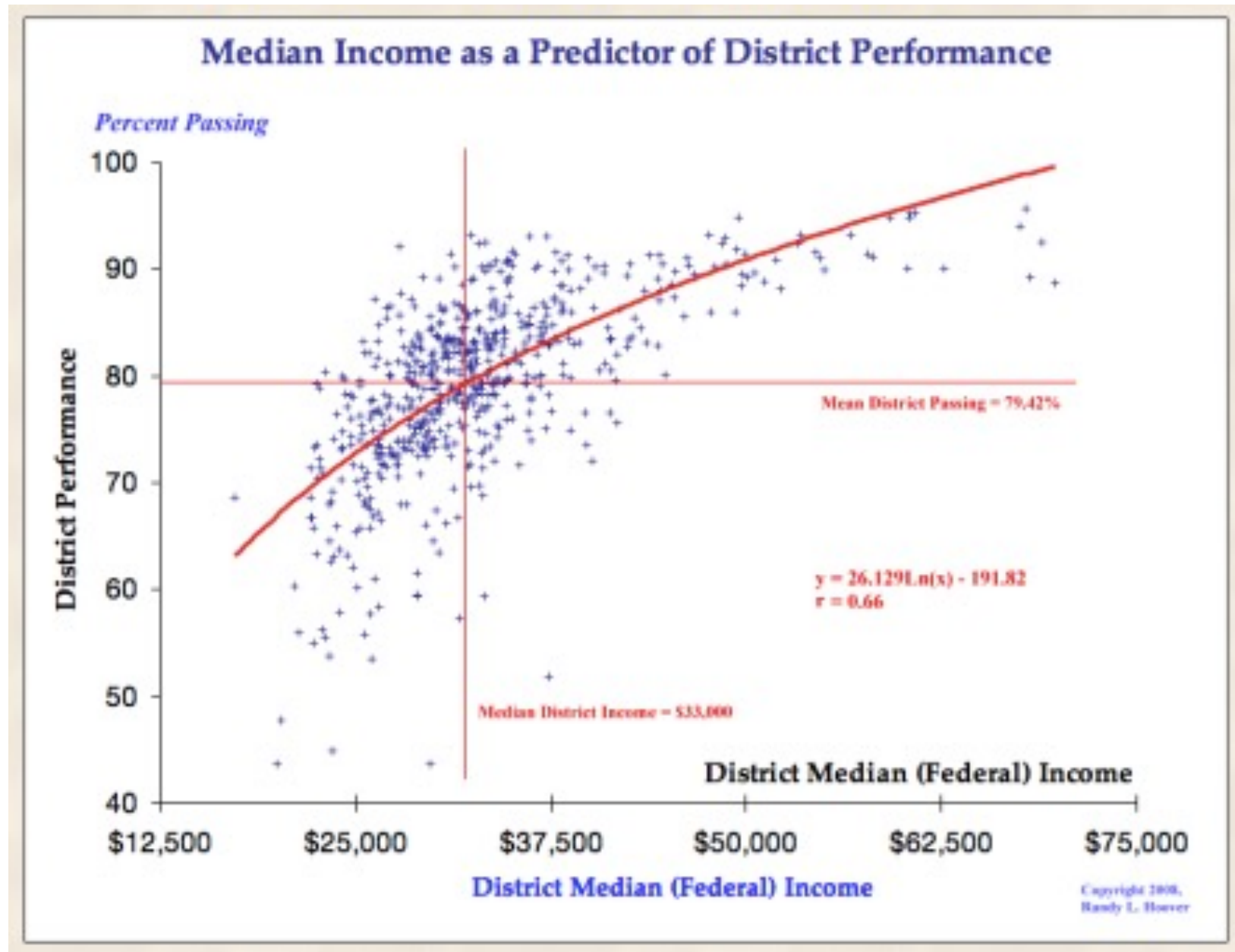
b = y intercept



Curvilinear Correlation



Curvilinear Correlation



Linear Correlation

Examines the relationship between two variables

Defined by “ r ”, the Correlation Coefficient

What is “ r ”: it’s a quantitative measure of the magnitude and direction of relationship

r is always a number between +1 and -1 with 0 implying no relationship whatsoever

Most common form, Pearson “ r ”

But what is “r”

COVARIANCE

But what is COVARIANCE really...

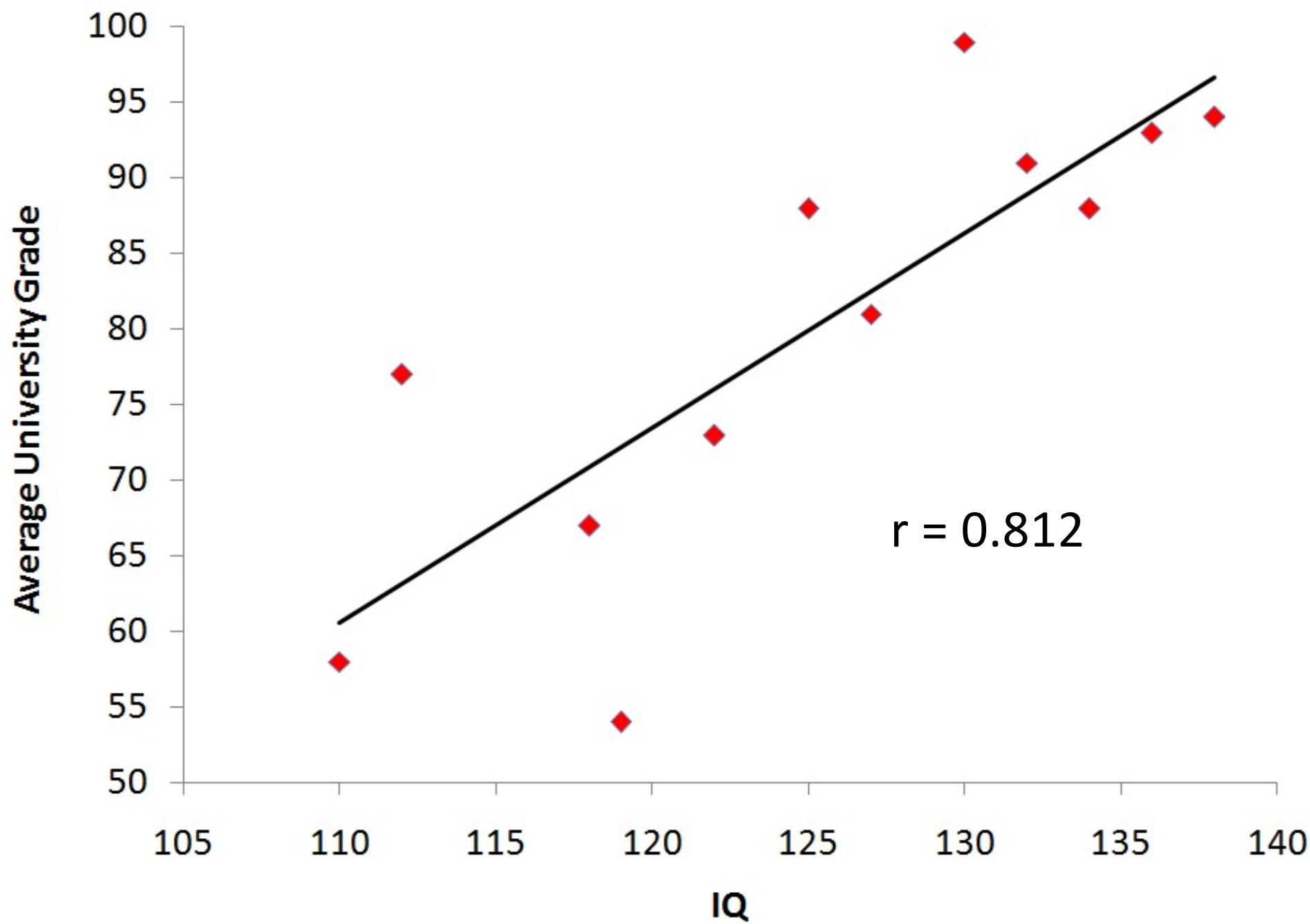
$$\text{cov}(x,y) = \frac{\sum (x - x_m)(y - y_m)}{(N - 1)}$$

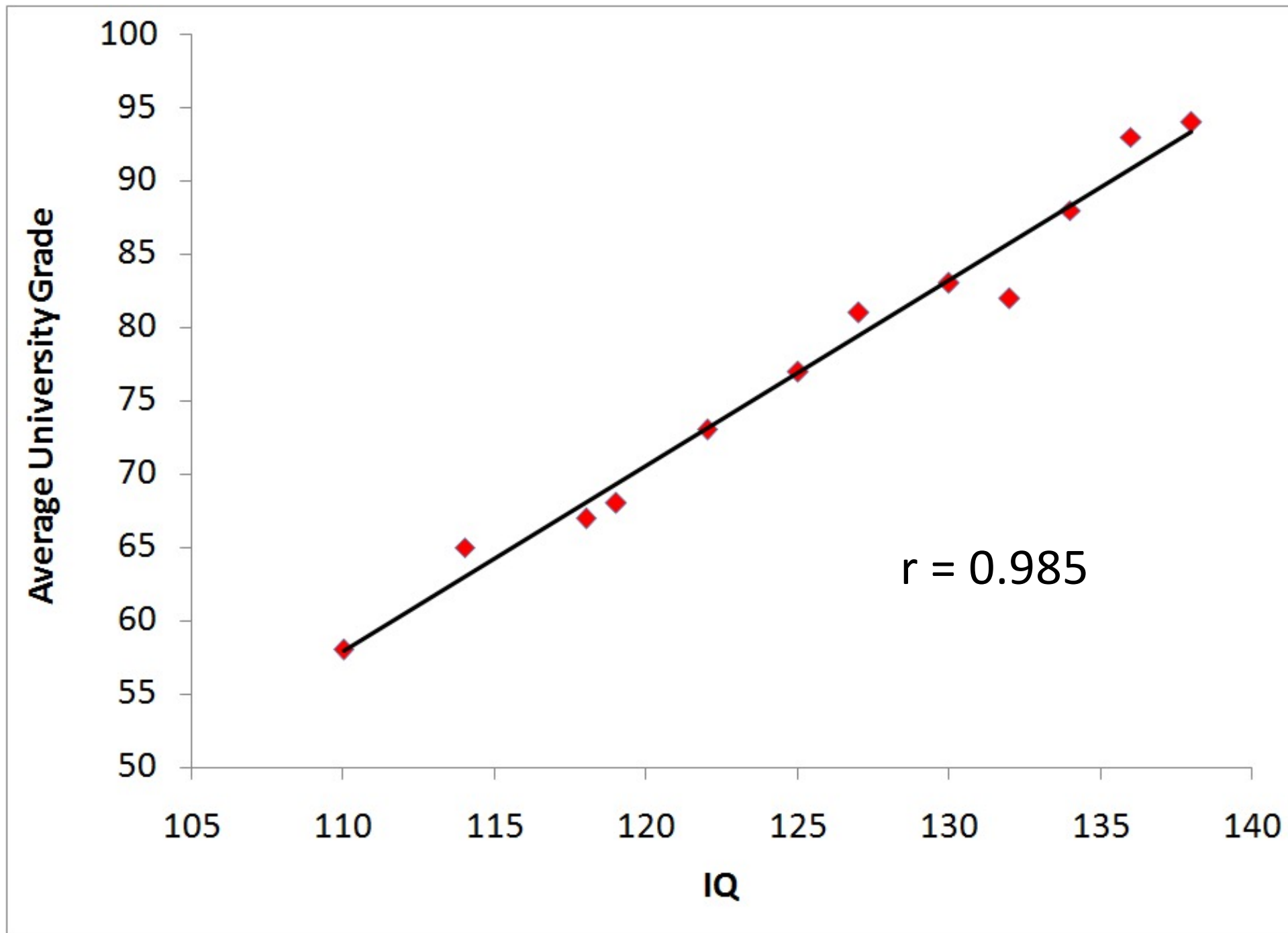
Covariance

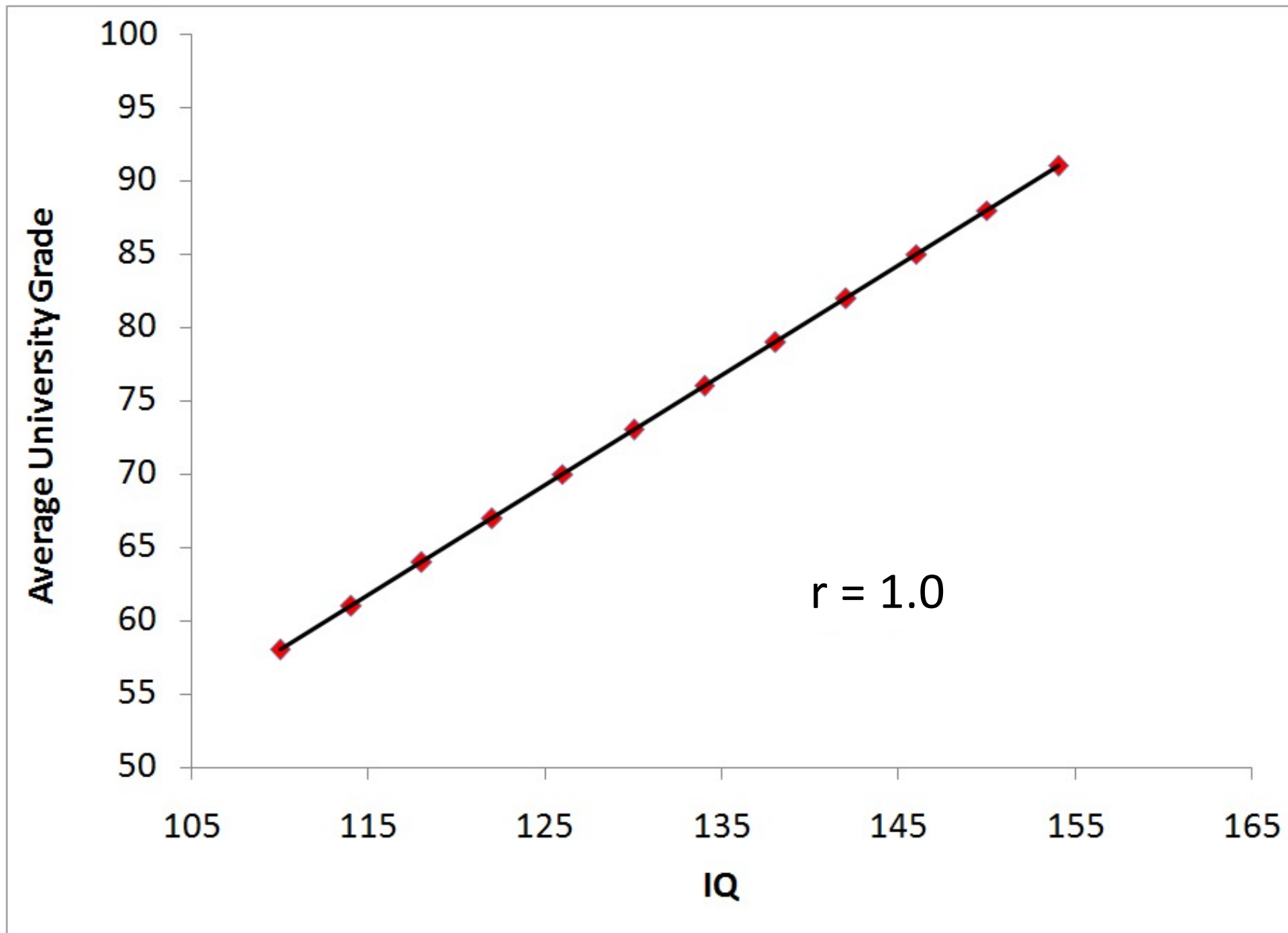
Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated.

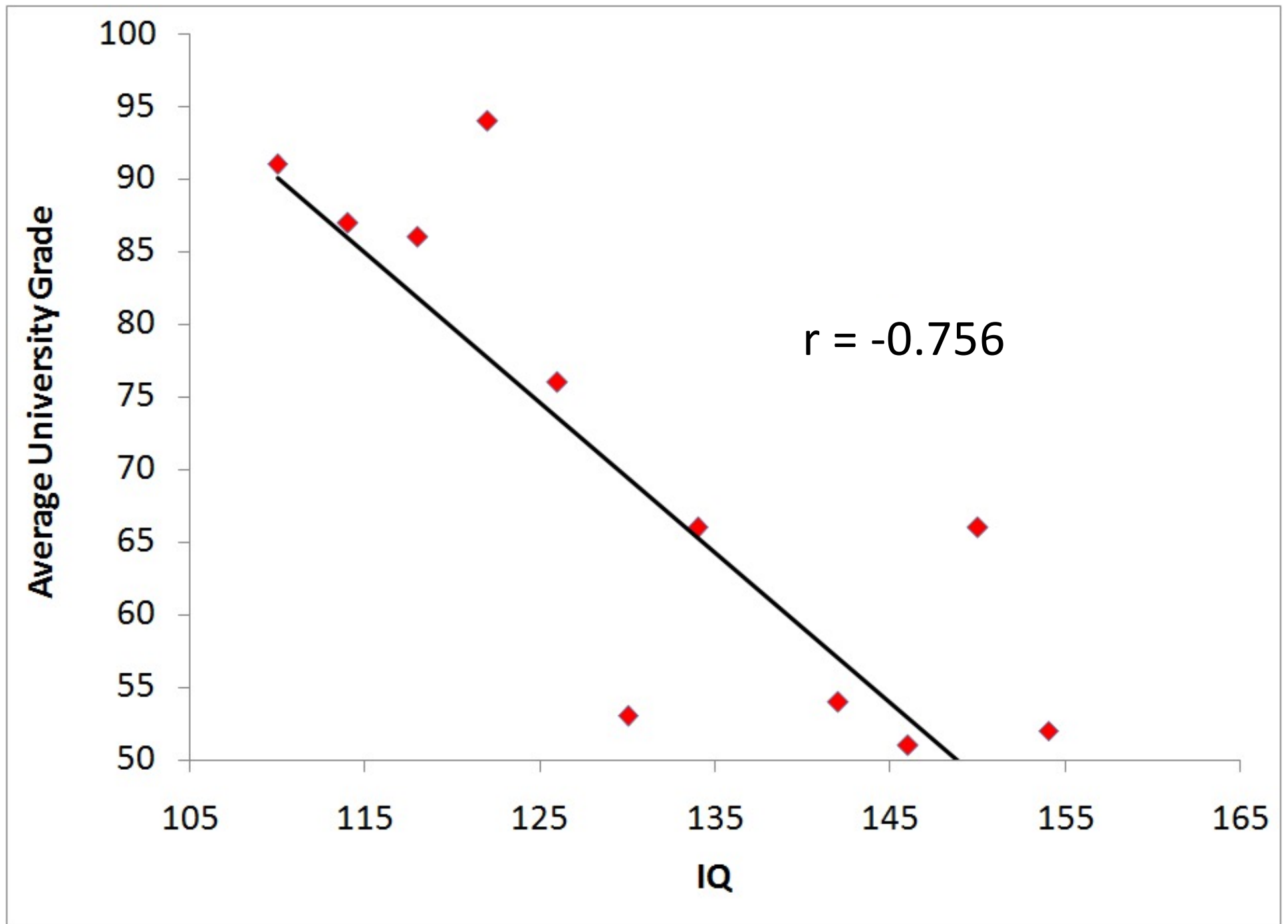
But what is “r”

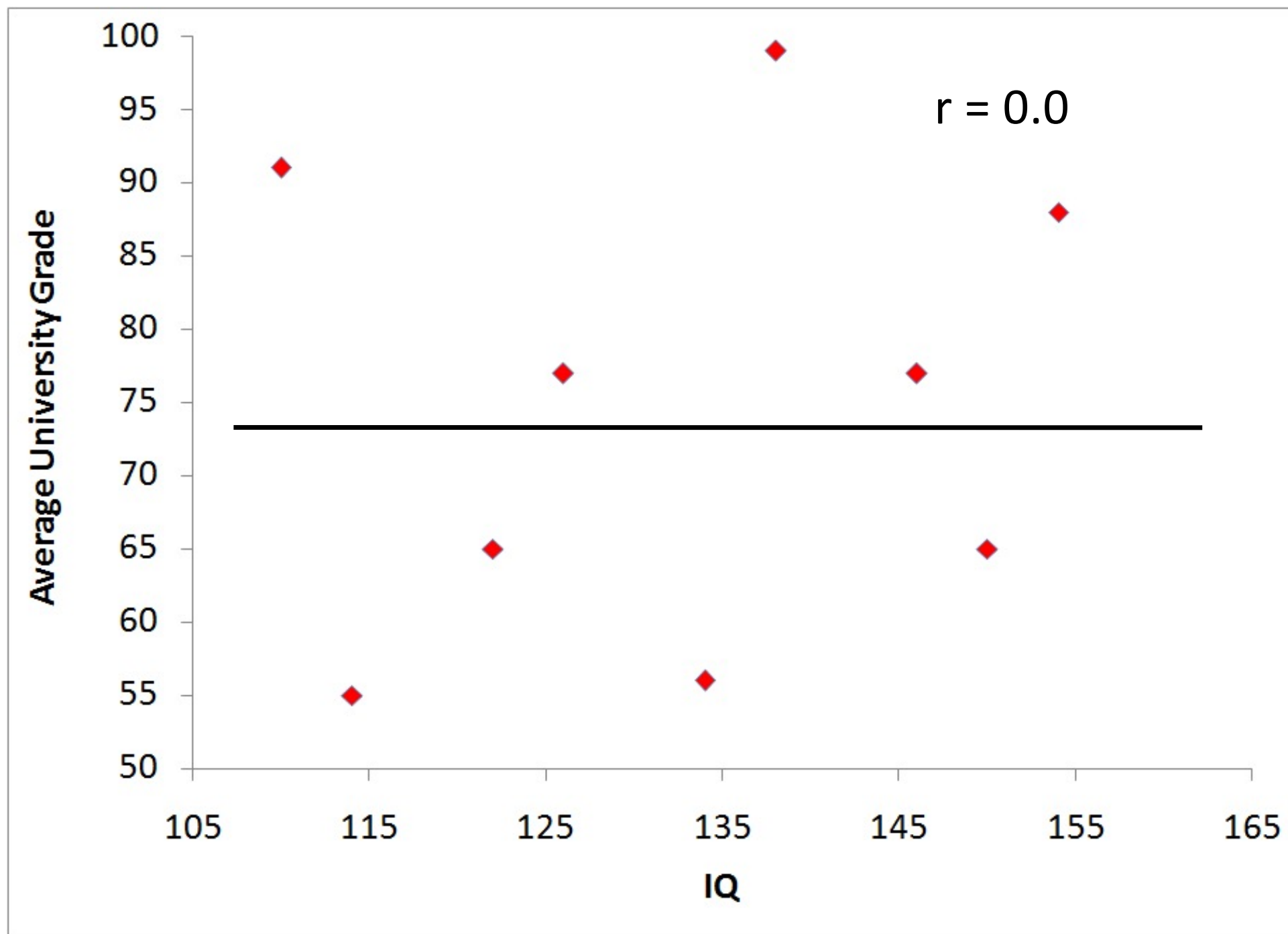
$$r = \frac{\text{cov}(x,y)}{s_x s_y}$$











Correlation \neq Causation

The Three Directions of Causality

- 1) X causes Y
- 2) Y causes X
- 3) Z causes X and Y

Issues in Interpreting the Correlation Coefficient

1. Statistical Significance

Calculating a t value from r

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Use the t distribution

$$df = N - 2$$

Issues in Interpreting the Correlation Coefficient

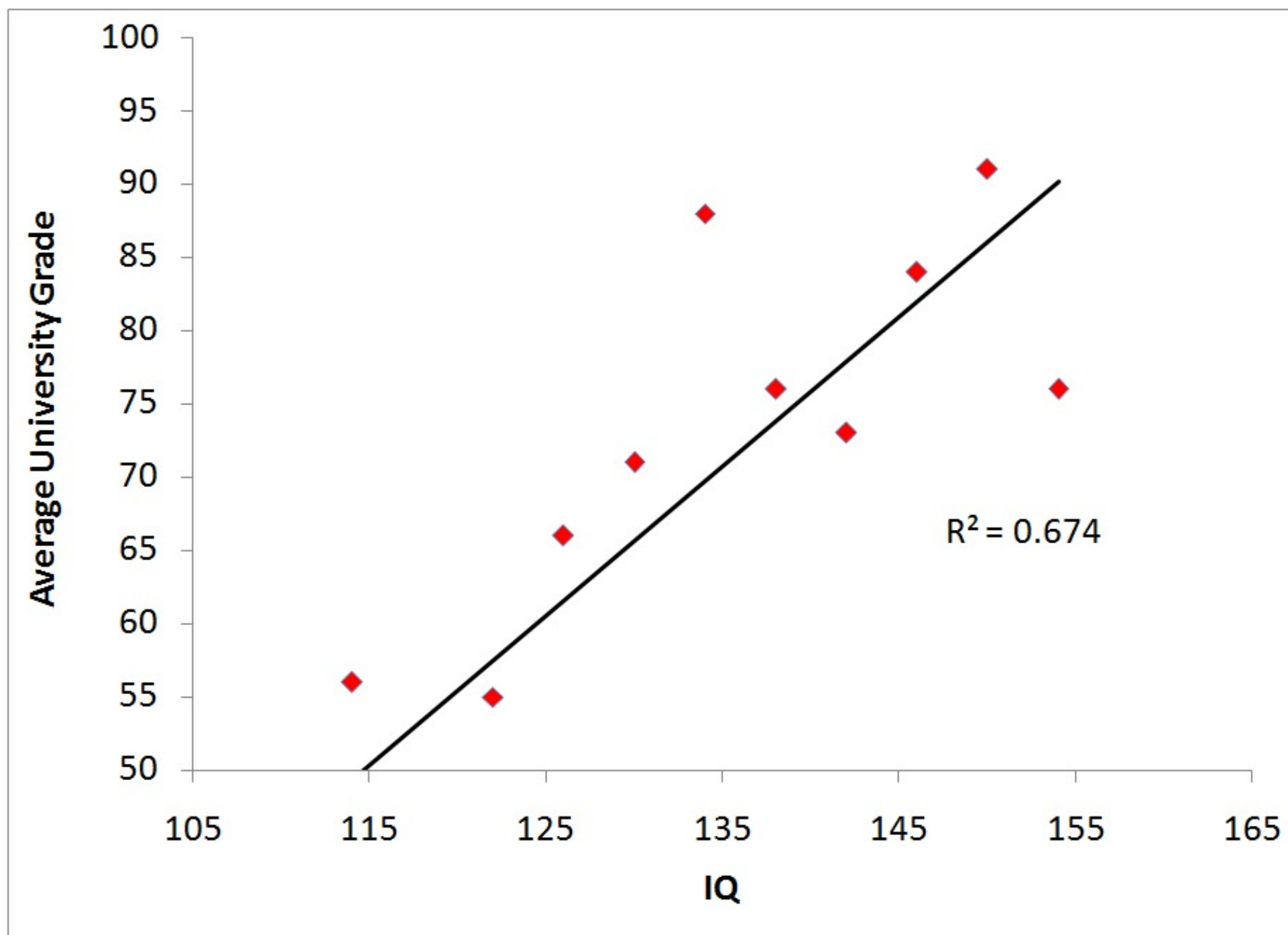
2. Proportion of Accounted Variance

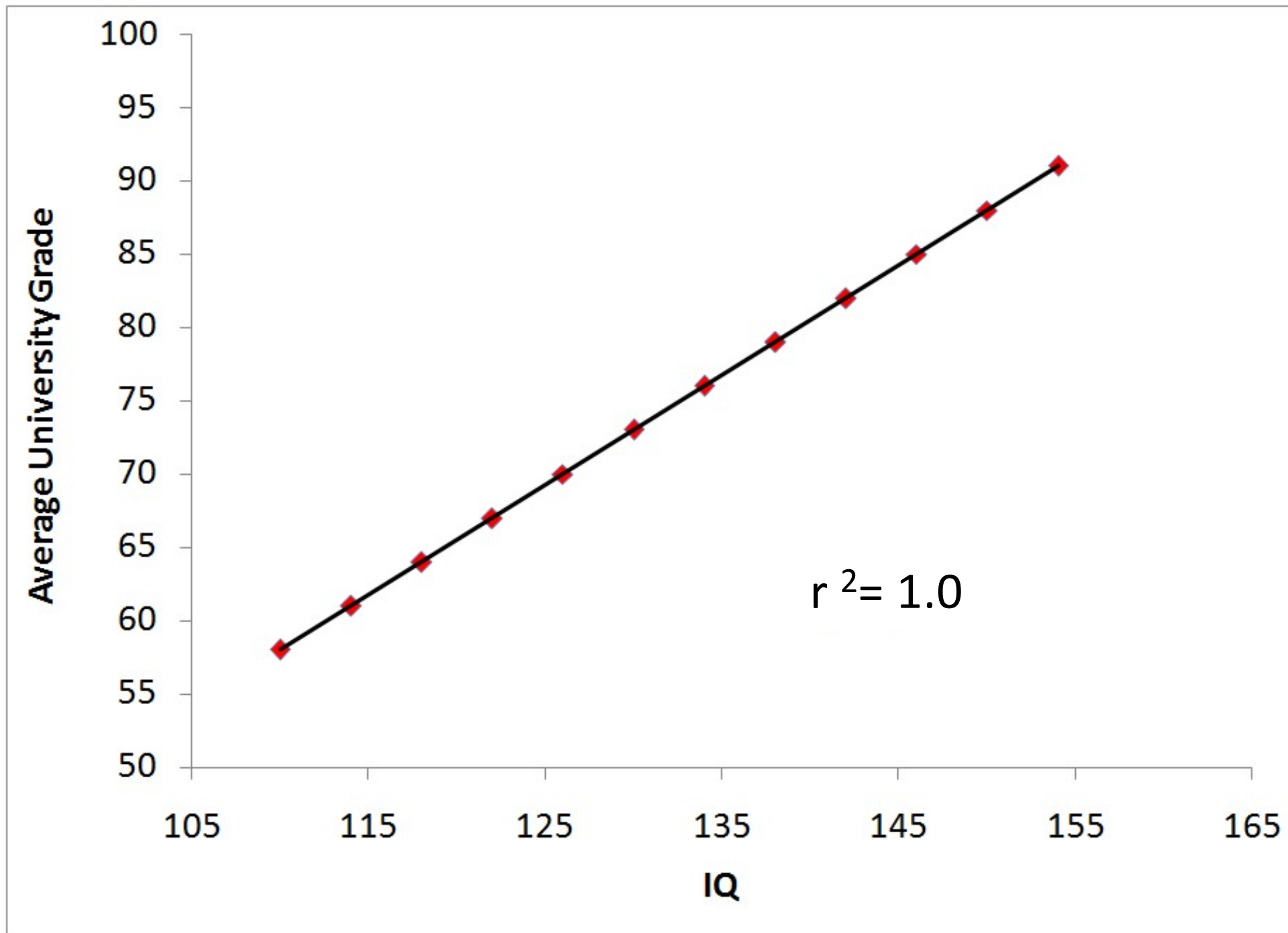
$$r = \sqrt{\text{Proportion of total variability of } y \text{ accounted for by } x}$$

$r^2 =$ Proportion of total variability of y accounted for by x

- coefficient of determination

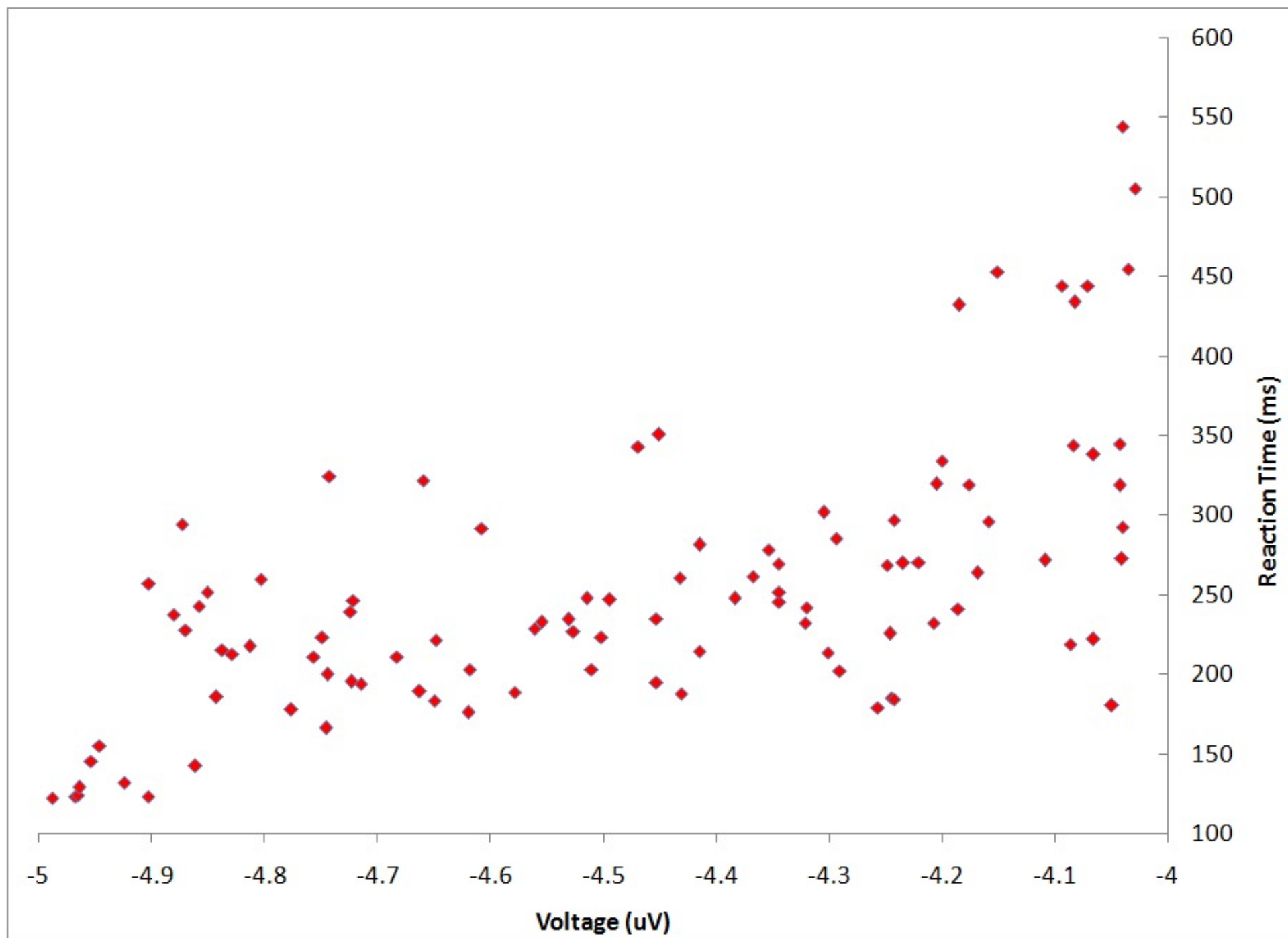
- proportion of explained variance



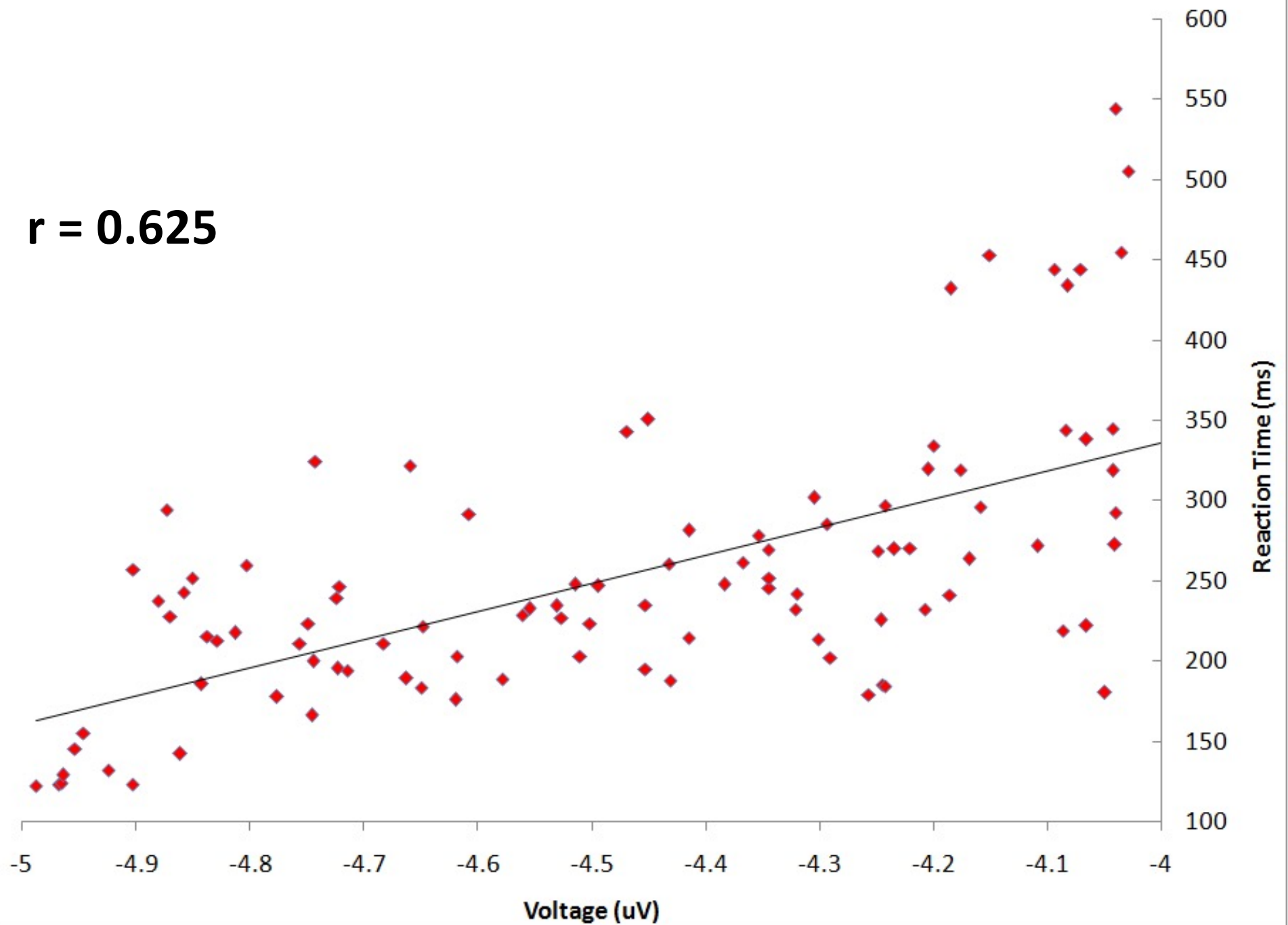


Issues in Interpreting the Correlation Coefficient

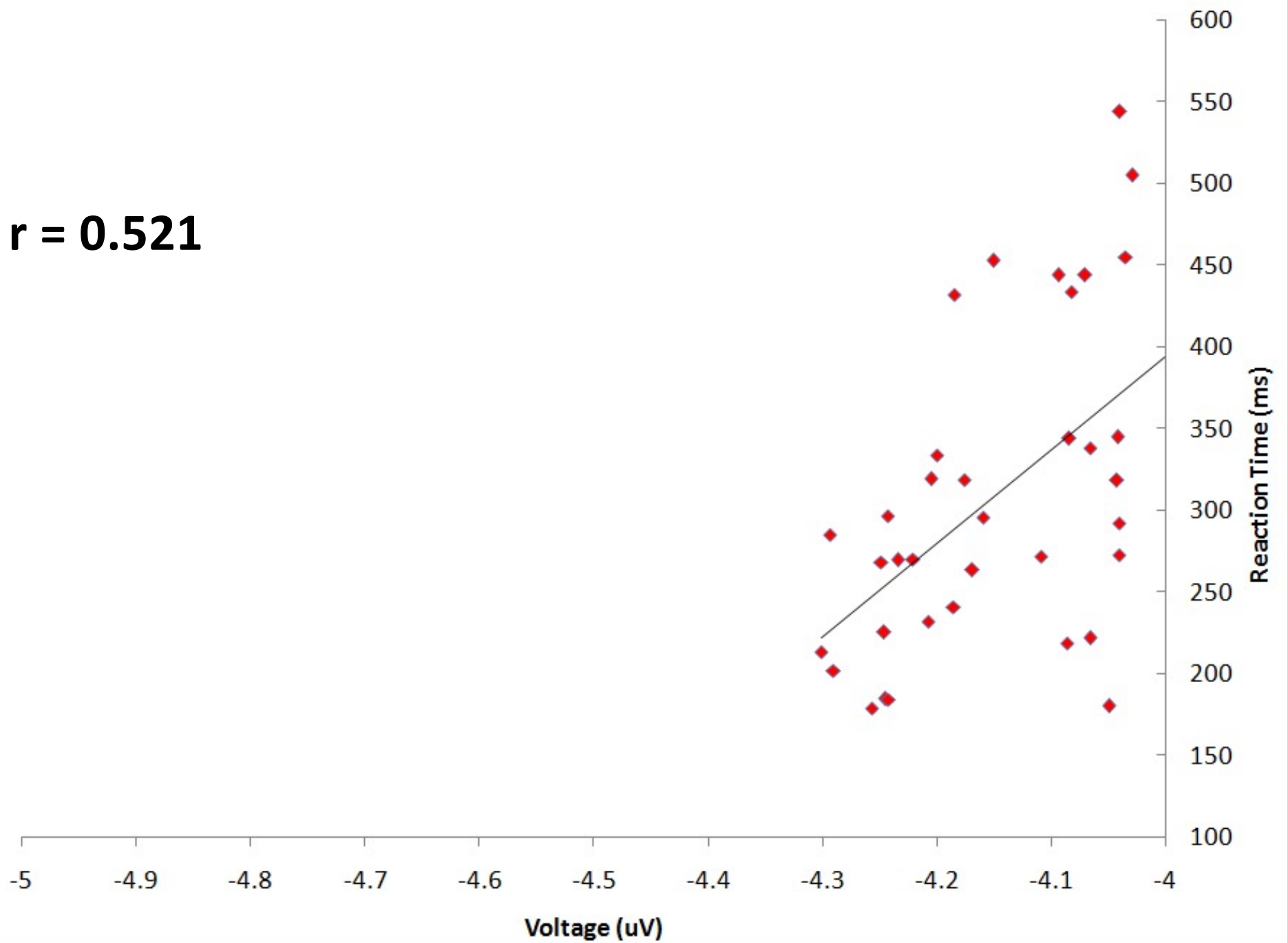
3. Restriction in Range



$r = 0.625$

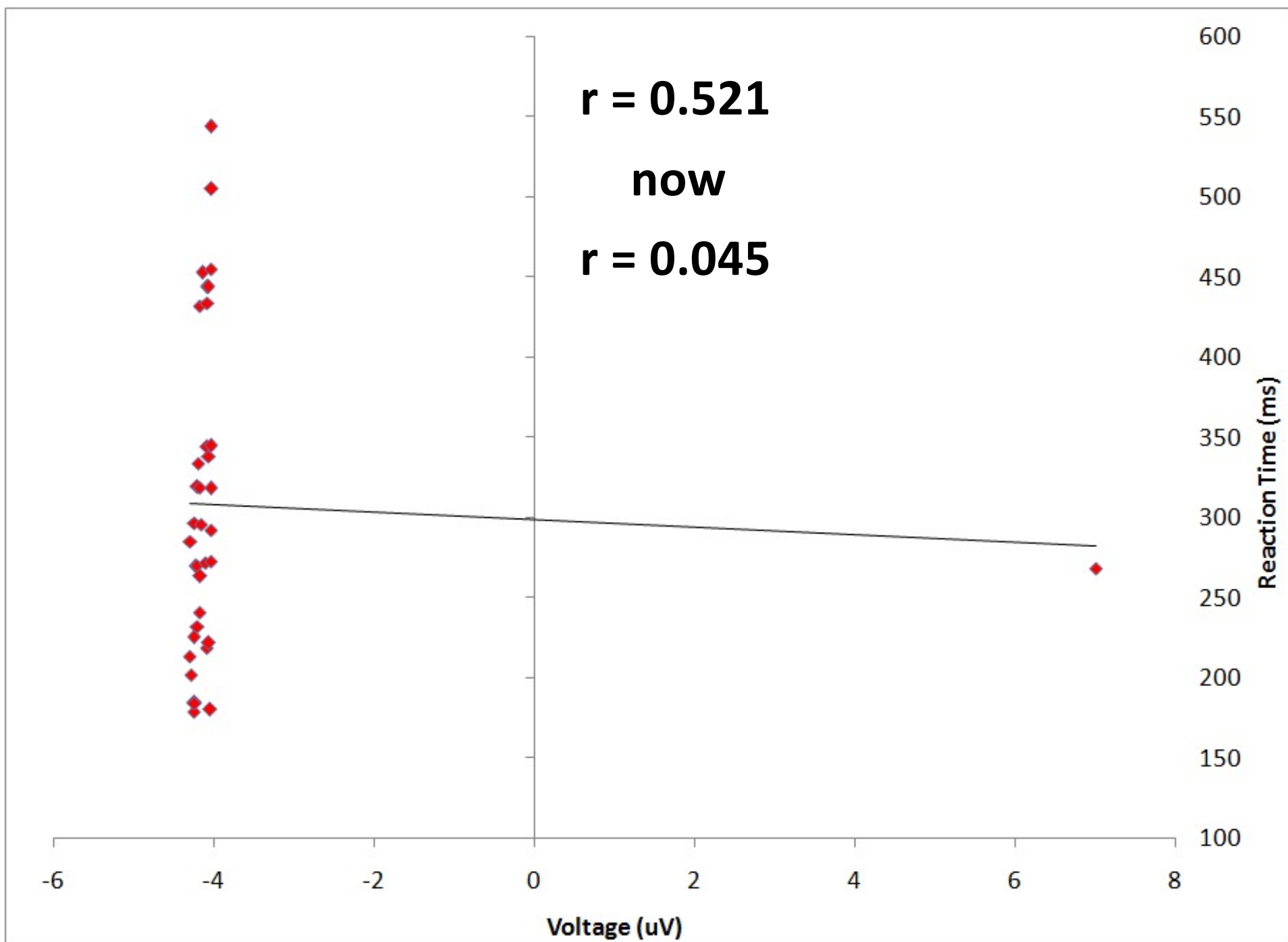


$r = 0.521$



Issues in Interpreting the Correlation Coefficient

4. Effect of Extreme Scores



Issues in Interpreting the Correlation Coefficient

5. Unreliability of Measurement

- our measurements are usually not accurate
- thus our correlations are typically less than they should be
- this is called attenuation

What is a “good” relation value?

Effect Size

The correlation coefficient itself is a measure of effect size

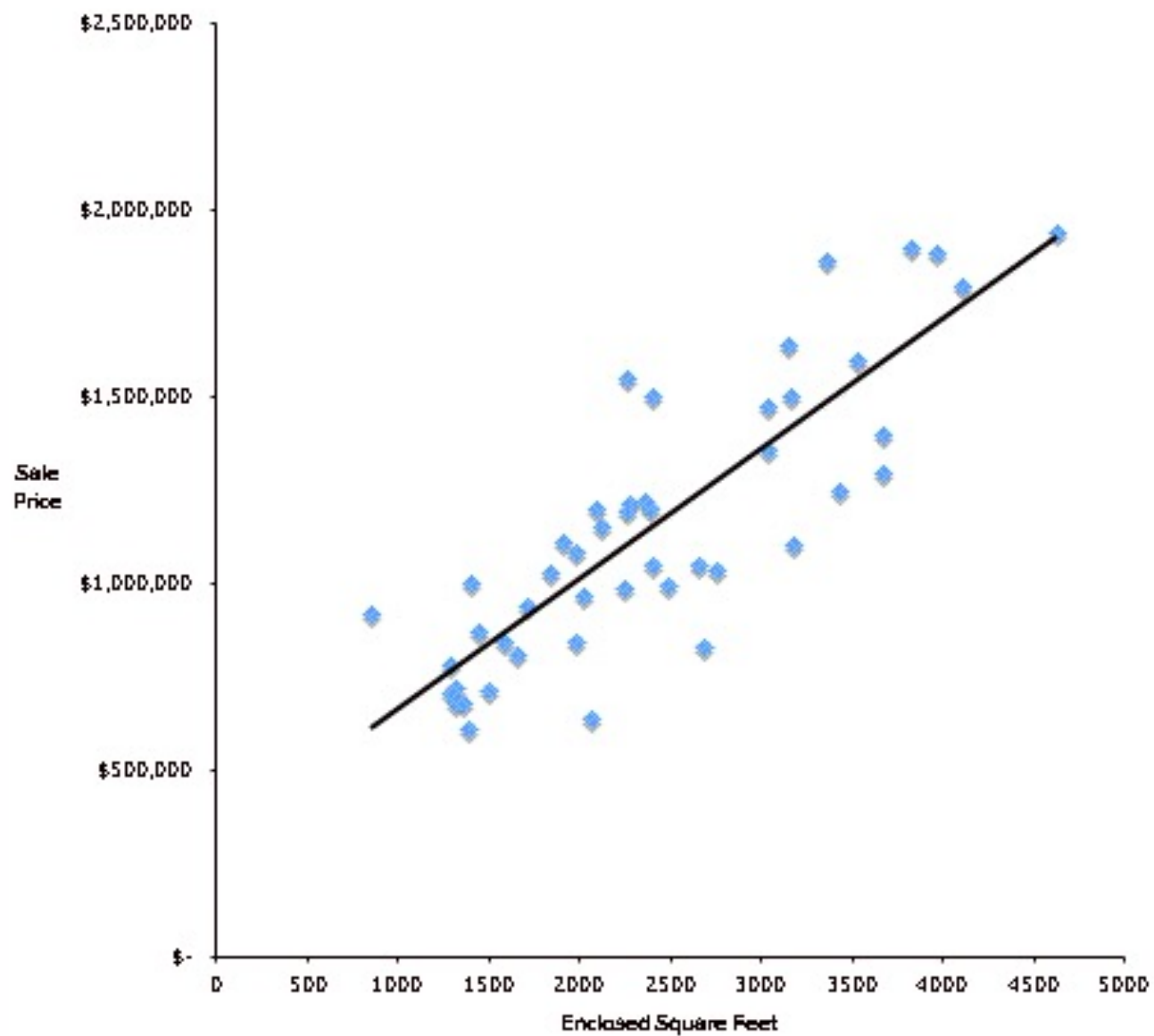
0.1 small

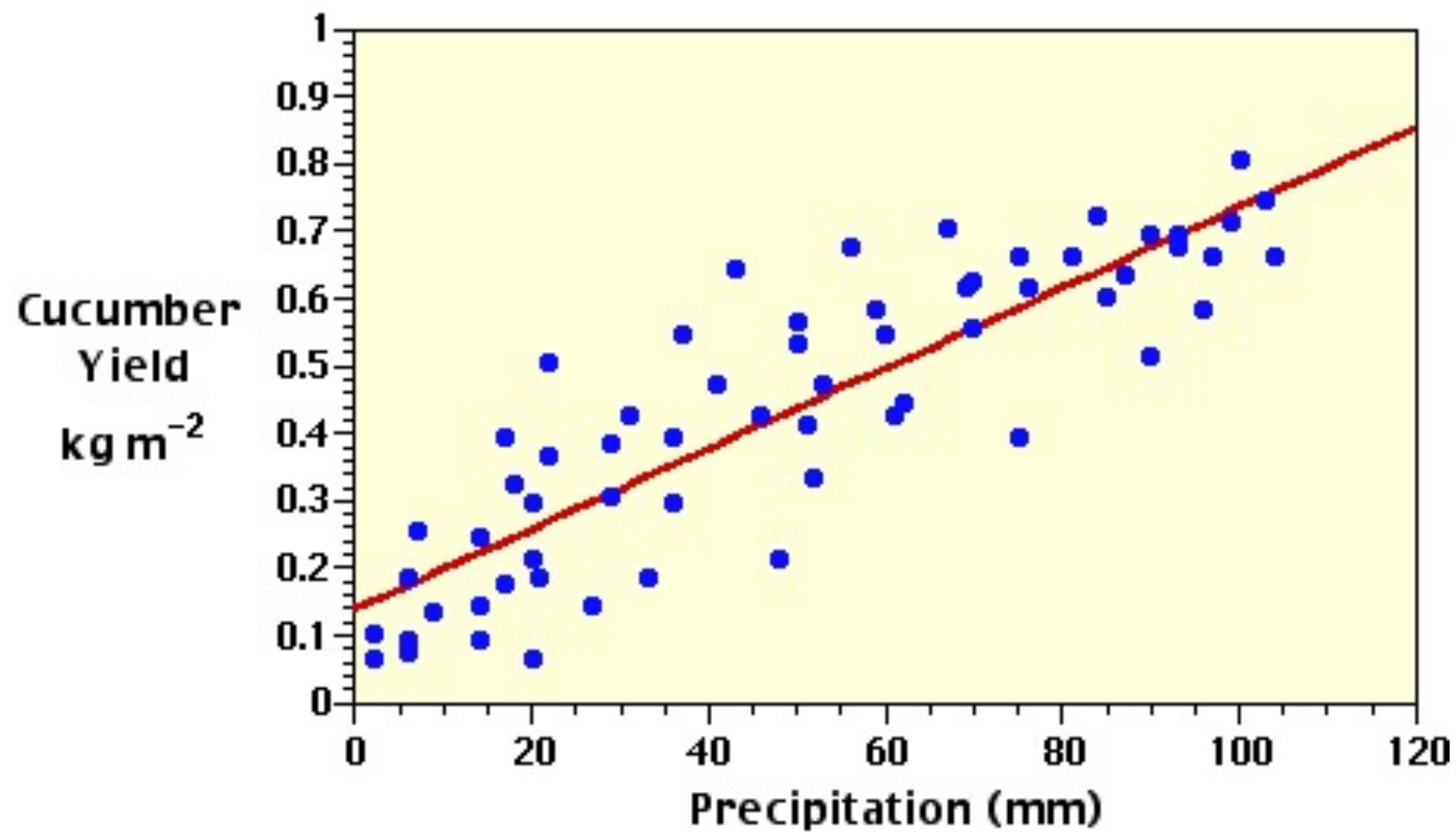
0.3 medium

0.5 large

Regression (Prediction)

Enclosed Area Regression Results

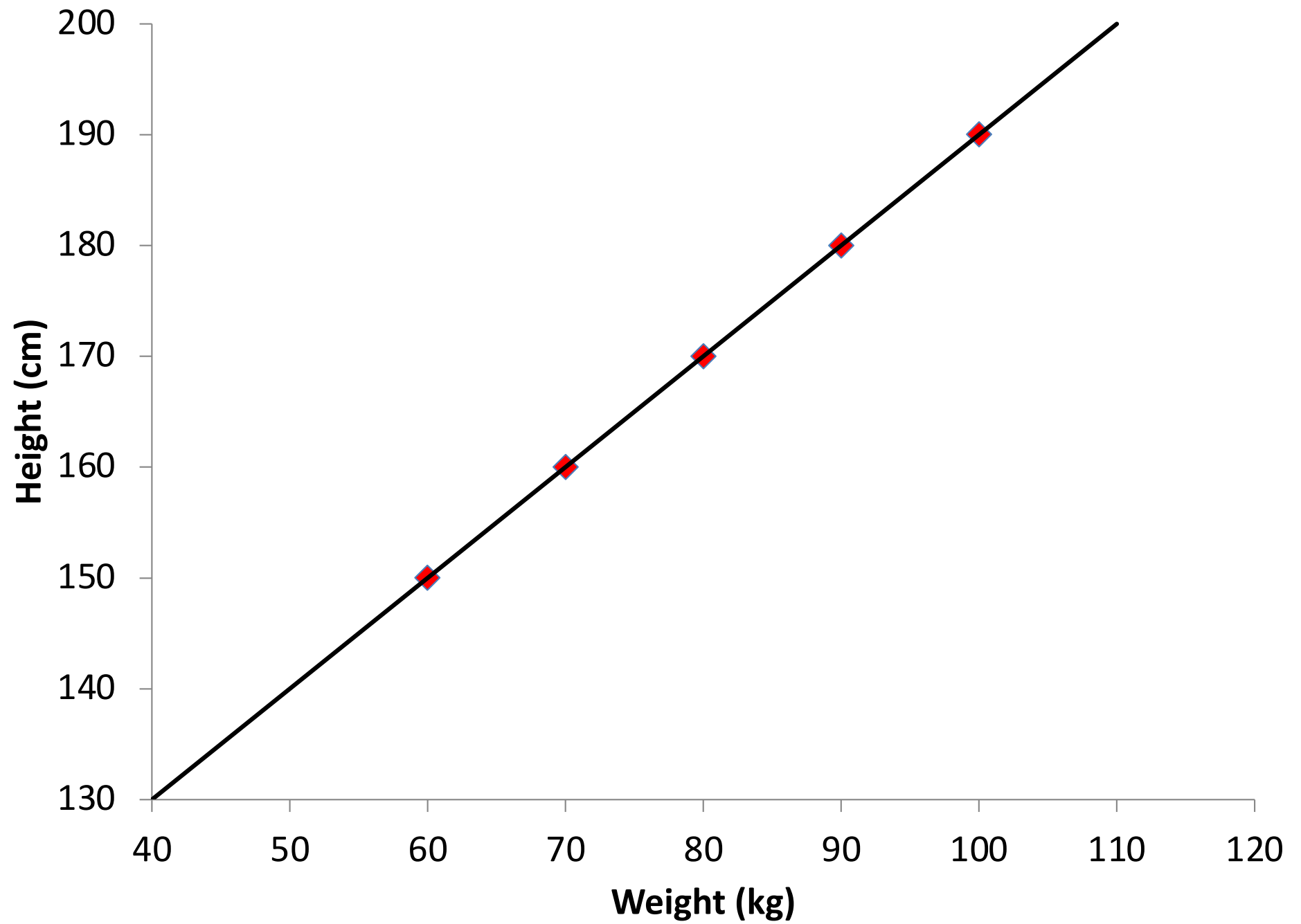


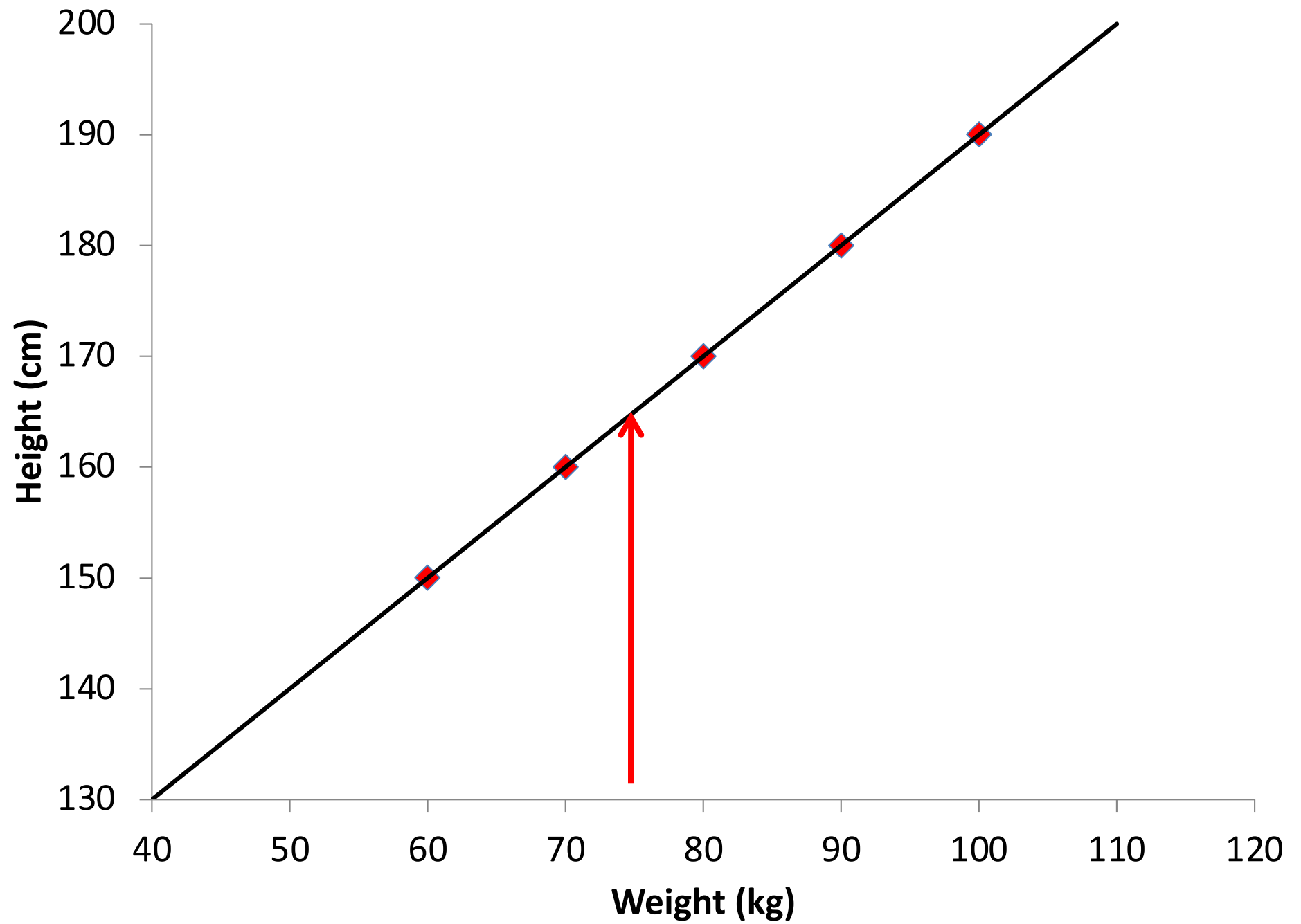


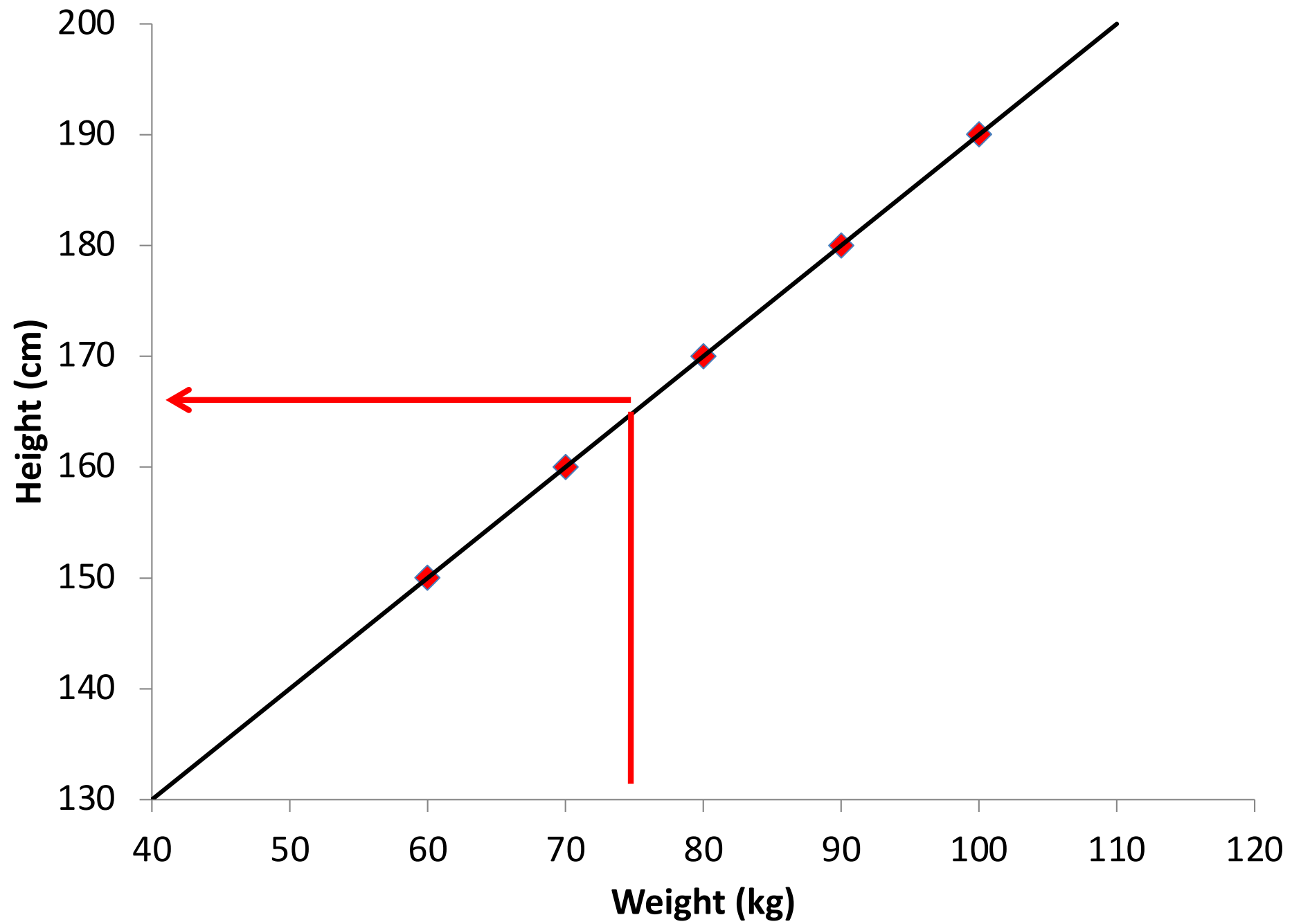
Linear Regression Line

Daily Chart - AT&T (T)









Finding the equation of the line

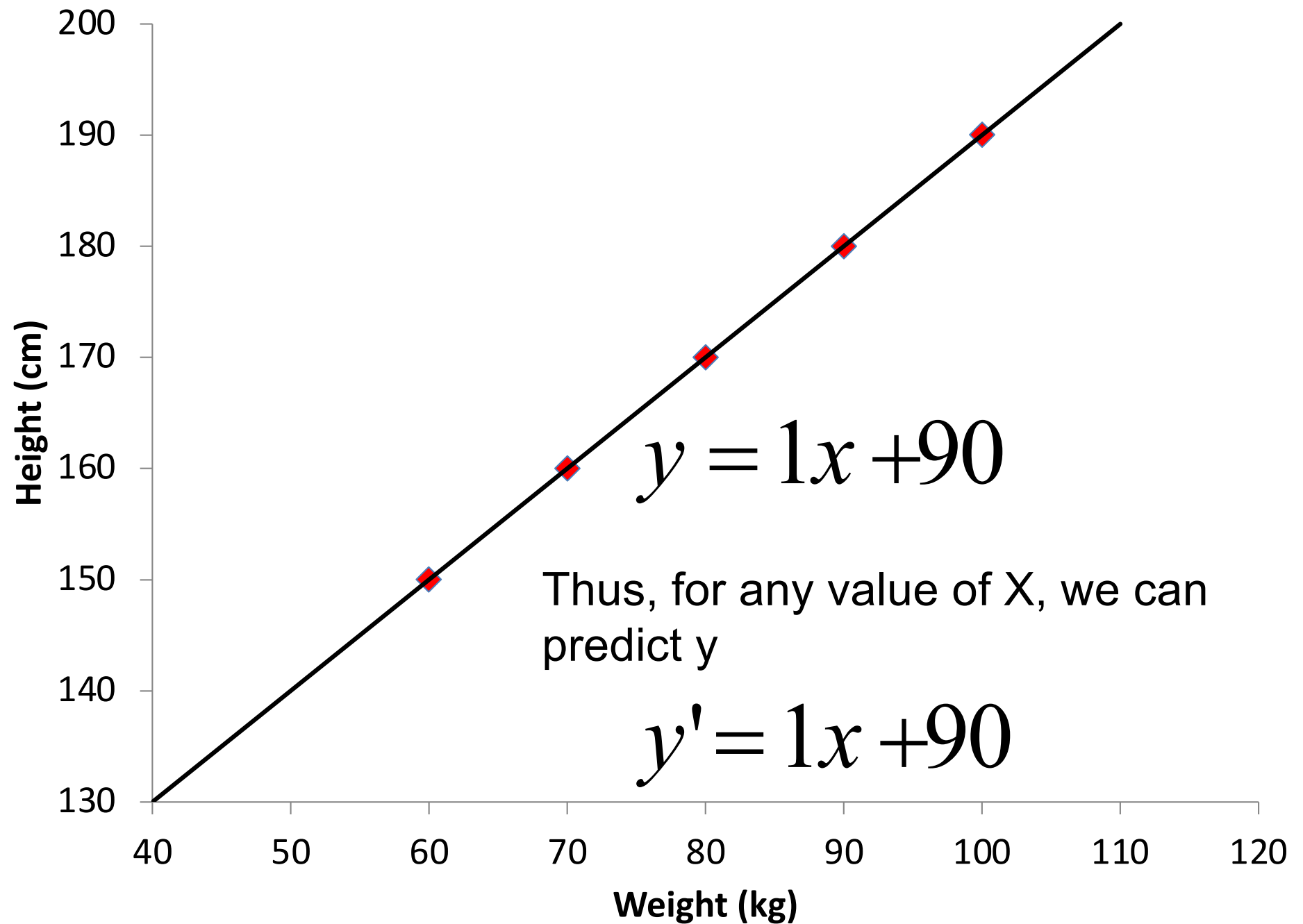
$$y = bx + a$$

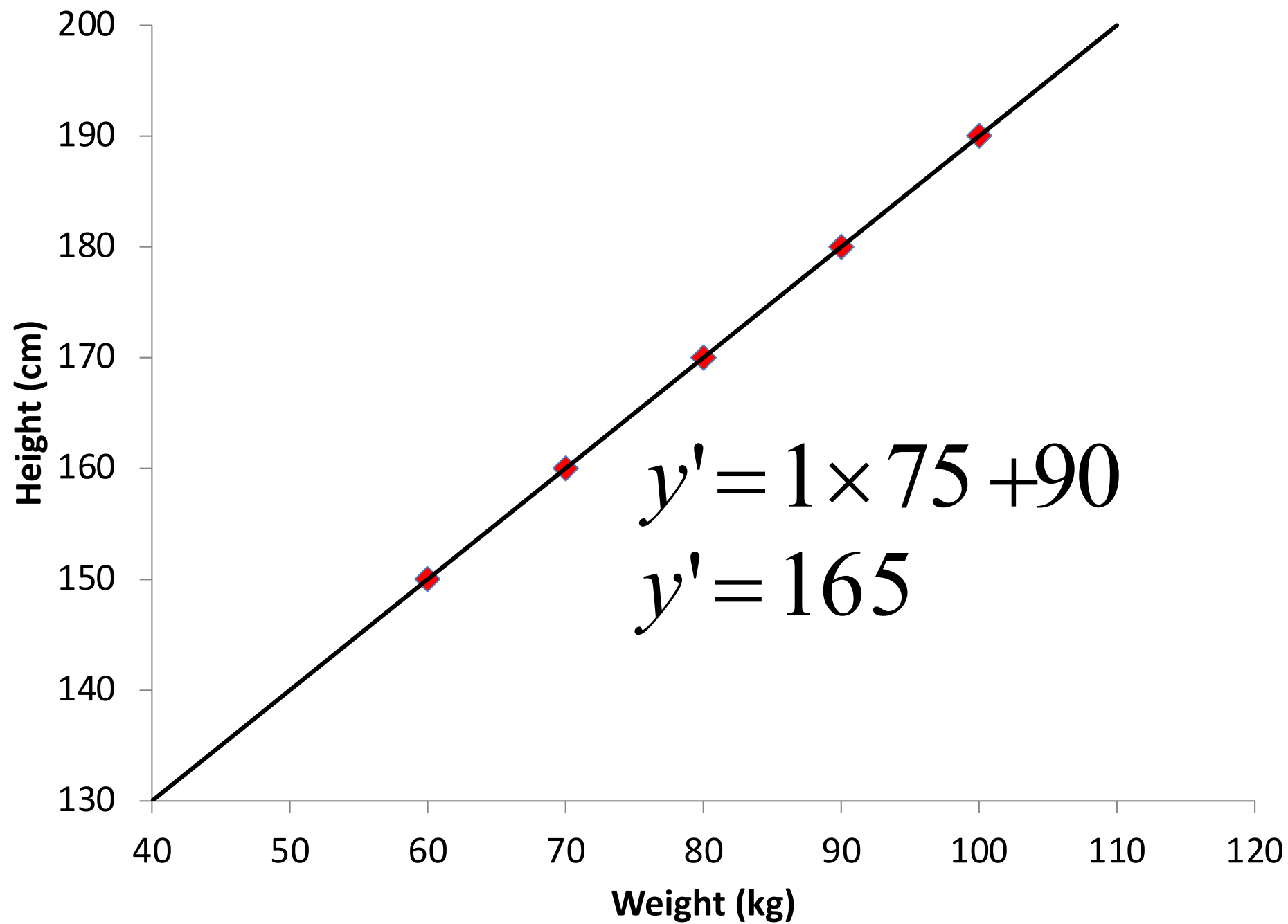
Where:

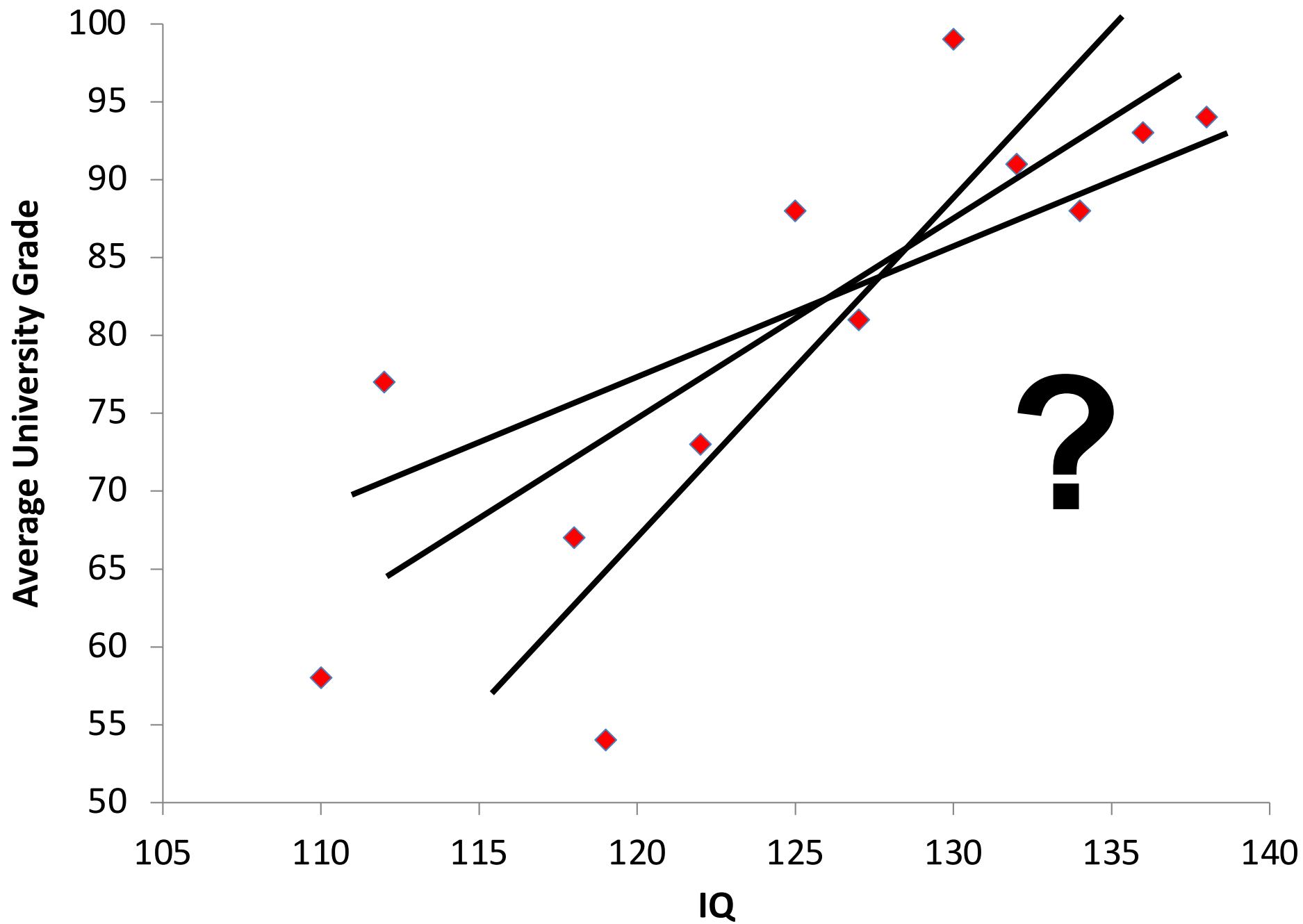
b = slope

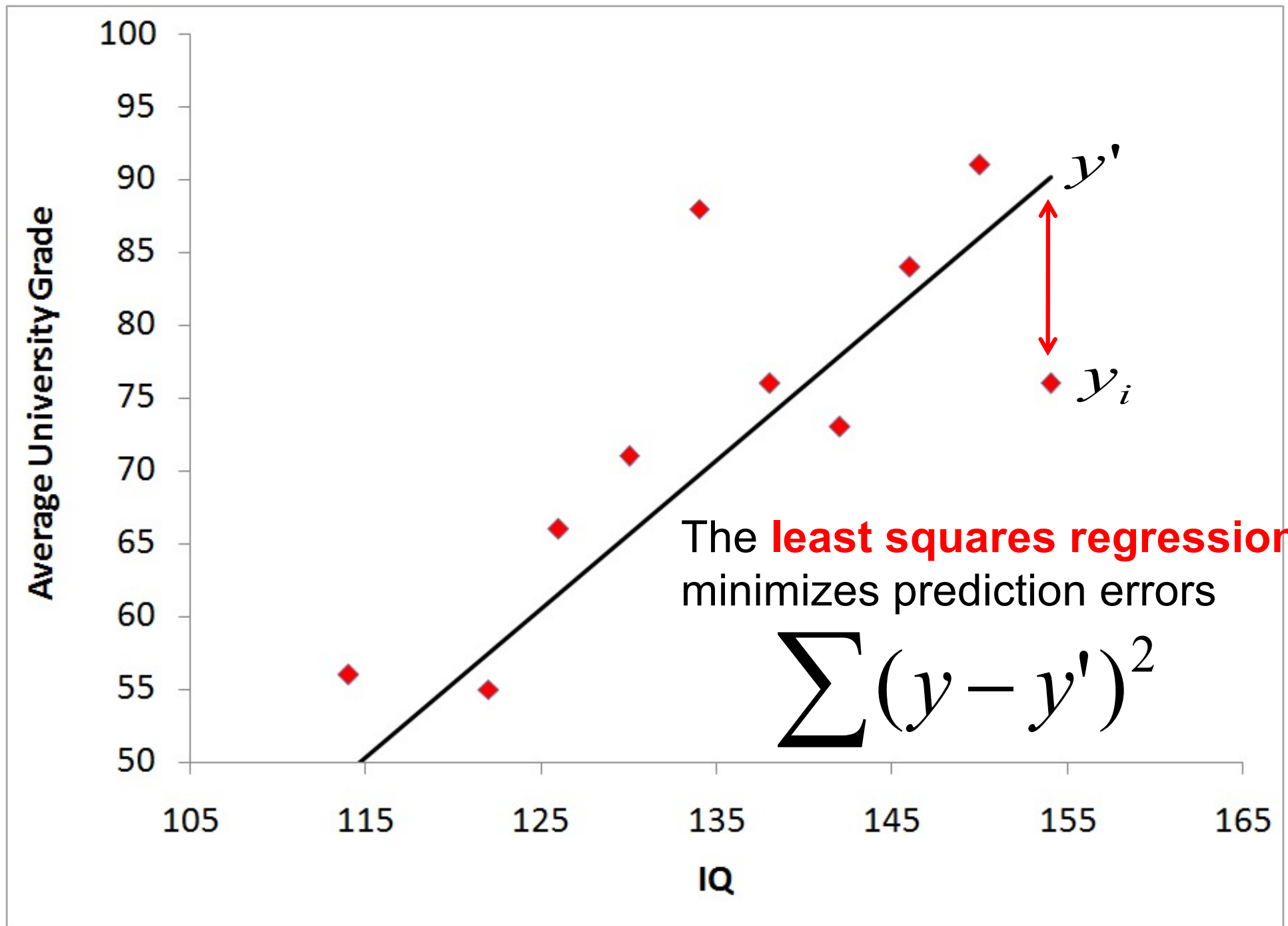
a = y intercept

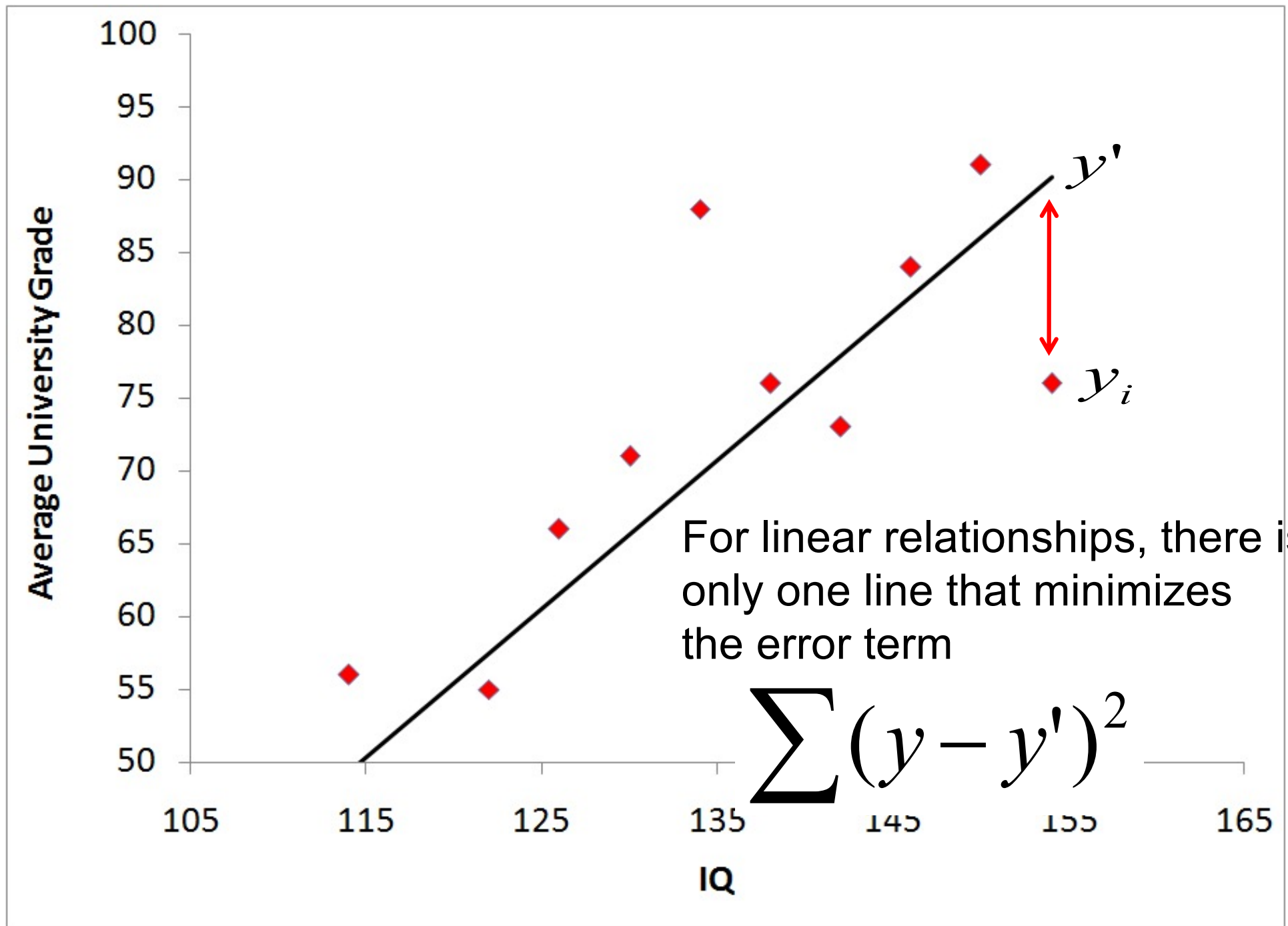
*** NOTE THE CHANGE IN LETTERING











So...

For a **given variable x** we want to
determine the **predicted variable y'**

Least squares equation

$$y' = 0.446x + 2.351$$

And we could now predict new values
of y given x

Formalizing This...

$$Y' = a + bX$$

Note, b , the slope is also called the:

Regression Coefficient

And is sometimes symbolized with Beta, β

Formalizing This...

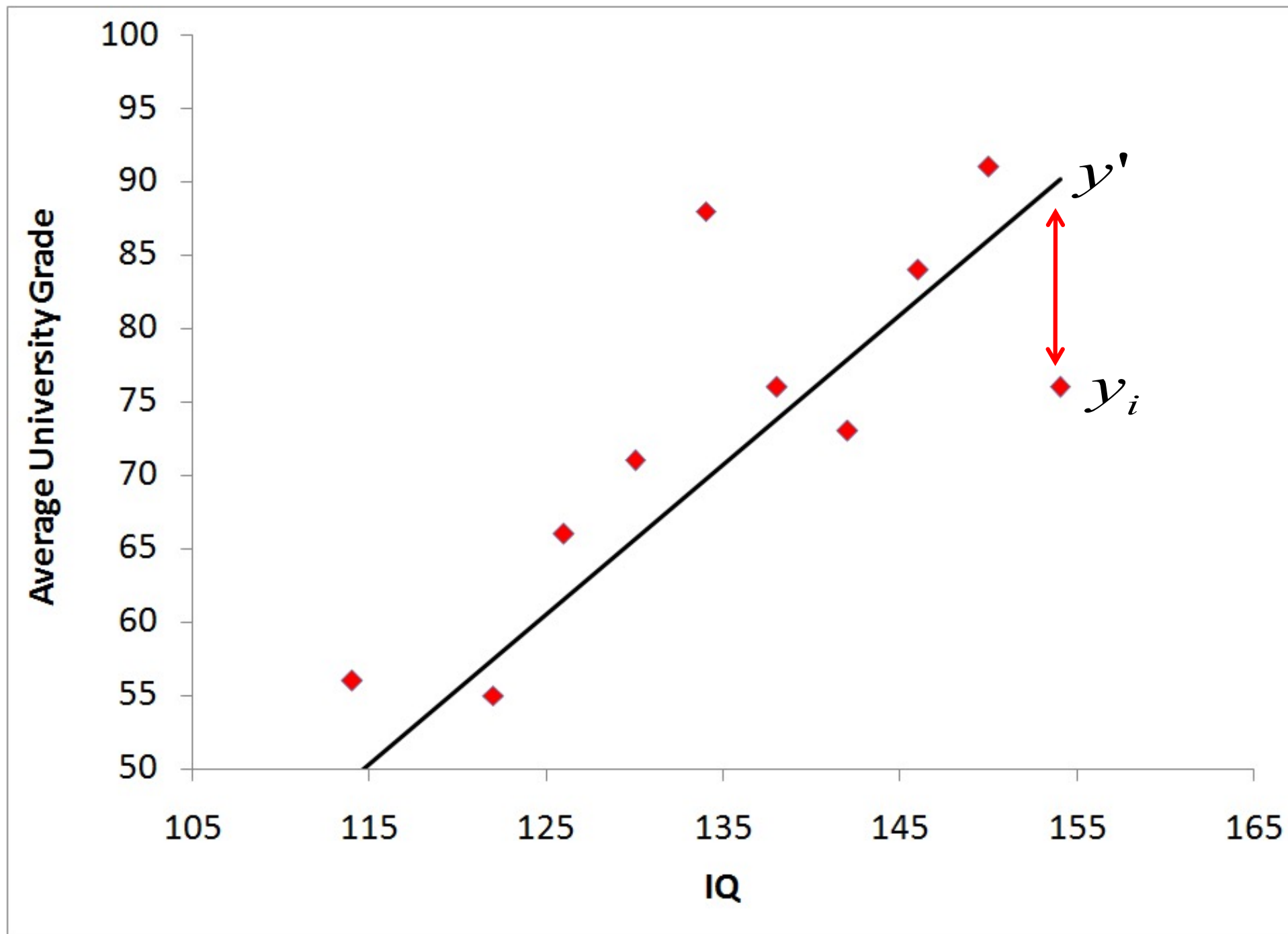
$$Y' = a + bX$$

And a is the value a person would have if they had a score of 0 on the predictor variable

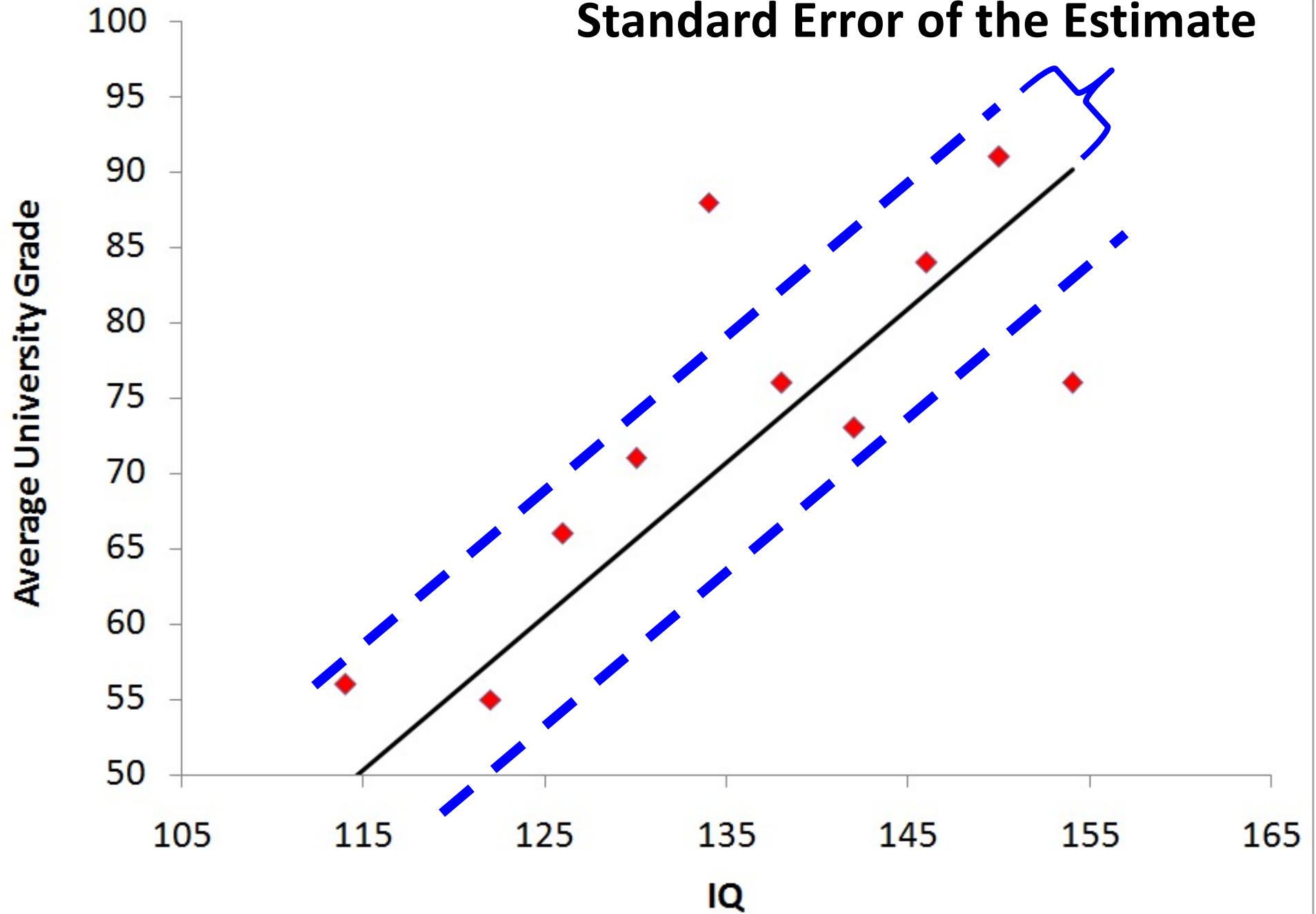
Measuring Prediction Errors

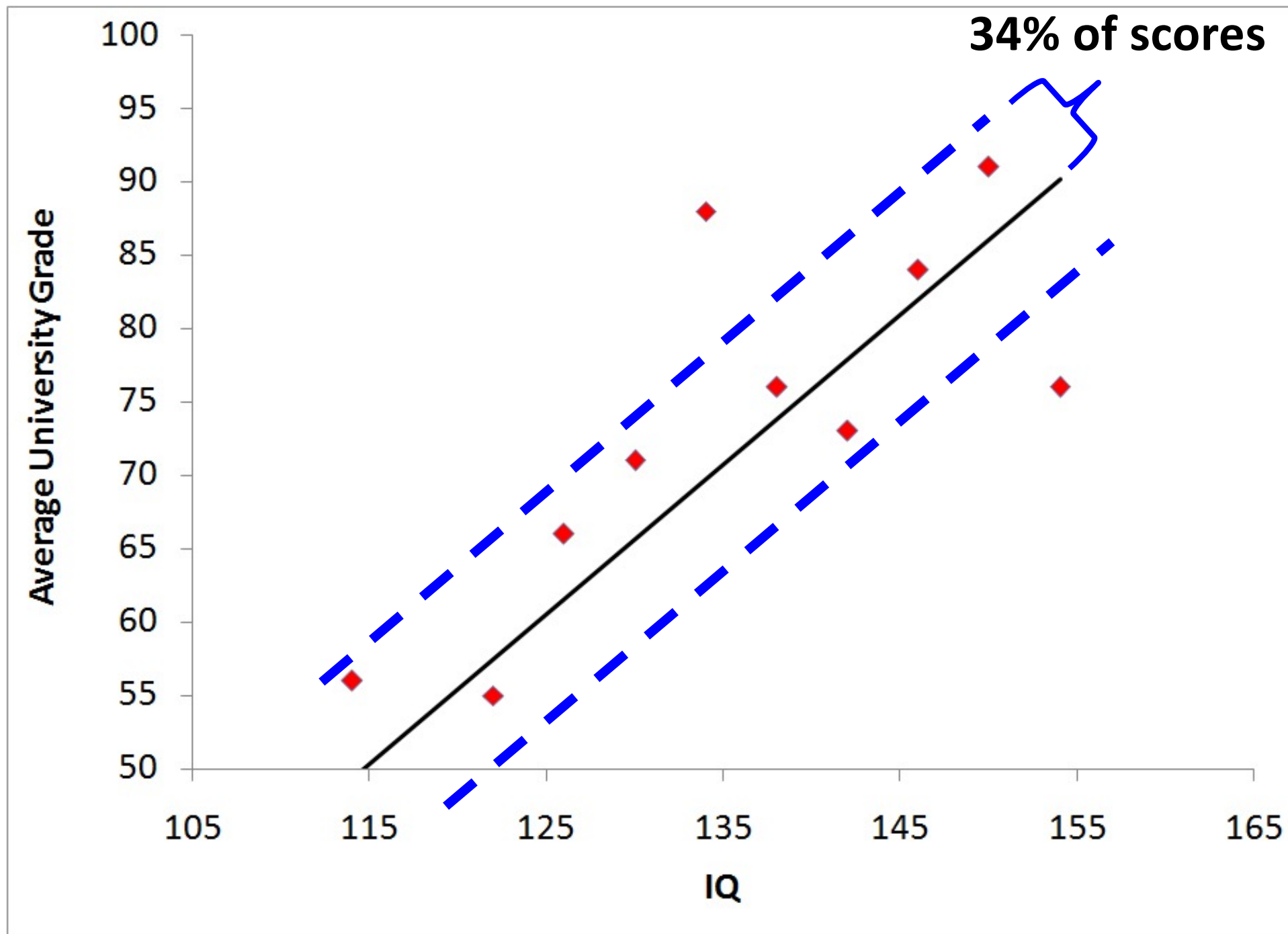
The Standard Error of the Estimate

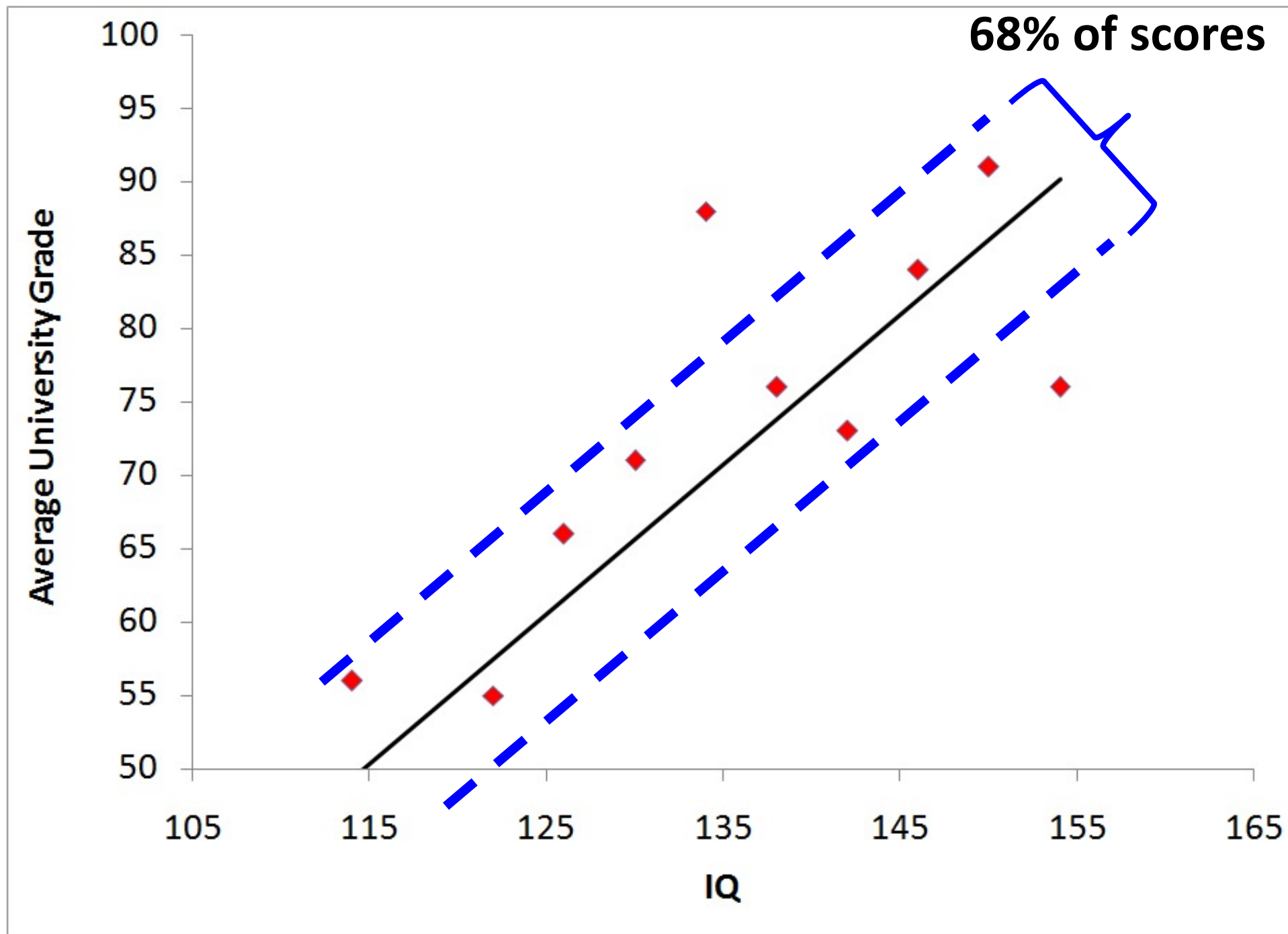
What is the error in prediction?

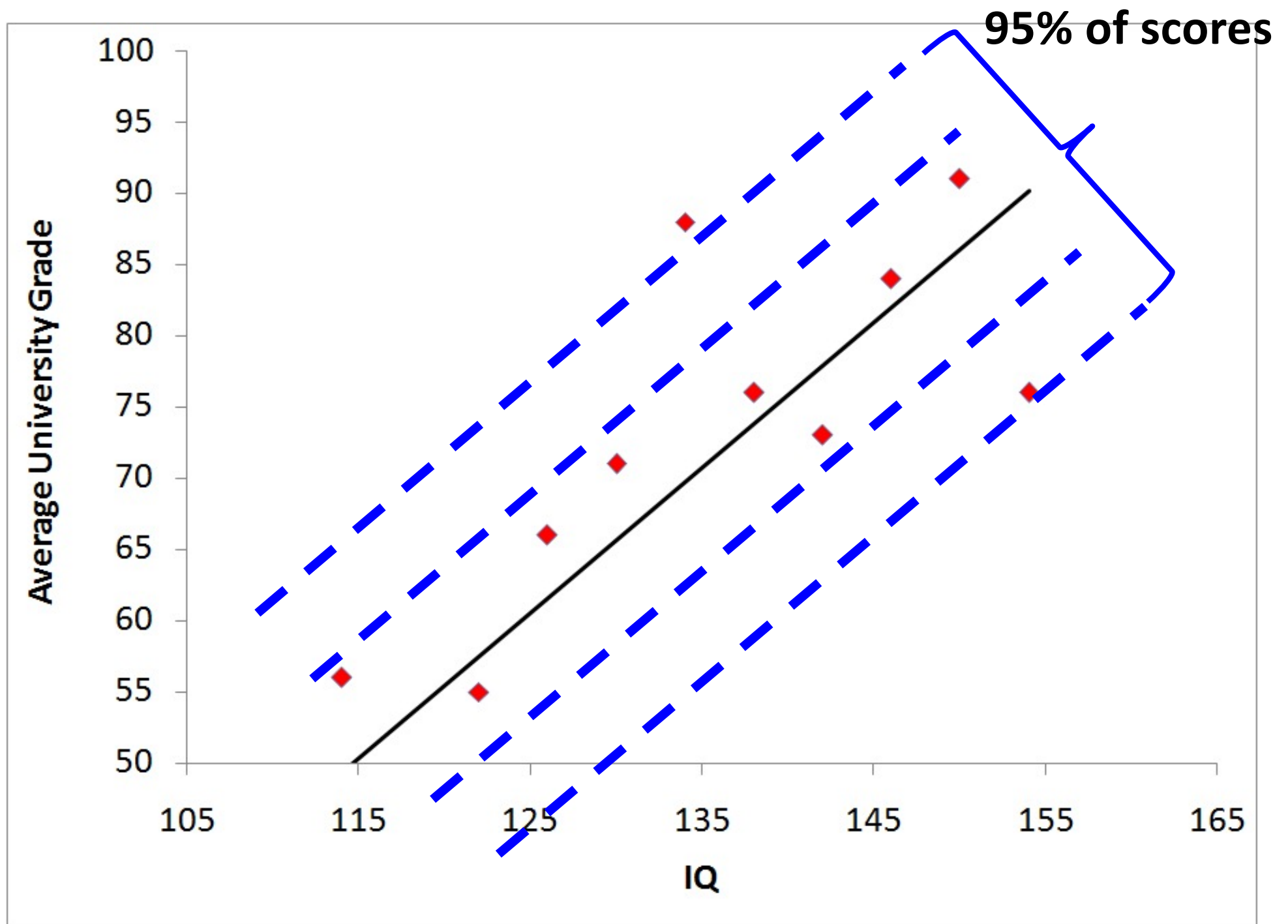


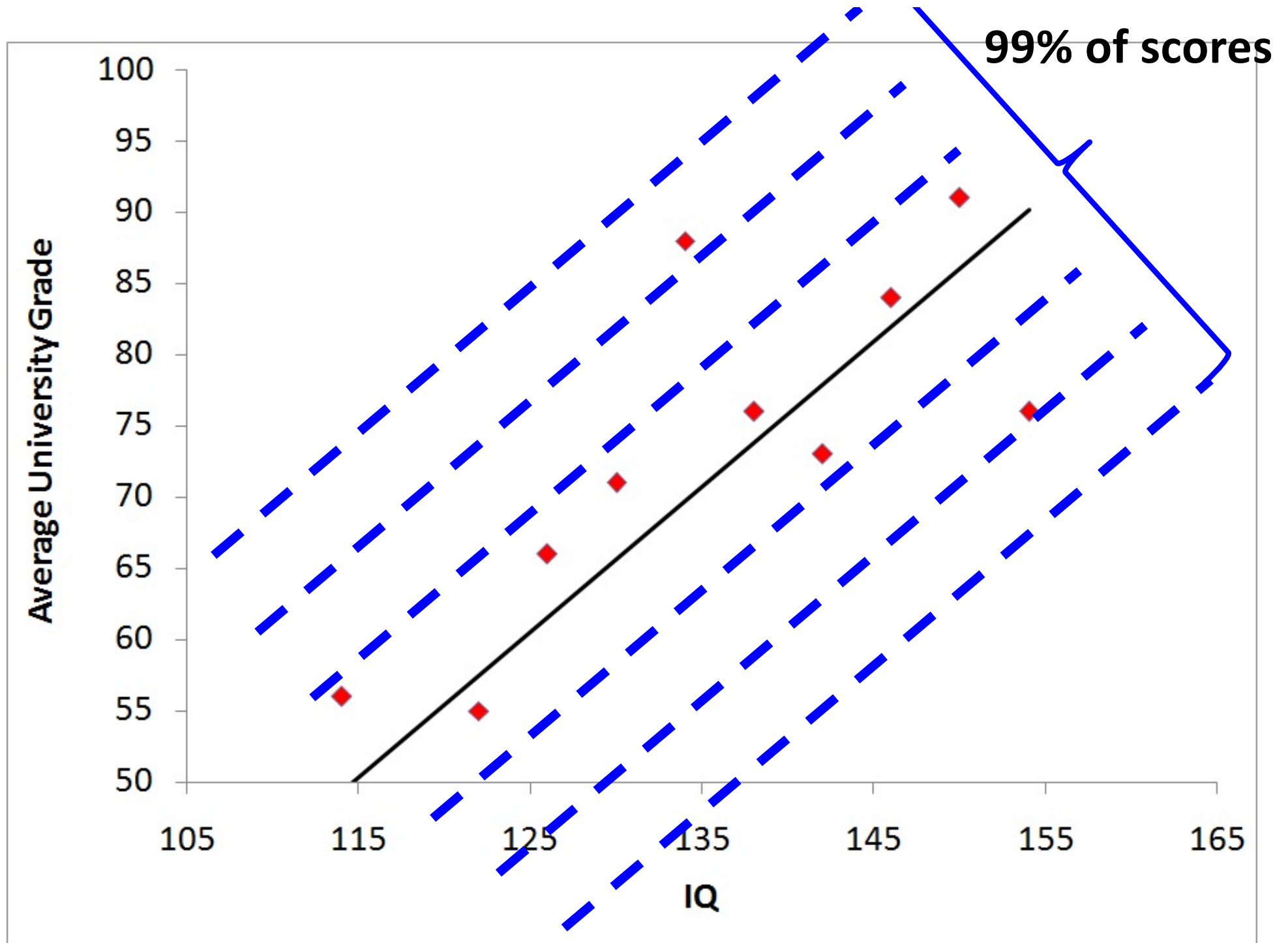
Standard Error of the Estimate











We can also think of this in terms of
error

$$\text{Error} = (Y - Y')$$

Or

$$\text{Error}^2 = (Y - Y')^2$$

And why think of error...

We do this because we can make an interesting comparison, we can compare the the amount of squared error our model has with the amount of squared error without the model...

We call this the PROPORTIONATE REDUCTION IN ERROR

The Proportionate Reduction in Error

$$SS_{error} = \sum (Y - Y')^2$$

$$SS_{total} = \sum (Y - \bar{Y})^2$$

$$PRE = \frac{SS_{total} - SS_{error}}{SS_{total}}$$

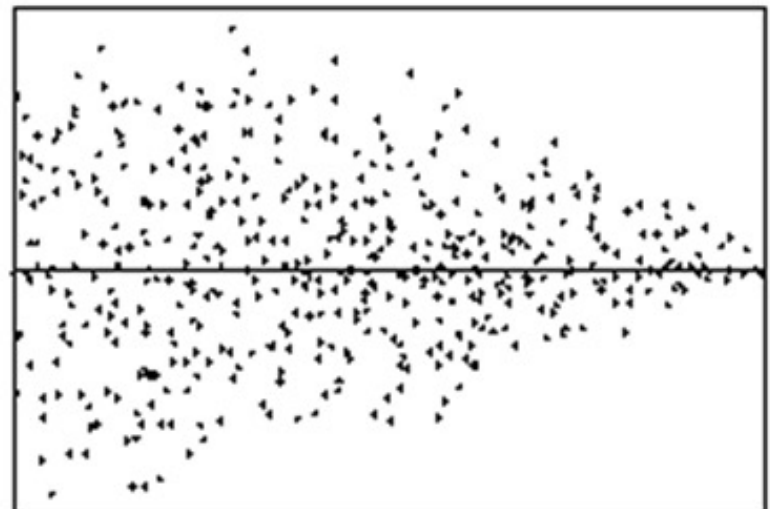
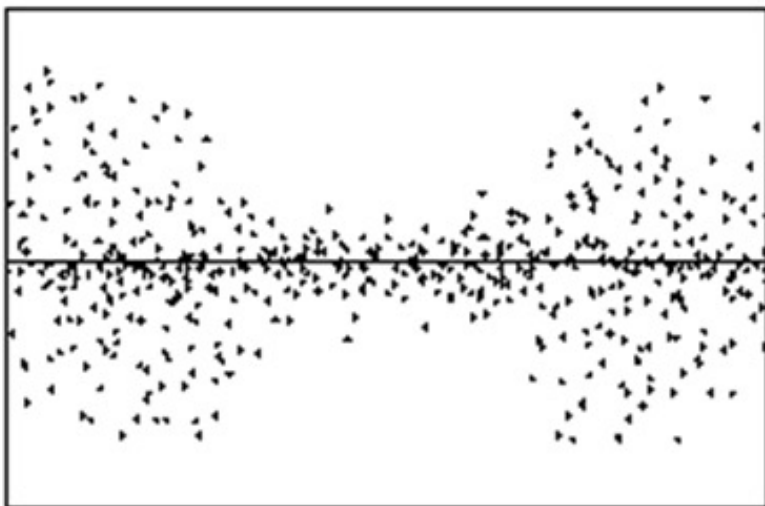
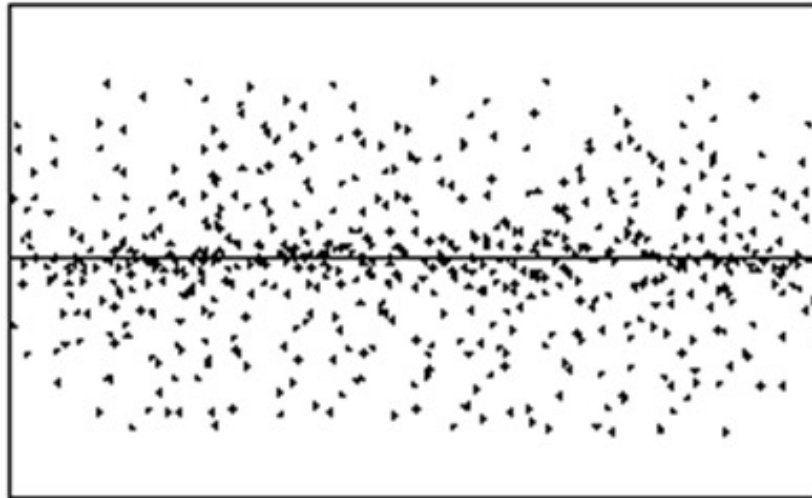
The Proportionate Reduction in Error

$$PRE = r^2$$

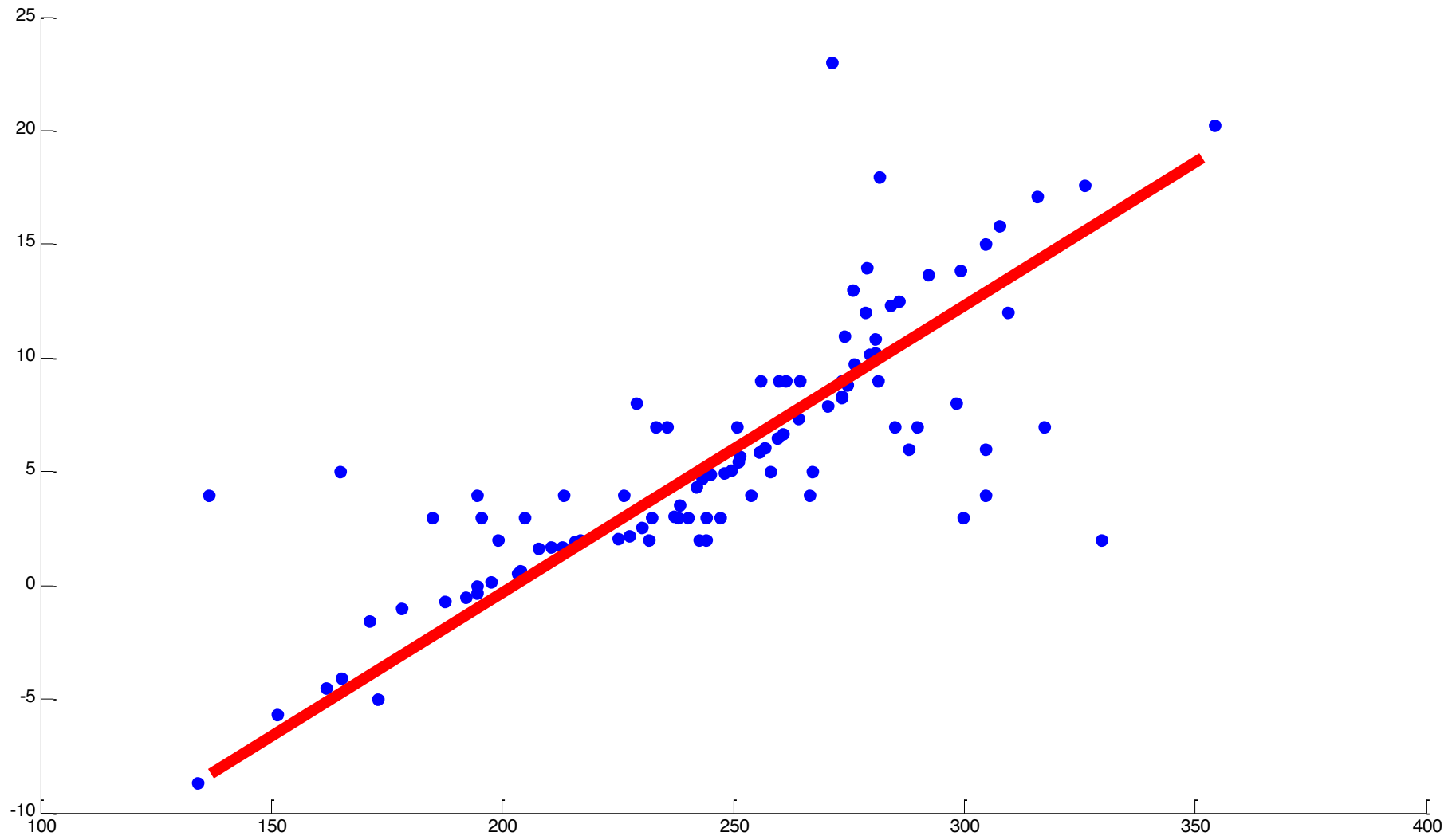
Thus the PRE is the same as the amount of variance accounted for by the regression model

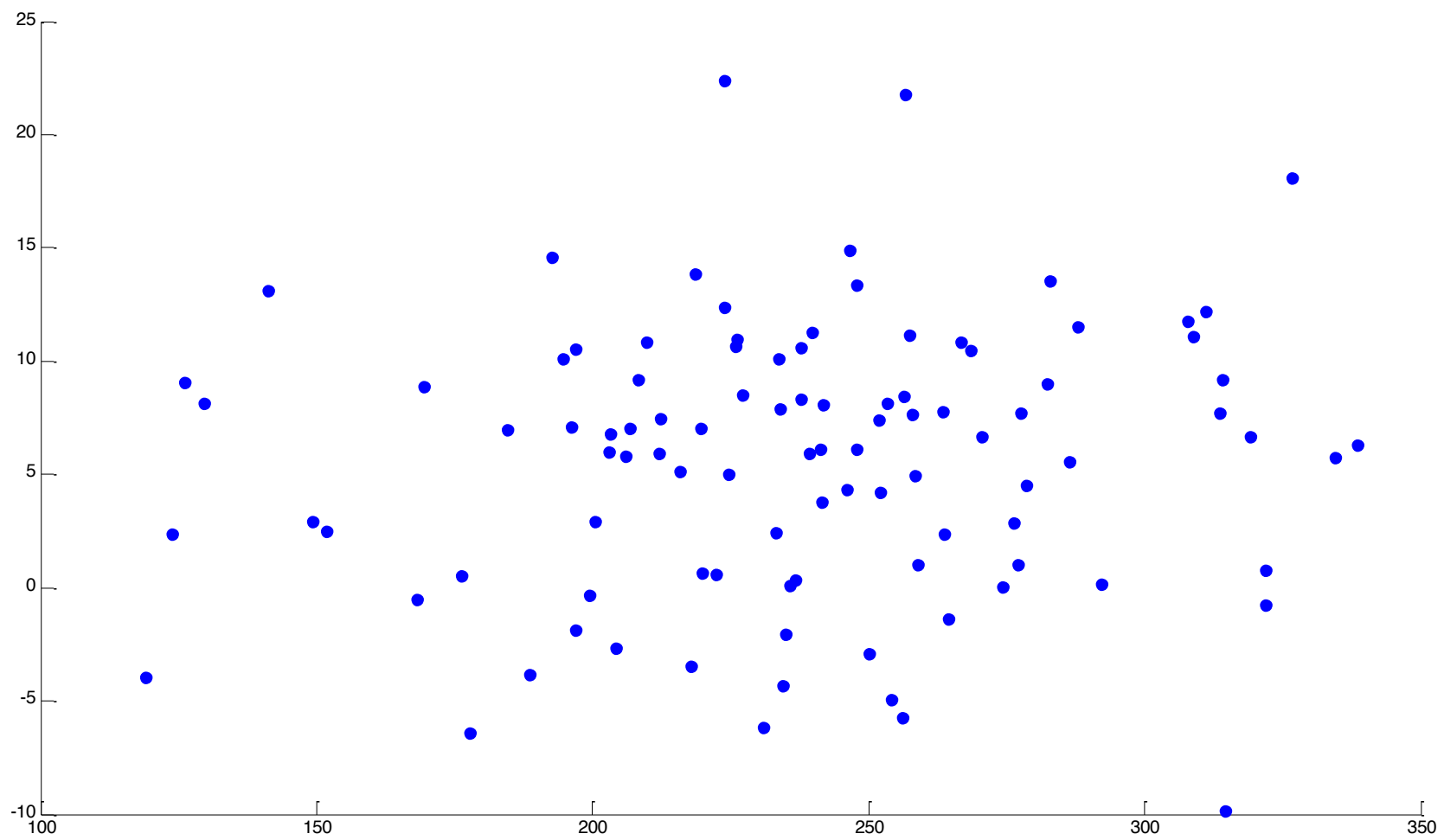
Assumptions of Regression

The Assumption of Homoscedasticity (Homogeneity of Variance)

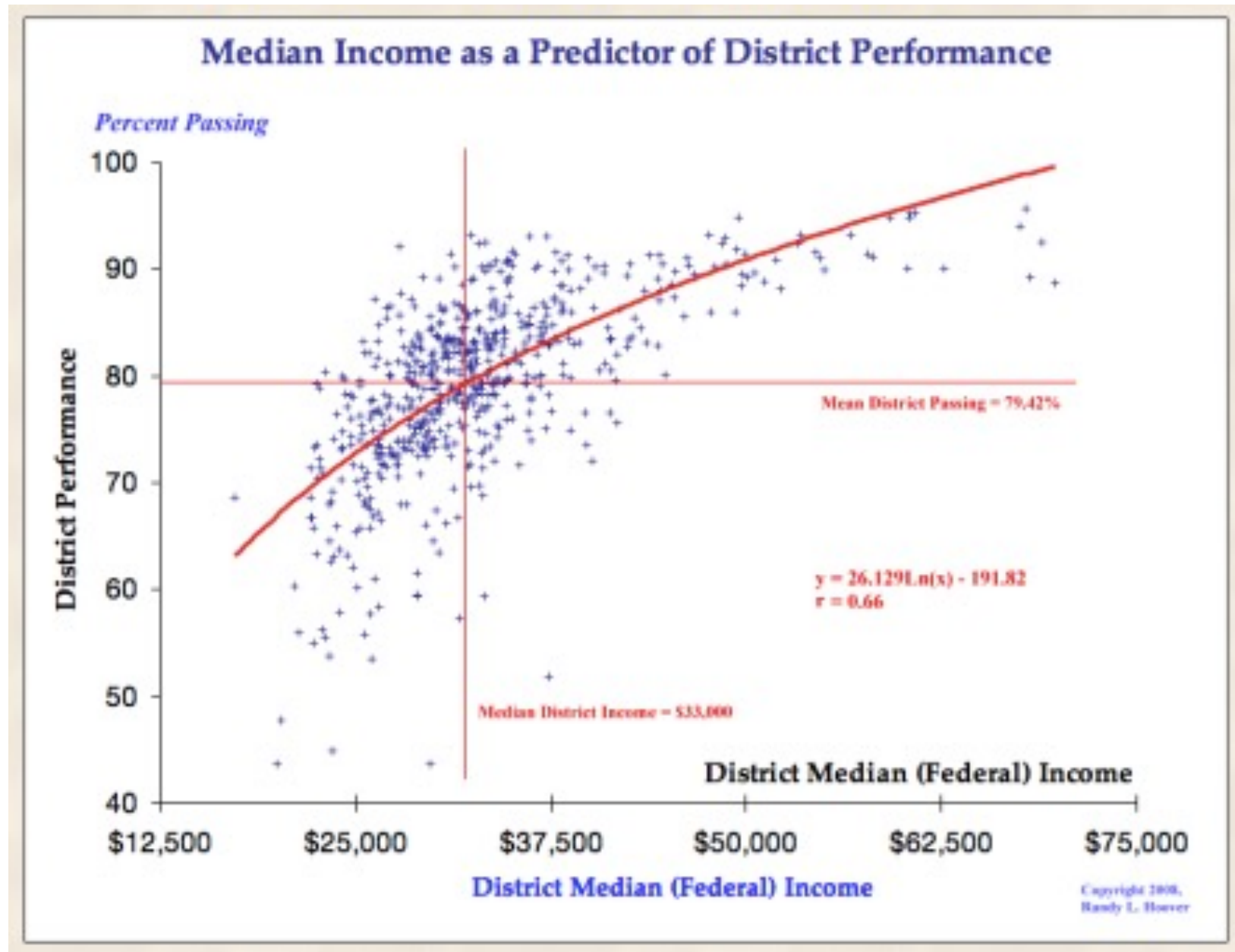


The Assumption of Linearity



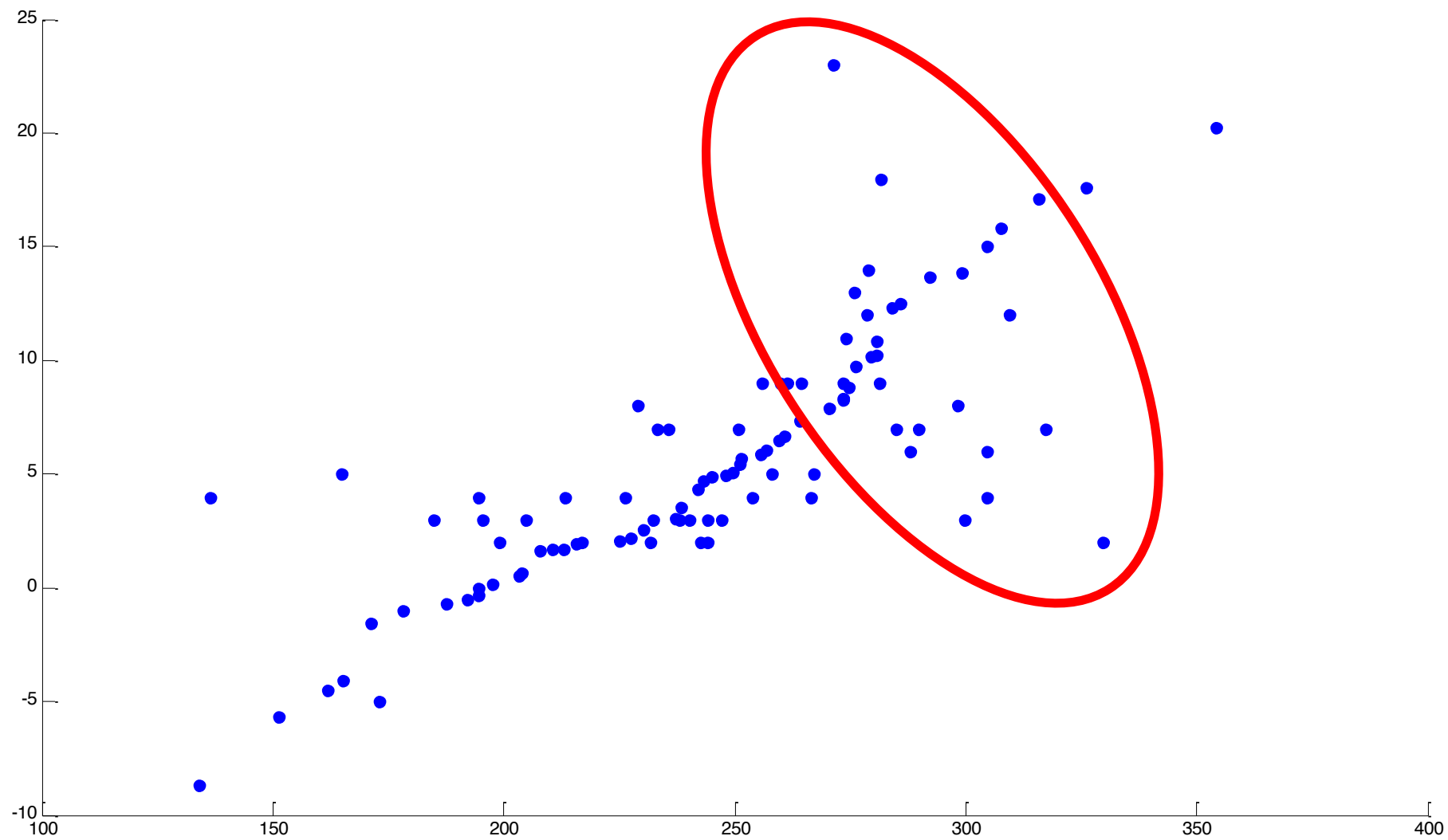


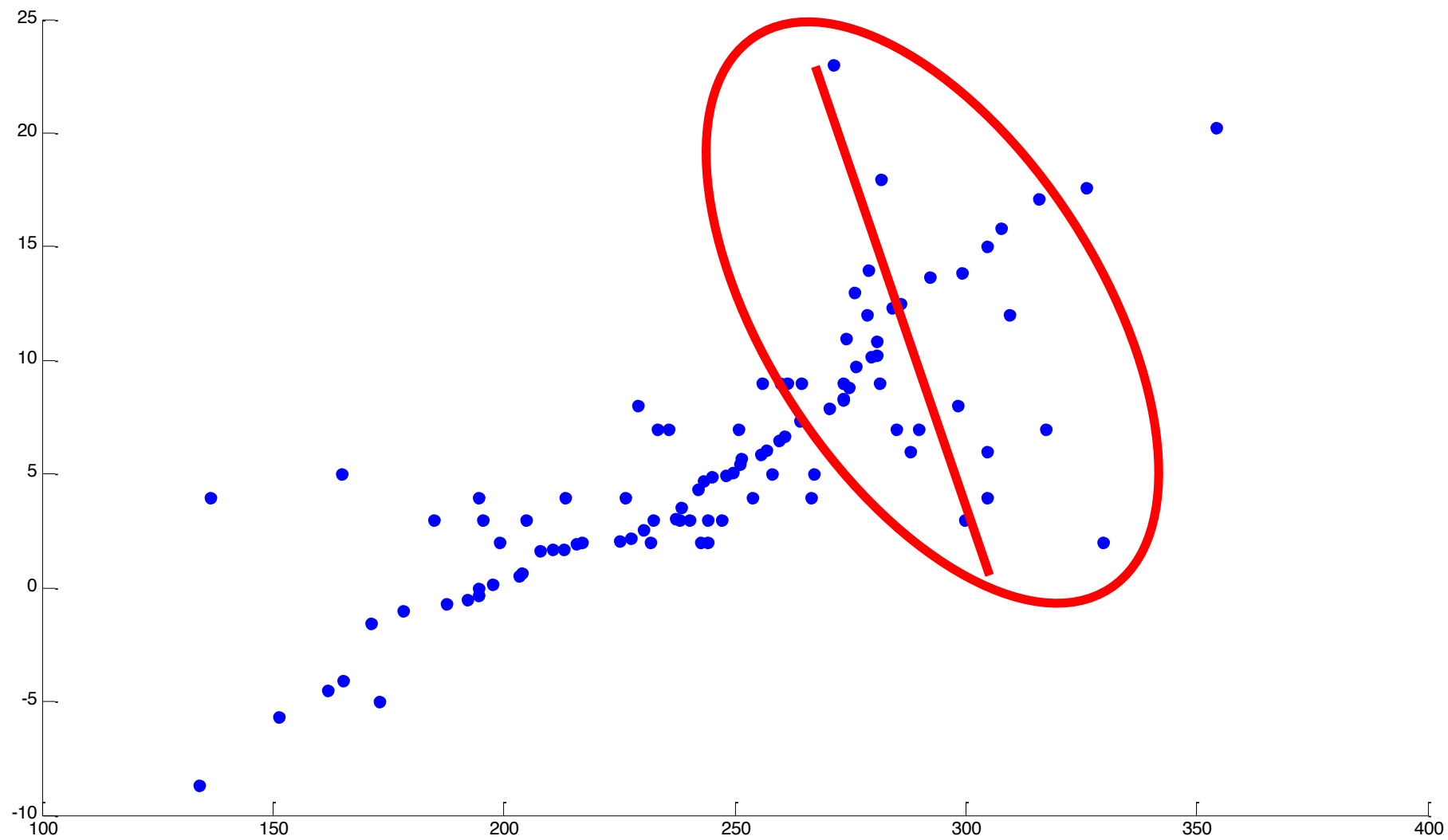
Curvilinear Correlation



Another Assumption

The group used to generate the prediction equation must reflect the population we want to predict





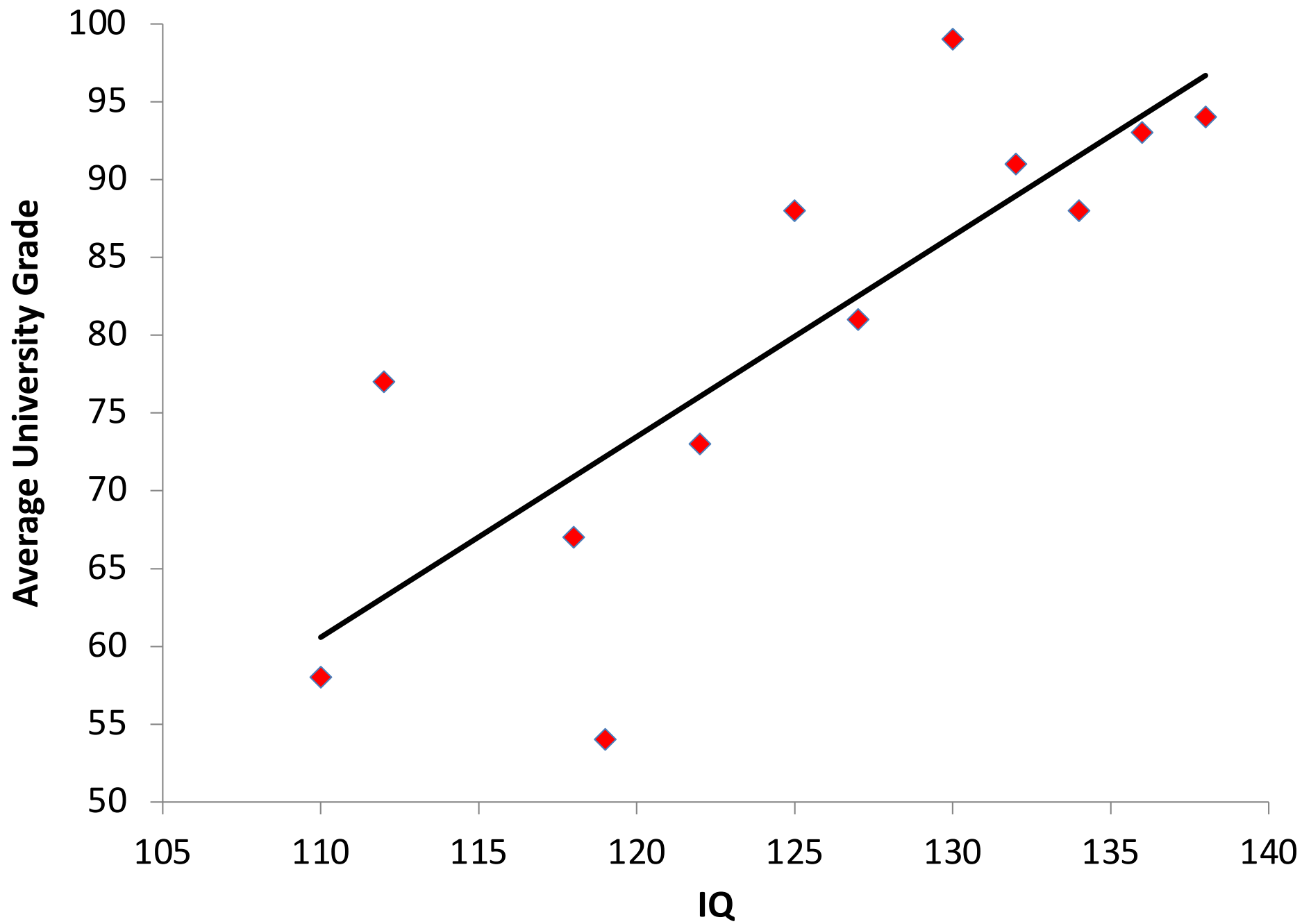
Multiple Regression

Regression

At heart, regression, be it simple or multiple, is all about prediction.

If we develop a model where we use “smoking” to predict “calories consumed” we are essentially saying we can use the number of cigarettes smoked in a day to predict the number of calories a person will consume.

Of course, this will not be a perfect prediction – it does not explain 100% of the variance. The regression will generate a “predicted score” and the difference between the “predicted score” and the “actual score” is the error in the regression model.



As you know...

The general equation for a simple regression is:

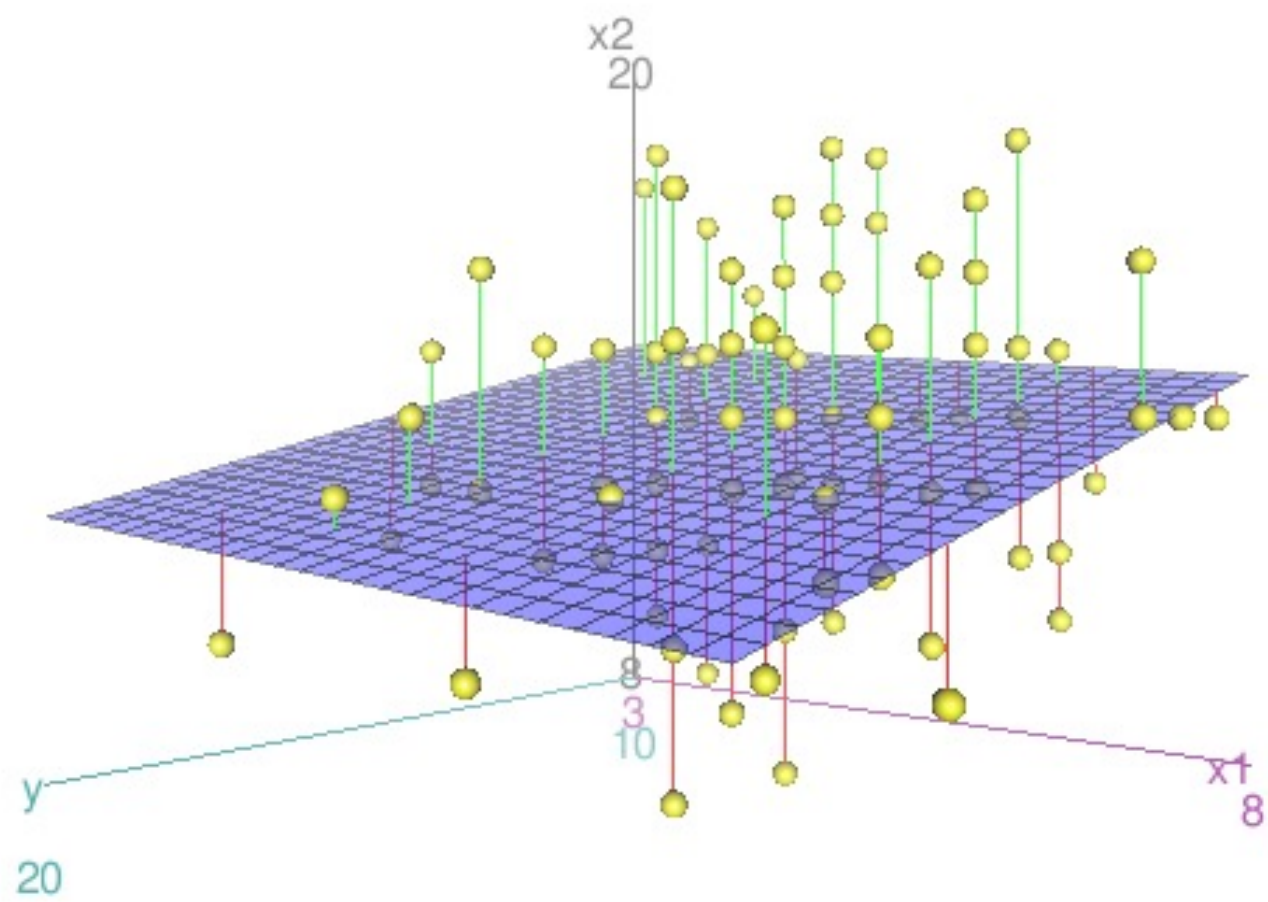
$$Y_i = B_0 + B_1X_1 + e_i$$

What does this mean?

You can also think of this as:

Output = Model + Error

Multiple Regression



What is MR?

$$Y_i = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \dots + e_i$$


Why we use Multiple Regression

1. Determine the degree of the relationship
2. Importance of individual IVs
3. Effect of adding or removing IVs
4. Effect of changing IVs
5. Relationships among IVs
6. Comparing sets of IVs
7. Predicting DV scores
8. Determining the model parameters

What is the output, and what does it mean?

The Output

Model Summary



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.231 ^a	.053	.042	2.026

a. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

What does R mean?

R is the multiple correlation coefficient – the fit of the model itself. Think of what R means in simple regression.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	113.914	6	18.986	4.626	.000 ^b
	Residual	2023.164	493	4.104		
	Total	2137.078	499			

a. Dependent Variable: Happiness

b. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger


Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.718	1.221		7.957	.000
	Anger	.014	.086	.007	.158	.875
	Spiritualism	-.021	.045	-.020	-.463	.644
	Income	.002	.005	.024	.535	.593
	Holidays	.110	.047	.103	2.353	.019
	Debt	.000	.001	-.014	-.328	.743
	Assets	.015	.003	.203	4.620	.000

a. Dependent Variable: Happiness

The Output

Model Summary



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.231 ^a	.053	.042	2.026

a. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

What does R^2 mean?

Yes, it is the proportion of explained variance. BUT, it thus is a measure of the goodness of fit of the model! A high R^2 value means your model fits the data – there is not a lot of error.

$$R^2 = SS_{\text{model}} / SS_{\text{total}}$$

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	113.914	6	18.986	4.626	.000 ^b
	Residual	2023.164	493	4.104		
	Total	2137.078	499			

a. Dependent Variable: Happiness

b. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger


Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.718	1.221		7.957	.000
	Anger	.014	.086	.007	.158	.875
	Spiritualism	-.021	.045	-.020	-.463	.644
	Income	.002	.005	.024	.535	.593
	Holidays	.110	.047	.103	2.353	.019
	Debt	.000	.001	-.014	-.328	.743
	Assets	.015	.003	.203	4.620	.000

a. Dependent Variable: Happiness

The Output

Model Summary



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.231 ^a	.053	.042	2.026

a. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

What does Adjusted R² mean?

Adjusted R² is simply a corrected value. For instance, as you add more variables to a model, there is an increasing effect of chance. Thus, the adjusted R² value reflects a more “honest” statement of the actual amount of accounted variance.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	113.914	6	18.986	4.626	.000 ^b
	Residual	2023.164	493	4.104		
	Total	2137.078	499			

a. Dependent Variable: Happiness

b. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger


Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.718	1.221		7.957	.000
	Anger	.014	.086	.007	.158	.875
	Spiritualism	-.021	.045	-.020	-.463	.644
	Income	.002	.005	.024	.535	.593
	Holidays	.110	.047	.103	2.353	.019
	Debt	.000	.001	-.014	-.328	.743
	Assets	.015	.003	.203	4.620	.000

a. Dependent Variable: Happiness

The Output

Model Summary



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.231 ^a	.053	.042	2.026

a. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	113.914	6	18.986	4.626	.000 ^b
	Residual	2023.164	493	4.104		
	Total	2137.078	499			

a. Dependent Variable: Happiness

b. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.718	1.221		7.957	.000
	Anger	.014	.086	.007	.158	.875
	Spiritualism	-.021	.045	-.020	-.463	.644
	Income	.002	.005	.024	.535	.593
	Holidays	.110	.047	.103	2.353	.019
	Debt	.000	.001	-.014	-.328	.743
	Assets	.015	.003	.203	4.620	.000


a. Dependent Variable: Happiness

The F Test

Is there a valid linear model?

Interpreting Regression Coefficients

Model Summary



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.231 ^a	.053	.042	2.026

a. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	113.914	6	18.986	4.626	.000 ^b
	Residual	2023.164	493	4.104		
	Total	2137.078	499			

a. Dependent Variable: Happiness

b. Predictors: (Constant), Assets, Holidays, Debt, Spiritualism, Income, Anger

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.718	1.221		7.957	.000
	Anger	.014	.086	.007	.158	.875
	Spiritualism	-.021	.045	-.020	-.463	.644
	Income	.002	.005	.024	.535	.593
	Holidays	.110	.047	.103	2.353	.019
	Debt	.000	.001	-.014	-.328	.743
	Assets	.015	.003	.203	4.620	.000

a. Dependent Variable: Happiness

Interpreting the Coefficients

What the coefficient is really telling us, is that there will be # unit amount of change in Y' for every unit change of the predictor variable if we hold the other predictor variables constant.

These slopes then of course can be tested significantly.

The coefficients are sometimes termed “partial slopes”

Interpreting the Coefficients

Note, that it is possible that a regression coefficient can be significant if tested on its own, but not as part of a larger regression – and vice versa.

It is extremely important that you do not draw too big of a conclusion of relative importance solely from these values – remember they reflect an effect derived from the regression as a whole.