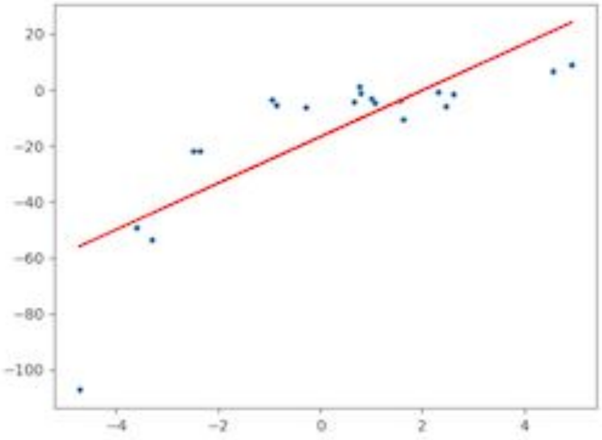
Regression Evaluation

Regression

Regression analysis - fitting lines to patterns of data



Regression terminology

Simple Linear Regression: One target and one independent variable

Multiple Linear Regression: One Target and >1 independent variables.

Univariate Linear Regression: Predicting one target variable.

Multivariate Linear Regression: Predicting multiple target variables.

Regression - Theoretical description

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Unexplained Error →

- **Independent**
- **Same variance (homoscedasticity)**
- **normally distributed**

Linear addition of ind variables

Intercept

Target

Regression Model - Estimate the coefficients

$$\hat{y}_i = b_0 + b_1\ x_i$$

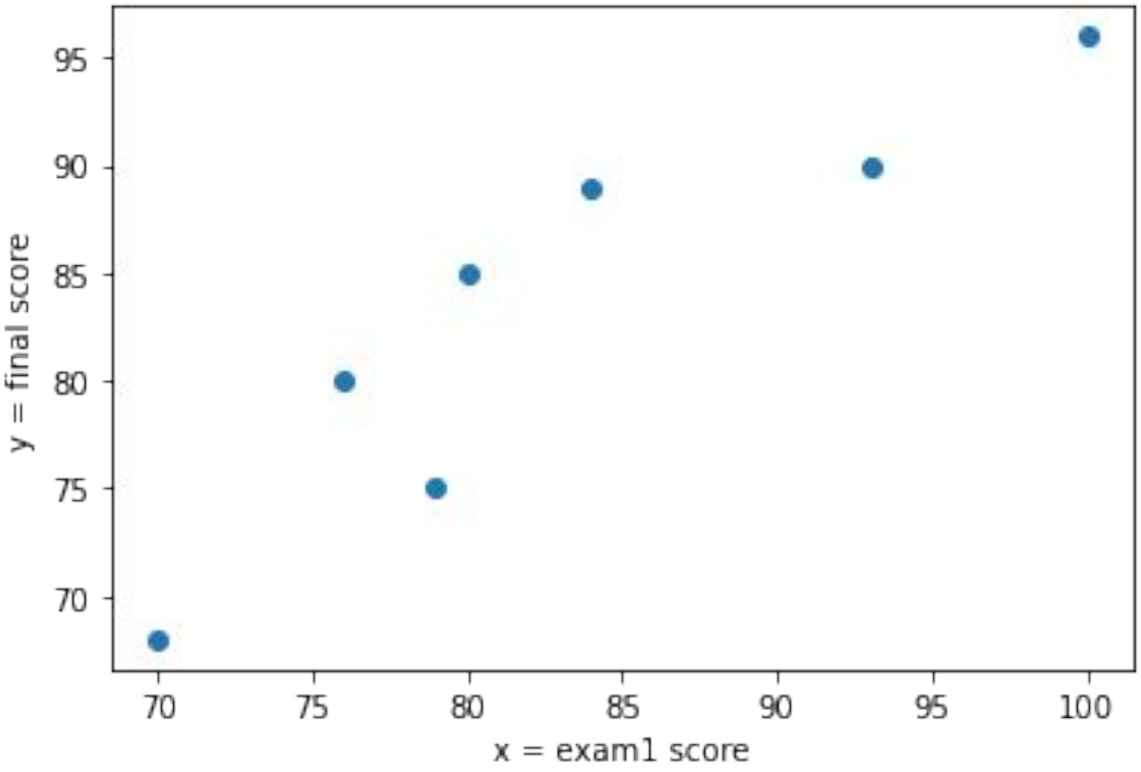
Estimated values of coefficients

Estimated Intercept

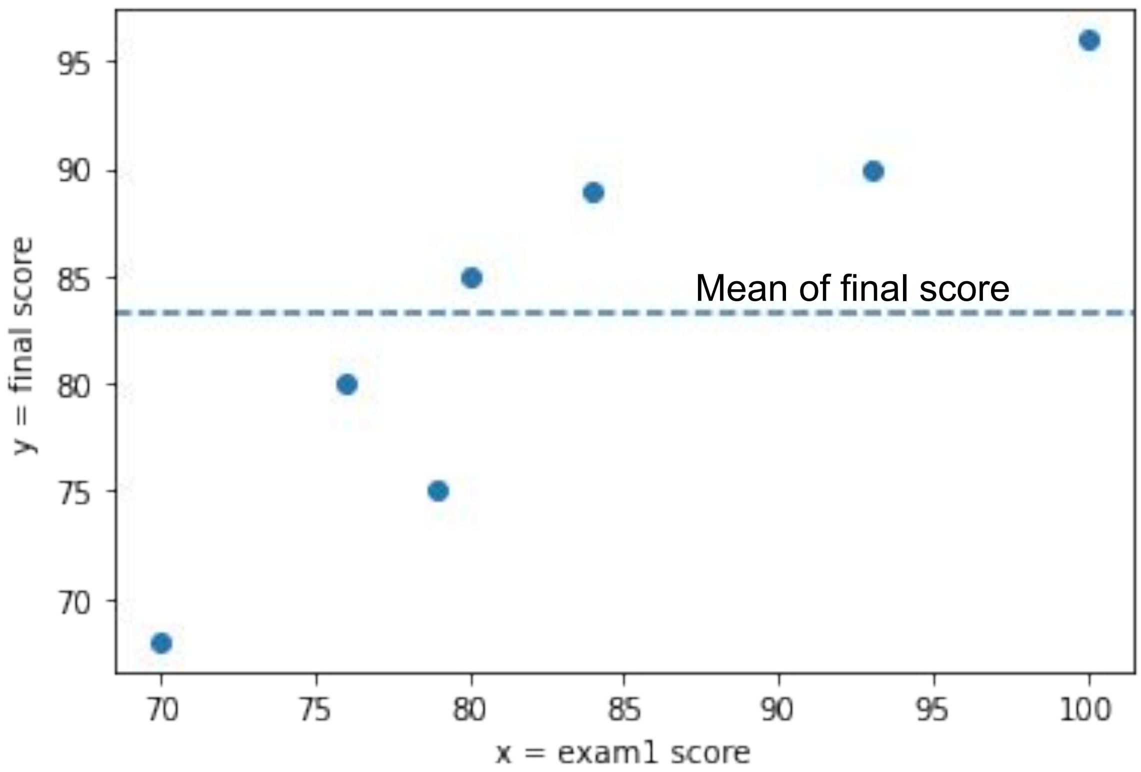
Estimate/Predicted Target

residuals: $e = y_i - \hat{y}_i$

Predict Final Exam Score



Baseline Model?



Regression Evaluation

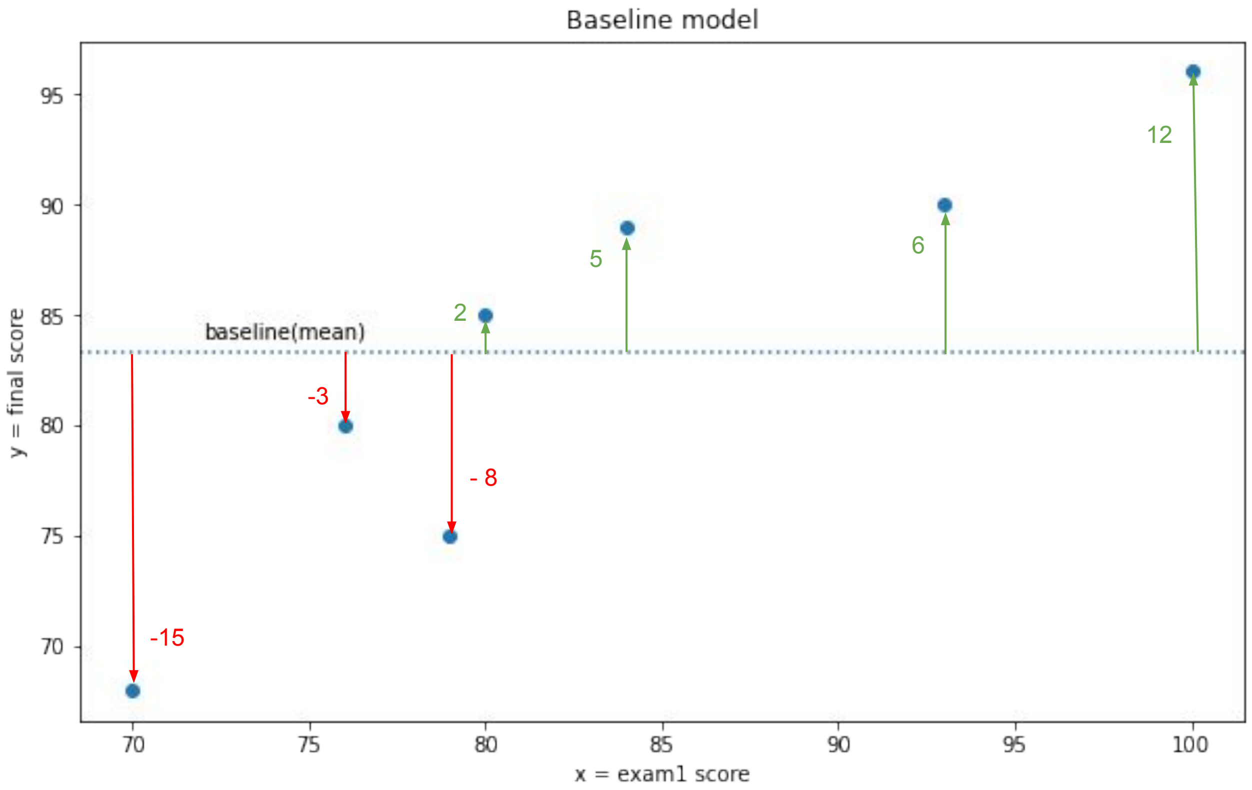
- Residual - error (actual minus prediction)

- SSE (sum of squared error)

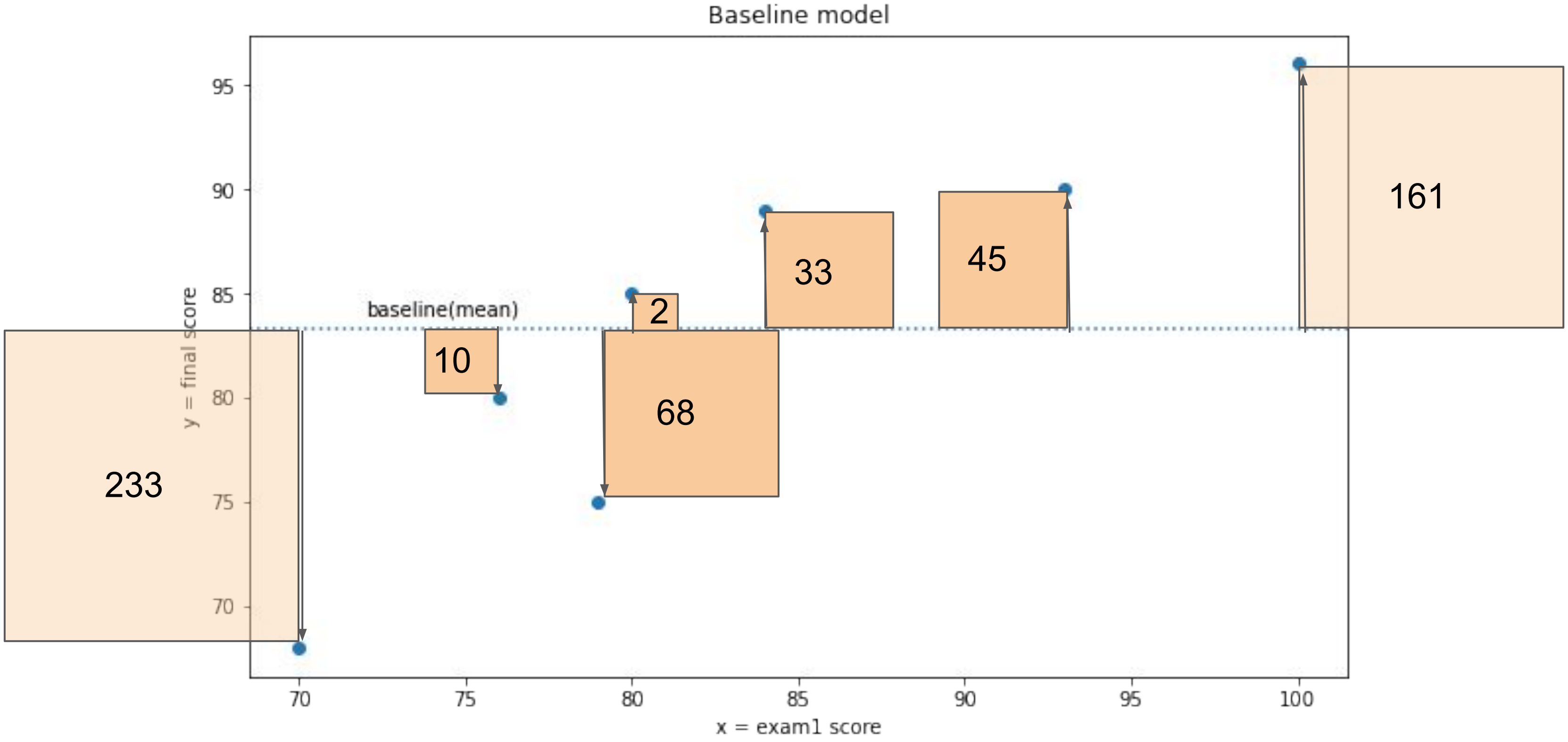
- MSE (mean squared error)

- RMSE (root mean squared error)

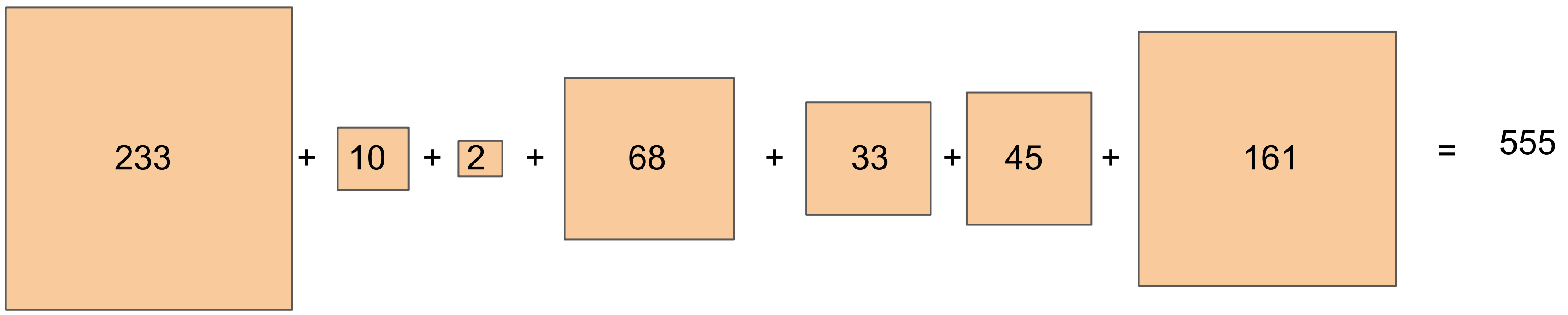
Baseline Residuals (Actual - Baseline Prediction)



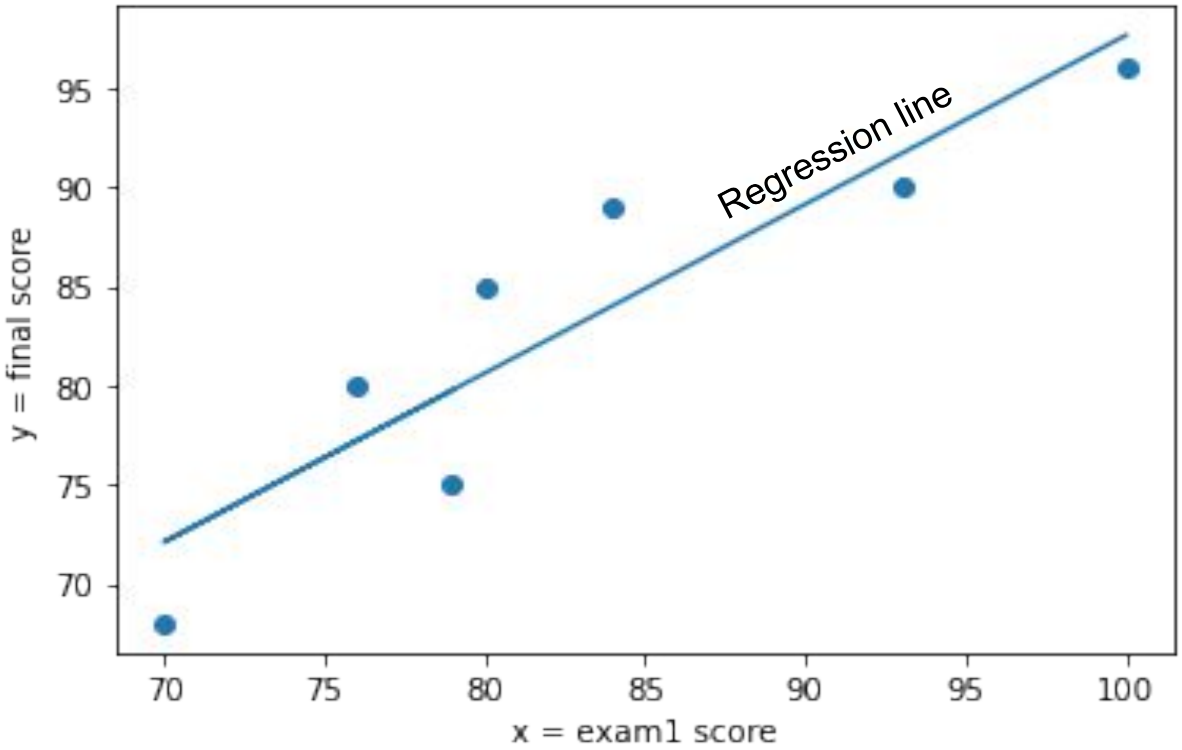
Baseline - Sum of Squared Errors



Baseline - Sum of Squared Errors (SSE)

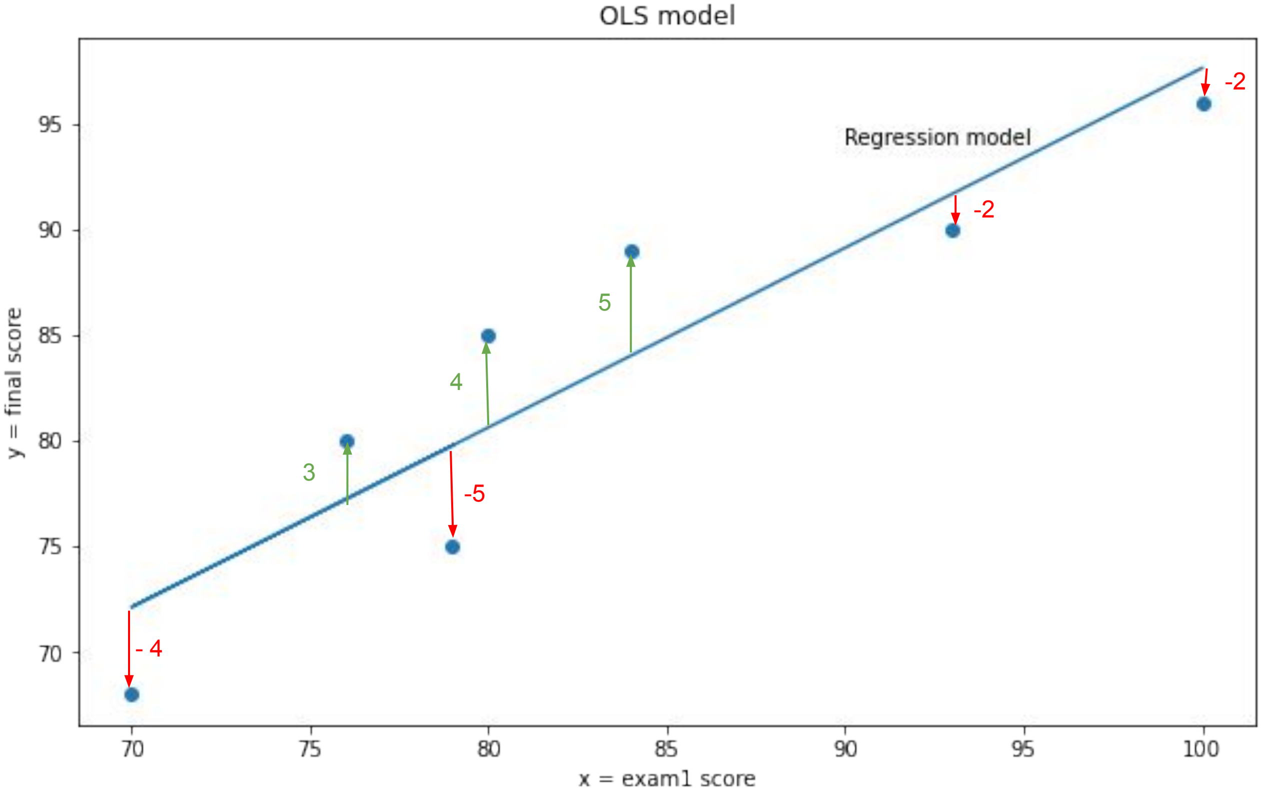


OLS model

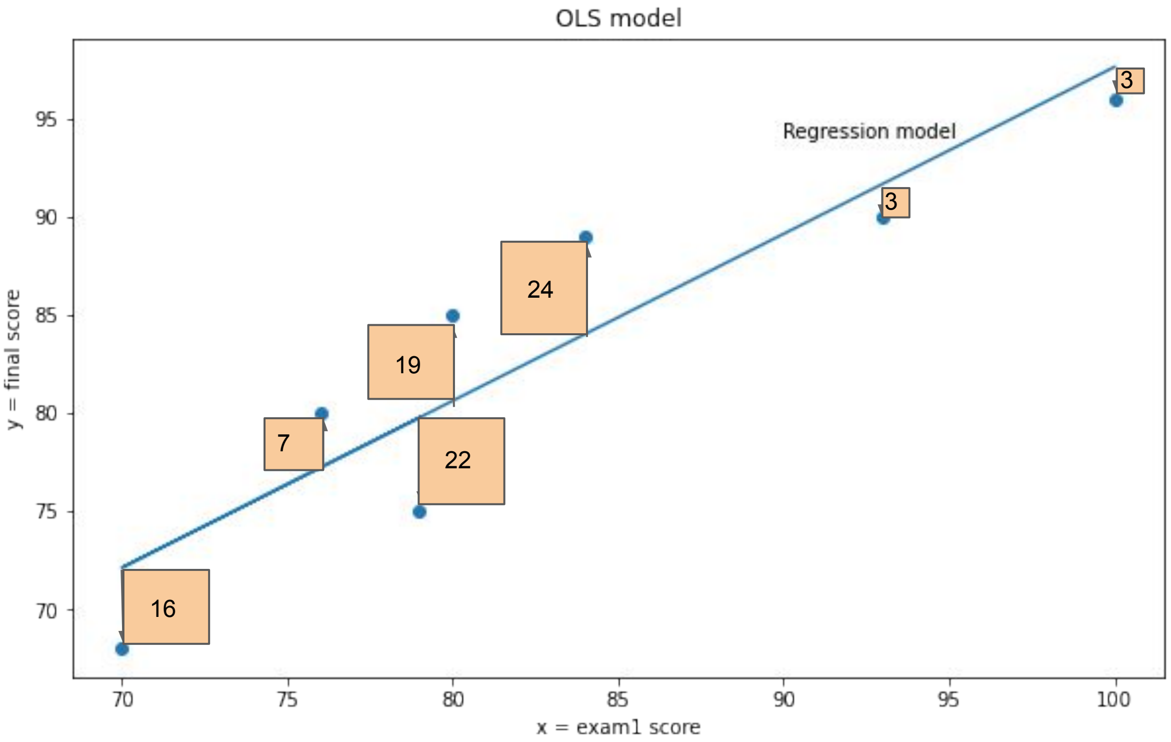


Regression line
y = final score
x = exam1 score
95
90
85
80
75
70
70
75
80
85
90
95
100

OLS model Residuals (Actual - OLS predictions)



OLS - Sum of Squared Errors (SSE)



OLS - Sum of Squared Errors (SSE)

$$\boxed{16} + \boxed{7} + \boxed{19} + \boxed{22} + \boxed{24} + \boxed{3} + \boxed{3} = 97$$

Baseline - Sum of Squared Errors (SSE)

233 + 10 + 2 + 68 + 33 + 45 + 161 = 555

OLS - Sum of Squared Errors (SSE)

16 + 7 + 19 + 22 + 24 + 3 + 3 = 97

Regressions Matrices

1. Residual
2. Sum of Square Error (SSE)
3. Mean Squared Error (MSE)
4. Root Mean Square Error (RMSE)

Regressions Matrices

Residual	$y_i - \hat{y}_i$
Sum of Squared Error (SSE)	$\sum (y_i - \hat{y}_i)^2$
Mean Squared Error (MSE)	$\frac{1}{n} \sum (y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$

R^2 - Explained Variance (aka coefficient of determination)

Fraction of variance/error explained by regression model

- $R^2 = 1.0$ (all of the data points fall perfectly on the regression line)
  - The predictor x accounts for all of the variation in y!

- $R^2 = 0$ (model performance is same as predicting baseline i.e. predicting mean)
  - The predictor x accounts for none of the variation in y!

R^2 - Explained Variance

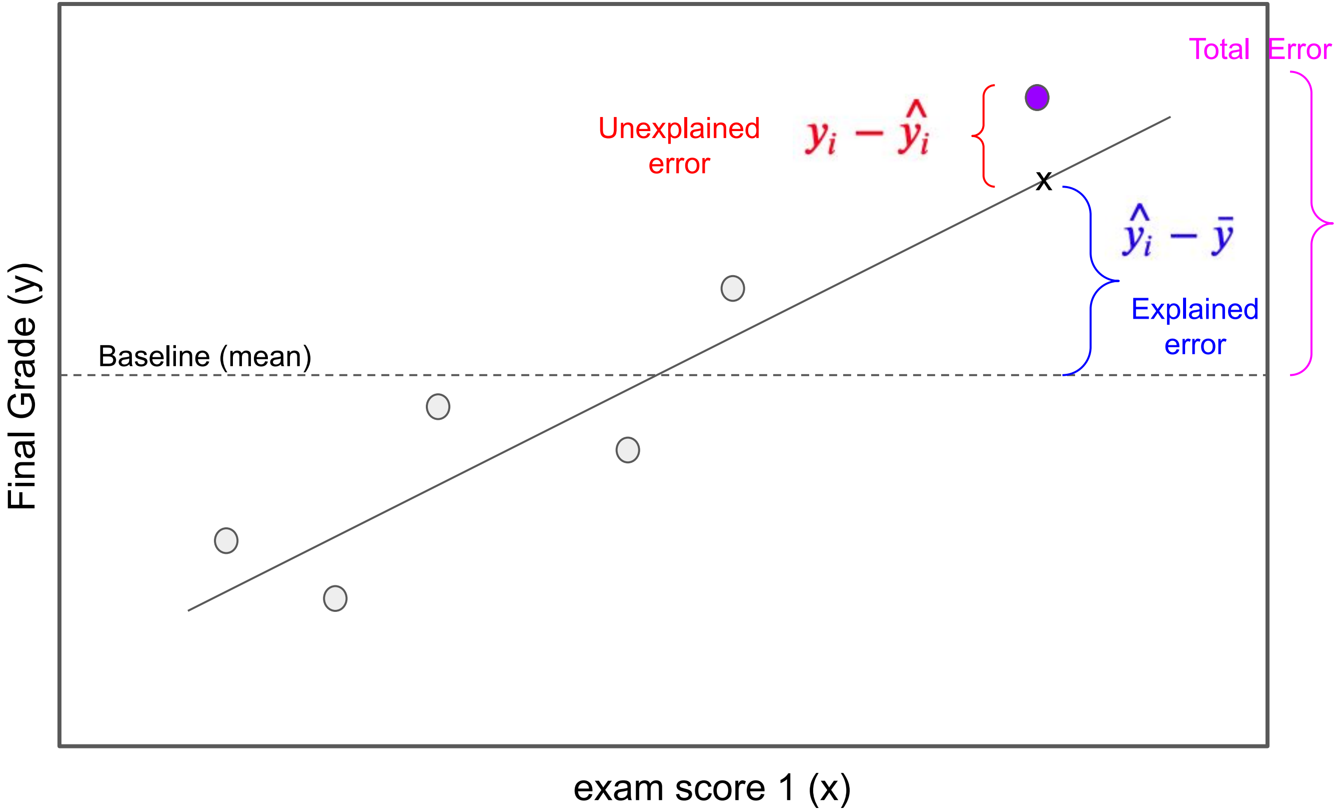
SSE_baseline = TSS

ESS = $\Sigma(\hat{y}_i - \bar{y} \;^2$

SSE = $\Sigma(y_i - \hat{y}_i)^2$

TSS = SSE + ESS

R^2 = ESS/TSS



Regression Metrics