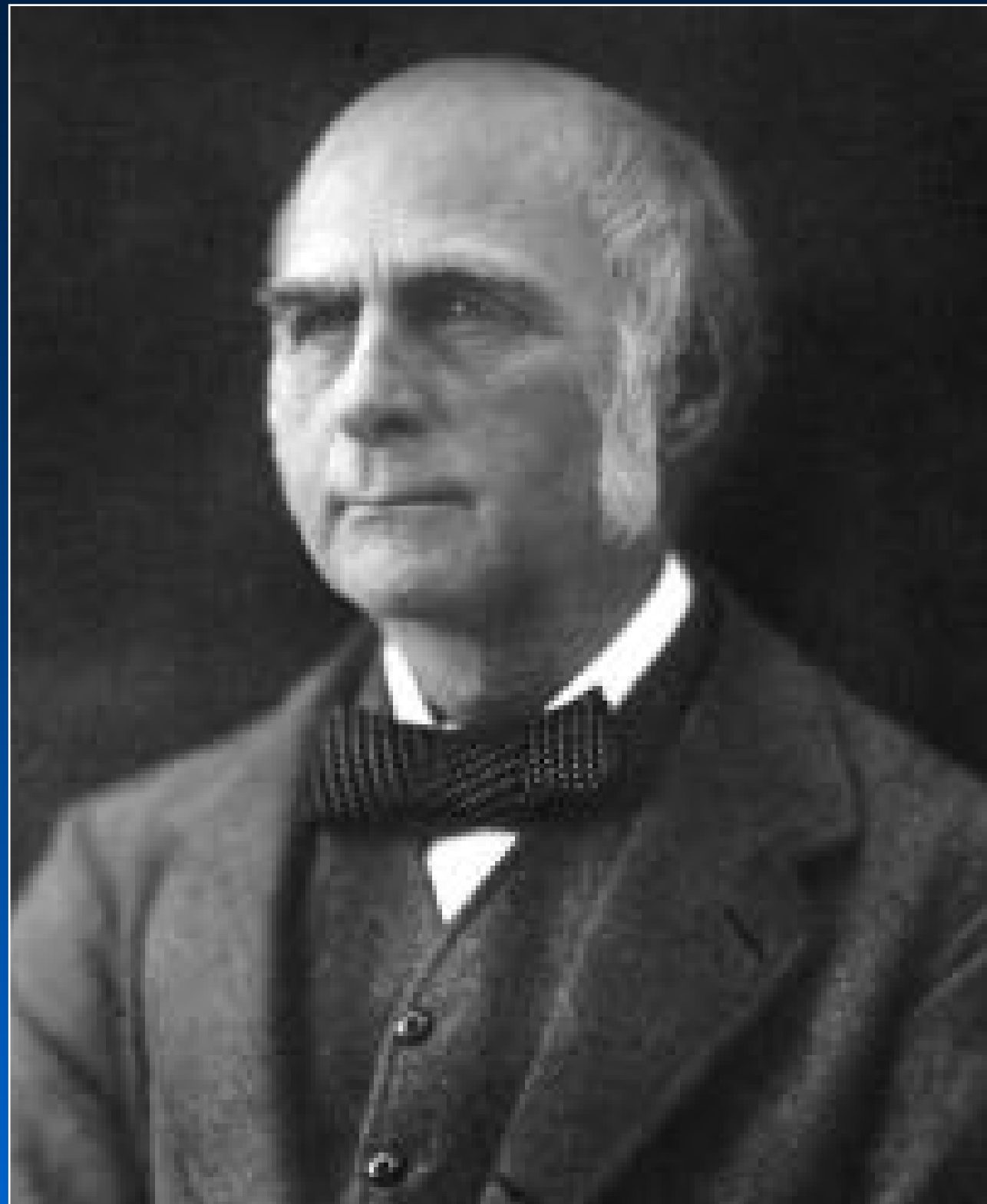
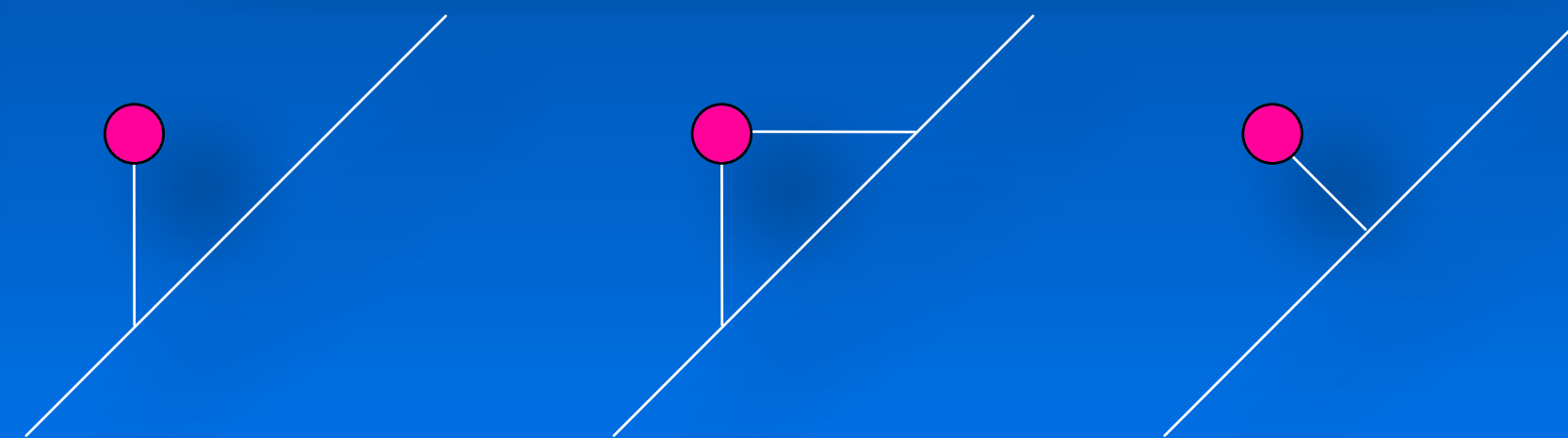
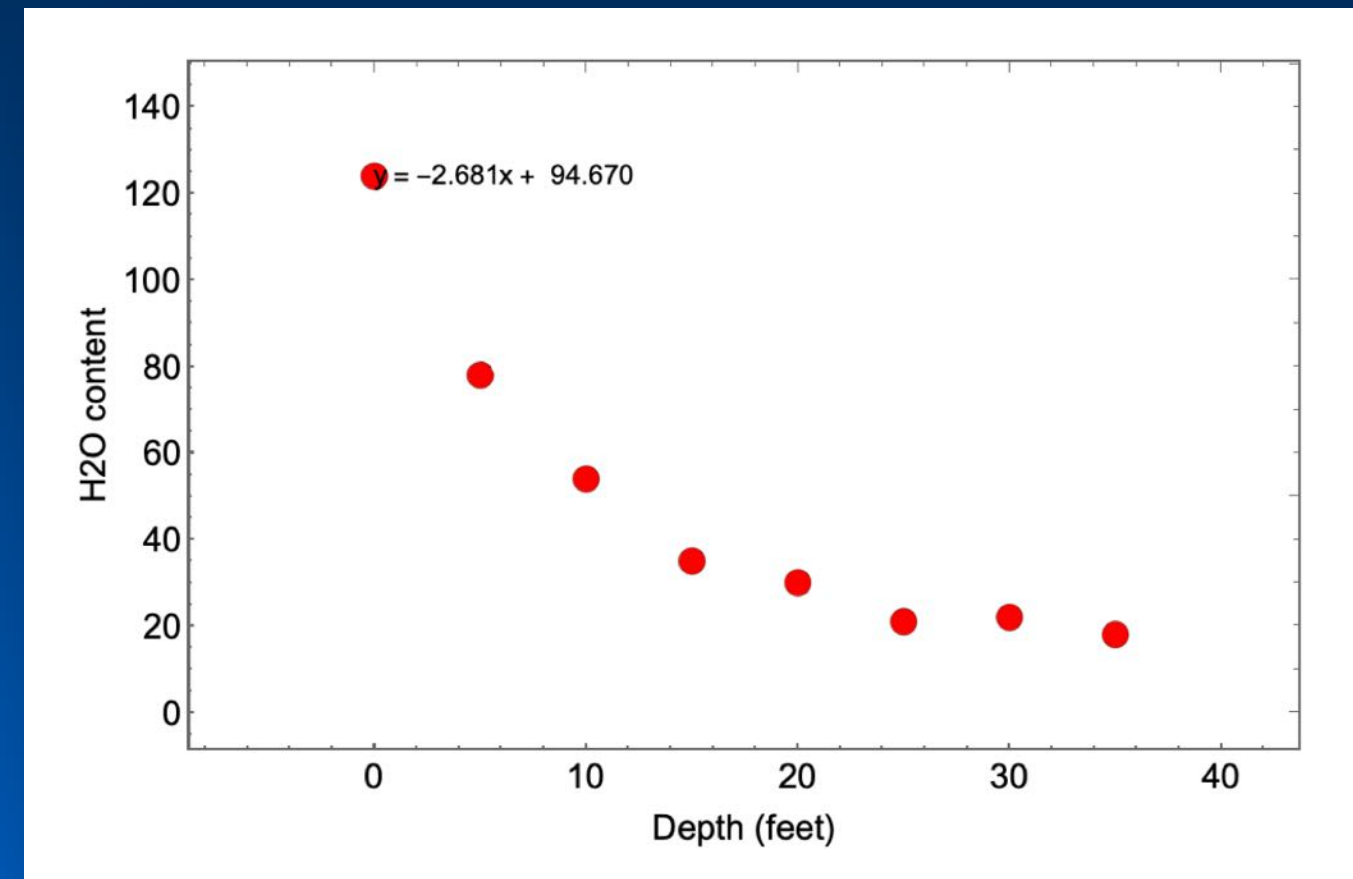


# Bivariate Regression, Modeling & Testing of Earth Science Data

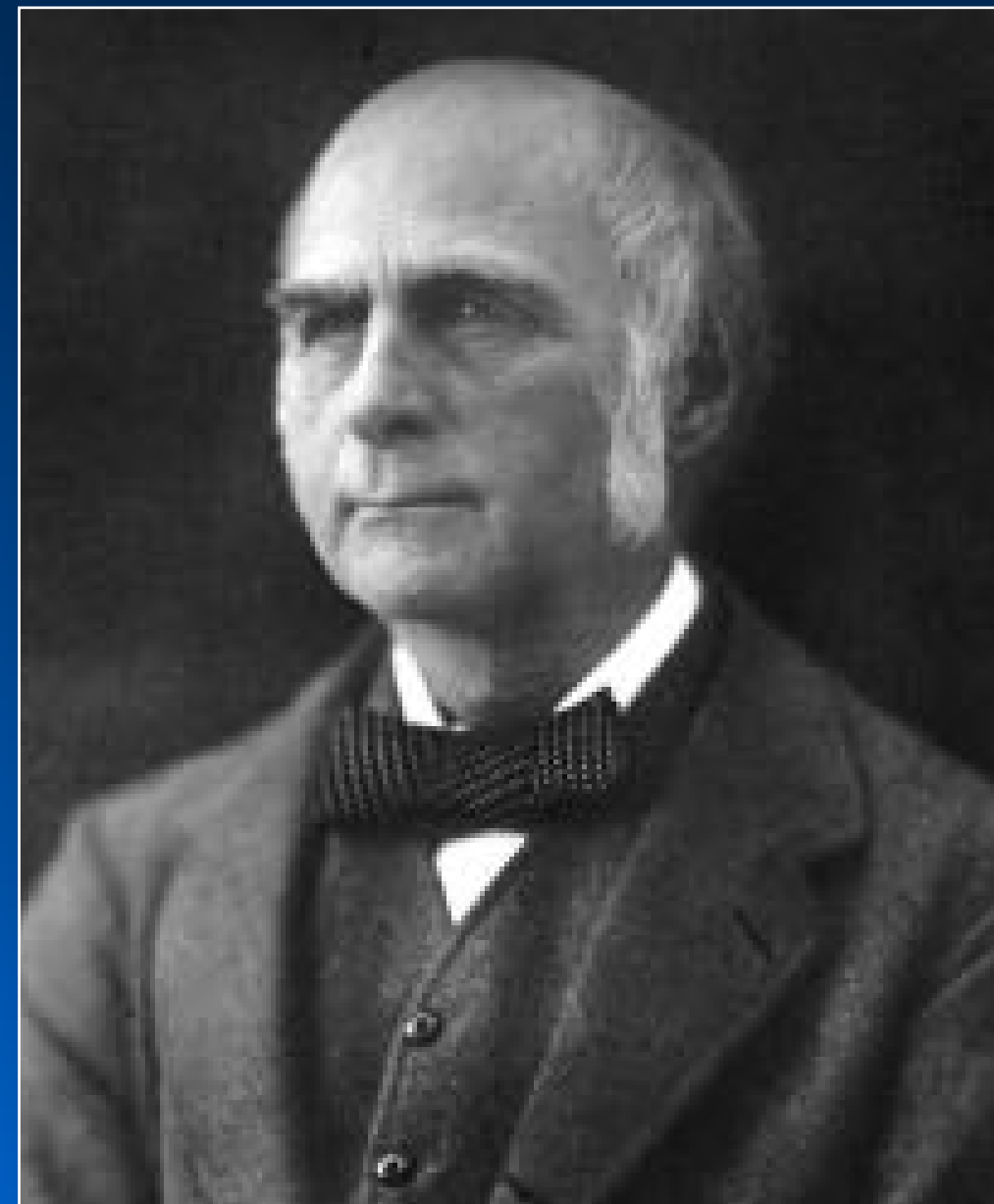


Prof. Norman MacLeod  
School of Earth Sciences & Engineering, Nanjing University

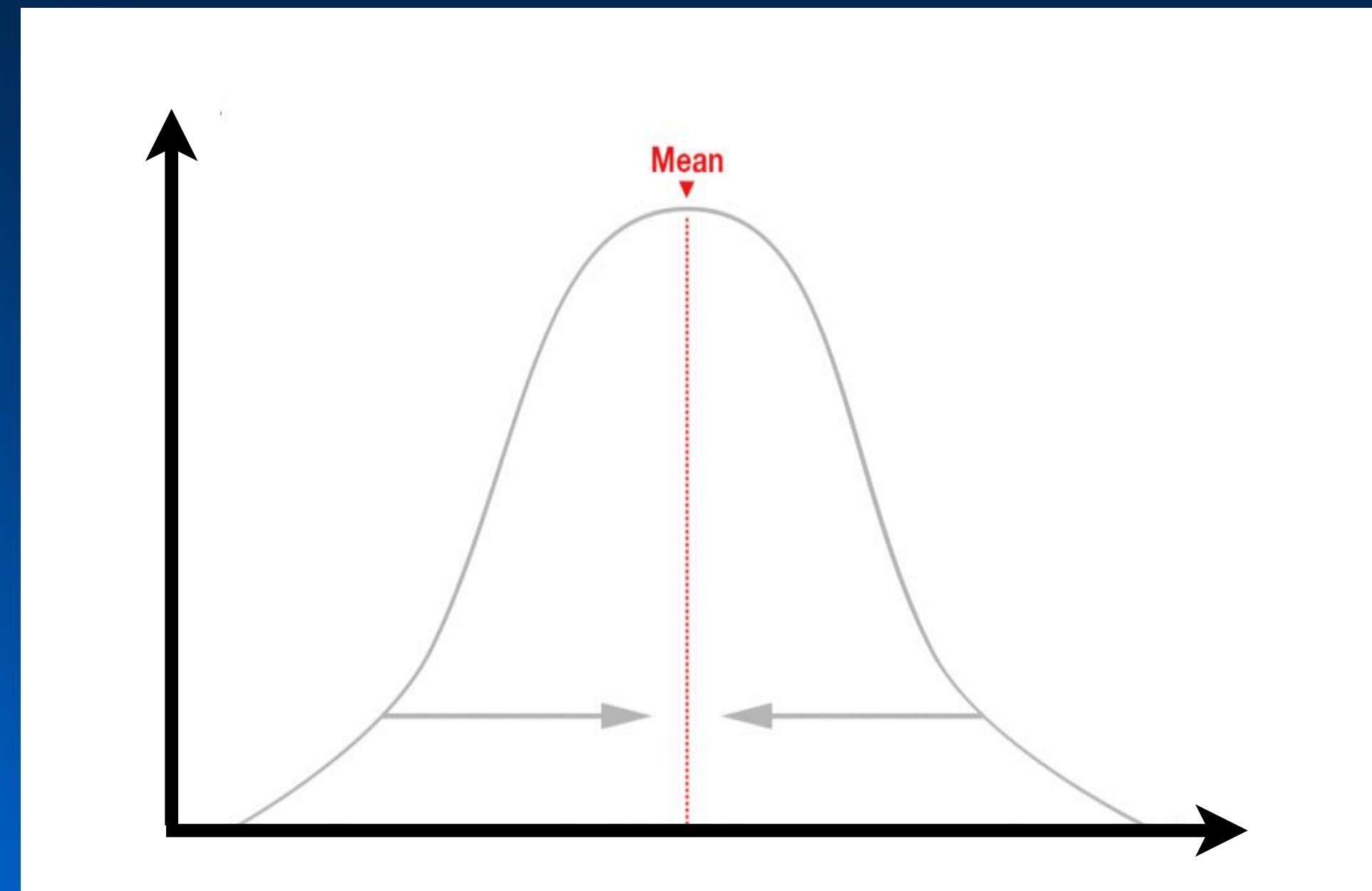


# Linear Regression, Modeling & Testing

**Regression Analysis** - The quantitative study of the manner in which variation in one (or more) variable(s) can be expressed in terms of variation in one (or more) other variable(s).



Francis Galton  
(1822 - 1911)

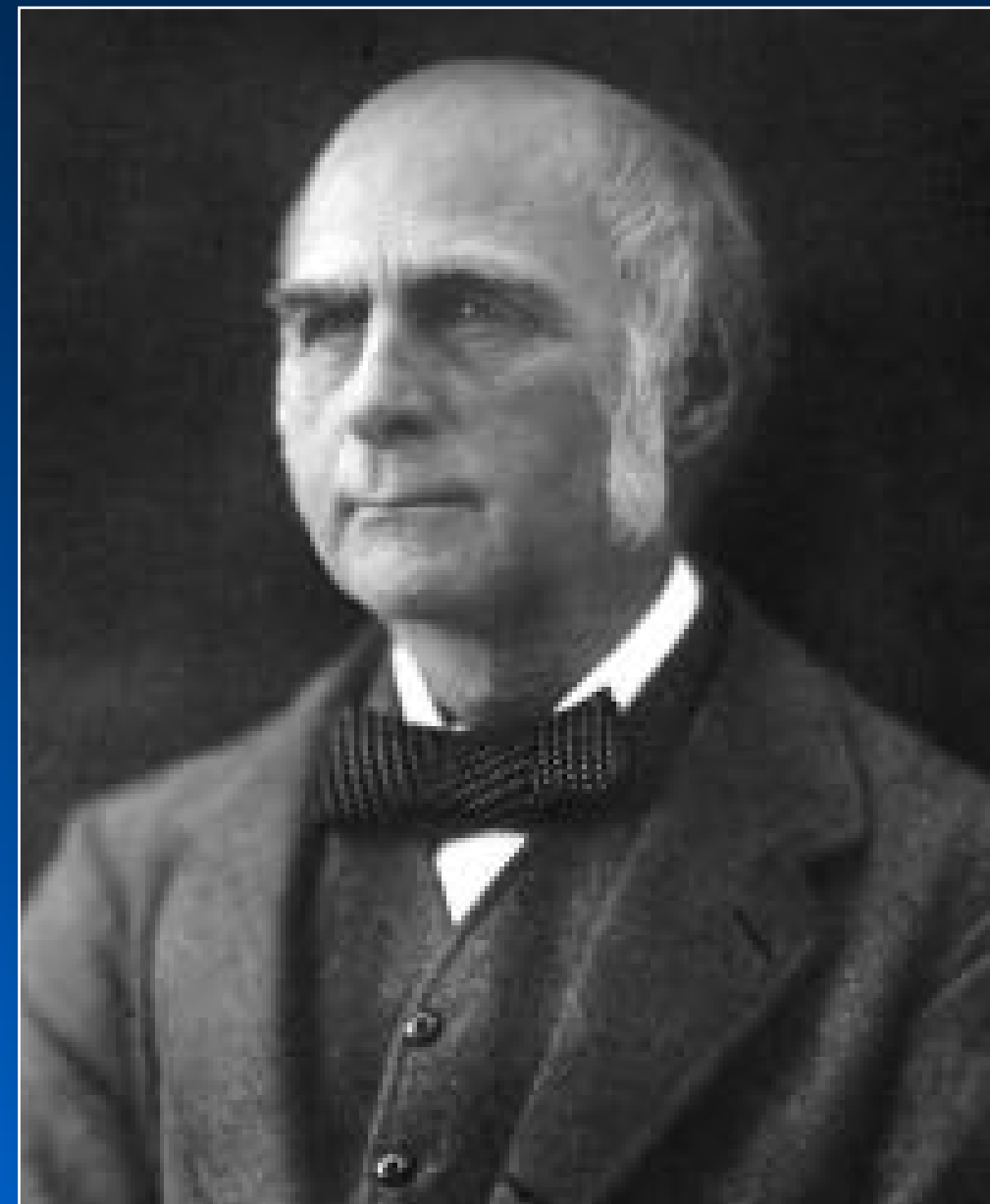


Regression to the Mean

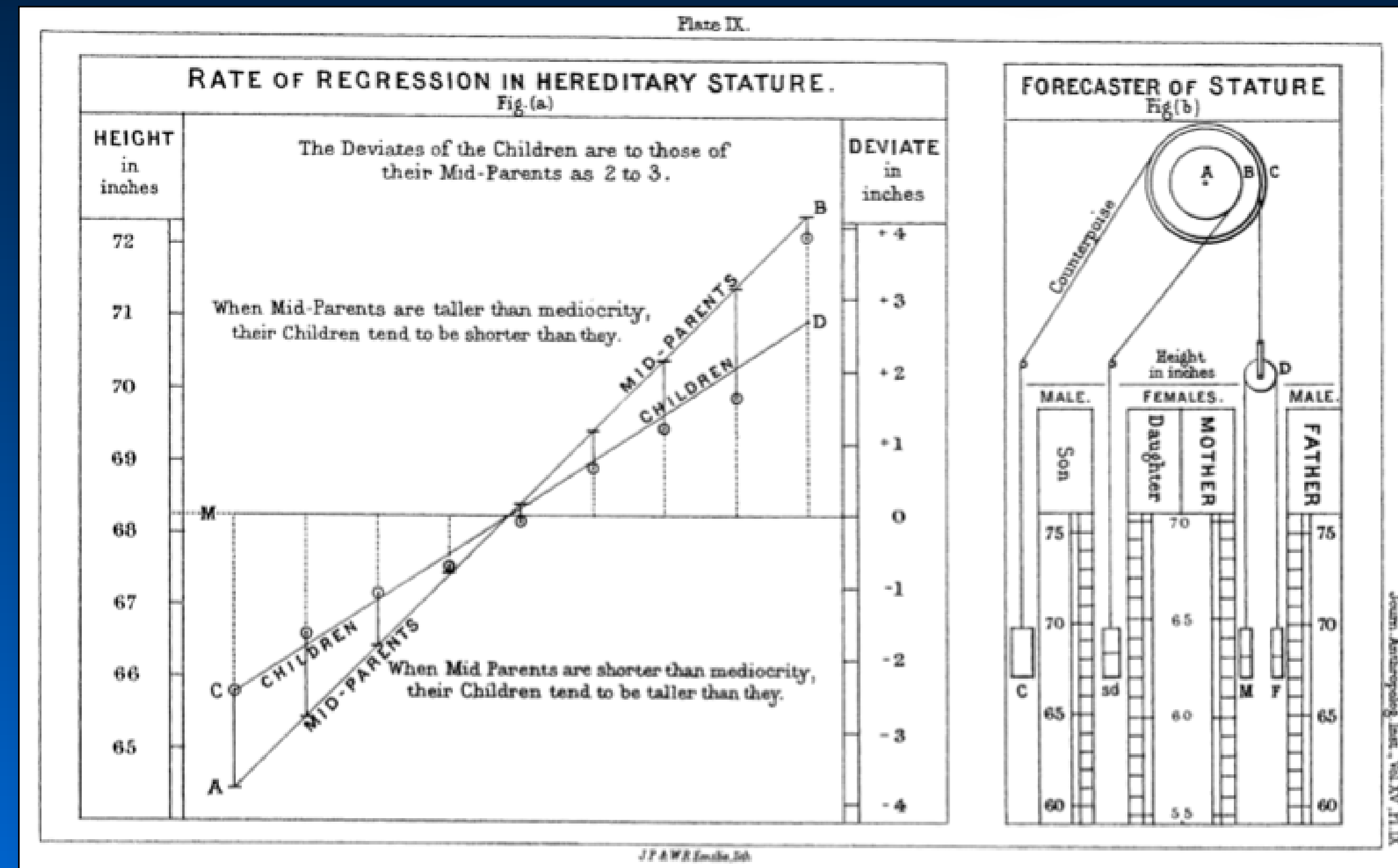
Developed originally by Galton to model heritability of physical dimensions between parents and children; a critical early demonstration of Darwin's principle of biological heredity.

# Linear Regression, Modeling & Testing

**Regression Analysis** - The quantitative study of the manner in which variation in one (or more) variable(s) can be expressed in terms of variation in one (or more) other variable(s).



Francis Galton  
(1822 - 1911)

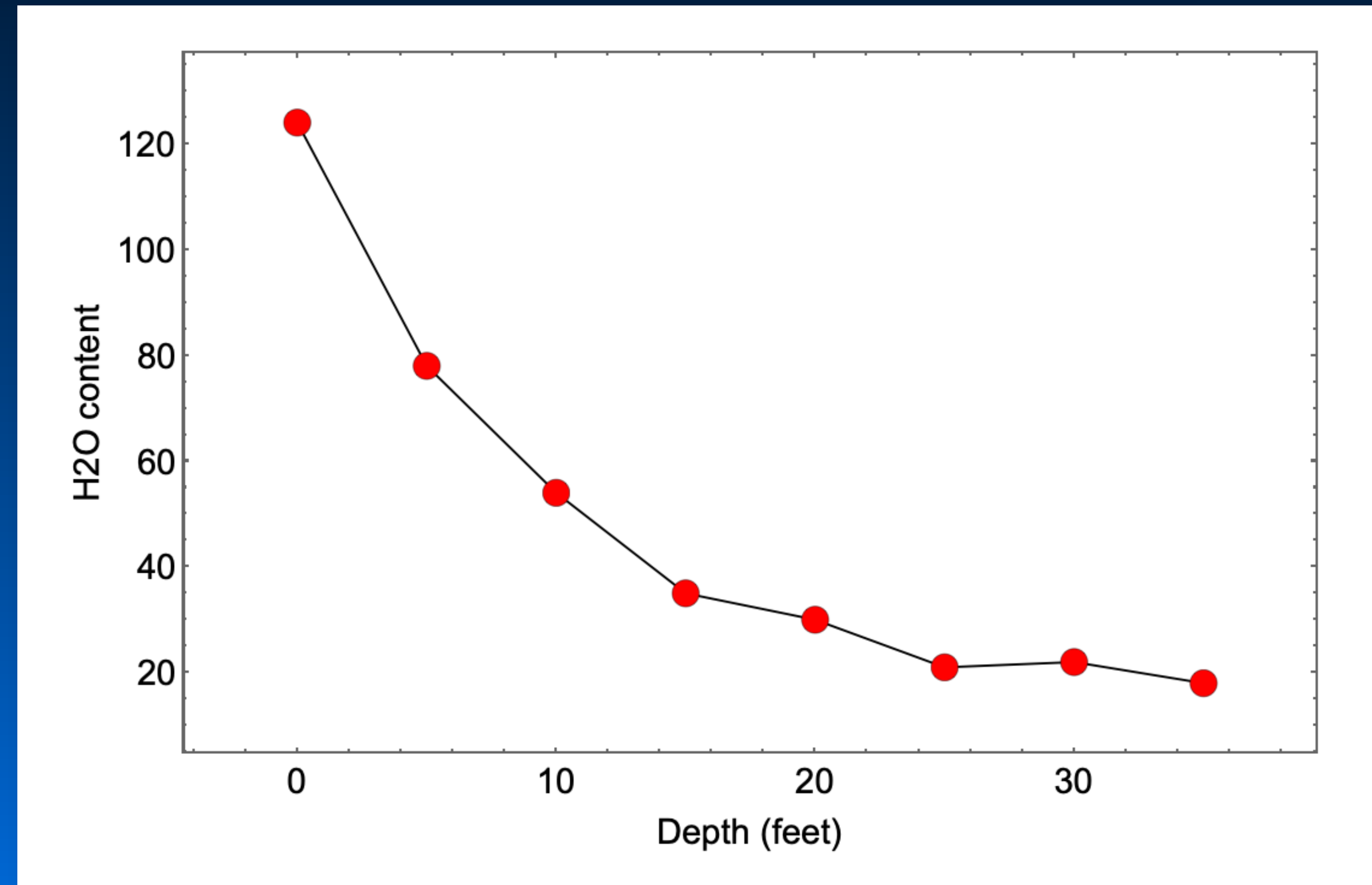


Developed originally by Galton to model heritability of physical dimensions between parents and children; a critical early demonstration of Darwin's principle of biological heredity.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

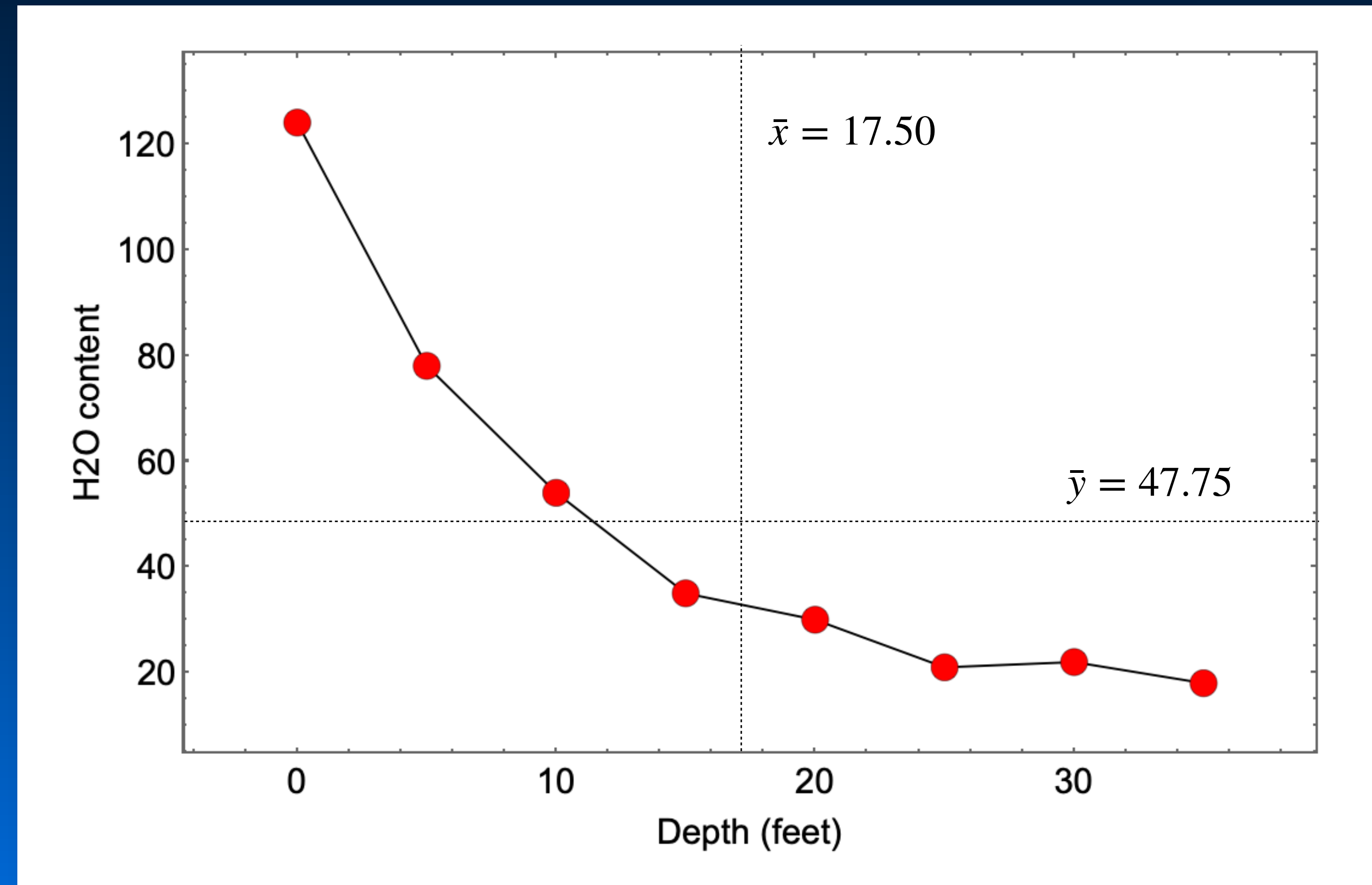


Obviously water content diminishes with depth, probably due to weight compaction. But can we model this relationship quantitatively and test its conformance to a strictly linear model?

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

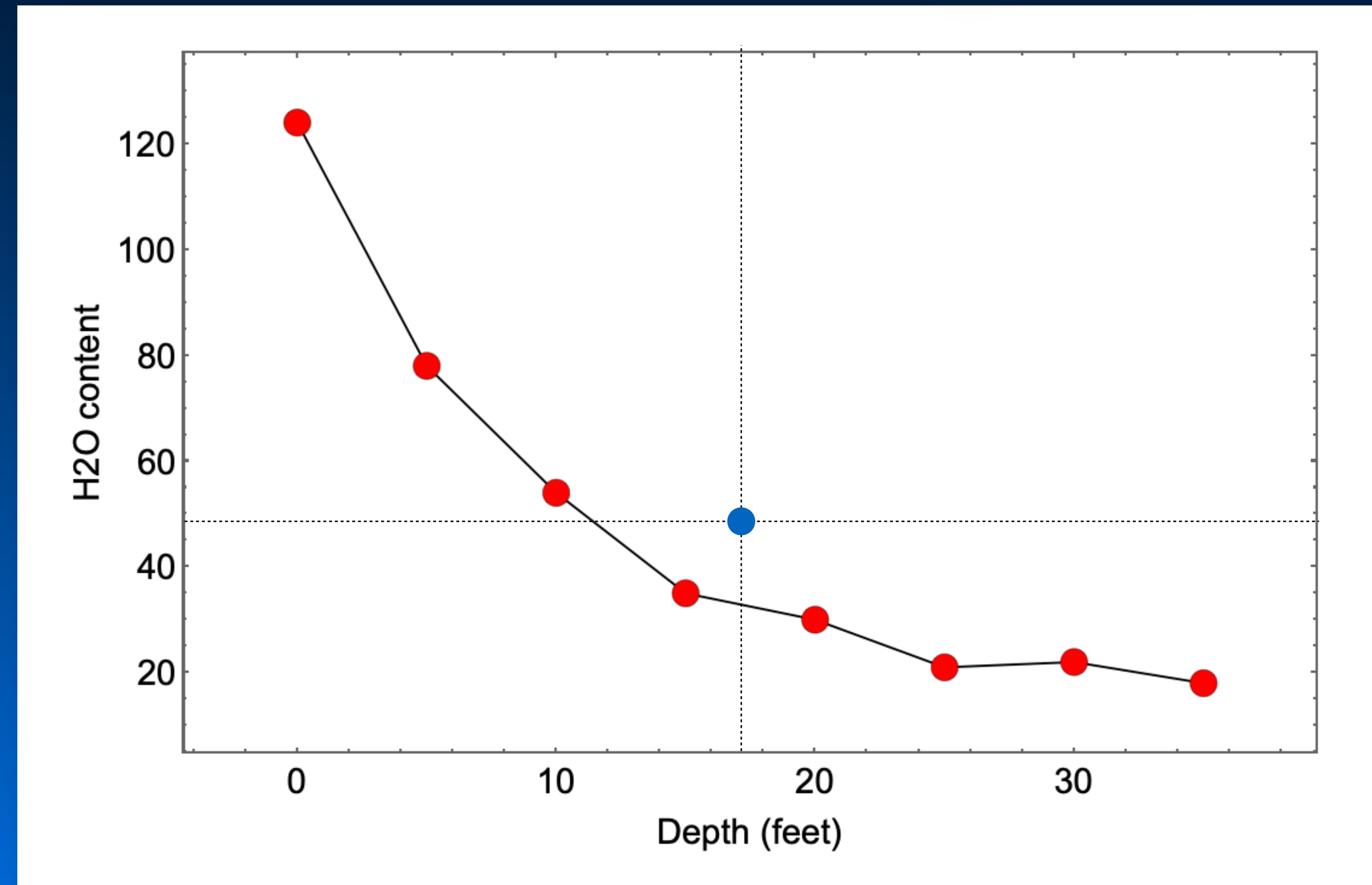


The first thing we might want to do is find a fixed point in the data. Since the arithmetic mean is a parameter we can use to locate the variable distributions relative to one another, that's a logical choice.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

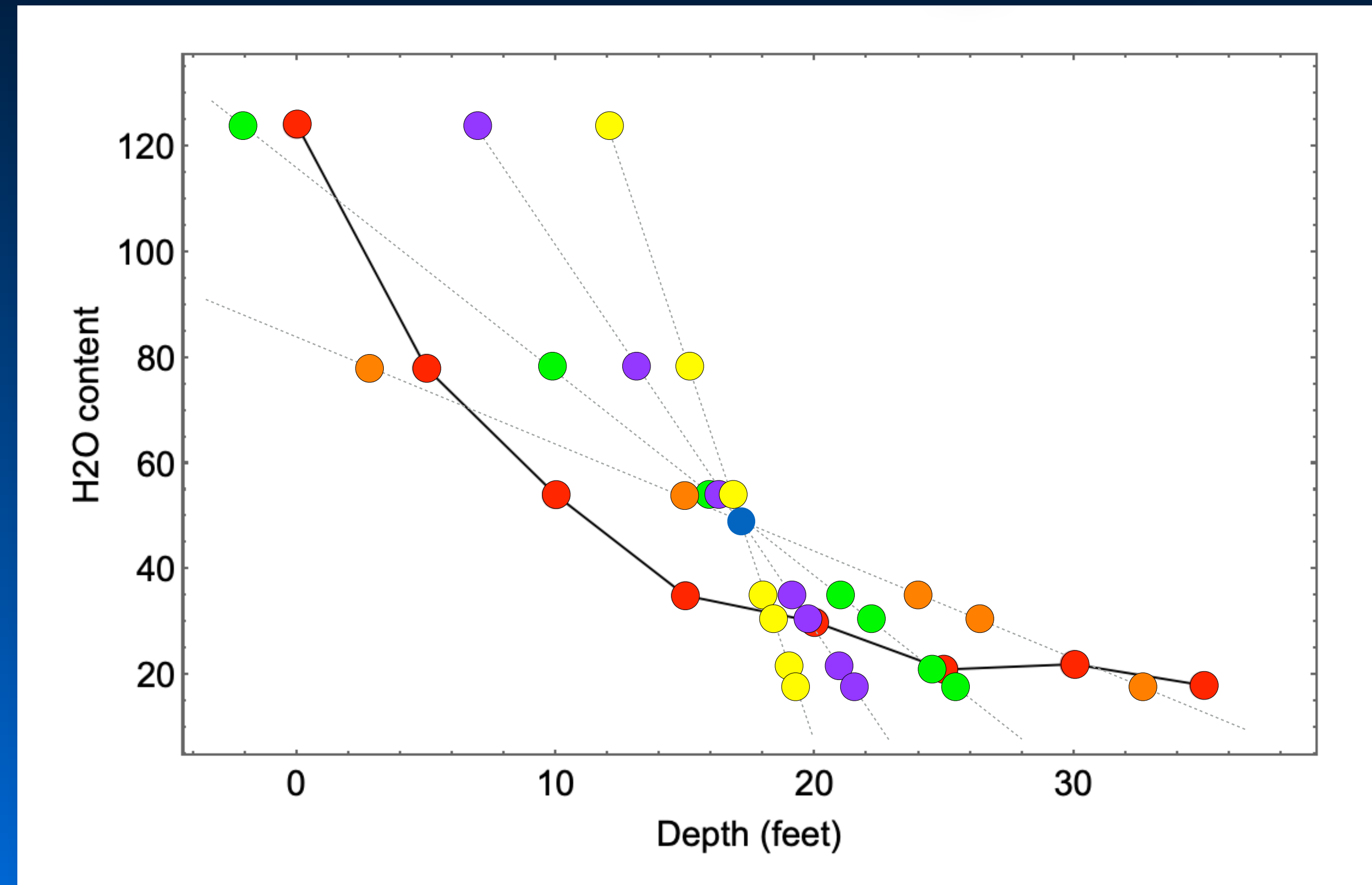


The first thing we might want to do is find a fixed point in the data. Since the arithmetic mean is a parameter we can use to locate the variable distributions relative to one another, that's a logical choice.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

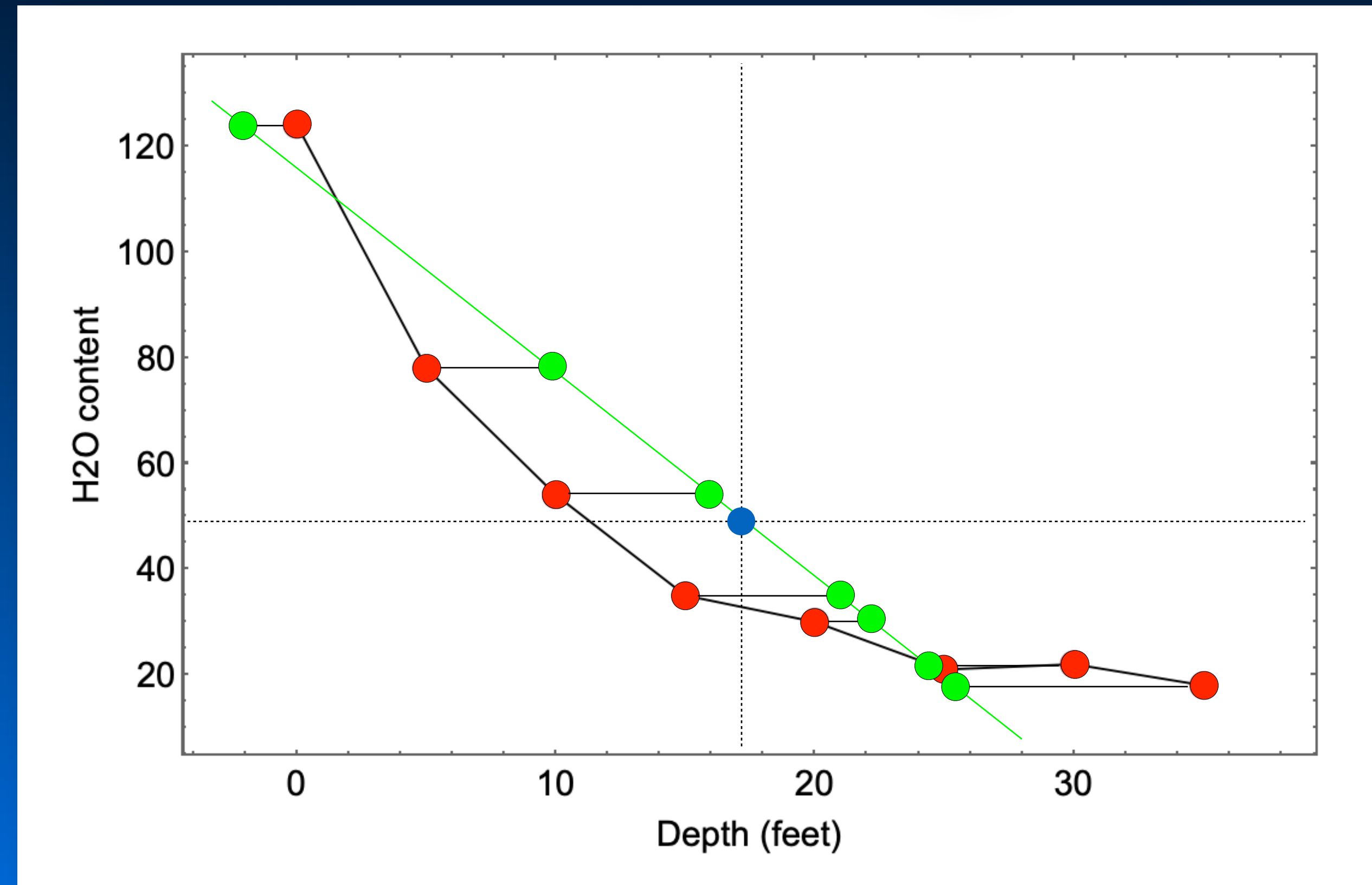


What we want is a straight line (=model) that minimizes its deviation from the data. We can find this by trial & error, and actually, this strategy has begun to be advocated.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

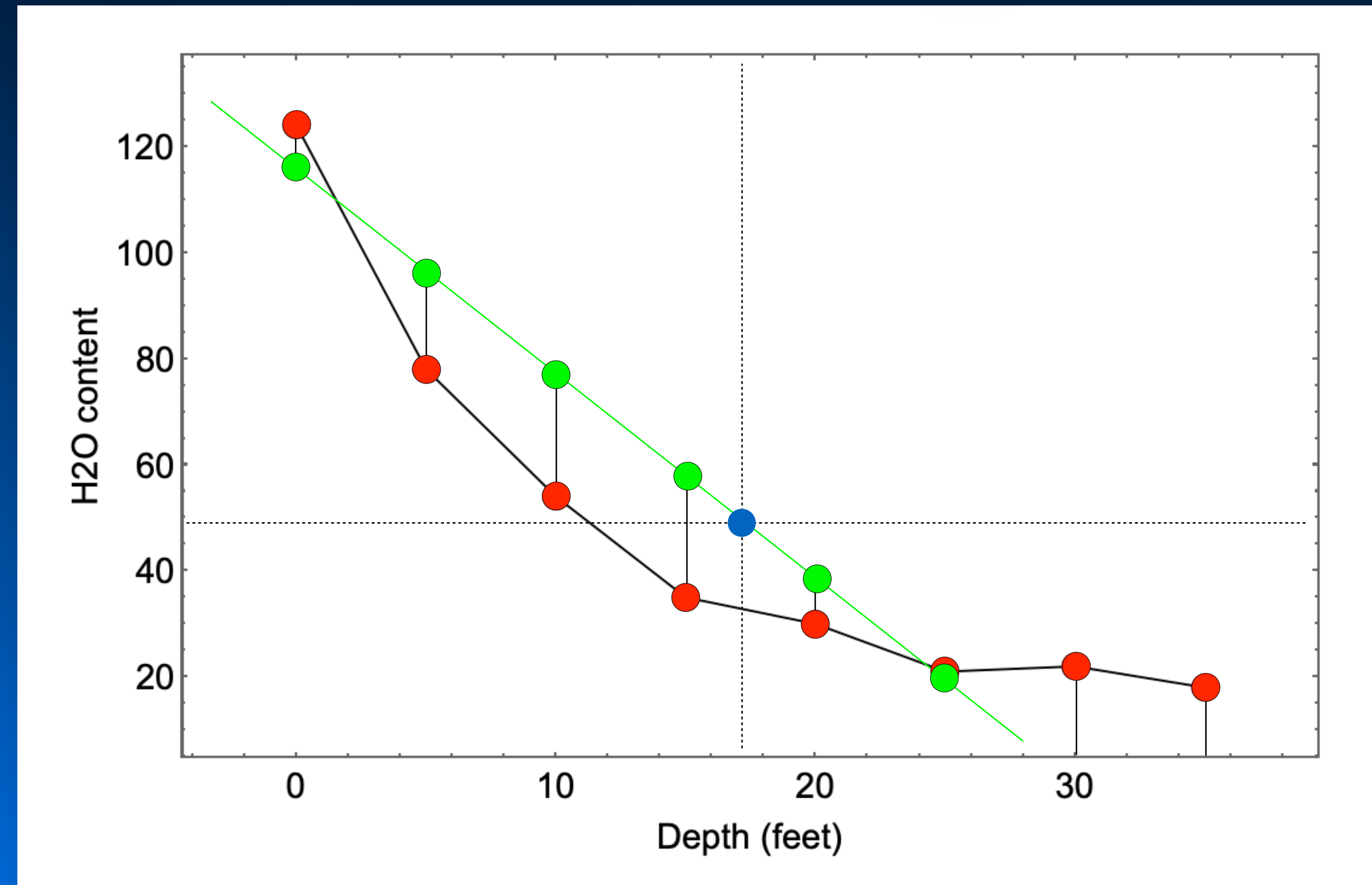


This model minimizes error in the  $x$ -value. It's called an "y on x" regression and it should be used when you want to estimate the values of  $x$  when you know the values of  $y$ .

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

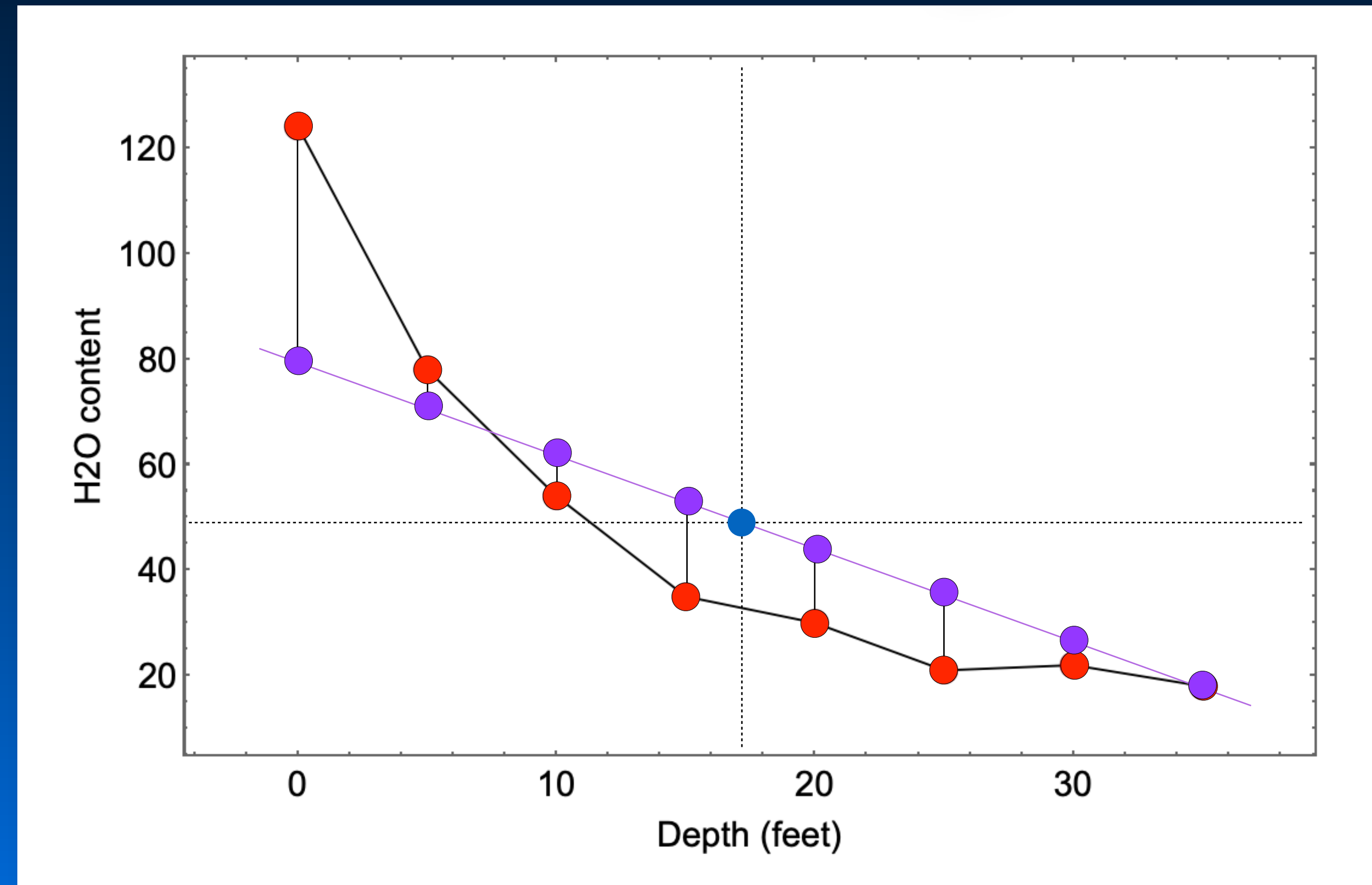


But this model doesn't seem to fit the data in the  $y$  direction very well.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

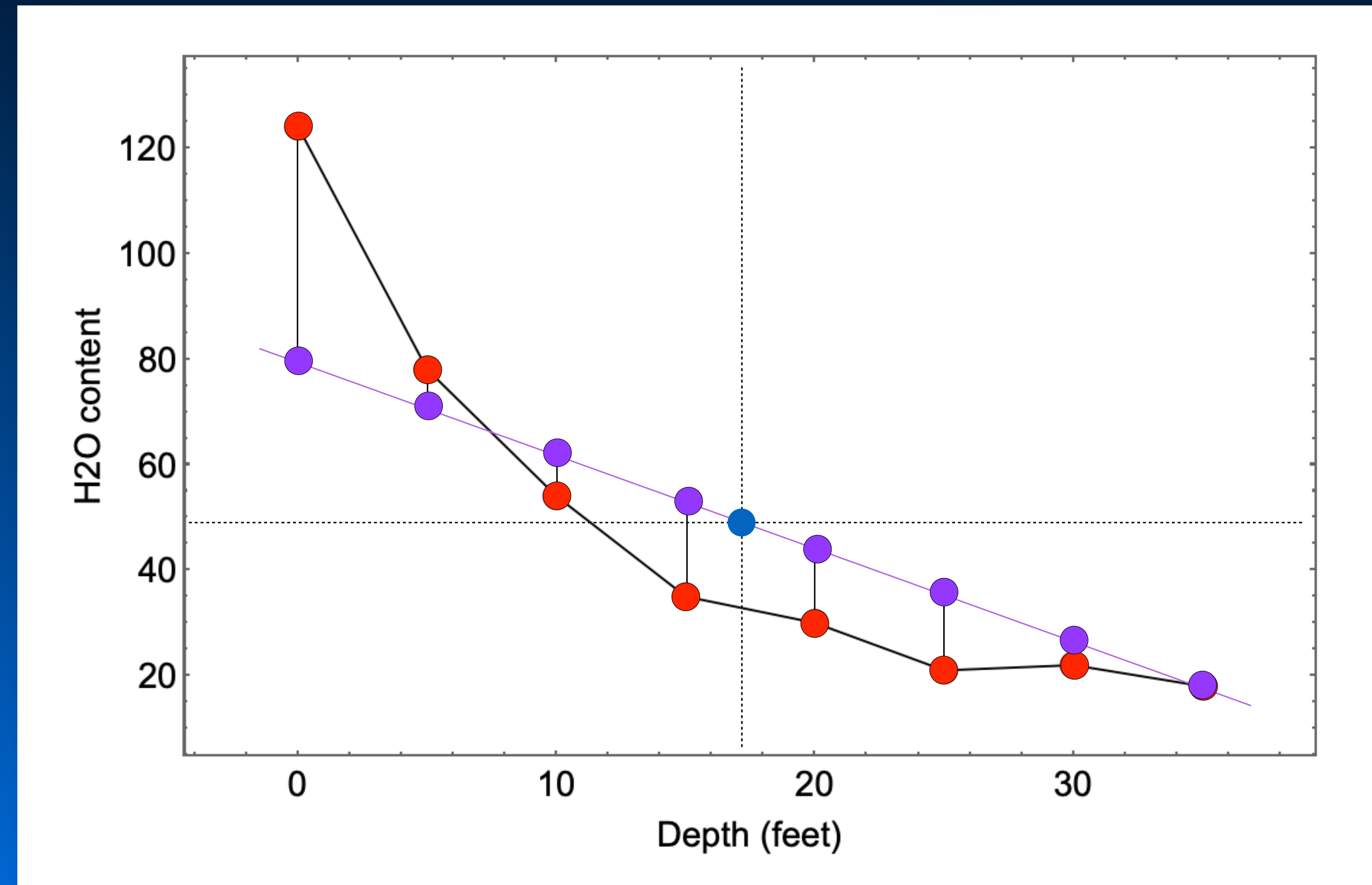


This model minimizes error in the  $y$ -value. It's called an "x on y" regression and it should be used when you want to estimate the values of  $y$  when you know the values of  $x$ .

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

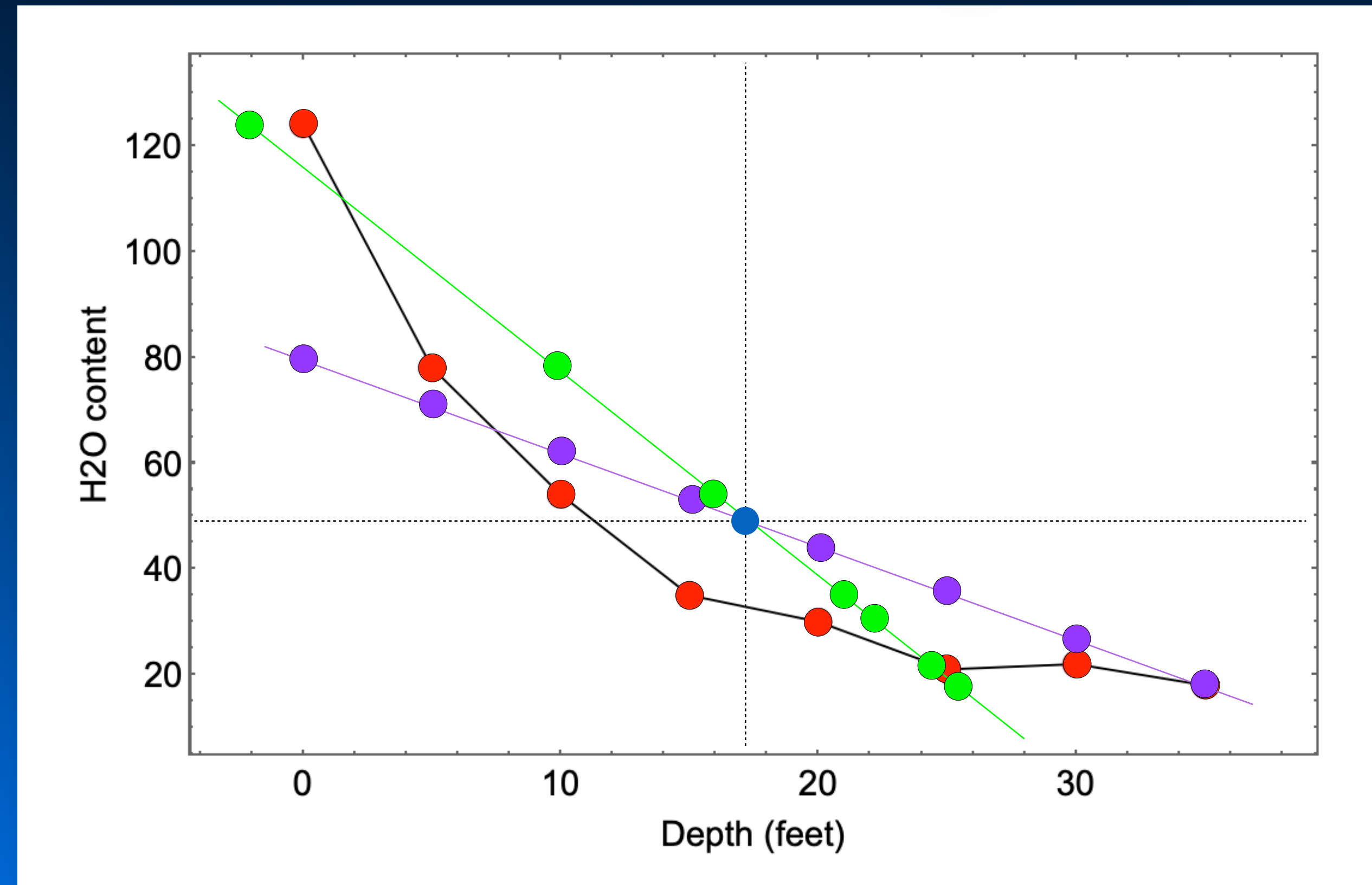


This orientation fits the data in the  $y$  direction better than the  $y$  on  $x$  model, but it doesn't fit the data in the  $x$  direction very well.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

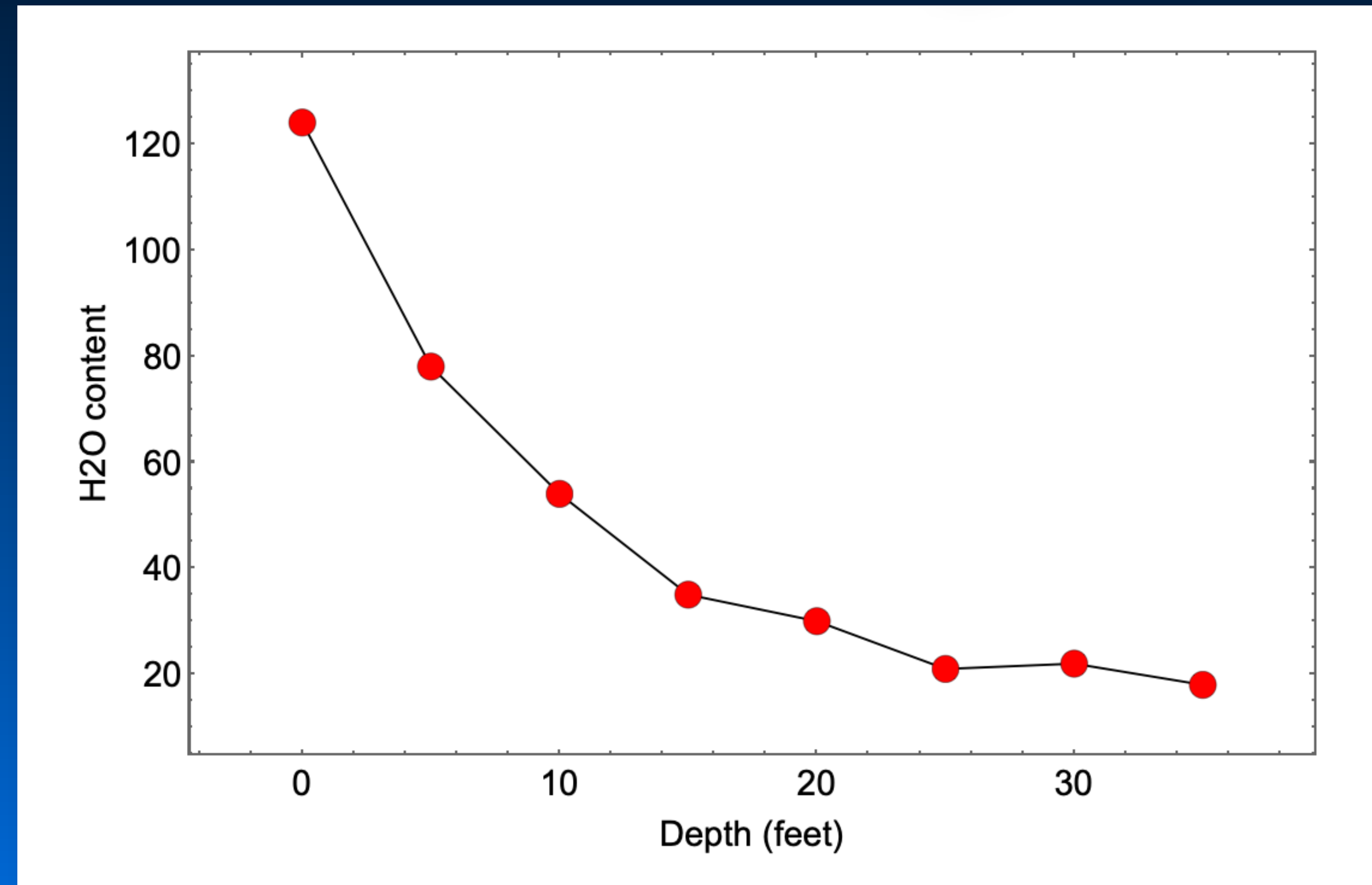


So, in order to choose between these two models we must decide which source of error we want to minimize.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |



**Independent variable** = variable you know (not interested in its error).

**Dependent variable** = variable you wish to estimate (minimize its error).

# Linear Regression, Modeling & Testing

## Ordinary Least-Squares Linear Regression

| i   | Depth | H <sub>2</sub> O Content | xy   | (x <sub>mean</sub> - x) <sup>2</sup> |
|-----|-------|--------------------------|------|--------------------------------------|
| 1   | 0     | 124                      | 0    | 306.25                               |
| 2   | 5     | 78                       | 390  | 156.25                               |
| 3   | 10    | 54                       | 540  | 56.25                                |
| 4   | 15    | 35                       | 525  | 6.25                                 |
| 5   | 20    | 30                       | 600  | 6.25                                 |
| 6   | 25    | 21                       | 525  | 56.25                                |
| 7   | 30    | 22                       | 660  | 156.25                               |
| 8   | 35    | 18                       | 630  | 306.25                               |
| Σ   | 140   | 382                      | 3870 | 1050                                 |
| Σ/n | 17.5  | 47.75                    |      |                                      |

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Where:  $i$  = observation no.

$\alpha$  = y-intercept

$\beta$  = slope

$\epsilon$  = error (residual)

$$\beta = \frac{\sum_{i=1}^n x_i y_i - ((\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i) / n)}{\sum_{i=1}^n (\bar{x} - x_i)^2}$$

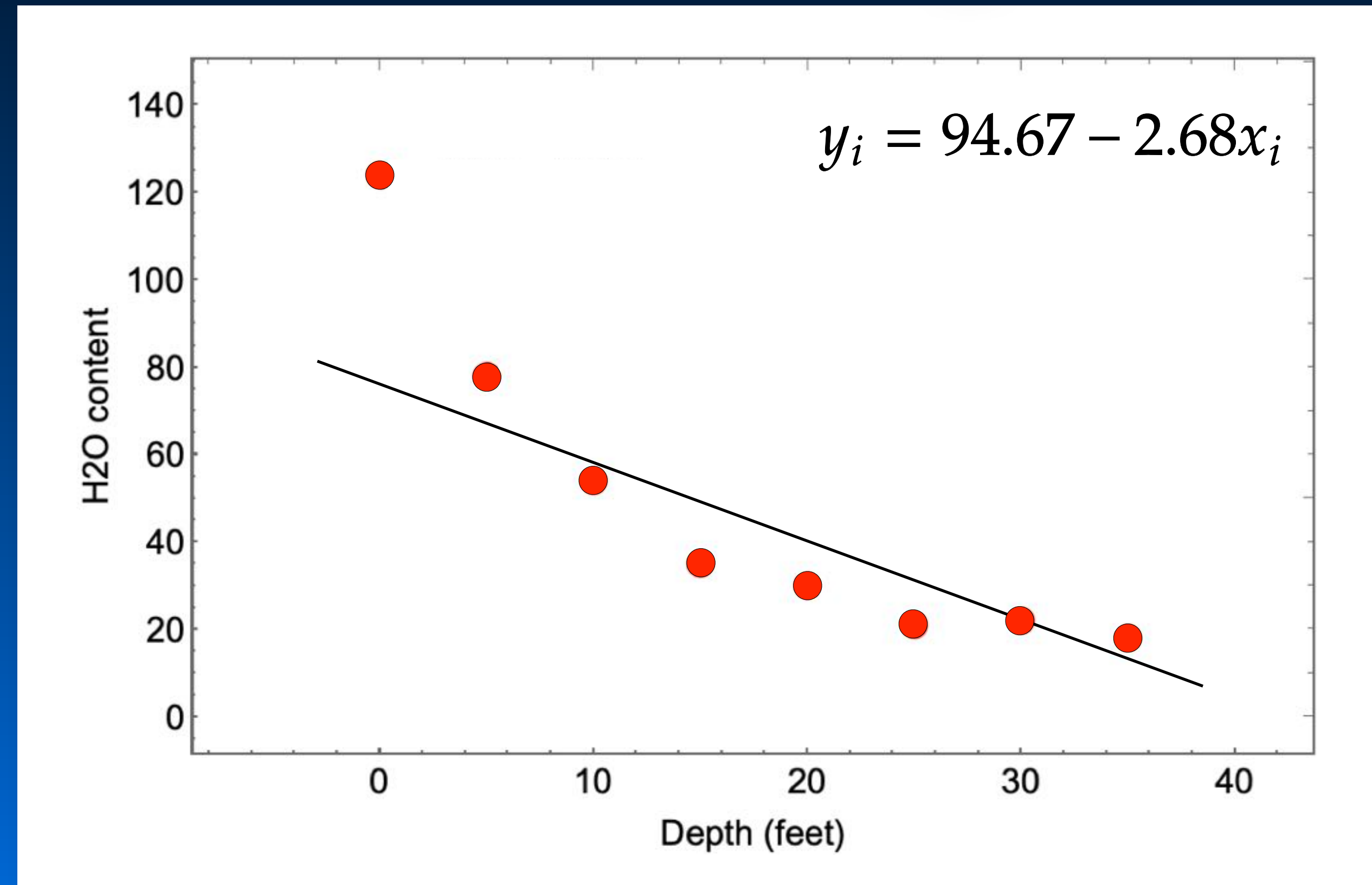
$$\alpha = \bar{y} - (\beta \cdot \bar{x})$$

When  $n$  = total number of observations

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

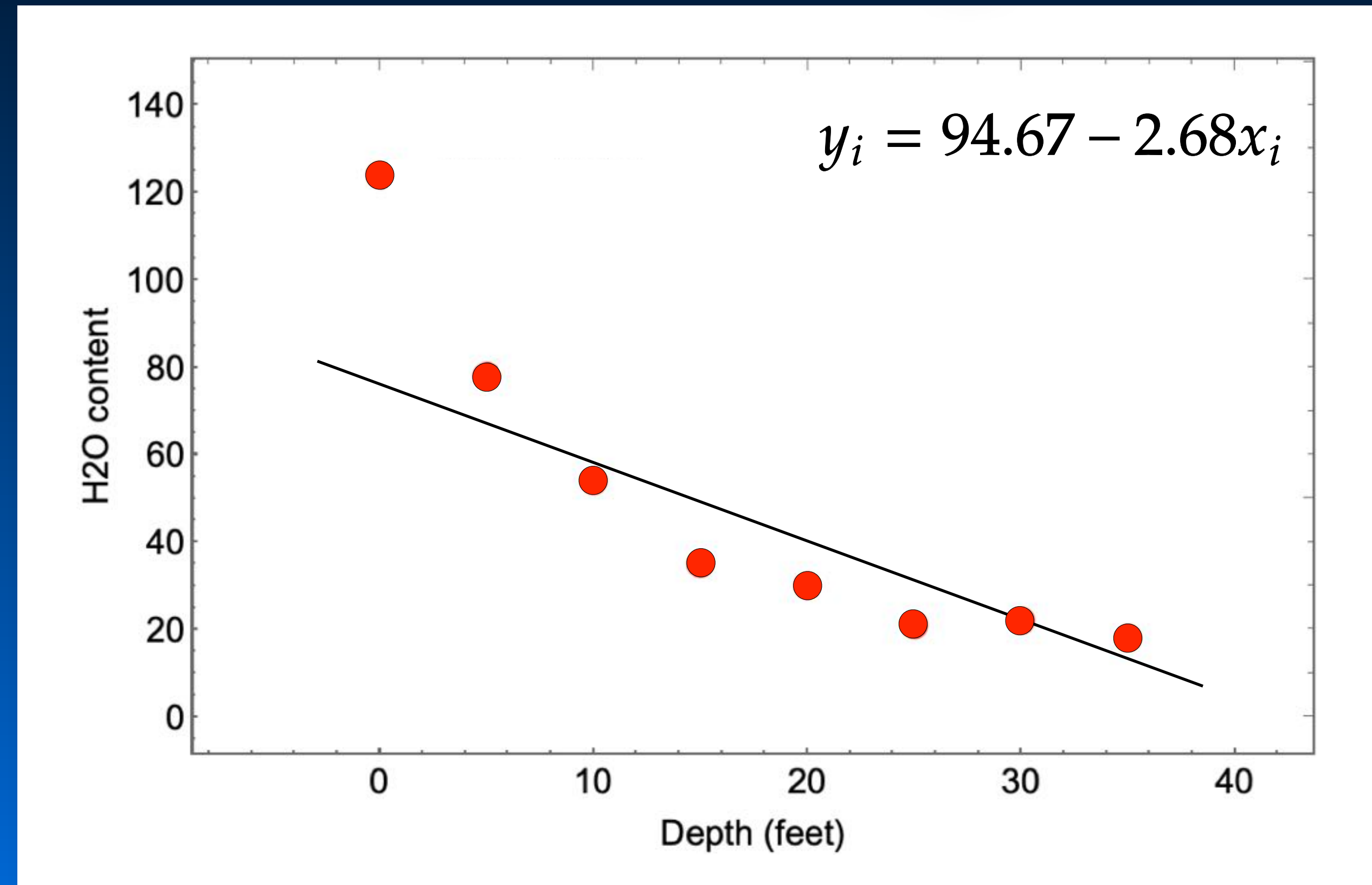


Independent variable =  $x$  variable. Dependent variable =  $y$  variable.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |



How well does this line (= linear model) fit these data? Is this fit significant statistically?

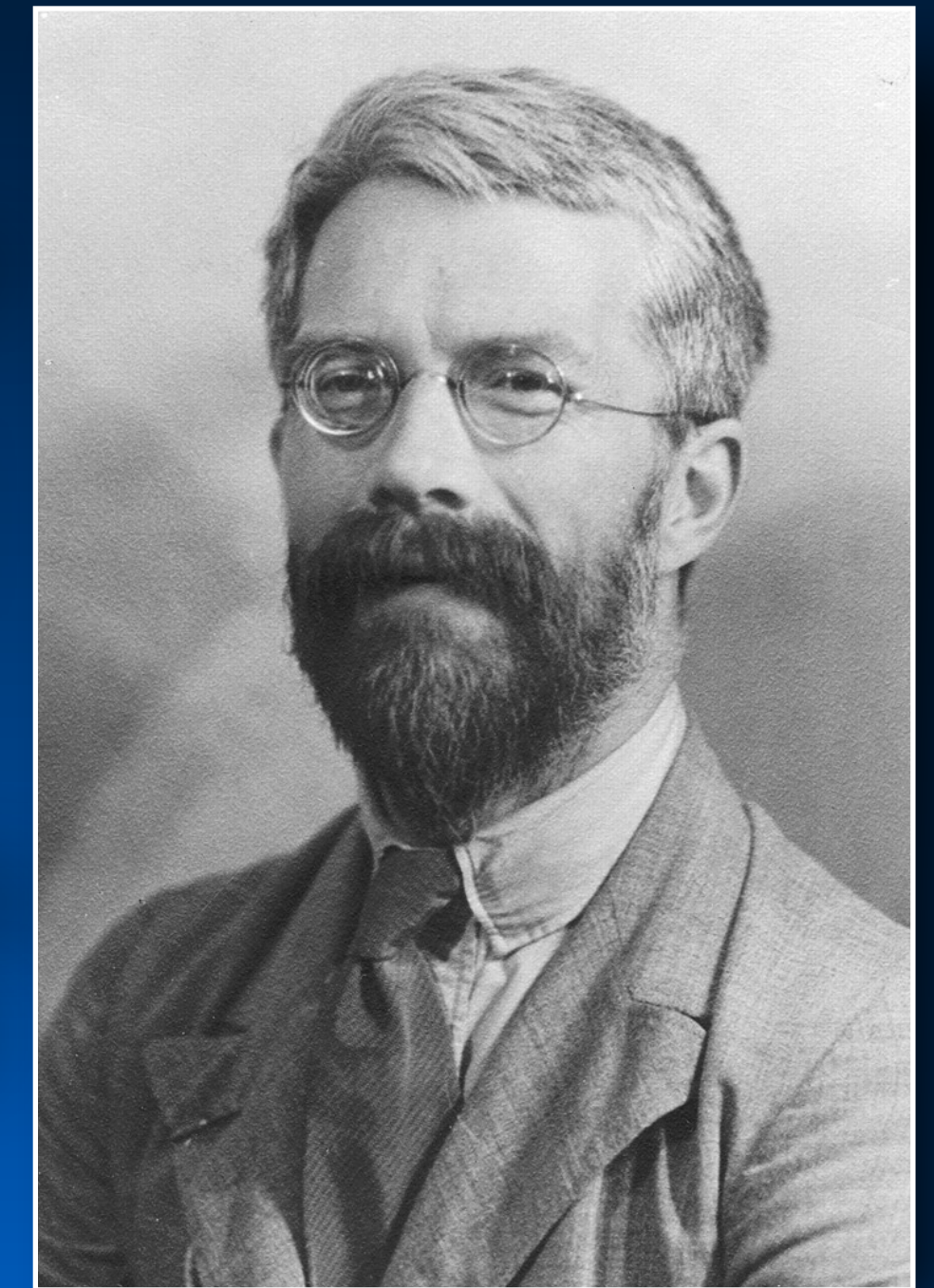
# Linear Regression, Modeling & Testing

---

## Analysis of Variance (ANOVA)

Developed originally by the English statistician Ronald A Fisher, ANOVA is based on the “law of total variance”. This holds the total variance of a dataset can be subdivided into components attributable to different sources. This is a very common principle that lies at the core of many statistical tests.

In the case of ANOVA determination of whether distinct components exist is made via comparison of the variance existing within data groups and the variance existing between data groups. So, unlike the  $t$ -Test, ANOVA tests the statements involving differences in the value of means by comparing variances.

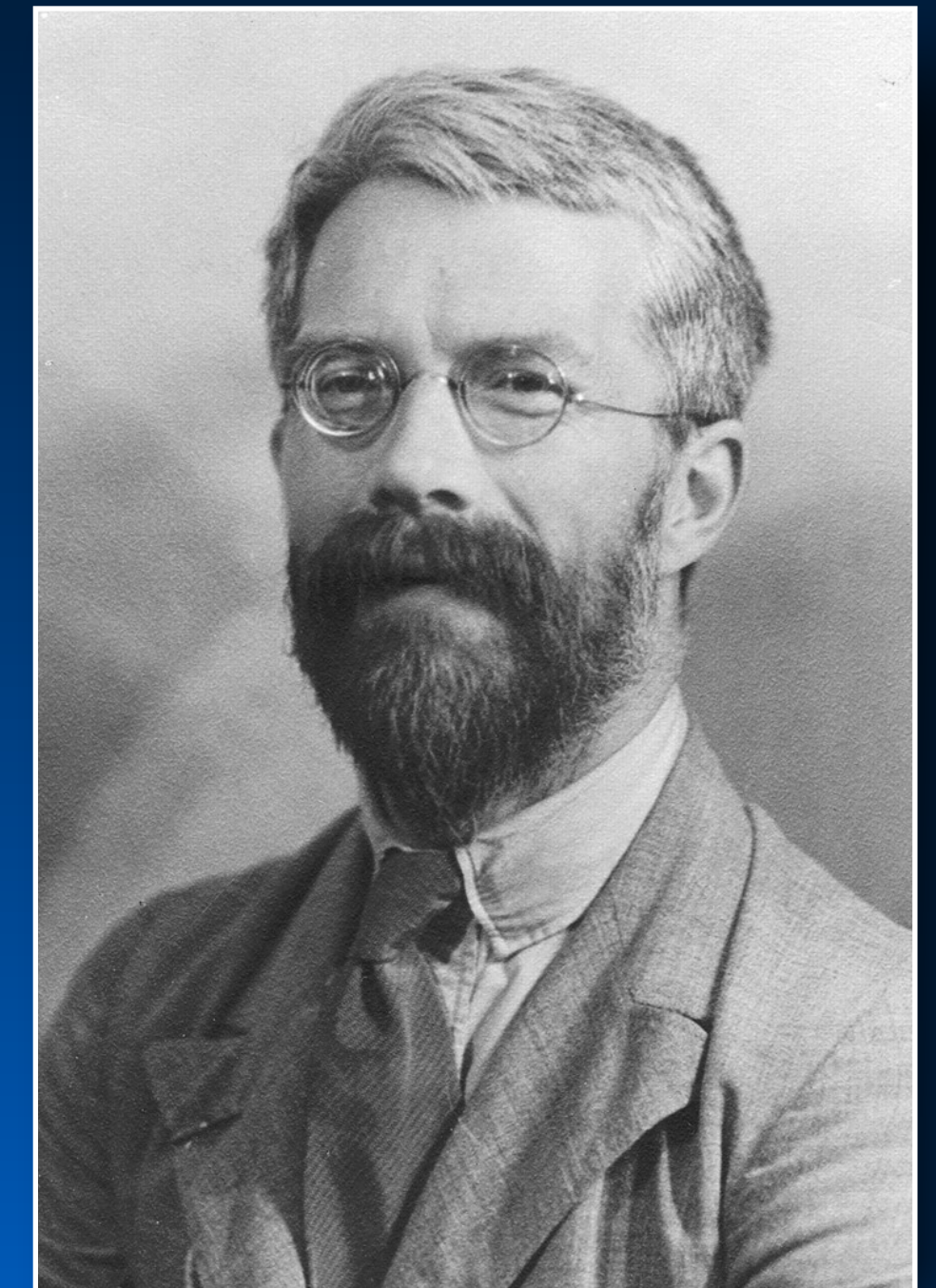
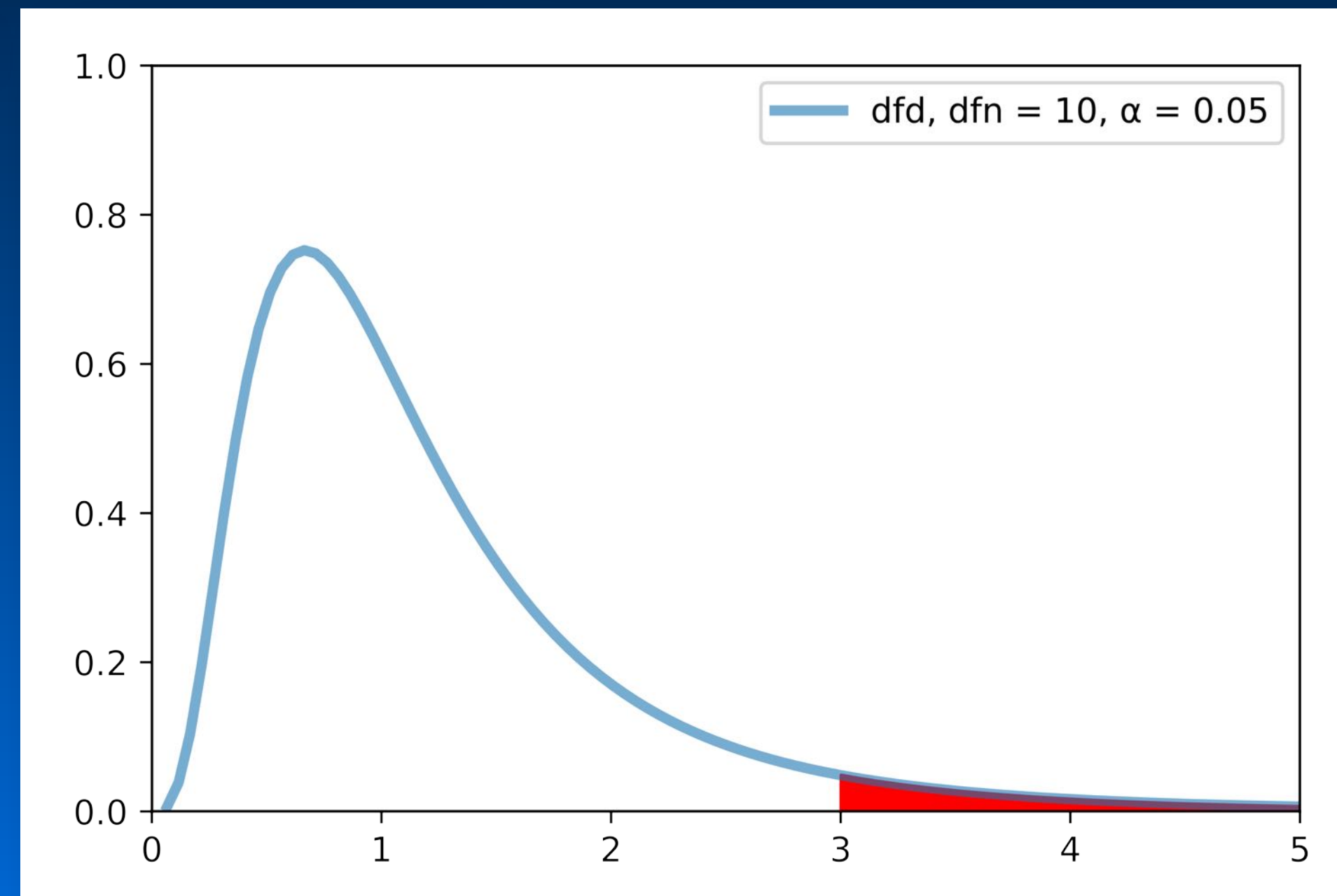


*R. A. Fisher*  
(1890-1962)

# Linear Regression, Modeling & Testing

## $F$ -Test

In order to implement his ANOVA, Fisher developed the  $F$ -Test, which is based on a class of statistical distributions, the  $F$ -distribution.



*R. A. Fisher*  
(1890-1962)

The  $F$ -distribution quantifies the probability of two independent variables having that same variances.

# Linear Regression, Modeling & Testing

---

## $F$ -Test Assumptions

Like all statistical tests, Fisher's  $F$ -Test makes a variety of assumptions concerning the data. By performing an  $F$ -Test the statistician – or researcher – is implicitly claiming that their data conforms to these assumptions.

- Both datasets have been sampled from normally distributed populations.
- Samples have been selected independently from other samples.
- Variances of the populations being compared are equal.
- Samples have been selected randomly such that the overall sample forms a valid representation of the populations in question.

# Linear Regression, Modeling & Testing

## Analysis of Variance (ANOVA)

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | F-Test      |
|---------------------|----------------|--------------------|--------------|-------------|
| Due to Model        | $SS_A$         | $m-1$              | $MS_A$       | $MS_A/MS_E$ |
| Unexplained (Error) | $SS_E$         | $N-m$              | $MS_E$       |             |
| Total Variation     | $SS_T$         | $N-1$              |              |             |

An ANOVA test is sometimes referred to as a “goodness of fit” test. It evaluates how much of the variance in a group or sample of observations can be represented by a particular explanation (or model) and how much is left unexplained (as “error”). The null hypotheses ( $H_0$ ) is that there is no difference between these two variance values. Fisher’s  $F$ -test is used to evaluate this contention.

# Linear Regression, Modeling & Testing

## Analysis of Variance (ANOVA)

| Source of Variation | Sum of Squares                                | Dof     | Mean Squares                | <i>F</i> Test       |
|---------------------|---|---------|-----------------------------|---------------------|
| Due to Model        | $SS_A = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | $m - 1$ | $MS_A = \frac{SS_A}{m - 1}$ | $\frac{MS_A}{MS_E}$ |
| Unexplained (Error) | $SS_E = \sum_{i=1}^n (\hat{y}_i - y_i)^2$     | $N - m$ | $MS_E = \frac{SS_E}{N - m}$ |                     |
| Total Variation     | $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$       | $N - 1$ |                             |                     |

Where:  $\hat{y}$  = the estimated value of  $y$ ;  
 $\bar{y}$  = the arithmetic mean of  $y$ ;  
 $m$  = number of variables;  
 $n$  = number of samples.

# Linear Regression, Modeling & Testing

## Analysis of Variance (ANOVA)

| Source of Variation | Sum of Squares | Dof | Mean Squares | $F$ Test |
|---------------------|----------------|-----|--------------|----------|
| Due to Model        | 7547.0         | 1   | 7547.0       | 23.072   |
| Unexplained (Error) | 1963.0         | 6   | 327.1        |          |
| Total Variation     | 9509.5         | 7   |              |          |

$$F_{0.05, \text{dof}:1,6} = 5.99$$

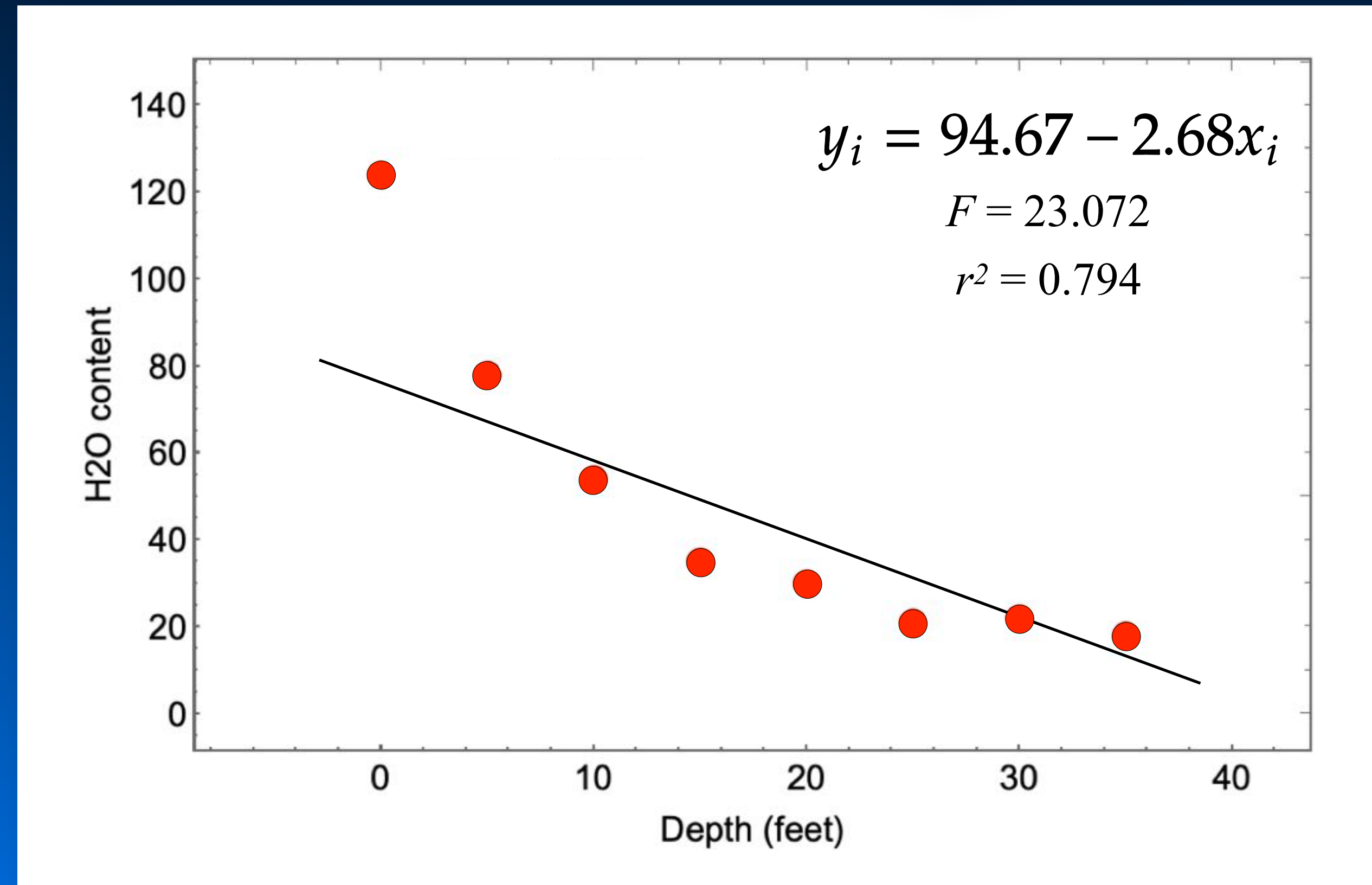
$$p_{F=23.072} = 0.003$$

Very unlikely the estimated and residual variances are equal.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |



So, we are very confident a linear trend exists in these data and estimate we only have a 0.3% chance of being wrong ... provided the assumptions we have made implicitly are correct.

# Linear Regression, Modeling & Testing

---

## Additional Linear Regression Statistics

**Confidence Interval on the Slope** - range of slopes within which the slope of the linear relationship of the population from which the sample was drawn lies to a set probability value (e.g., 95%).

**Confidence Interval on the  $y$ -Intercept** - range of  $y$ -intercepts within which the  $y$ -intercept of the linear relationship of the population from which the sample was drawn lies to a set probability value (e.g., 95%).

**Coefficient of Dispersion ( $r^2$ )** - used to determine the degree to which data are clustered about the regression line; useful for comparing among different regression results.

**Multiple Correlation Coefficient ( $r$ )** - a generalization of the standard or Pearson correlation coefficient; used to assess the quality of the prediction of the independent variable's (estimated) values.

# Linear Regression, Modeling & Testing

## Standard Error & Confidence Intervals

$$s_{\beta}^2 = \frac{MS_E}{\sum_{i=1}^n x_i - \frac{(\sum_{i=1}^n x)^2}{n}}$$

$$b_{CI} = \pm t_{criticalvalue} \cdot \sqrt{s_p^2}$$

$$\alpha_{CI} = \pm t_{criticalvalue} \cdot \sqrt{MS_E \cdot \frac{\sum_{i=1}^n x^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Where:  $s_{\beta}^2$  = standard error of regression slope;

$b_{CI}$  = confidence interval on regression slope;

$\alpha_{CI}$  = confidence interval on y-intercept.

# Linear Regression, Modeling & Testing

---

## Coefficients of Determination & Correlation

$$r^2 = \frac{SS_A}{SS_T}$$

$$r_{adj}^2 = \frac{SS_A}{SS_T} \cdot N \frac{(n-1) - 1}{(N-1) - 2}$$

$$r = \sqrt{r^2}$$

$$r_{adj} = \sqrt{r_{adj}^2}$$

Where:  $r^2$  = coefficient of determination;

$r_{adj}^2$  = coefficient of determination (adjusted);

$r$  = multiple correlation coefficient;

$r_{adj}$  = multiple correlation coefficient (adjusted).

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

### Confidence Intervals (95%)

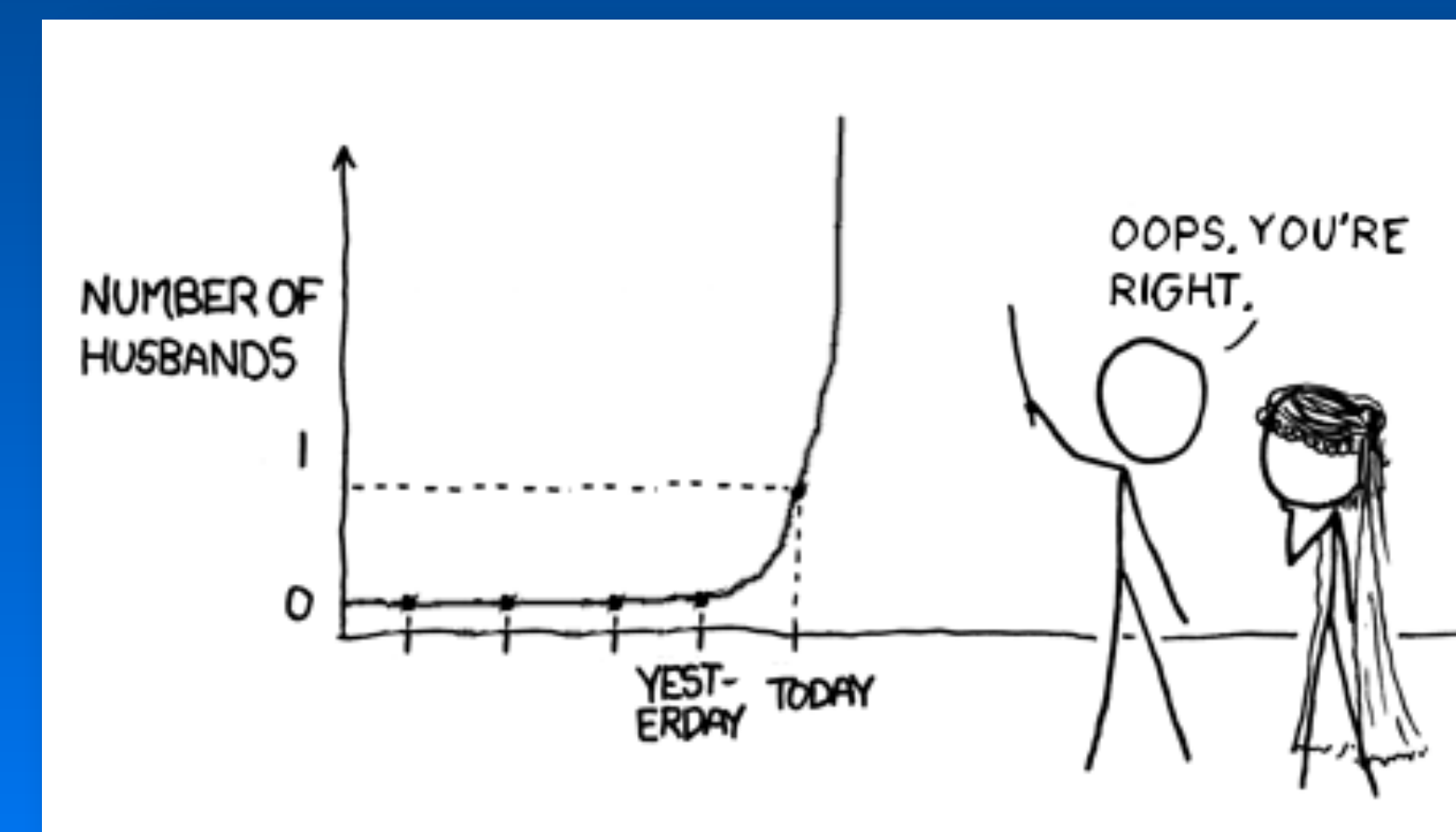
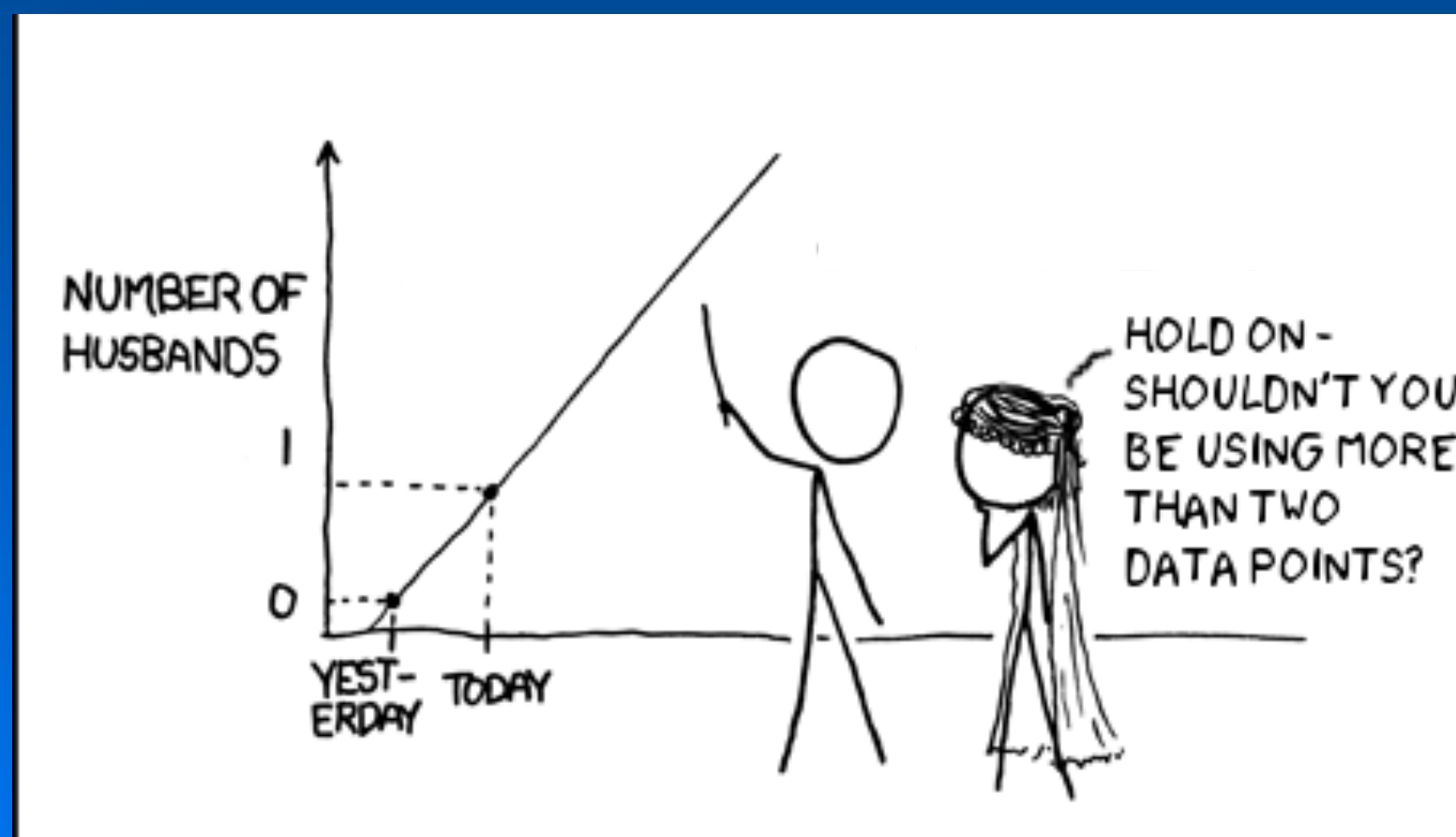
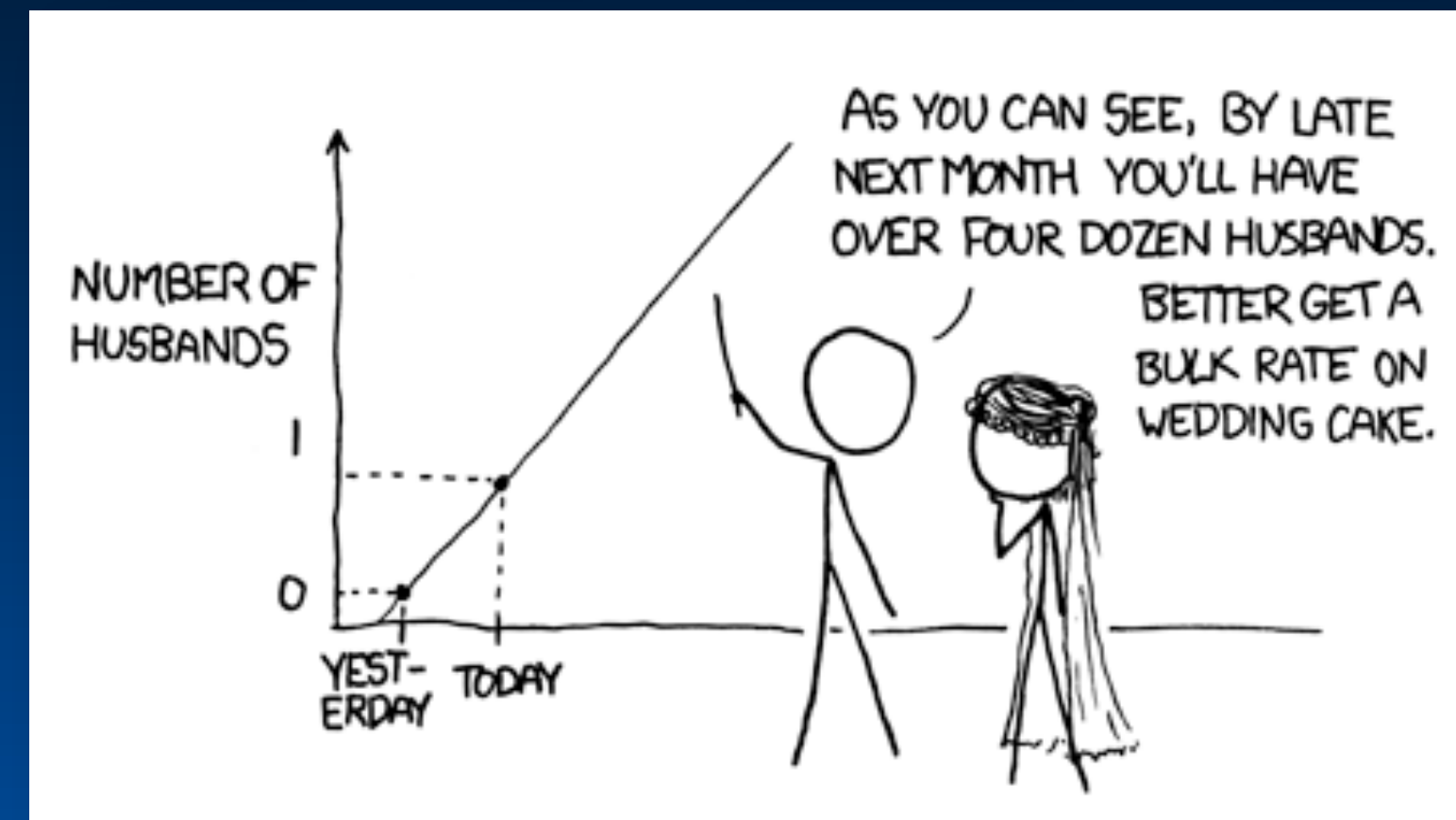
|       | <i>y</i> -Intercept ( $\alpha$ ) | Slope ( $\beta$ ) |
|-------|----------------------------------|-------------------|
| Upper | 113.60                           | -3.77             |
| Mean  | 94.67                            | -2.68             |
| Lower | 75.69                            | -1.60             |

### Measures of Dispersion

|          | Coefficient of Dispersion ( $r^2$ ) | Multiple Correlation Coefficient |
|----------|-------------------------------------|----------------------------------|
| Raw      | 0.794                               | 0.891                            |
| Adjusted | 0.926                               | 0.962                            |

# Linear Regression, Modeling & Testing

So, is that all there is to know about linear regression analysis?



# Linear Regression, Modeling & Testing

What about trilobites?

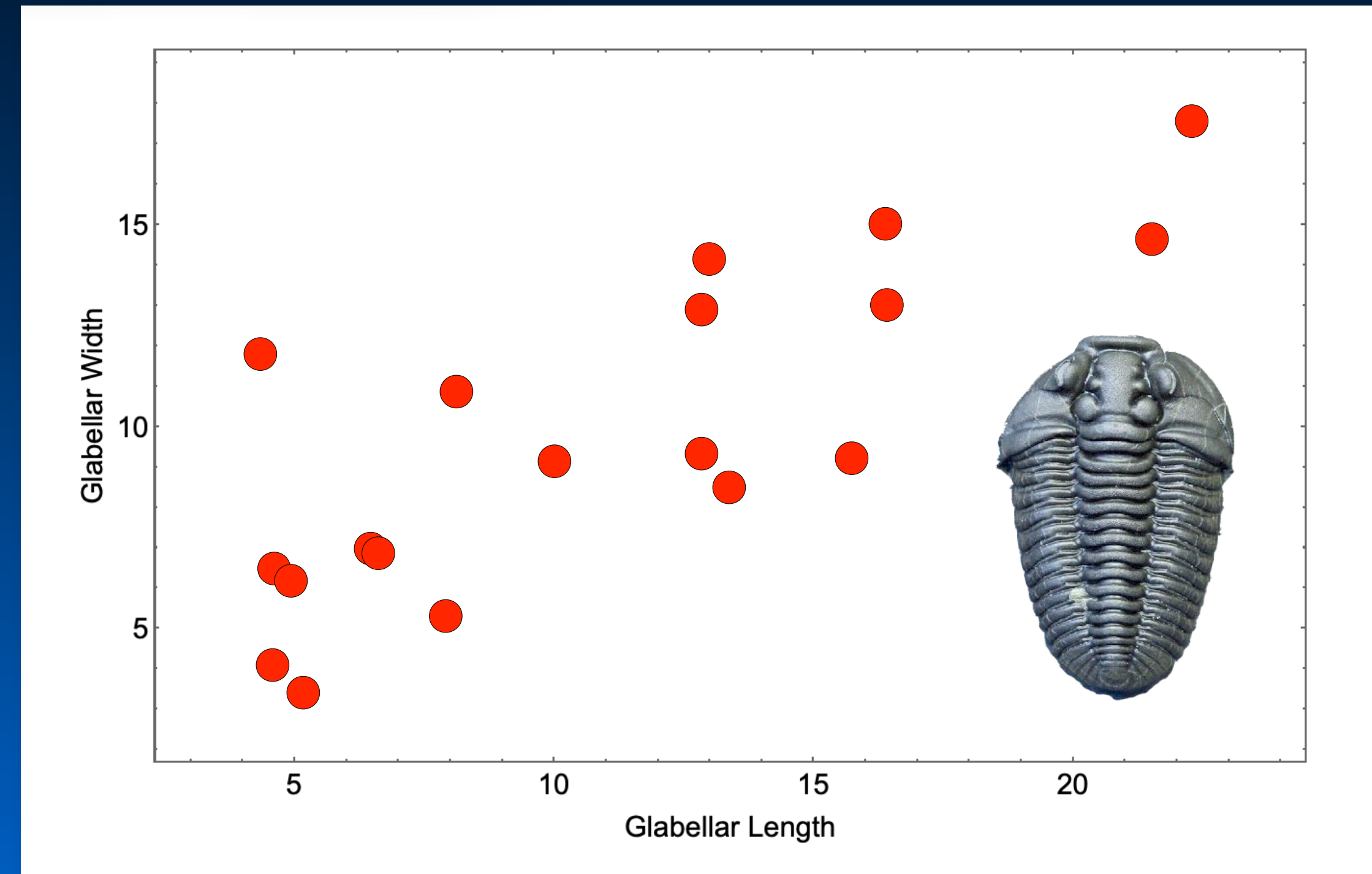
| Genus                | Length | Width |
|----------------------|--------|-------|
| <i>Acaste</i>        | 5.10   | 3.46  |
| <i>Balizoma</i>      | 4.60   | 6.53  |
| <i>Calymene</i>      | 12.98  | 14.15 |
| <i>Ceraurus</i>      | 7.90   | 5.32  |
| <i>Cheirurus</i>     | 12.83  | 12.96 |
| <i>Cybantyx</i>      | 16.41  | 13.08 |
| <i>Cybeloides</i>    | 6.60   | 6.84  |
| <i>Dalmanites</i>    | 10.00  | 9.12  |
| <i>Deiphon</i>       | 8.08   | 10.77 |
| <i>Narroia</i>       | 15.67  | 9.25  |
| <i>Ormathops</i>     | 4.53   | 4.11  |
| <i>Phacopidina</i>   | 6.44   | 6.94  |
| <i>Pricyclopyge</i>  | 21.53  | 14.64 |
| <i>Ptychoparia</i>   | 12.82  | 9.36  |
| <i>Rhenops</i>       | 22.27  | 17.56 |
| <i>Sphaerexochus</i> | 4.93   | 6.21  |
| <i>Trimerus</i>      | 16.35  | 15.02 |
| <i>Zachanthoides</i> | 13.41  | 8.51  |



# Linear Regression, Modeling & Testing

What about trilobites?

| Genus                | Length | Width |
|----------------------|--------|-------|
| <i>Acaste</i>        | 5.10   | 3.46  |
| <i>Balizoma</i>      | 4.60   | 6.53  |
| <i>Calymene</i>      | 12.98  | 14.15 |
| <i>Ceraurus</i>      | 7.90   | 5.32  |
| <i>Cheirurus</i>     | 12.83  | 12.96 |
| <i>Cybantyx</i>      | 16.41  | 13.08 |
| <i>Cybeloides</i>    | 6.60   | 6.84  |
| <i>Dalmanites</i>    | 10.00  | 9.12  |
| <i>Deiphon</i>       | 8.08   | 10.77 |
| <i>Narroia</i>       | 15.67  | 9.25  |
| <i>Ormathops</i>     | 4.53   | 4.11  |
| <i>Phacopidina</i>   | 6.44   | 6.94  |
| <i>Pricyclopyge</i>  | 21.53  | 14.64 |
| <i>Ptychoparia</i>   | 12.82  | 9.36  |
| <i>Rhenops</i>       | 22.27  | 17.56 |
| <i>Sphaerexochus</i> | 4.93   | 6.21  |
| <i>Trimerus</i>      | 16.35  | 15.02 |
| <i>Zachanthoides</i> | 13.41  | 8.51  |

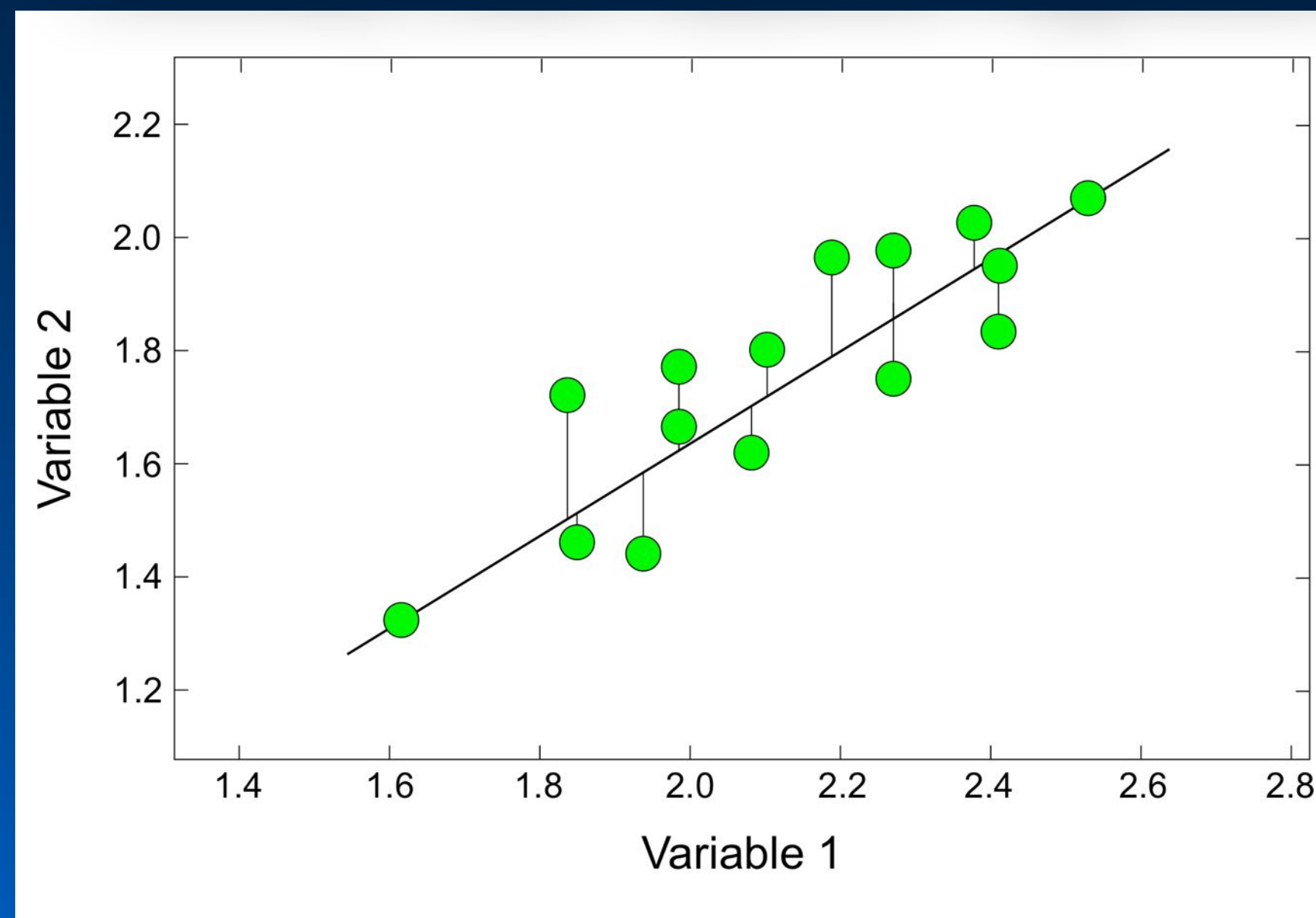


Which one's the independent variable?  
Glabellar length?  
Glabellar width?

# Linear Regression, Modeling & Testing

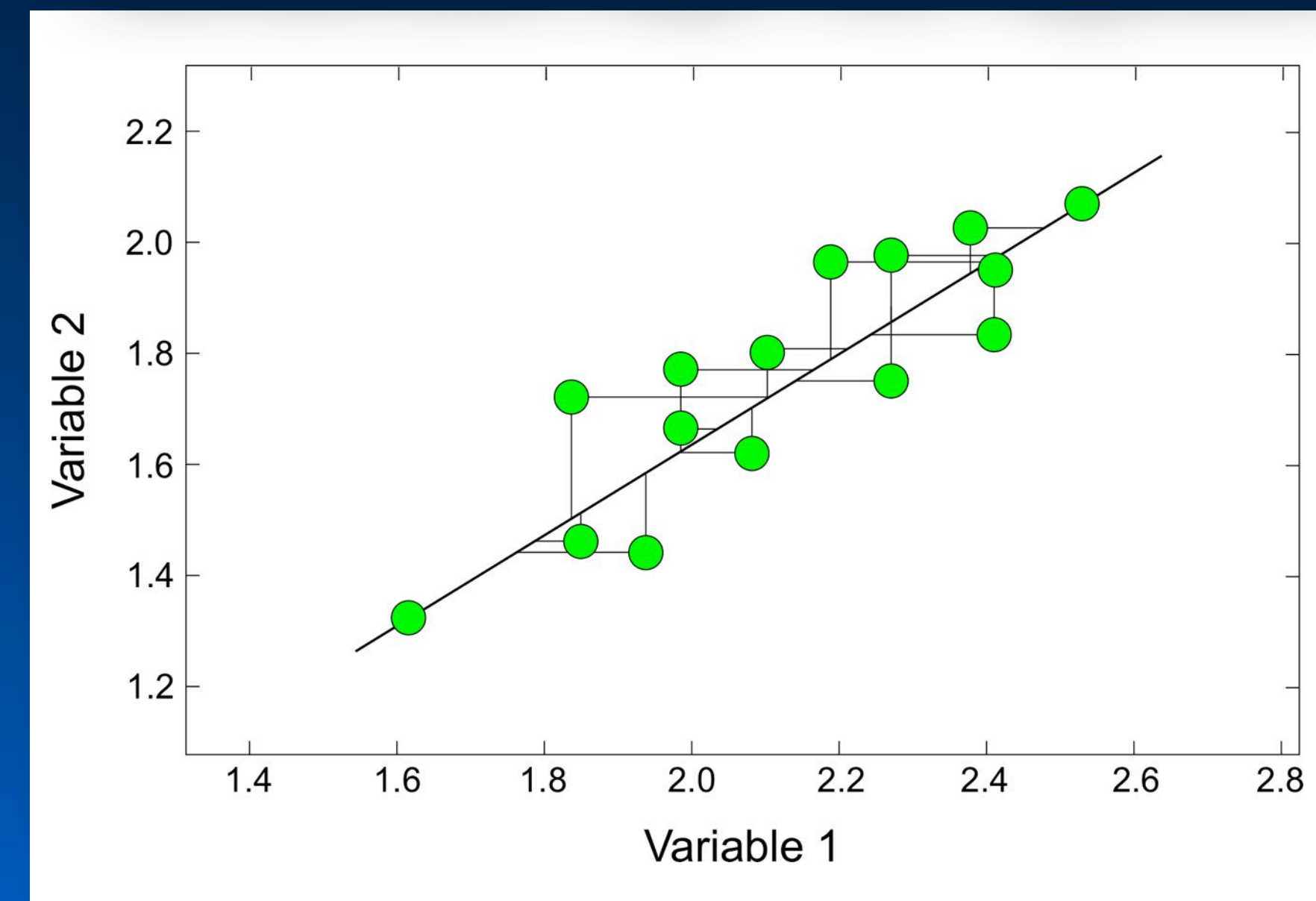
## Alternative Regression Error Minimization Strategies

Ordinary Least Squares Regression



Also called:  
Regression  
Least Squares Regression  
Model I Regression

Standardized Major Axis Regression



Also called:  
Reduced Major Axis Regression  
Geometric Mean Regression  
Model II Regression

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

| Genus                | Length | Width  | Dev <sub>L</sub> | Dev <sub>w</sub> | SDP   |
|----------------------|--------|--------|------------------|------------------|-------|
| <i>Acaste</i>        | 5.10   | 3.46   | -6.15            | -6.20            | 38.10 |
| <i>Balizoma</i>      | 4.60   | 6.53   | -6.65            | -3.13            | 20.79 |
| <i>Calymene</i>      | 12.98  | 14.15  | 1.73             | 4.49             | 7.78  |
| <i>Ceraurus</i>      | 7.90   | 5.32   | -3.35            | -4.34            | 14.52 |
| <i>Cheirurus</i>     | 12.83  | 12.96  | 1.58             | 3.30             | 5.23  |
| <i>Cybantyx</i>      | 16.41  | 13.08  | 5.16             | 3.42             | 17.67 |
| <i>Cybeloides</i>    | 6.60   | 6.84   | -4.65            | -2.82            | 13.09 |
| <i>Dalmanites</i>    | 10.00  | 9.12   | -1.25            | -0.54            | 0.67  |
| <i>Deiphon</i>       | 8.08   | 10.77  | -3.17            | 1.11             | -3.52 |
| <i>Narroia</i>       | 15.67  | 9.25   | 4.42             | -0.41            | -1.80 |
| <i>Ormathops</i>     | 4.53   | 4.11   | -6.72            | -5.55            | 37.26 |
| <i>Phacopidina</i>   | 6.44   | 6.94   | -4.81            | -2.72            | 13.06 |
| <i>Pricyclopyge</i>  | 21.53  | 14.64  | 10.28            | 4.98             | 51.24 |
| <i>Ptychoparia</i>   | 12.82  | 9.36   | 1.57             | -0.30            | -0.47 |
| <i>Rhenops</i>       | 22.27  | 17.56  | 11.02            | 7.90             | 87.11 |
| <i>Sphaerexochus</i> | 4.93   | 6.21   | -6.32            | -3.45            | 21.78 |
| <i>Trimerus</i>      | 16.35  | 15.02  | 5.10             | 5.36             | 27.37 |
| <i>Zachanthoides</i> | 13.41  | 8.51   | 2.16             | -1.15            | -2.48 |
| Sum                  | 202.45 | 173.83 | 0.00             | 0.00             |       |
| Mean                 | 11.25  | 9.66   | 0.00             | 0.00             |       |
| Std. Dev.            | 5.67   | 4.11   | 5.67             | 4.11             |       |

$$\beta = \frac{4.11}{5.67}$$

$$\alpha = \bar{y} - (\beta \cdot \bar{x})$$

When  $n$  = total number of observations

$$SDP = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Sign of slope ( $\beta$ ) = sign of SDP

# Linear Regression, Modeling & Testing

---

## Standardized Major Axis Linear Regression

$$\beta = \frac{s_y}{s_x}$$

$$\beta = \frac{s_y}{s_x} = 0.72$$

$$\alpha = \bar{y} - (\beta \cdot \bar{x})$$

$$\alpha = 9.66 - (0.72 \cdot 11.25)$$

When  $n$  = total number  
of observations

$$a = 1.51$$

$$SDP = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

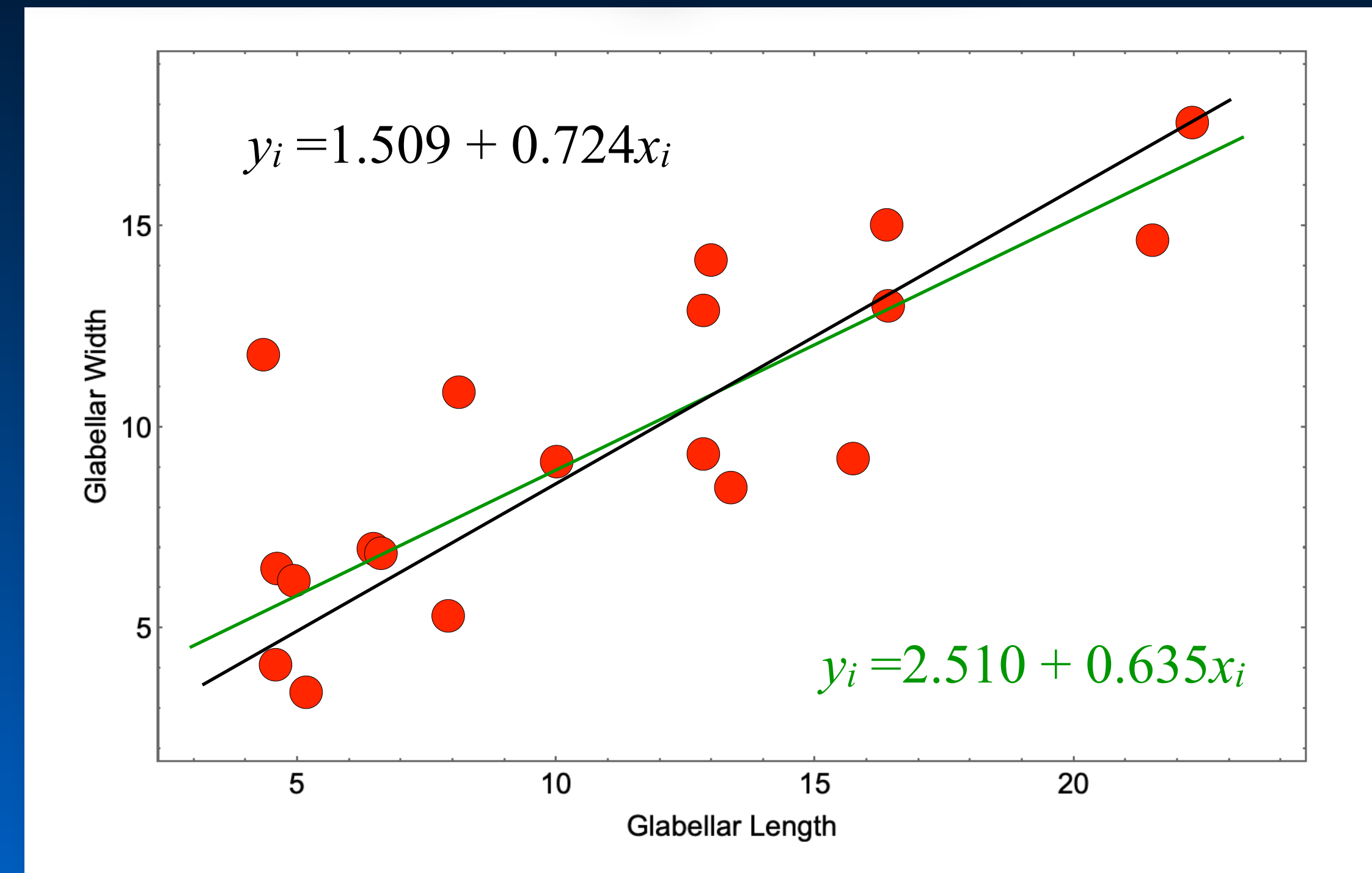
$$SDP = 347.39$$

Sign of slope ( $\beta$ ) = sign of SDP

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

| Genus                | Length | Width |
|----------------------|--------|-------|
| <i>Acaste</i>        | 5.10   | 3.46  |
| <i>Balizoma</i>      | 4.60   | 6.53  |
| <i>Calymene</i>      | 12.98  | 14.15 |
| <i>Ceraurus</i>      | 7.90   | 5.32  |
| <i>Cheirurus</i>     | 12.83  | 12.96 |
| <i>Cybantyx</i>      | 16.41  | 13.08 |
| <i>Cybeloides</i>    | 6.60   | 6.84  |
| <i>Dalmanites</i>    | 10.00  | 9.12  |
| <i>Deiphon</i>       | 8.08   | 10.77 |
| <i>Narroia</i>       | 15.67  | 9.25  |
| <i>Ormathops</i>     | 4.53   | 4.11  |
| <i>Phacopidina</i>   | 6.44   | 6.94  |
| <i>Pricyclopyge</i>  | 21.53  | 14.64 |
| <i>Ptychoparia</i>   | 12.82  | 9.36  |
| <i>Rhenops</i>       | 22.27  | 17.56 |
| <i>Sphaerexochus</i> | 4.93   | 6.21  |
| <i>Trimerus</i>      | 16.35  | 15.02 |
| <i>Zachanthoides</i> | 13.41  | 8.51  |



Note: the standardized major axis regression line is usually closer to the line we'd draw by eye than the least-squares regression line.

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

| Source of Variation | Sum of Squares  | DoF     | Mean Squares                | $F$ Test            |
|---------------------|---|---------|-----------------------------|---------------------|
| Due to Model        | $SS_A = \sum_{i=1}^n ((\hat{x}_i, \hat{y}_i) - (\bar{x}_i, \bar{y}_i))^2$ | $m - 1$ | $MS_A = \frac{SS_A}{m - 1}$ | $\frac{MS_A}{MS_E}$ |
| Unexplained (Error) | $SS_A = \sum_{i=1}^n ((x_i, y_i) - (\hat{x}_i, \hat{y}_i))^2$             | $N - m$ | $MS_E = \frac{SS_E}{N - m}$ |                     |
| Total Variation     | $SS_T = \sum_{i=1}^n ((x_i, y_i) - (\bar{x}_i, \bar{y}_i))^2$             | $N - 1$ |                             |                     |

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

| Source of Variation | Sum of Squares | DoF | Mean Squares | $F$ Test |
|---------------------|----------------|-----|--------------|----------|
| Due to Model        | 269.3          | 1   | 269.3        | 244.5    |
| Unexplained (Error) | 17.6           | 16  | 1.01         |          |
| Total Variation     | 286.9          | 17  |              |          |

$$F_{0.05, \text{dof}:1,6} = 4.49$$

$$p_{F=244.5} = 0.0000$$

Very, very unlikely the estimated and residual variances are equal.

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

Confidence Intervals (95%)

|       | <i>y</i> -Intercept ( $\alpha$ ) | Slope ( $\beta$ ) |
|-------|----------------------------------|-------------------|
| Upper | 3.323                            | 0.93              |
| Mean  | 1.51                             | 0.72              |
| Lower | -0.825                           | 0.56              |

Measures of Dispersion

|          | Coefficient of Dispersion ( $r^2$ ) | Multiple Correlation Coefficient |
|----------|-------------------------------------|----------------------------------|
| Raw      | 0.939                               | 0.969                            |
| Adjusted | 0.997                               | 0.999                            |

# Linear Regression, Modeling & Testing

---

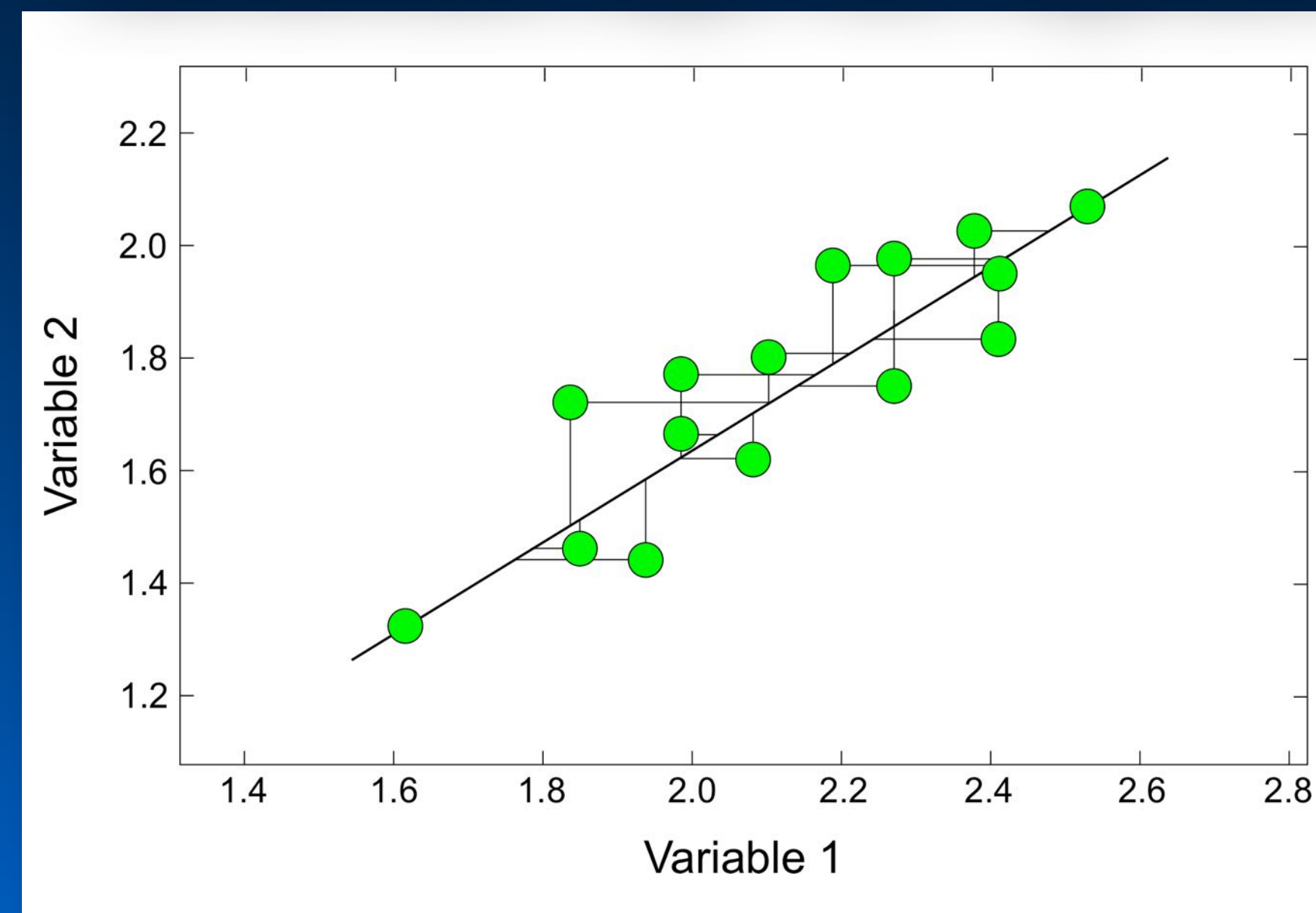
SMA regression is too simple. Don't you have anything more complex?



# Linear Regression, Modeling & Testing

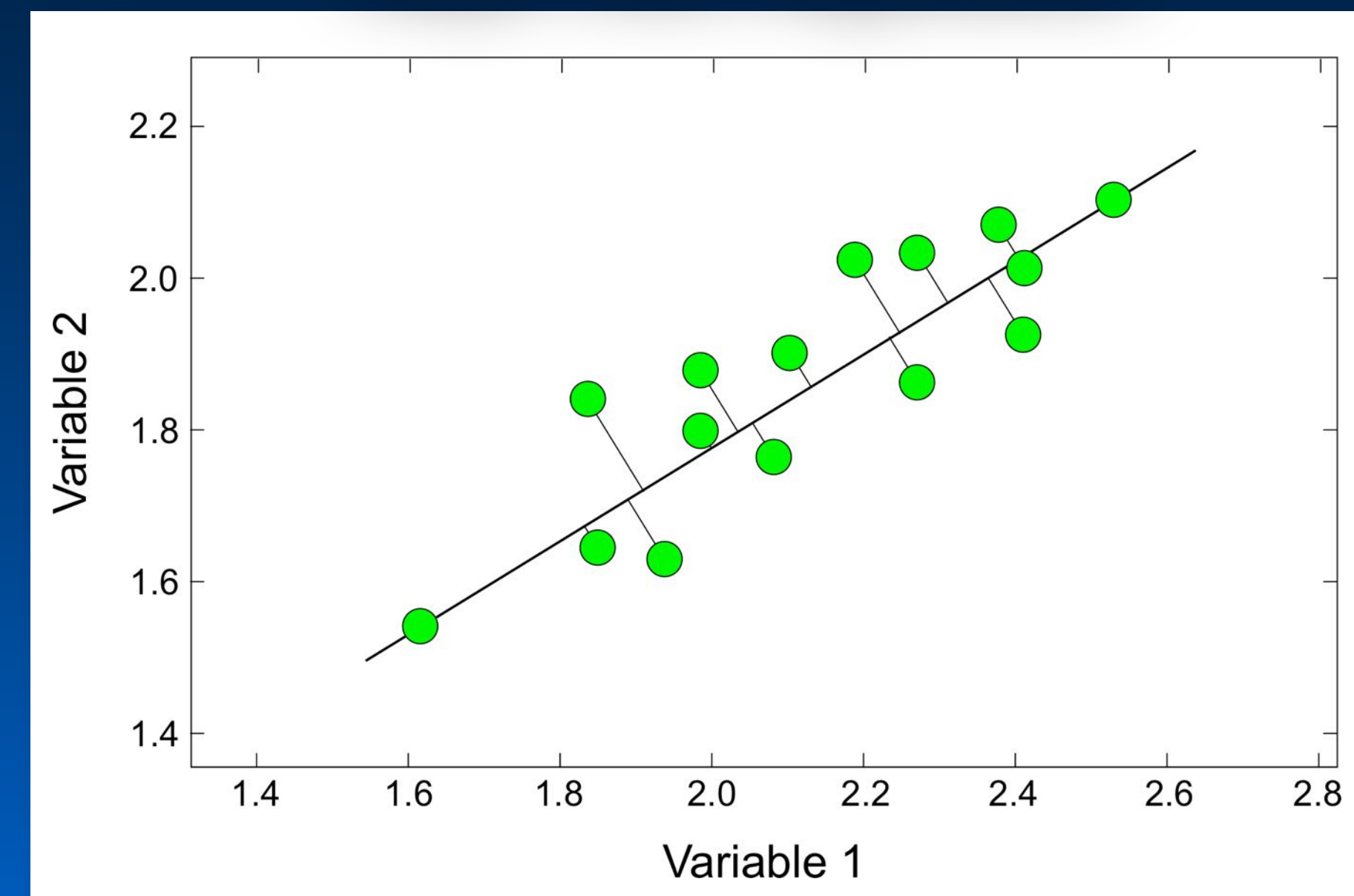
## Additional Regression Error Minimization Strategies

Standardized Major Axis Regression



Also called:  
Reduced Major Axis Regression  
Geometric Mean Regression  
Model II Regression

Major Axis Regression



Also called:  
Structural Analysis, Orthogonal  
Regression, Bivariate PCA  
Model II Regression

# Linear Regression, Modeling & Testing

## Major Axis Linear Regression

$$s_x^2 = \frac{\sum_{i=1}^n x_i^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=1}^n y_i^2}{n-1}$$

$$s_{xy} = \frac{\sum_{i=1}^n (\bar{x} - x_i) \cdot (\bar{y} - y_i)}{n-1}$$

$$D = \sqrt{(s_x^2 + s_y^2)^2 - 4(s_x^2 \cdot s_y^2)}$$

$$\lambda = \frac{(s_x^2 + s_y^2 + D)}{2}$$

$$\beta = \frac{s_{xy}}{(\lambda - s_y^2)}$$

$$\alpha = \bar{y} - (\beta \cdot \bar{x})$$

# Linear Regression, Modeling & Testing

---

## Major Axis Linear Regression

$$s_x^2 = 32.16$$

$$s_y^2 = 16.88$$

$$s_{xy} = 20.43$$

$$D = 43.63$$

$$\lambda = 46.33$$

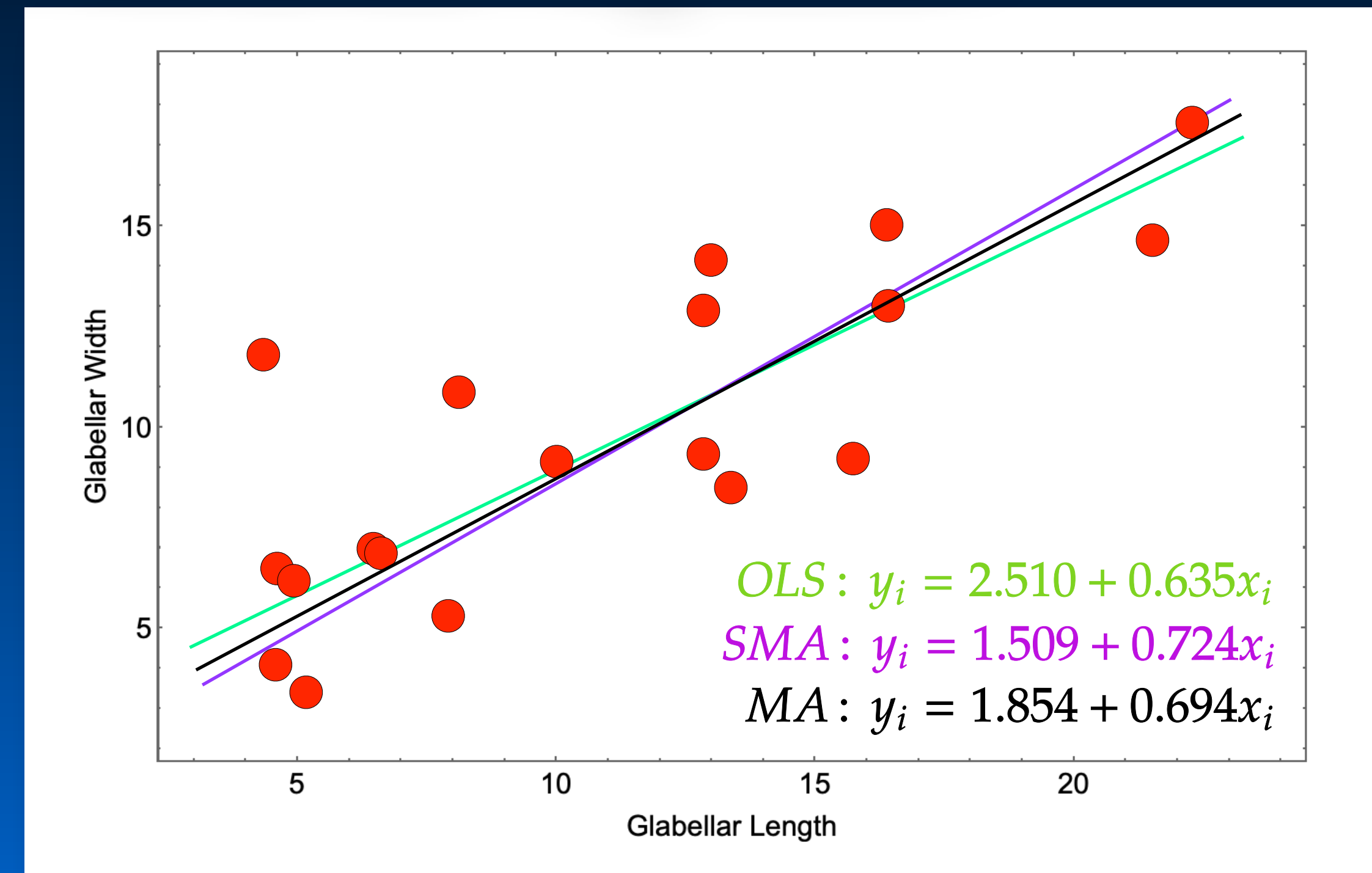
$$\beta = 0.693$$

$$\alpha = 1.854$$

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

| Genus                | Length | Width |
|----------------------|--------|-------|
| <i>Acaste</i>        | 5.10   | 3.46  |
| <i>Balizoma</i>      | 4.60   | 6.53  |
| <i>Calymene</i>      | 12.98  | 14.15 |
| <i>Ceraurus</i>      | 7.90   | 5.32  |
| <i>Cheirurus</i>     | 12.83  | 12.96 |
| <i>Cybantyx</i>      | 16.41  | 13.08 |
| <i>Cybeloides</i>    | 6.60   | 6.84  |
| <i>Dalmanites</i>    | 10.00  | 9.12  |
| <i>Deiphon</i>       | 8.08   | 10.77 |
| <i>Narroia</i>       | 15.67  | 9.25  |
| <i>Ormathops</i>     | 4.53   | 4.11  |
| <i>Phacopidina</i>   | 6.44   | 6.94  |
| <i>Pricyclopyge</i>  | 21.53  | 14.64 |
| <i>Ptychoparia</i>   | 12.82  | 9.36  |
| <i>Rhenops</i>       | 22.27  | 17.56 |
| <i>Sphaerexochus</i> | 4.93   | 6.21  |
| <i>Trimerus</i>      | 16.35  | 15.02 |
| <i>Zachanthoides</i> | 13.41  | 8.51  |



Note: the major axis regression line usually splits the difference between the least squares regression and the standardized major axis regression lines.

# Linear Regression, Modeling & Testing

## Analysis of Variance (ANOVA)

| Source of Variation | Sum of Squares  | DoF     | Mean Squares                | $F$ Test            |
|---------------------|---|---------|-----------------------------|---------------------|
| Due to Model        | $SS_A = \sum_{i=1}^n ((\hat{x}_i, \hat{y}_i) - (\bar{x}_i, \bar{y}_i))^2$ | $m - 1$ | $MS_A = \frac{SS_A}{m - 1}$ | $\frac{MS_A}{MS_E}$ |
| Unexplained (Error) | $SS_A = \sum_{i=1}^n ((x_i, y_i) - (\hat{x}_i, \hat{y}_i))^2$             | $N - m$ | $MS_E = \frac{SS_E}{N - m}$ |                     |
| Total Variation     | $SS_T = \sum_{i=1}^n ((x_i, y_i) - (\bar{x}_i, \bar{y}_i))^2$             | $N - 1$ |                             |                     |

# Linear Regression, Modeling & Testing

## Standardized Major Axis Linear Regression

| Source of Variation | Sum of Squares | DoF | Mean Squares | $F$ Test |
|---------------------|----------------|-----|--------------|----------|
| Due to Model        | 255.9          | 1   | 255.9        | 132.1    |
| Unexplained (Error) | 31.0           | 16  | 1.937        |          |
| Total Variation     | 286.9          | 17  |              |          |

$$F_{0.05,dof:1,6} = 5.99$$

$$P_{F=132.1,dof:1,16} = 0.0000261$$

Very, very unlikely the estimated and residual variances are equal.

# Linear Regression, Modeling & Testing

## Major Axis Linear Regression

Confidence Intervals (95%)

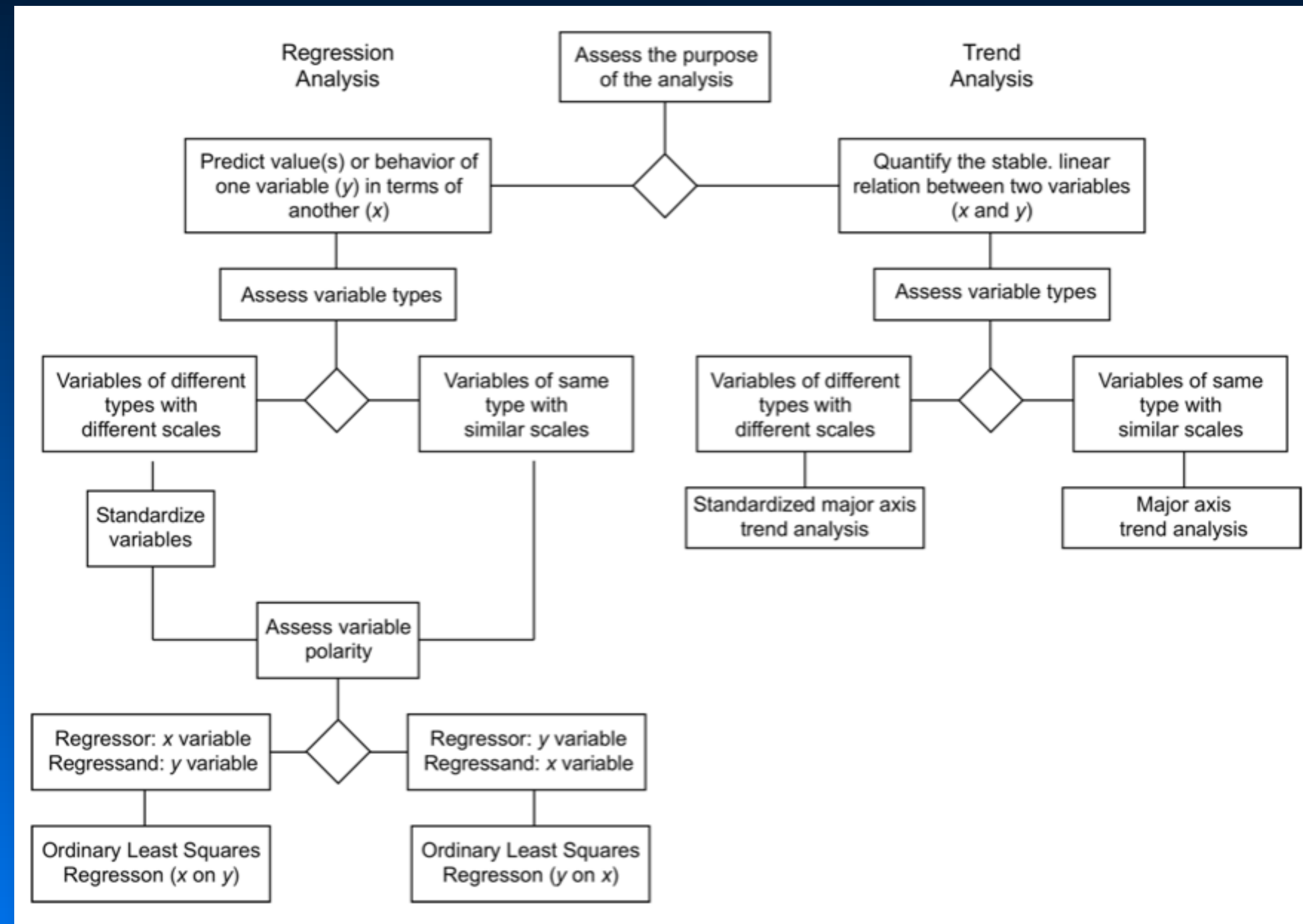
|       | <i>y</i> -Intercept ( $\alpha$ ) | Slope ( $\beta$ ) |
|-------|----------------------------------|-------------------|
| Upper | 3.959                            | 0.921             |
| Mean  | 1.854                            | 0.694             |
| Lower | -0.698                           | 0.507             |

Measures of Dispersion

|          | Coefficient of Dispersion ( $r^2$ ) | Multiple Correlation Coefficient |
|----------|-------------------------------------|----------------------------------|
| Raw      | 0.892                               | 0.944                            |
| Adjusted | 0.948                               | 0.974                            |

# Linear Regression, Modeling & Testing

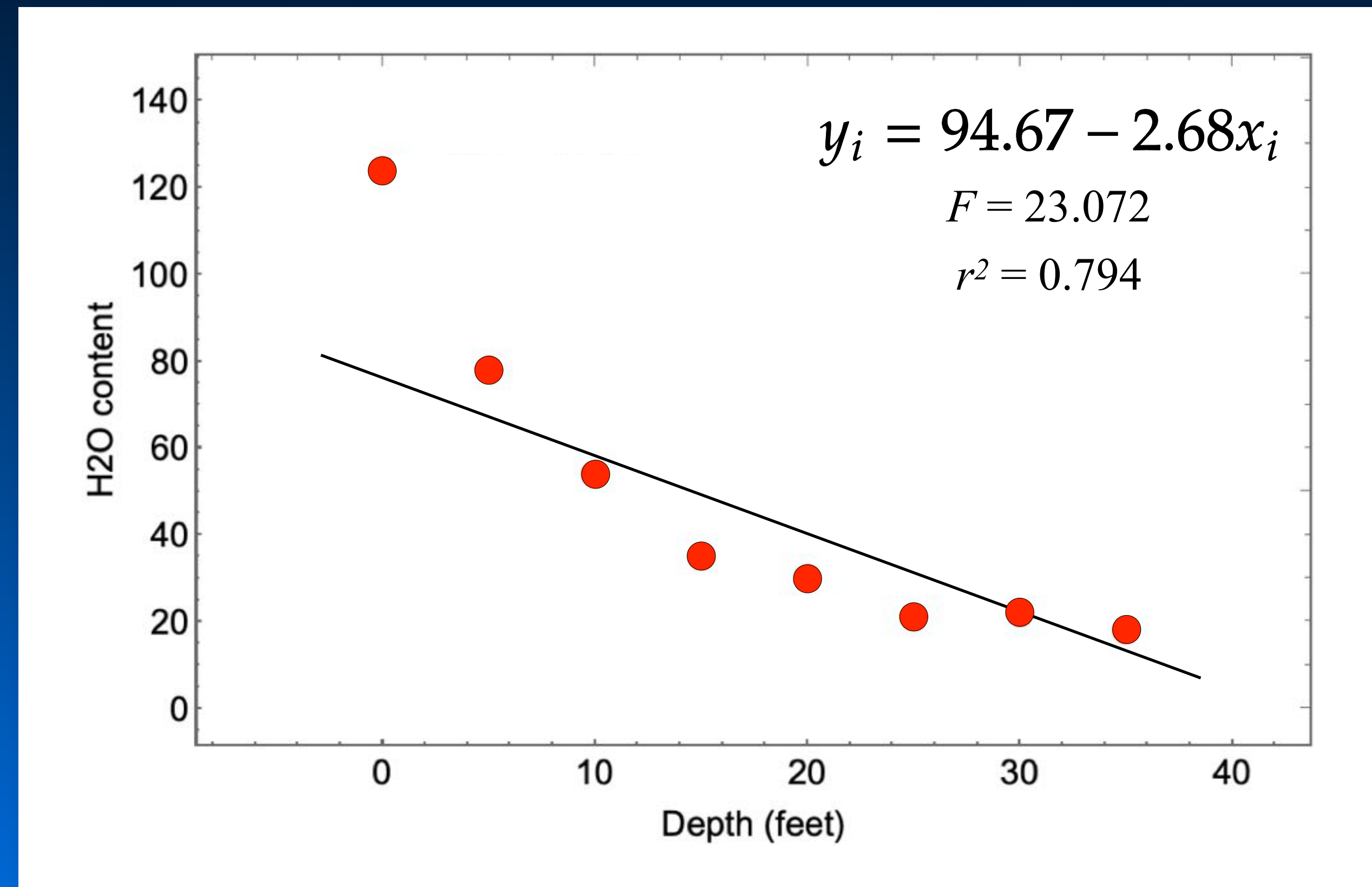
## Flow Chart Evaluating a Regression Problem



# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

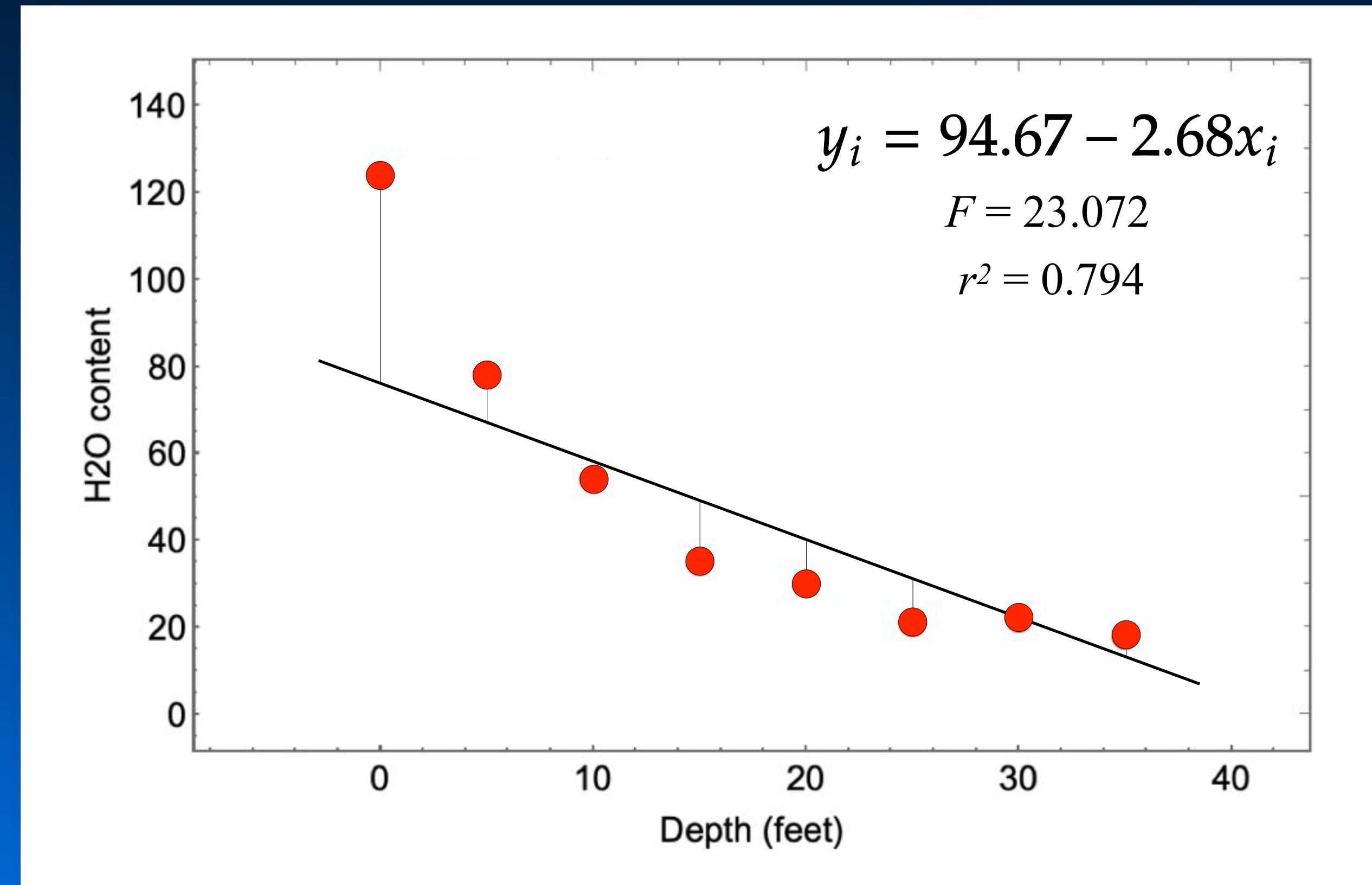


So, we are very confident a linear trend exists in these data. But, does this look like a very good model for the data to you?

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

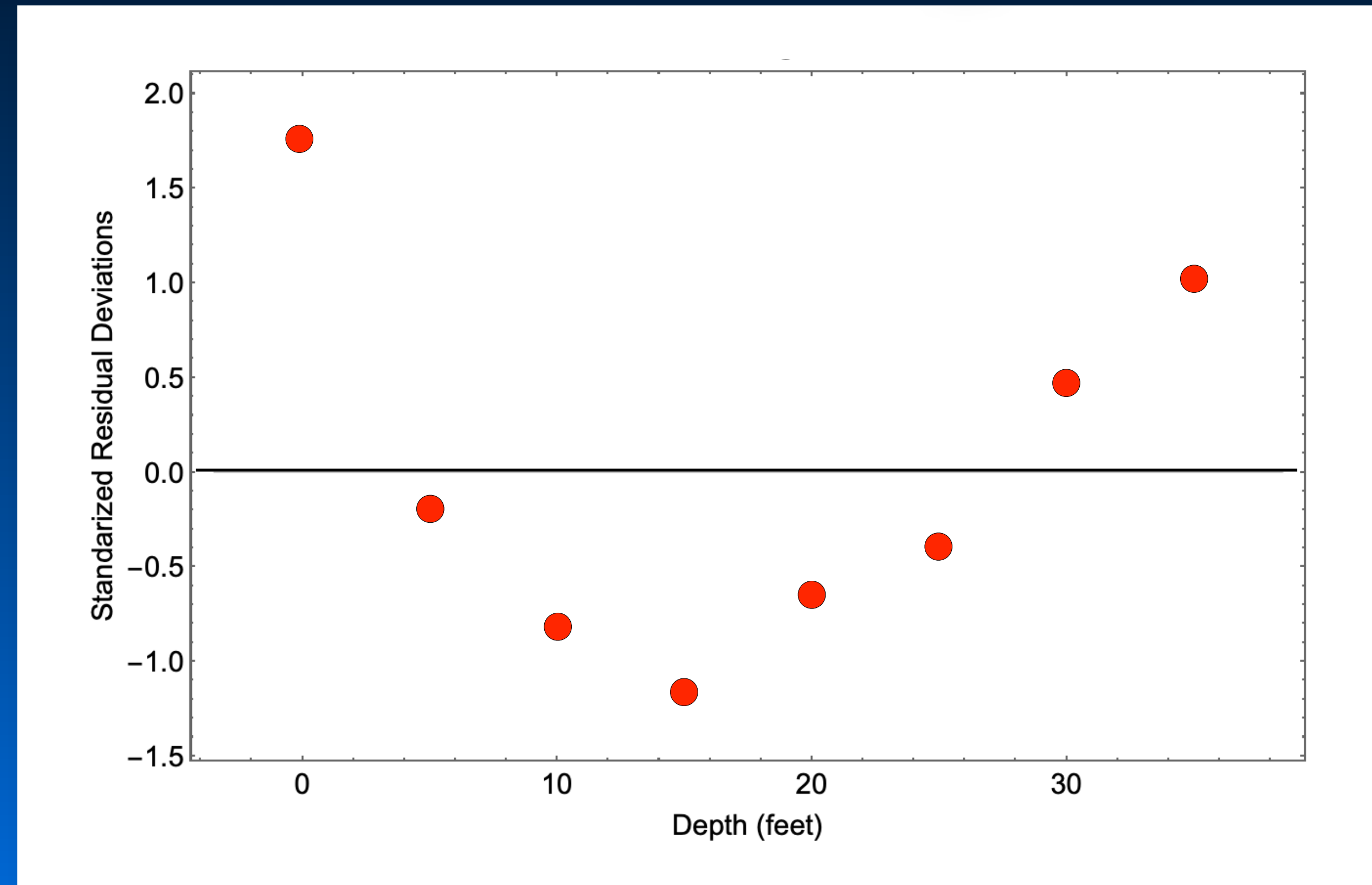


Our analysis has minimized the square of the distances between the data-point and the linear regression model. These distances are referred to as “residuals”.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

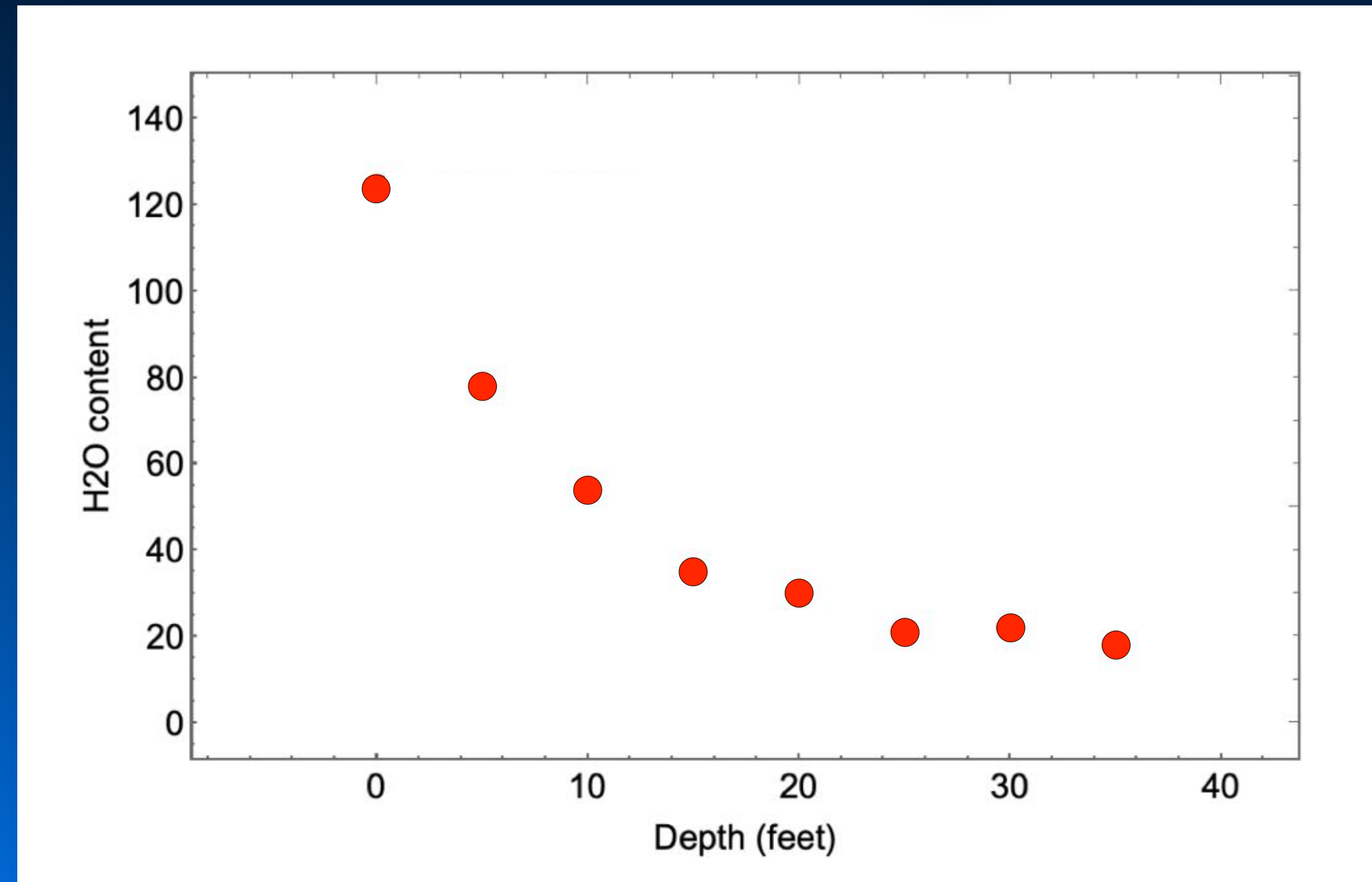


A plot of the standardized residual values reveals non-random structure. This indicates that, despite its statistical significance, the linear model is suboptimal.

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

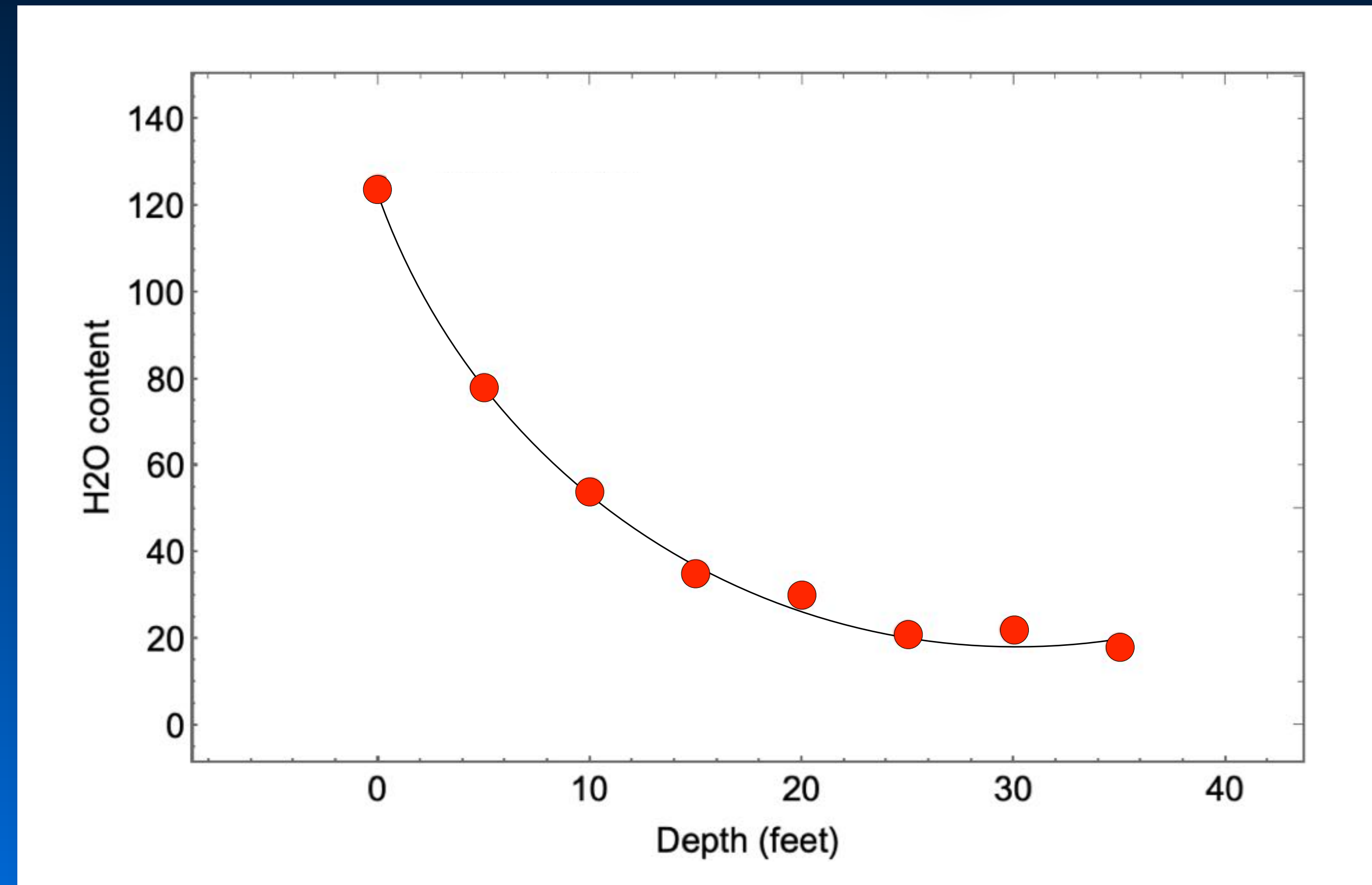


We modeled these variable's relationship as linear, but is there a better model we could apply?

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |

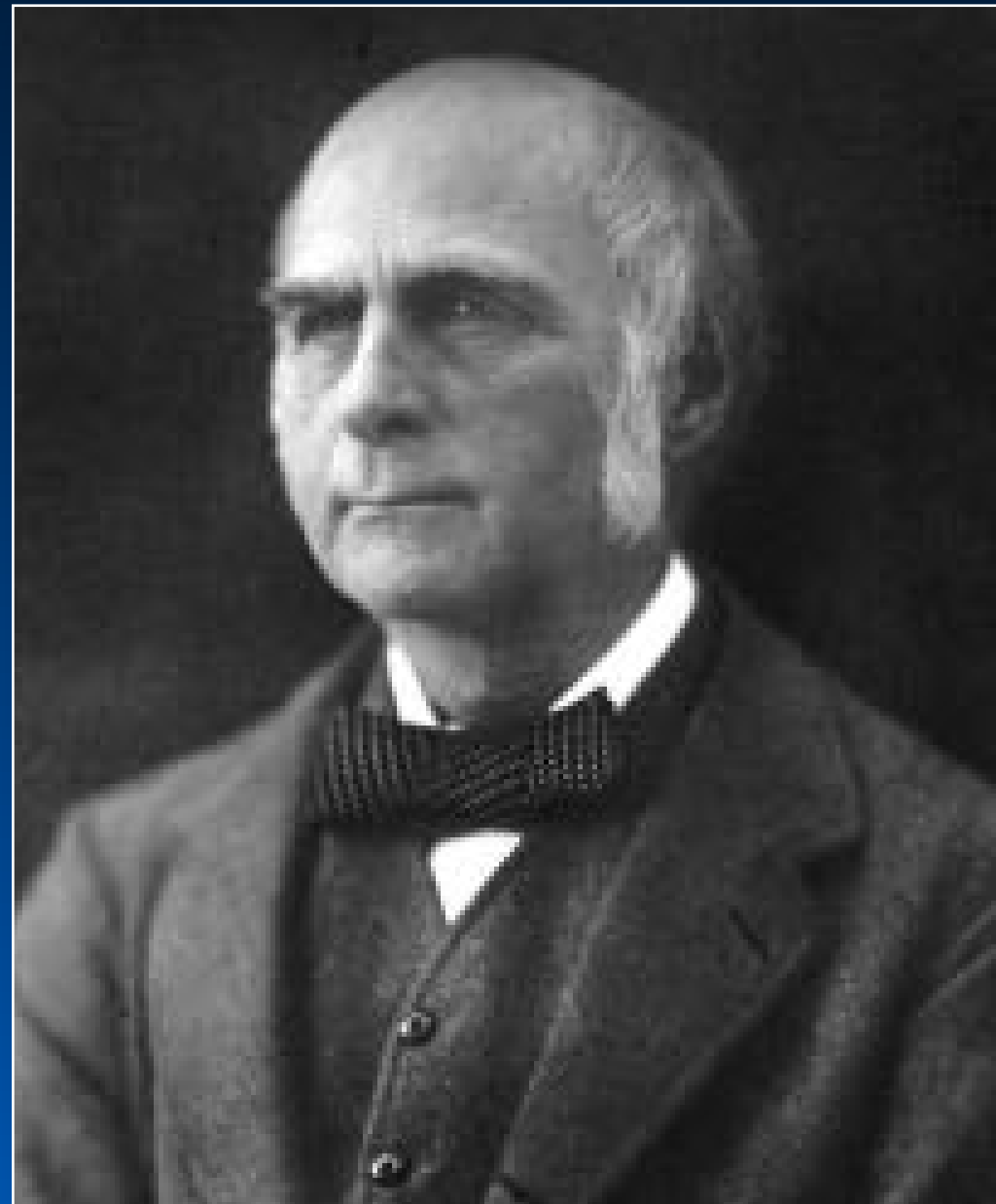


By hand we can easily first a curved line to these data. Such a line would be a much closer fit. But how can we do use mathematics to fit the curved line while minimizing deviations from it?

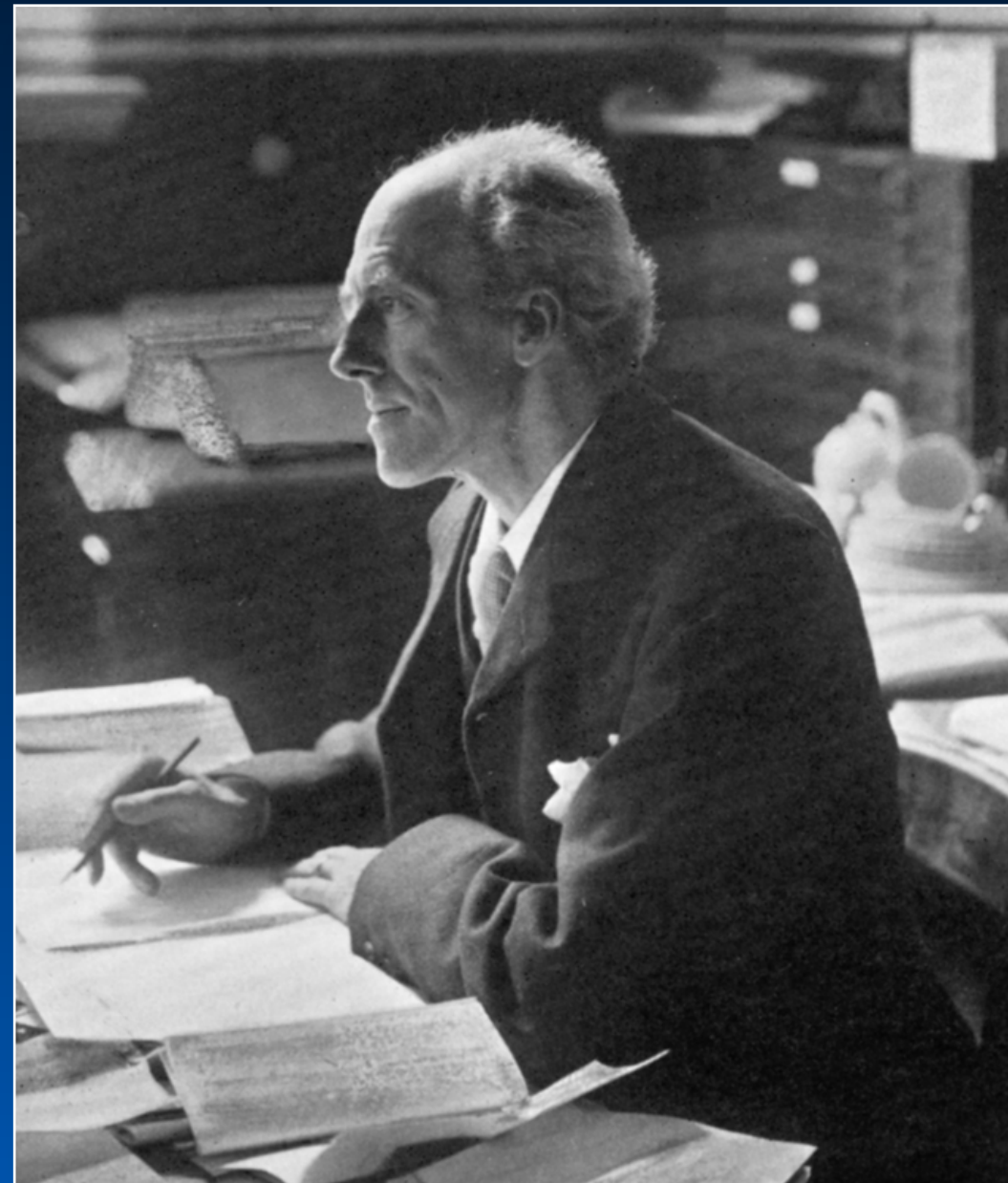
# Linear Regression, Modeling & Testing

---

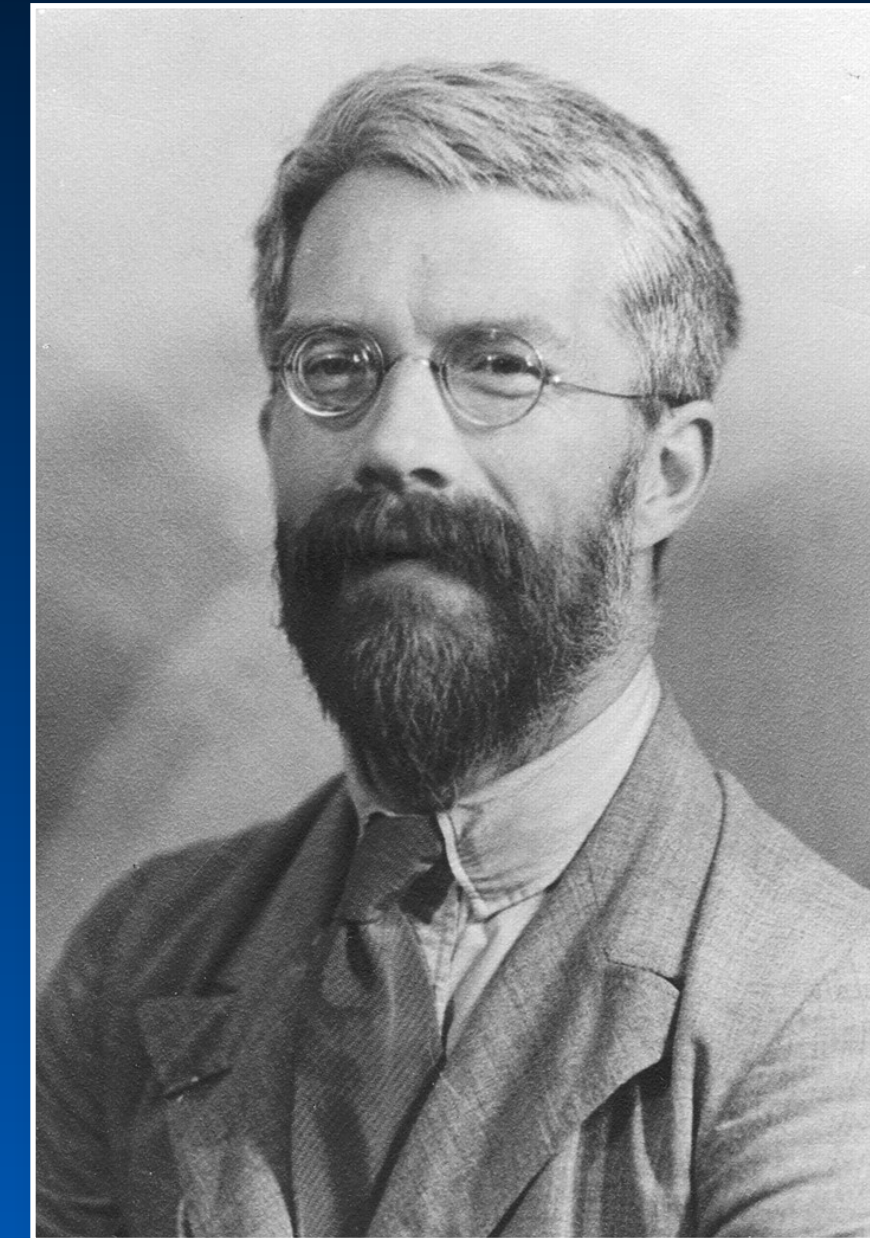
## Curvilinear Least-Squares Regression



Francis Galton  
(1822 - 1911)



Karl Pearson  
(1857 – 1926)



*R. A. Fisher*  
(1890-1962)

Fitting polynomial equations to data based based on Galton's foundational work, defined by his student, Karl Pearson, and further developed – essentially to its present state – by Ronald Fisher and by George U. Yule.

# Linear Regression, Modeling & Testing

---

## Curvilinear Least-Squares Regression

$$y_i = a + (\beta_1 x_i) + (\beta_2 x_i^2) + (\beta_3 x_i^3) + \cdots + \beta_m x_i^m + \epsilon_i$$

Where:  $i$  = observation no.

$a$  = y-intercept

$\beta$  = slope

$\epsilon$  = error

# Linear Regression, Modeling & Testing

---

## Curvilinear Least-Squares Regression

$$y_i = a + (\beta_1 x_i) + (\beta_2 x_i^2) + (\beta_3 x_i^3) + \cdots + \beta_m x_i^m + \epsilon_i$$

$$\sum_{i=1}^n y_i = \alpha \cdot n + \beta_1 \sum_{i=1}^n x_i + \beta_2 \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n x_i y_i = \alpha \cdot \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 + \beta_2 \sum_{i=1}^n x_i^3$$

$$\sum_{i=1}^n x_i^2 y_i = \alpha \cdot \sum_{i=1}^n x_i^2 + \beta_1 \sum_{i=1}^n x_i^3 + \beta_2 \sum_{i=1}^n x_i^4$$

# Linear Regression, Modeling & Testing

## Curvilinear Least-Squares Regression

$$y_i = \alpha + (\beta_1 x_i) + (\beta_2 x_i^2) + \epsilon_i$$

$$\begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}$$

The matrix formulation for this problem is more efficient and more easily generalized.

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{pmatrix} \cdot \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}^{-1}$$

# Linear Regression, Modeling & Testing

## Curvilinear Least-Squares Regression

$$y_i = \alpha + (\beta_1 x_i) + (\beta_2 x_i^2) + \epsilon_i$$

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{pmatrix} \cdot \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}^{-1}$$

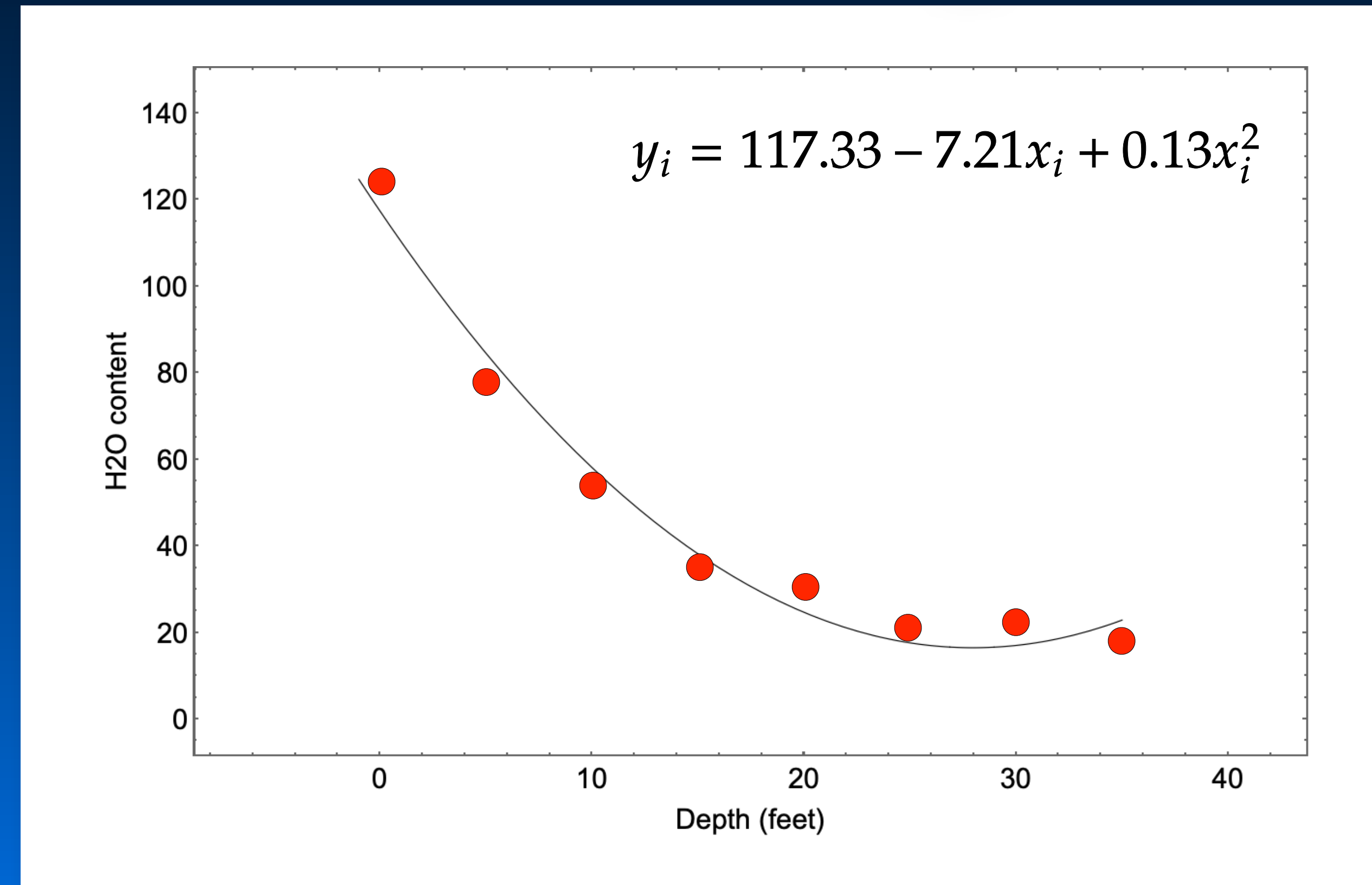
$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 382 \\ 3870 \\ 82200 \end{pmatrix} \cdot \begin{pmatrix} 8 & 382 & 3500 \\ 382 & 3500 & 98000 \\ 3500 & 98000 & 2922500 \end{pmatrix}^{-1}$$

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 117.32 \\ -7.21 \\ 0.13 \end{pmatrix}$$

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |



Our curvilinear OLS regression has minimized the square of the distances between the data points and the curvilinear regression model. Obviously the “fit” of the model to the data is much closer.

# Linear Regression, Modeling & Testing

## Analysis of Variance (ANOVA)

| Source of Variation | Sum of Squares                                | Dof     | Mean Squares                | <i>F</i> Test       |
|---------------------|---|---------|-----------------------------|---------------------|
| Due to Model        | $SS_A = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | $m - 1$ | $MS_A = \frac{SS_A}{m - 1}$ | $\frac{MS_A}{MS_E}$ |
| Unexplained (Error) | $SS_E = \sum_{i=1}^n (\hat{y}_i - y_i)^2$     | $N - m$ | $MS_E = \frac{SS_E}{N - m}$ |                     |
| Total Variation     | $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$       | $N - 1$ |                             |                     |

Where:  $\hat{y}$  = the estimated value of  $y$ ;  
 $\bar{y}$  = the arithmetic mean of  $y$ ;  
 $m$  = number of variables;  
 $n$  = number of samples.

# Linear Regression, Modeling & Testing

## Analysis of Variance (ANOVA)

| Source of Variation | Sum of Squares | Dof | Mean Squares | <i>F</i> Test |
|---------------------|----------------|-----|--------------|---------------|
| Due to Model        | 9308.4         | 1   | 9308.4       | 277.7         |
| Unexplained (Error) | 201.1          | 6   | 33.5         |               |
| Total Variation     | 9309.5         | 7   |              |               |

$$F_{0.05,dof:1,6} = 5.99$$

$$P_{F=277.7,dof:1,6} = 2.98 \times 10^{-6}$$

Very, very, very unlikely the estimated and residual variances are equal.

# Linear Regression, Modeling & Testing

## Depth vs H2O Content of Louisiana Estuary Mud Linear Least-Squares Regression

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | <i>F</i> Test |
|---------------------|----------------|--------------------|--------------|---------------|
| Due to Model        | 7547.0         | 1                  | 7547.0       | 23.07         |
| Unexplained (Error) | 1963.0         | 6                  | 327.1        |               |
| Total Variation     | 9309.5         | 7                  |              |               |

$$F_{0.05,dof:1,6} = 5.99$$

## Curvilinear Least-Squares Regression

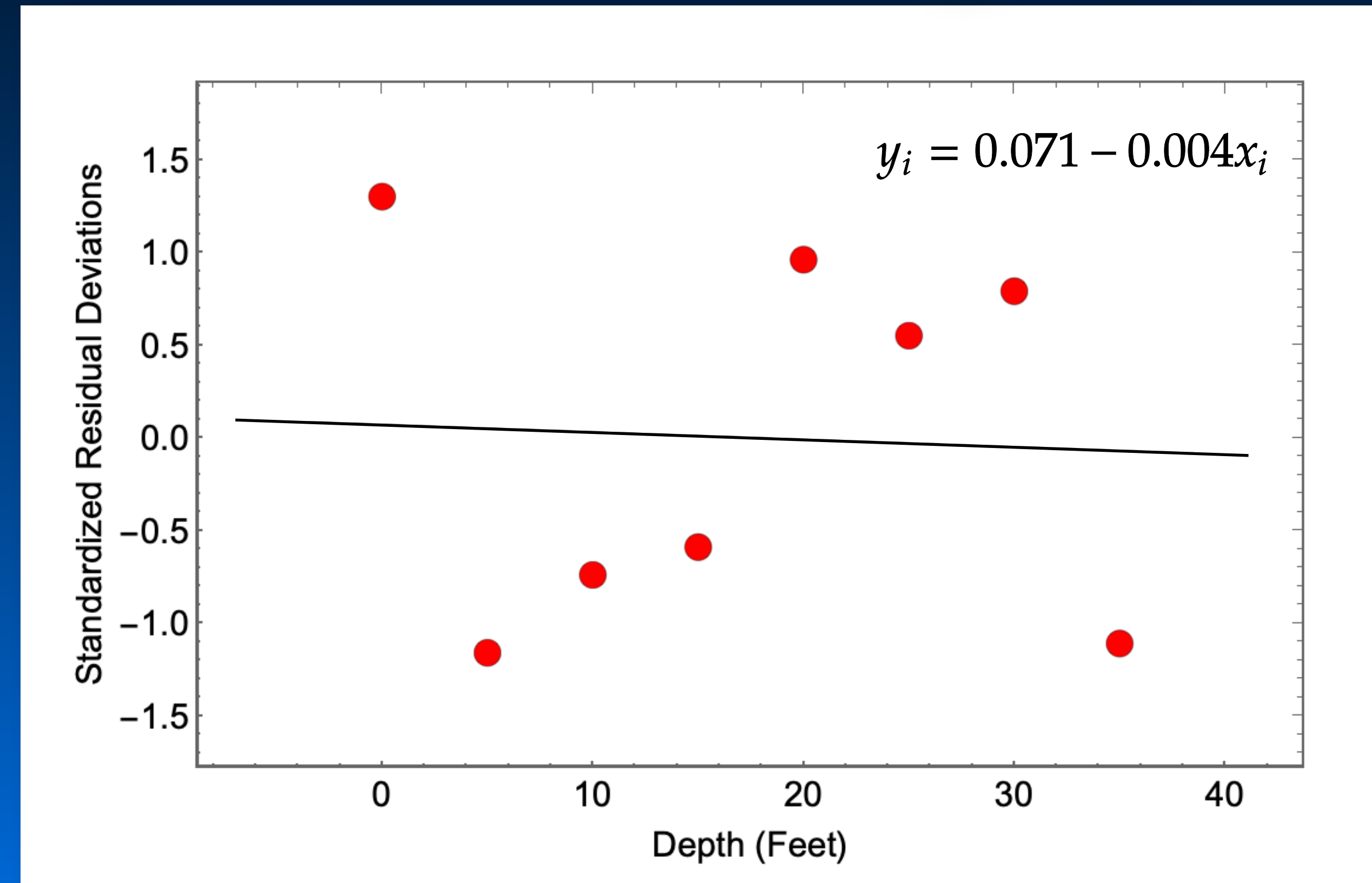
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | <i>F</i> Test |
|---------------------|----------------|--------------------|--------------|---------------|
| Due to Model        | 9308.4         | 1                  | 9308.4       | 277.7         |
| Unexplained (Error) | 201.1          | 6                  | 33.5         |               |
| Total Variation     | 9309.5         | 7                  |              |               |

$$F_{0.05,dof:1,6} = 5.99$$

# Linear Regression, Modeling & Testing

## Depth vs H<sub>2</sub>O Content of Louisiana Estuary Mud

| Depth (Feet) | H <sub>2</sub> O Content (g/100g solids) |
|--------------|--|
| 0            | 124                                      |
| 5            | 78                                       |
| 10           | 54                                       |
| 15           | 35                                       |
| 20           | 30                                       |
| 25           | 21                                       |
| 30           | 22                                       |
| 35           | 18                                       |



A plot of the standardized residual values reveals random structure. This indicates that the curvilinear linear model is close to optimal.

# Linear Regression, Modeling & Testing

## Test of Curvilinear Regression Residuals

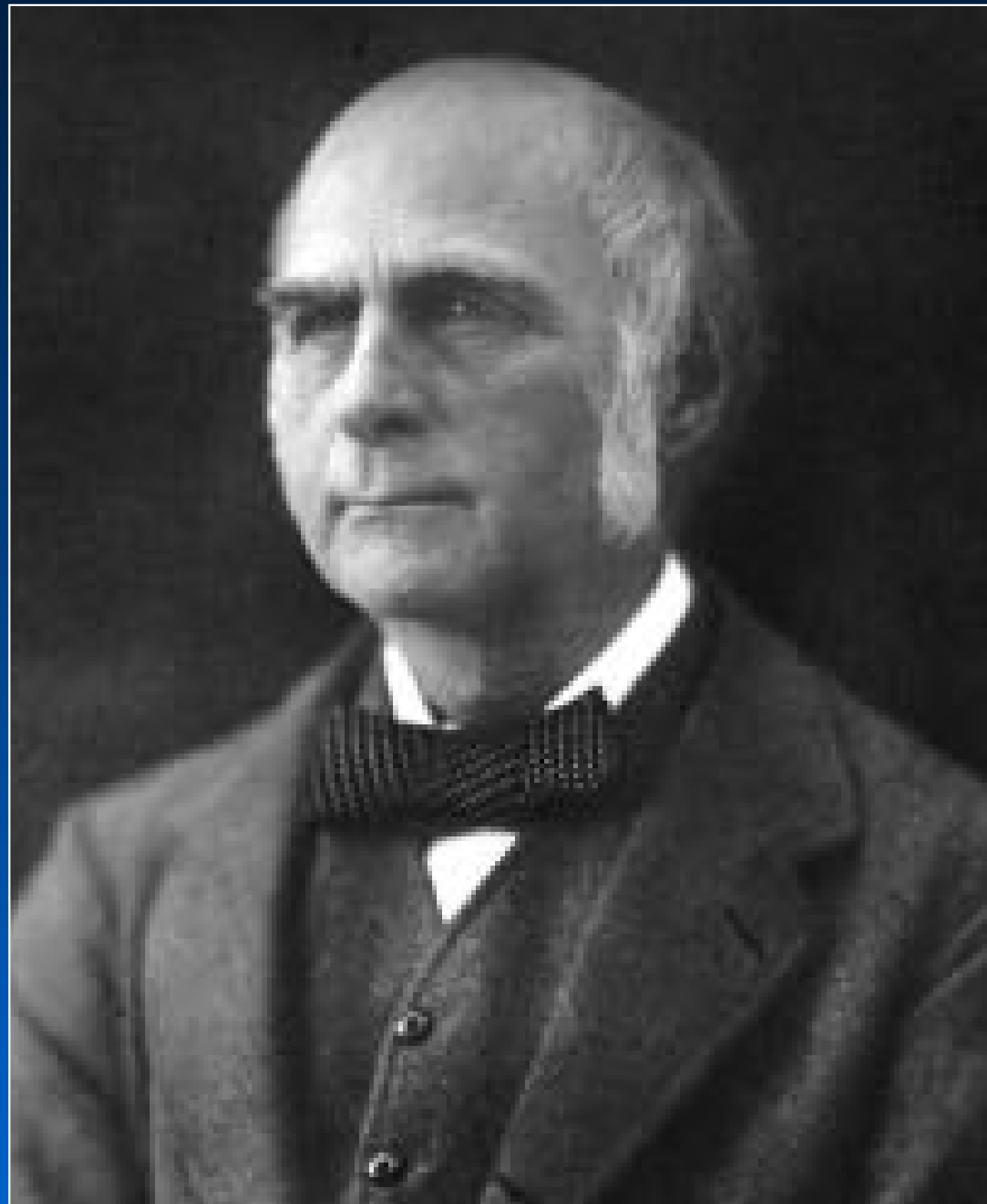
| Source of Variation | Sum of Squares | Dof | Mean Squares | <i>F</i> Test |
|---------------------|----------------|-----|--------------|---------------|
| Due to Model        | 0.65           | 1   | 0.65         | 0.02          |
| Unexplained (Error) | 200.93         | 6   | 33.49        |               |
| Total Variation     | 201.58         | 7   |              |               |

$$F_{0.05, \text{dof}:1,6} = 5.987$$

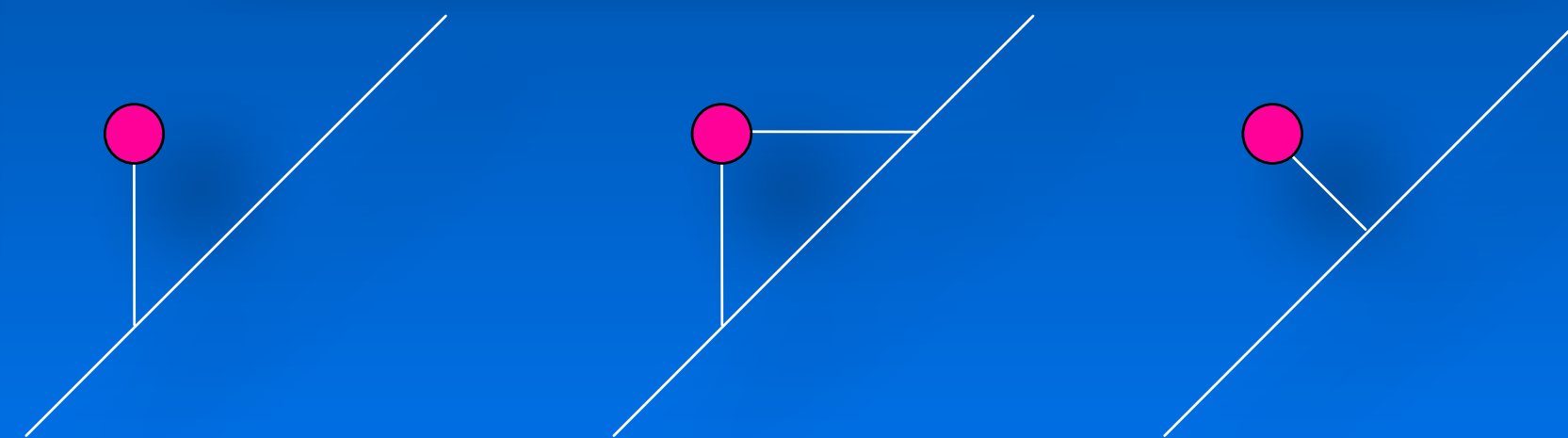
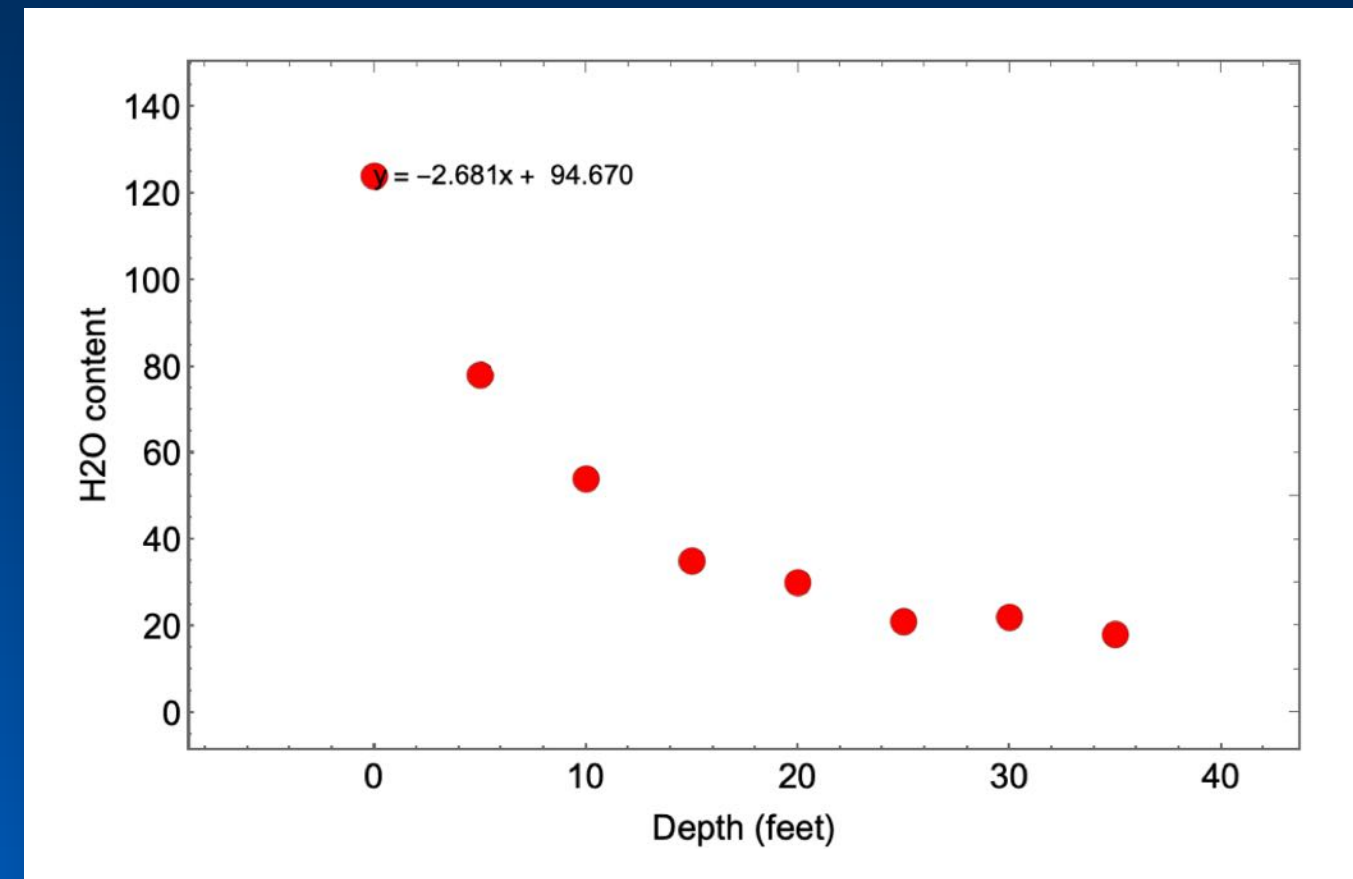
$$p_{F=0.02} = 0.9999$$

Very, very, very likely the estimated and residual variances are equal.

# Bivariate Regression, Modeling & Testing of Earth Science Data



Prof. Norman MacLeod  
School of Earth Sciences & Engineering, Nanjing University



# Linear Regression, Modeling & Testing

---

## Curvilinear Least-Squares Regression

$$y_i = \alpha + (\beta_1 x_i) + (\beta_2 x_i^2) + \epsilon_i$$

$$\sum_{i=1}^n y_i = \alpha \cdot n + \beta_1 \sum_{i=1}^n x_i + \beta_2 \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n x_i y_i = \alpha \cdot \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 + \beta_2 \sum_{i=1}^n x_i^3$$

$$\sum_{i=1}^n x_i^2 y_i = \alpha \cdot \sum_{i=1}^n x_i^2 + \beta_1 \sum_{i=1}^n x_i^3 + \beta_2 \sum_{i=1}^n x_i^4$$