Simple linear regression

E0420

Week 6

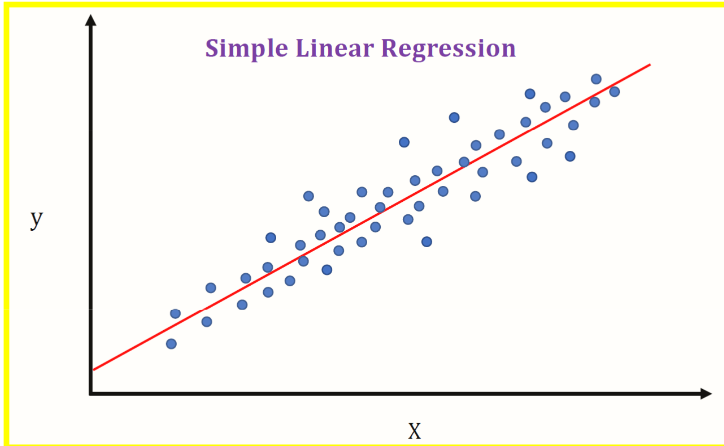
Constant/Intercept

Independent Variable

$$Y_i = \beta_0 + \beta_1 X_i$$

Dependent Variable

Slope/Coefficient



What is it good for?

- Testing associations between:
- 1 or more independent variables (IVs)
  - Categorical/binary
  - Ordinal
  - Continuous
- 1 dependent variable (DV, outcome)
- Possibility to add covariates

Goal of regression

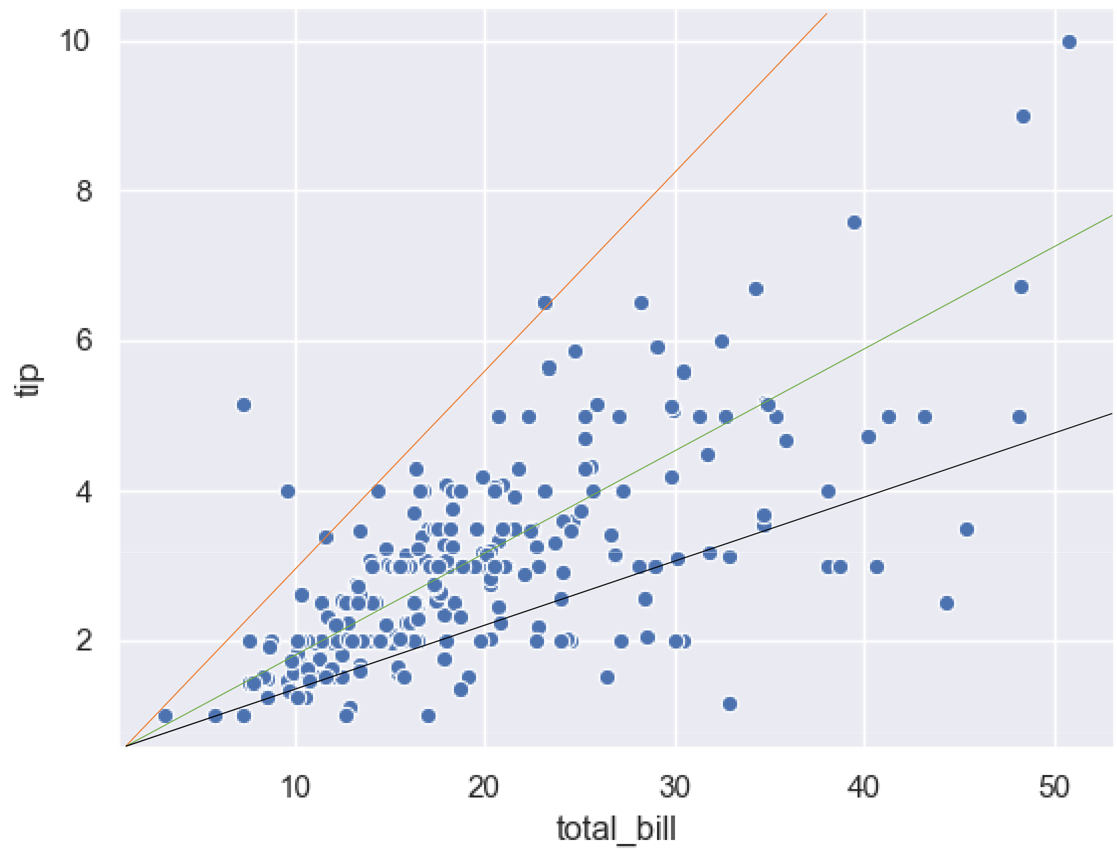
- Prediction
  - Use known IV(s) to predict DV
  - Correlation ≠ causation still applies!
- Explanation
  - Explain the DV’s variability by partitioning a “chunk” that is explained by the IV, and a “chunk” that is left unexplained

How regression works

- Finding a matematical function, or model that best describes the association between the variables
- Simple linear regression - a straight line, or linear equation
- The regression line is obtained that provides the best possible description of the relationship between X (IV) and Y (DV)

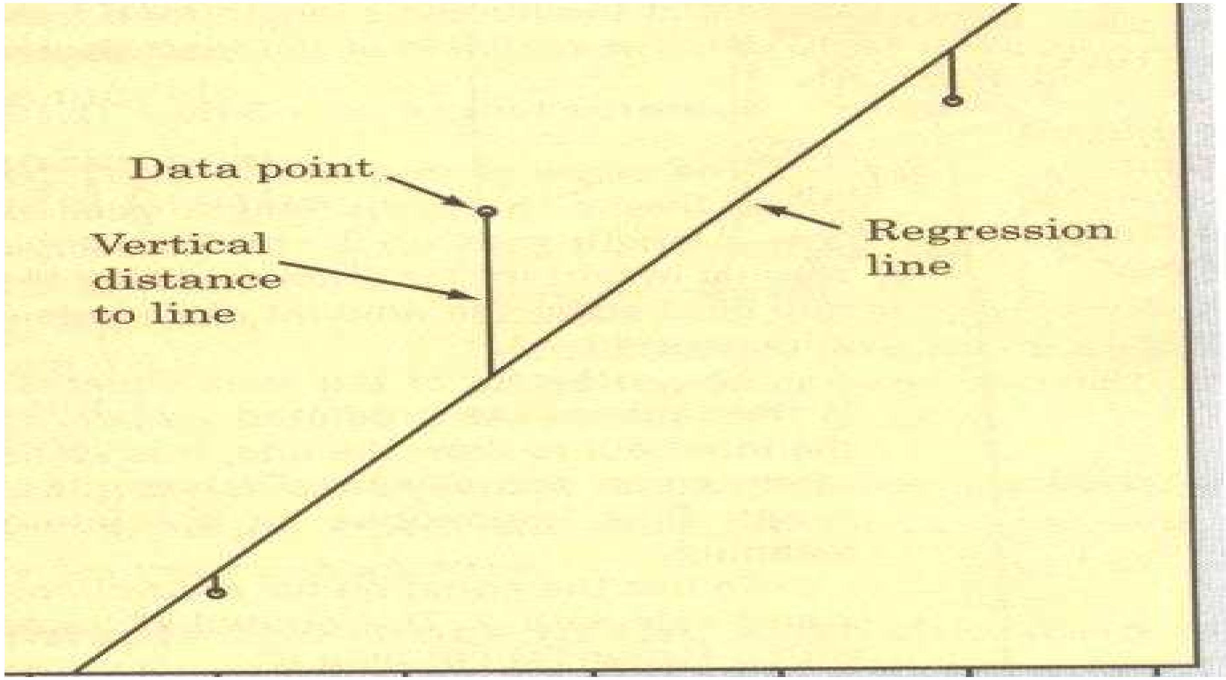
- If the association is not linear, we can model also quadratic function

Which line is the best fitting?



Which line?

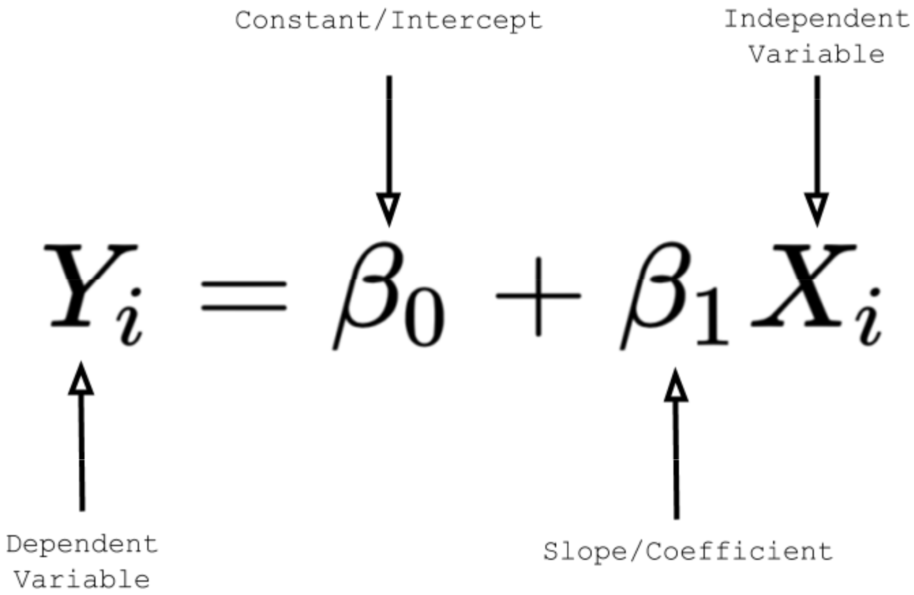
- The regression line selected is the one that minimizes the sum of the squared vertical distances to the data points



The regression line

- The **slope** of the line is given by a constant value used for everyone in the sample, β_1
  - How much of a change in Y (DV) is expected for every one-unit increase in X (IV)
  - Unstandardized (B) or standardized (β)
- The point at which the line crosses the Y axis is also a constant, β_0
  - **Intercept**
  - This is also the value of Y (DV) when X equals zero (IV)
- X_i and Y_i are variable scores for each observation

$$Y_i = \beta_0 + \beta_1 X_i$$



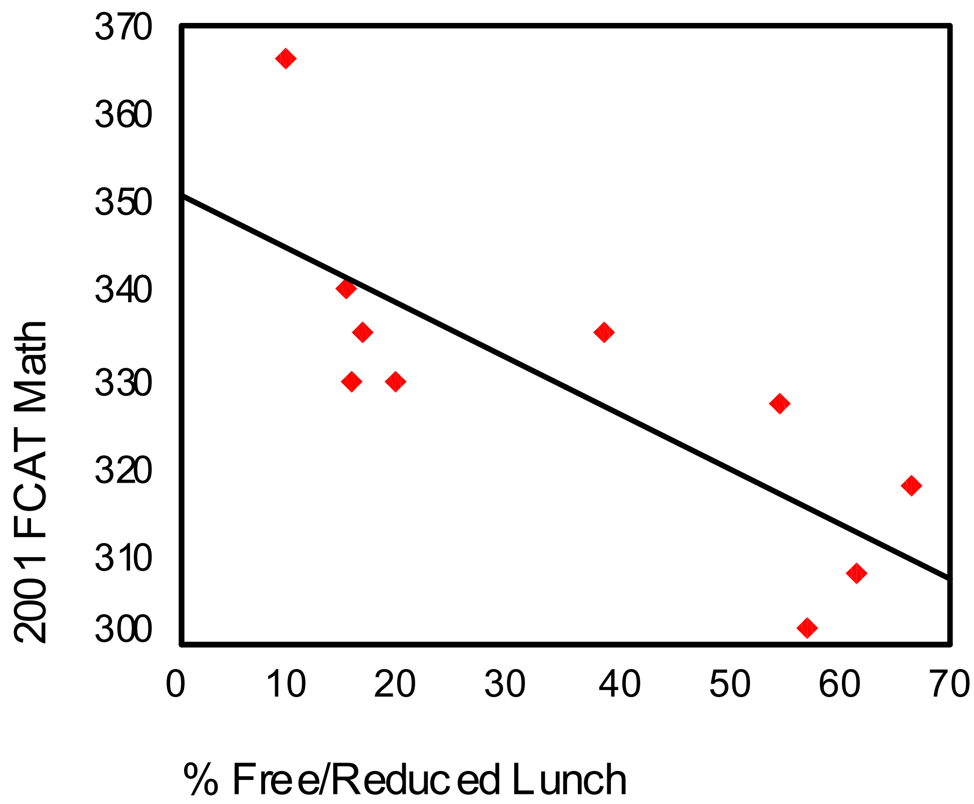
FCAT example

- The following data show 2001 FCAT math scores and the percentage of free/reduced lunch students for 10 elementary schools
- What is the IV? And DV?

% Free Lunch	2001 FCAT Math
9.4	366
57.2	302
54.6	327
15.7	330
19.5	330
38.5	335
61.3	308
66.3	318
16.4	335
14.9	340

FCAT example

- The values of B_1 and B_0 are:
  - $B_1 = -.618$
  - $B_0 = 350.969$
- The regression equation is
  - $Y_i = 350.969 - .618(X_i)$
- Interpreting the unstandardized regression coefficient:
  - For every 1% increase in the free/reduced lunch rate, a .618 ***decrease*** is predicted in FCAT scores

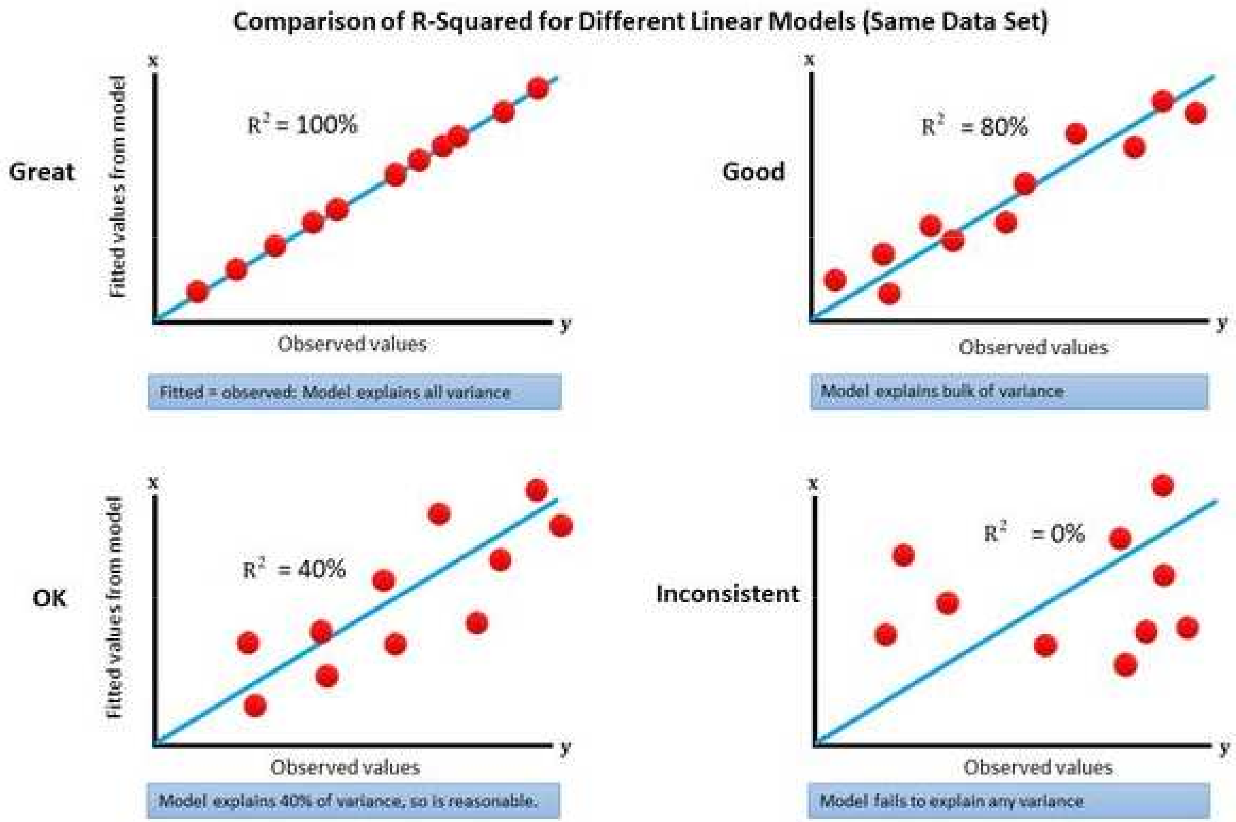


Predicting scores (DV from IV)

- Consider the school with a F/R lunch rate of 38.5
- $Y'_i = 350.969 - .618(38.5) = 327.176$
- In regression terms, Y' is the *predicted score*, or predicted value of Y
- Y' would be the same for every school with F/R % (i.e, X) = 38.5
- However, the predictions come with an error!

How much of an error?

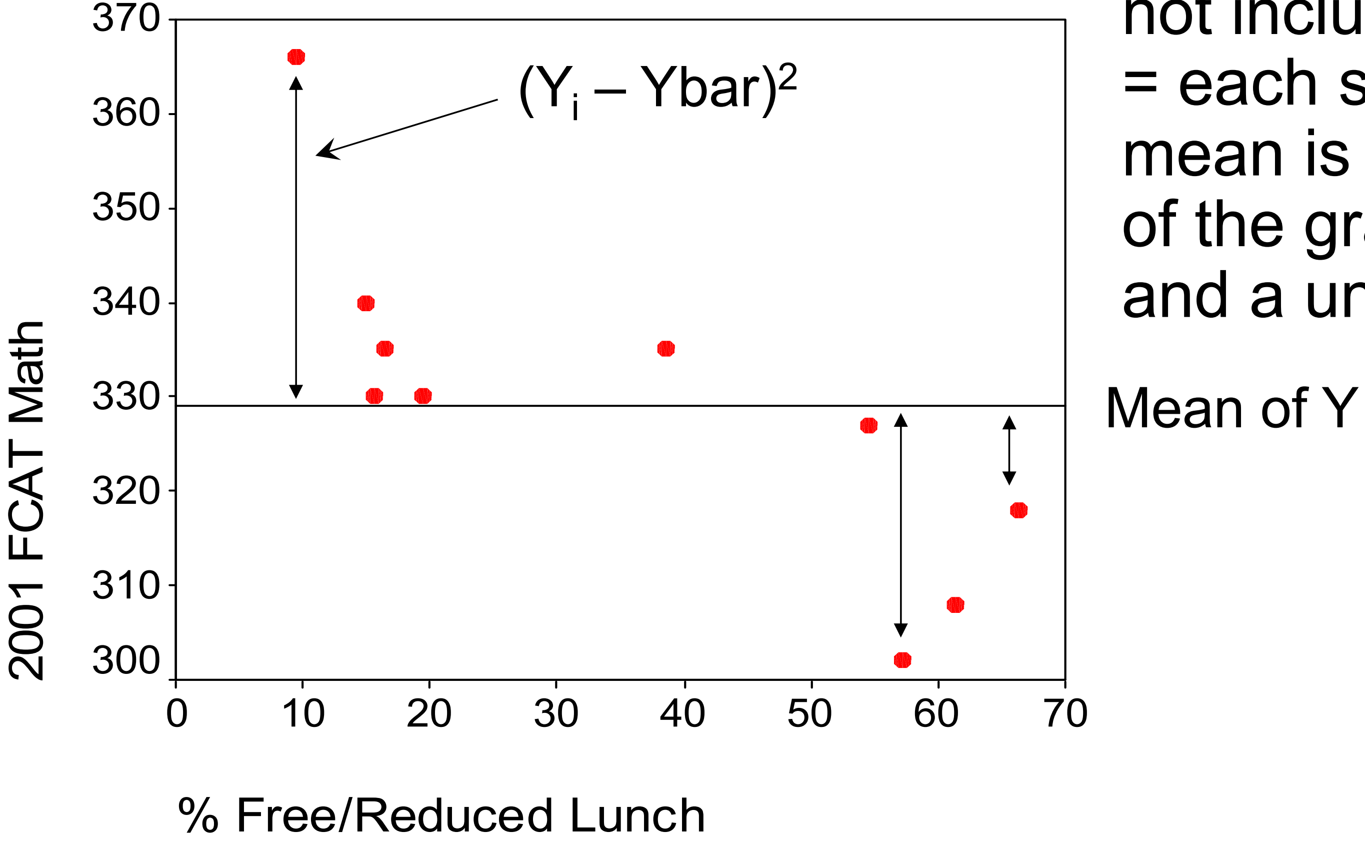
- How well does our line fit the data?
- How much variability in DV is explained by IV?



Explaining variability

- Consider the school with a free/reduced lunch rate (X_1) of 9.4 and an FCAT mean (Y_1) of 366
- This school was 36.9 FCAT points above the *grand* mean of Y (329.10)
- This distance of 36.9 points represents school 1's contribution to the Y variability
- It is the goal of the regression procedure to explain this total variation (SS_{TOTAL})

Graphic of SS_{TOTAL}

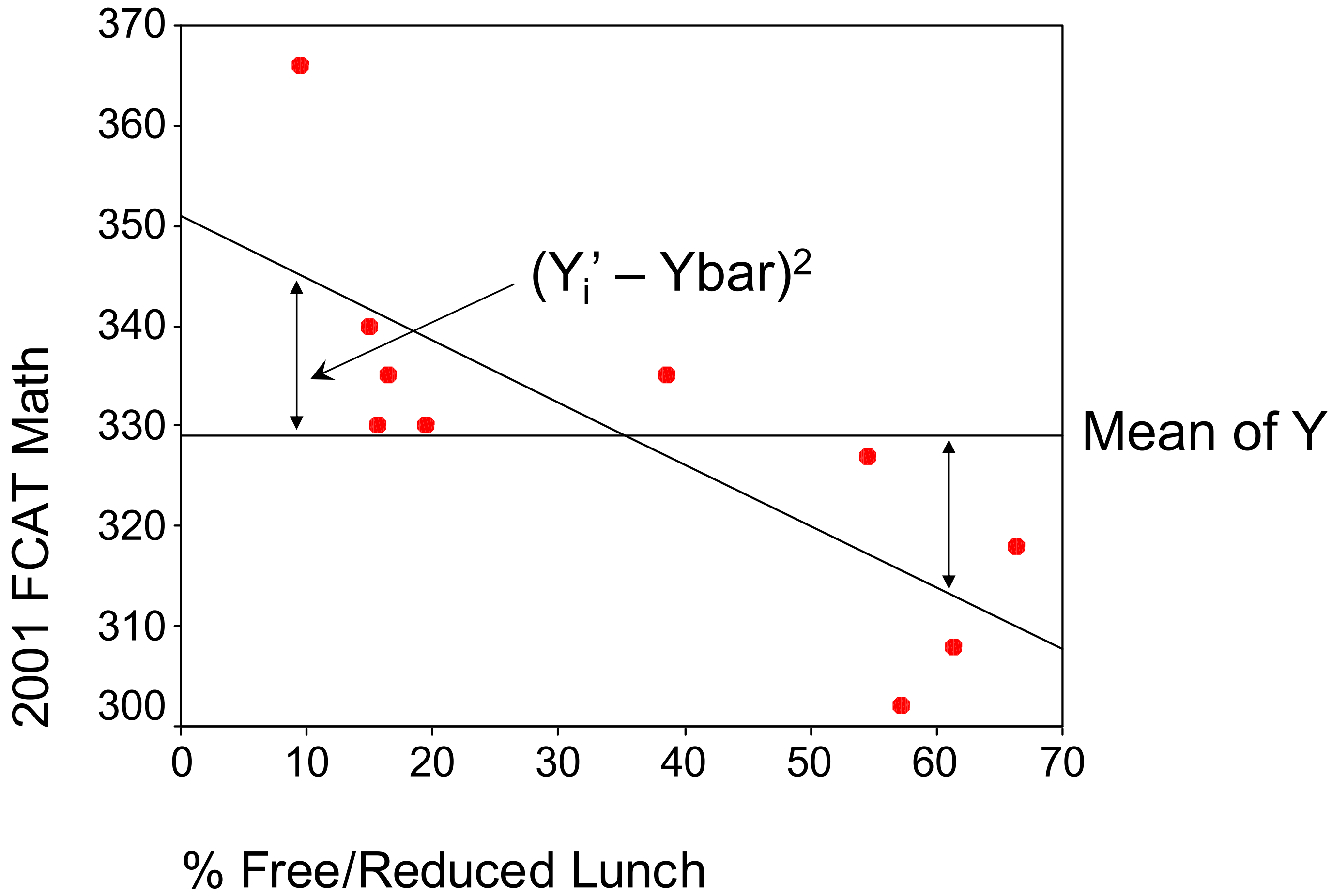


This model does not include the IV = each school's mean is a function of the grand mean and a unique error

The effect of IV

- Quantification of the shift in scores from the overall mean that can be attributed to the IV
- This shift can be computed for each value of X (the IV)
- This is found using the predicted Y scores from the regression equation (line)
- The variability attributed to the IV for the entire sample is computed by
  - Squaring each expected distance from the mean (the positive and negative distances would cancel out otherwise)
  - Summing these values across the entire sample (SS_{REG})

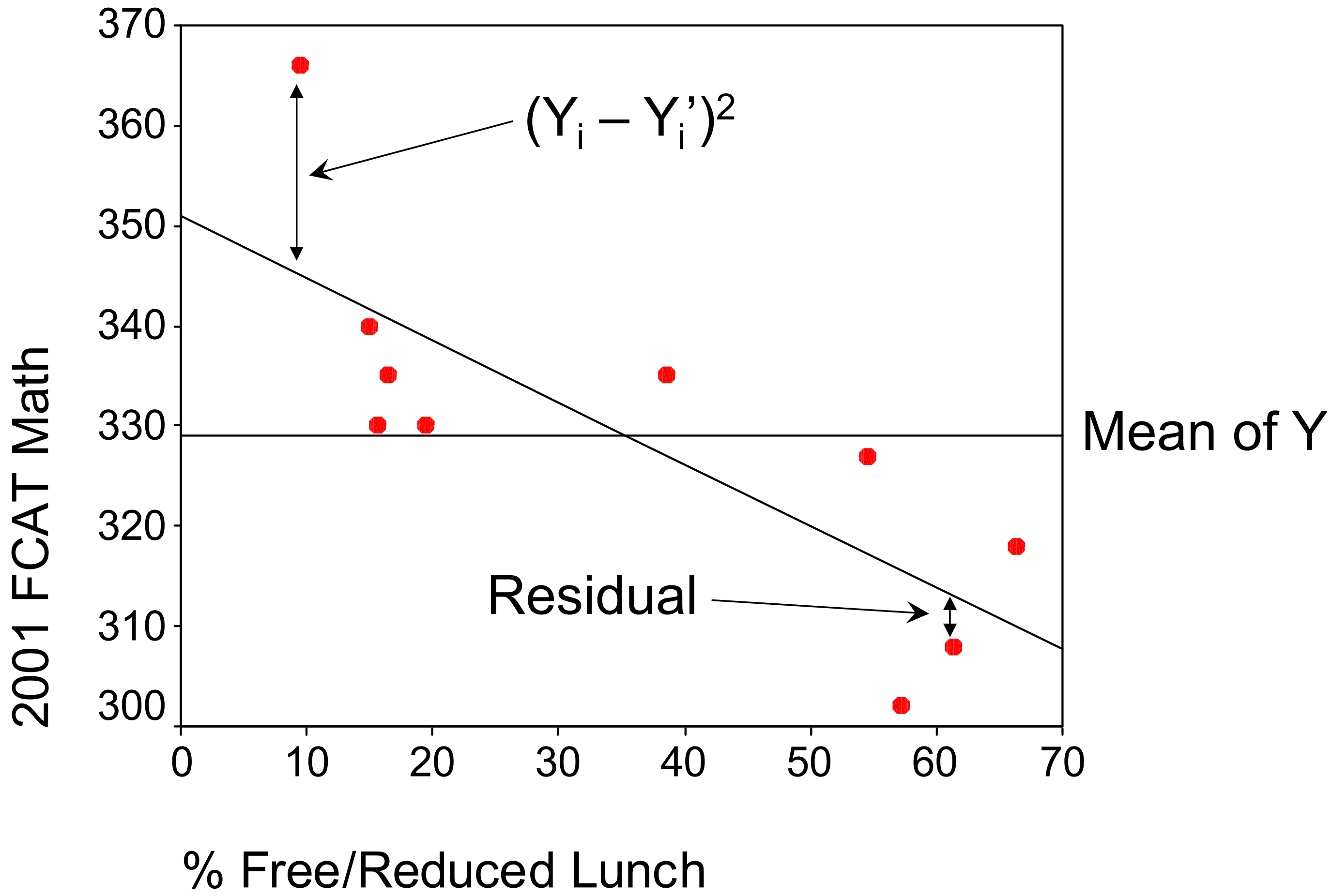
Graphic of SS_{REG}



Residual (error) variability

- The IV does not completely explain the variation in Y scores
- The portion of the variation around the mean that is not captured by the IV is called *residual* variability
- This is defined as the difference between the observed Y values and those predicted by the regression line
- $e_i = Y_i – Y'_I$
- The residual variability for the entire sample is computed by
  - Squaring each person's residual (the positive and negative errors would cancel out otherwise)
  - Summing these values across the entire sample (SS_{RES})

Graphic of SS_{RES}



FCAT example

- The residual for the school with a 9.4% F/R lunch rate would be:
  - $e_1 = Y_1 – Y'_1 = 366 – 345.16 = 20.84$
- Thus, the school's actual performance was 20.84 FCAT points higher than what would be predicted using %F/R
- 20.84 is the portion of that school’s FCAT variation that is <u>not</u> explained by the IV

Summary of FCAT example

- The 1st school's total distance, or variation from the mean of Y was 36.9
- Of this variation, 16.06 can be attributed to the IV, while 20.84 is unexplained
- Thus, the total variation for this school has been partitioned into two components that sum to the total variation for that school
- 36.9 = 16.06 + 20.84

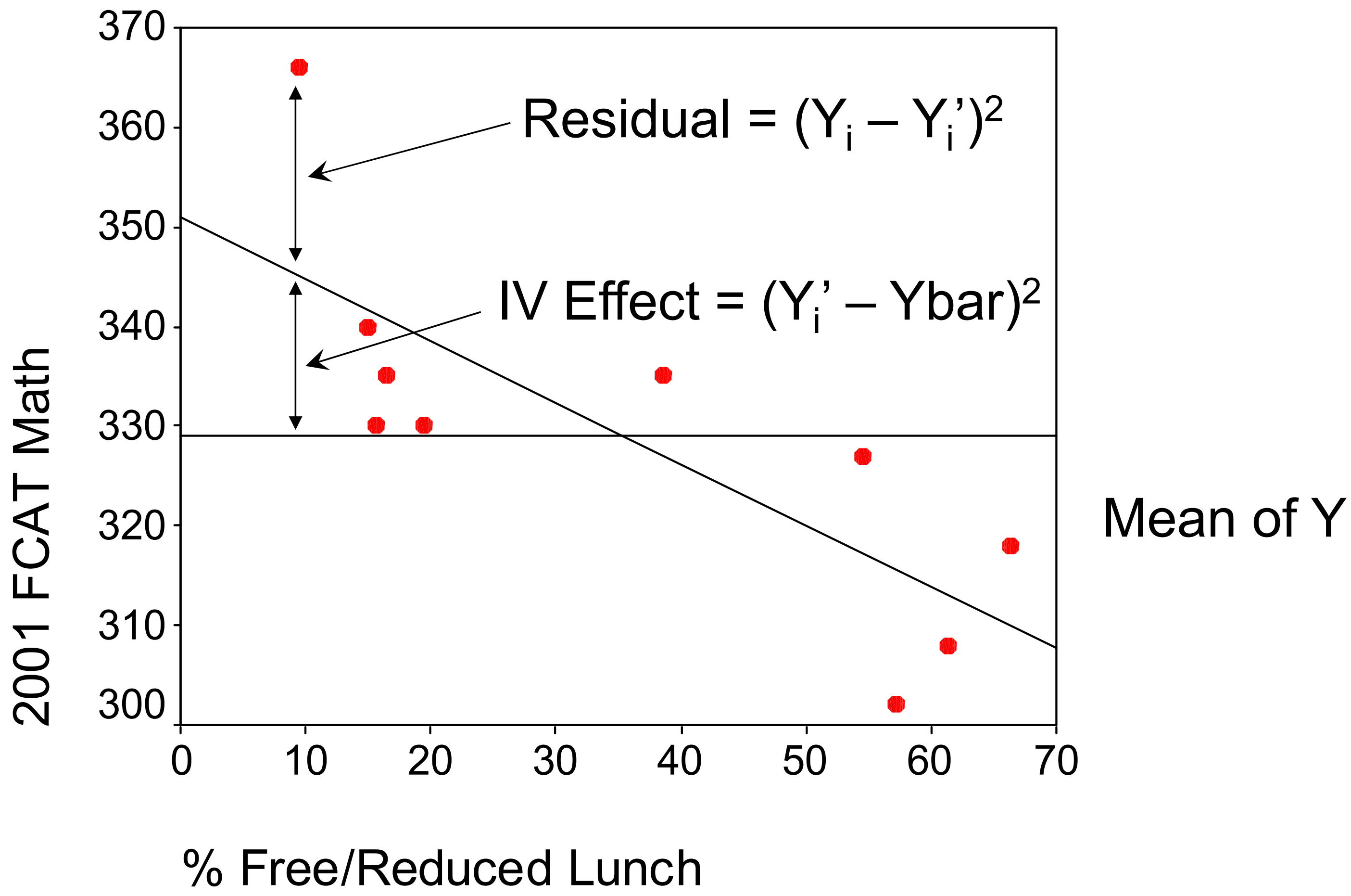
Partitioning total variability

- For the entire sample, the total variation in Y can be partitioned into two components:
  - Variability attributed to the IV (SS_{REG})
  - Variability not accounted for by the IV (SS_{RES})

$$SS_Y = SS_{REG} + SS_{RES}$$

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum \left(Y_i' - \overline{Y}\right)^2 + \sum \left(Y_i - Y_i'\right)^2$$

Graphic of variance partitioning



FCAT example

- The following quantities are obtained from the ANOVA summary table
  - SS_{TOT} = 2858.9
  - SS_{REG} = 1748.826
  - SS_{RES} = 1110.074

$$SS_Y = SS_{REG} + SS_{RES}$$

$$2858.9 = 1748.826 + 1110.074$$

Coefficient of determination (R^2)

- The total proportion (or %) of the DV variability that is explained by knowing X is called the coefficient of determination

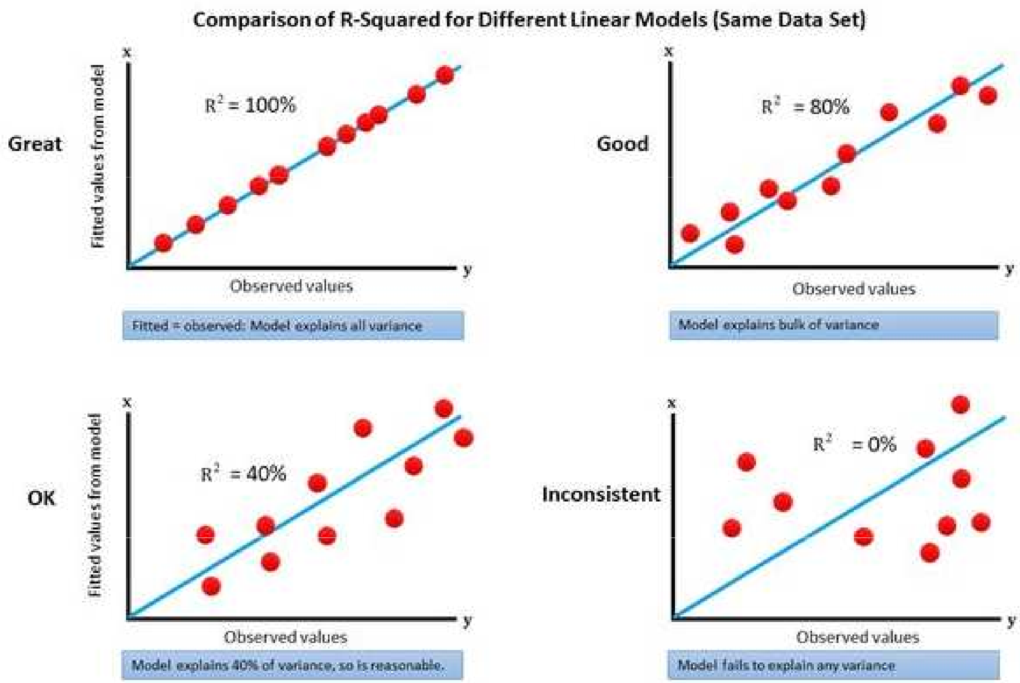
$$R^2 = \frac{SS_{REG}}{SS_{TOT}}$$

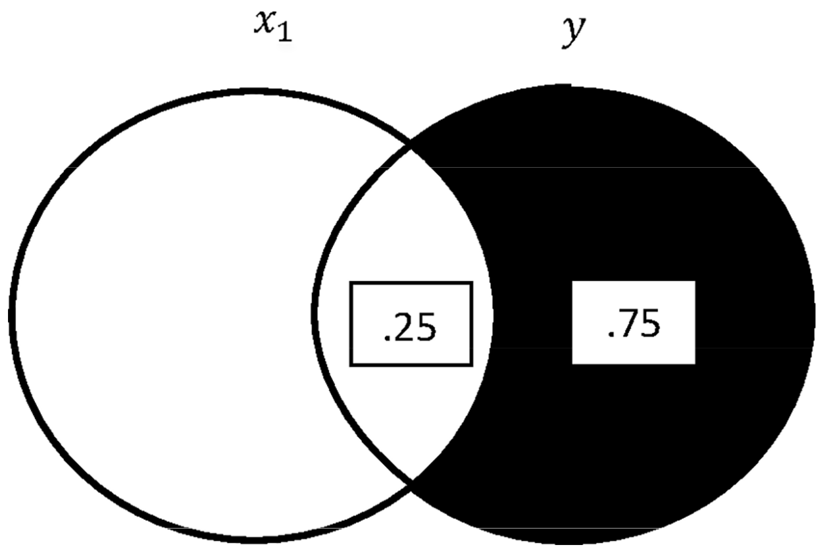
- FCAT example:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{1748.826}{2858.9} = .612$$

- Squaring Pearson's *r* yields $.782^2 = .612$

R^2


Comparison of R-Squared for Different Linear Models (Same Data Set)
Great
R² = 100%
Fitted values from model
Observed values
Fitted = observed: Model explains all variance
Good
R² = 80%
Observed values
Model explains bulk of variance
OK
R² = 40%
Fitted values from model
Observed values
Model explains 40% of variance, so is reasonable.
Inconsistent
R² = 0%
Observed values
Model fails to explain any variance


x_1
y
.25
.75

Significance testing of R^2

- RQ: Does the IV (IVs) account for variability in DV?
- H_0: R^2 is no larger than 0
- Test this assumption via F statistic, reject H0 if F statistic is ≥ critical F ($p \leq .05$)
- F represents a comparison of the variance explained by the IV and the residual variance

$$F = \frac{MS_{REG}}{MS_{RES}}$$

$$MS_{REG} = \frac{SS_{REG}}{df_{REG}} \qquad df_{REG} = k$$

$$MS_{RES} = \frac{SS_{RES}}{df_{RES}} \qquad df_{RES} = N - k - 1$$

- F test tells you if a *group* of variables are jointly significant

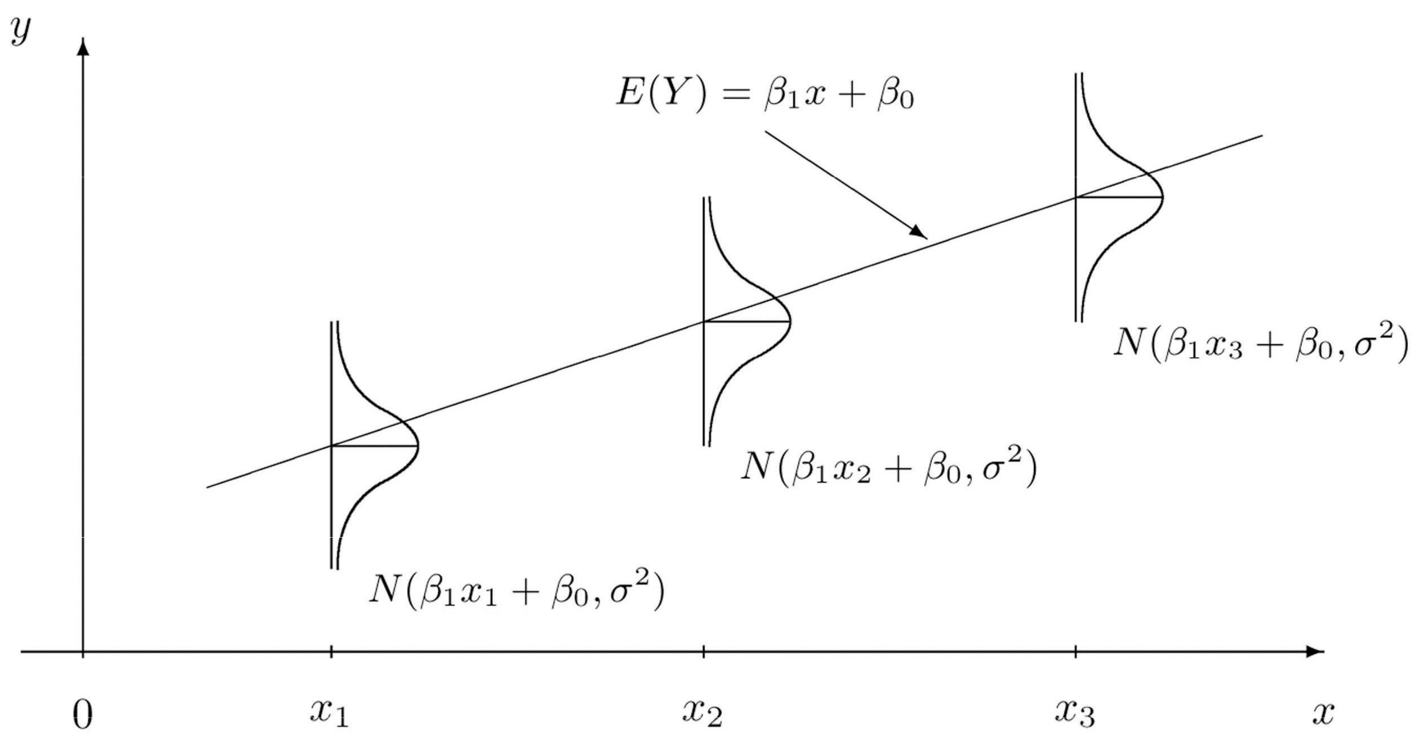
Significance testing of regression (*b*) coefficients

- Upon finding a significant R^2 value, determine which IV is contributing most to the significant R^2
- Unstandardized regression coefficients are tested using a t statistic
  - T-test tells you if a *single* variable is statistically significant
- This tests whether or not the slope is different from 0
  - H_0: $\beta = 0$, H_1: $\beta \neq 0$

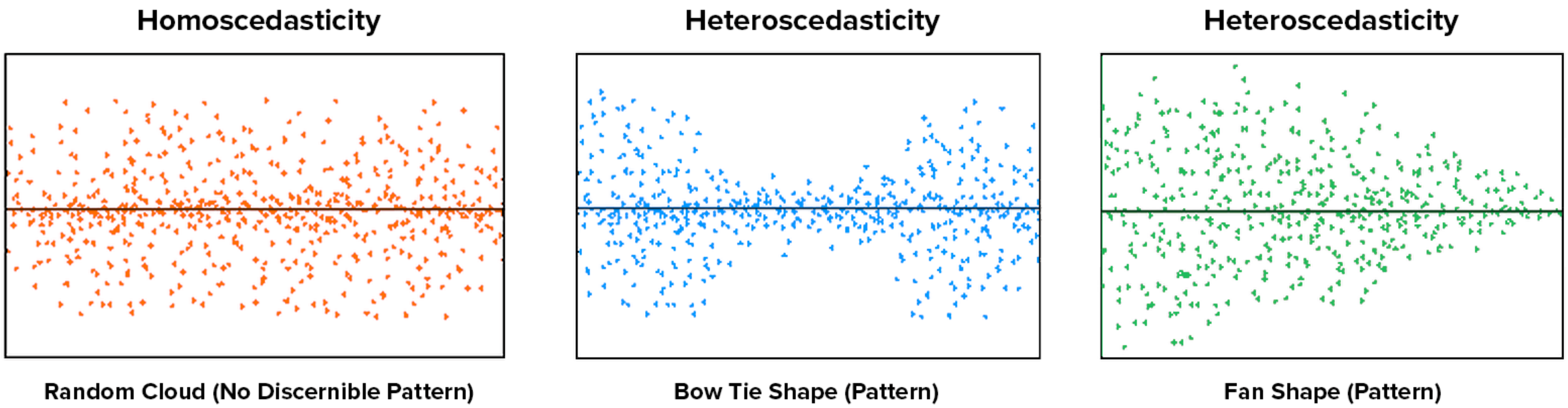
Simple linear regression assumptions

- **Linearity:** The relationship between X and the mean of Y is linear
- **Independent errors:** Residuals of observations should be uncorrelated
- **Homoscedasticity:** The variance of residual is the same for any value of X
- **Normally distributed errors**: Residuals in the model should be random, normaly distributed values with a mean of 0

Normality of residuals



Homoscedascity



Regression write-up

- The results of regression analysis showed that extraversion explained 35.8% of the variance (R^2 =.38, $F(2,55)=5.56$, $p<.01$) in aggressive tendencies (β = .56, $p<.001$).

Regression analysis steps

1. Run the analysis in SPSS
2. Check the assumptions
3. Determine the magnitude and significance of R^2
4. If R^2 significant, determine the magnitude and significance of regression coefficients (B, β)
5. Interpret R^2, B, β
6. Write-up the results