MACHINE LEARNING PART 2

Lecture
SOURCES OF ERROR

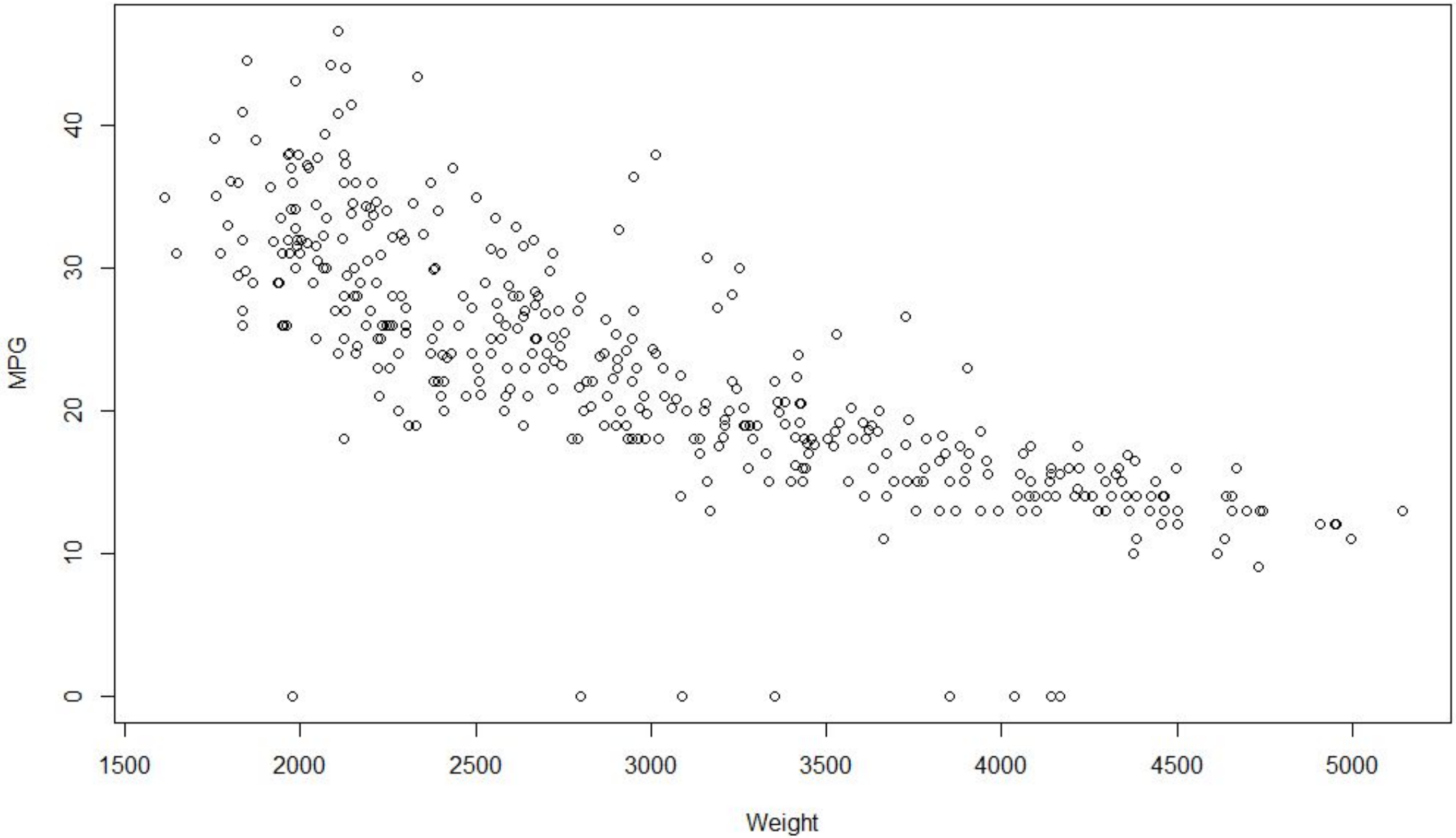
By

Gourab Nath

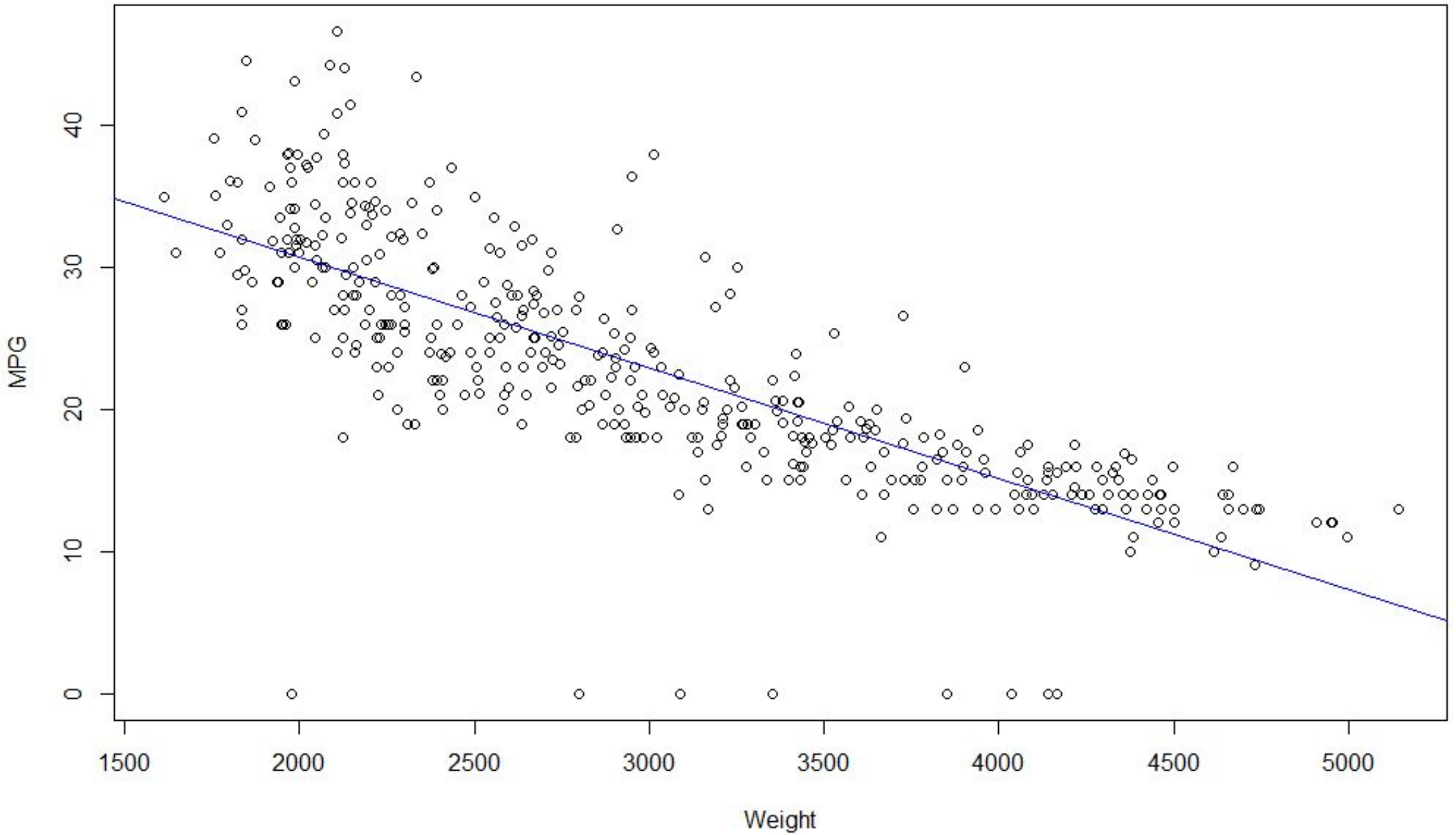
Assistant Professor of Data Science

Praxis Business School

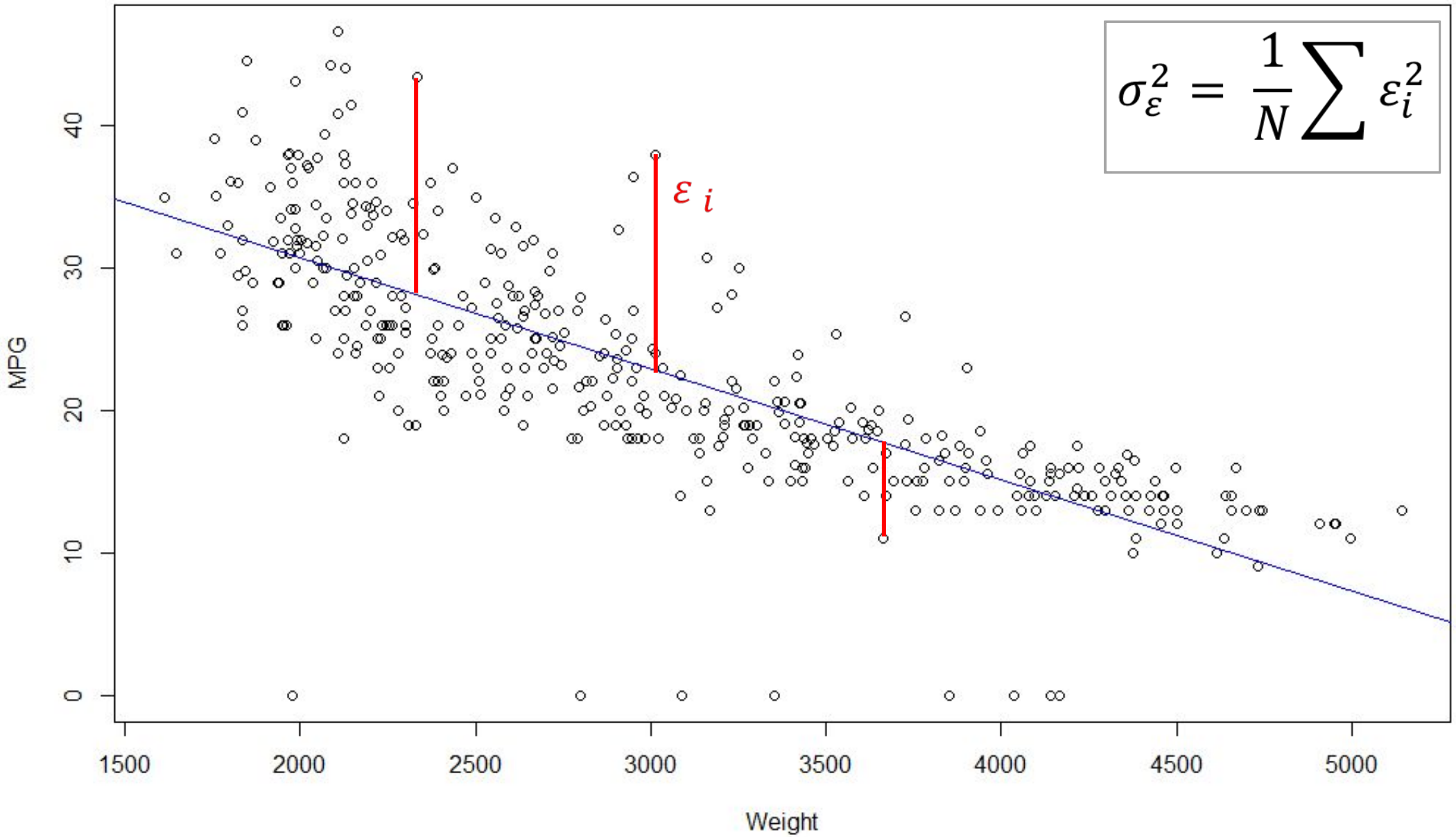
POPULATION DATA



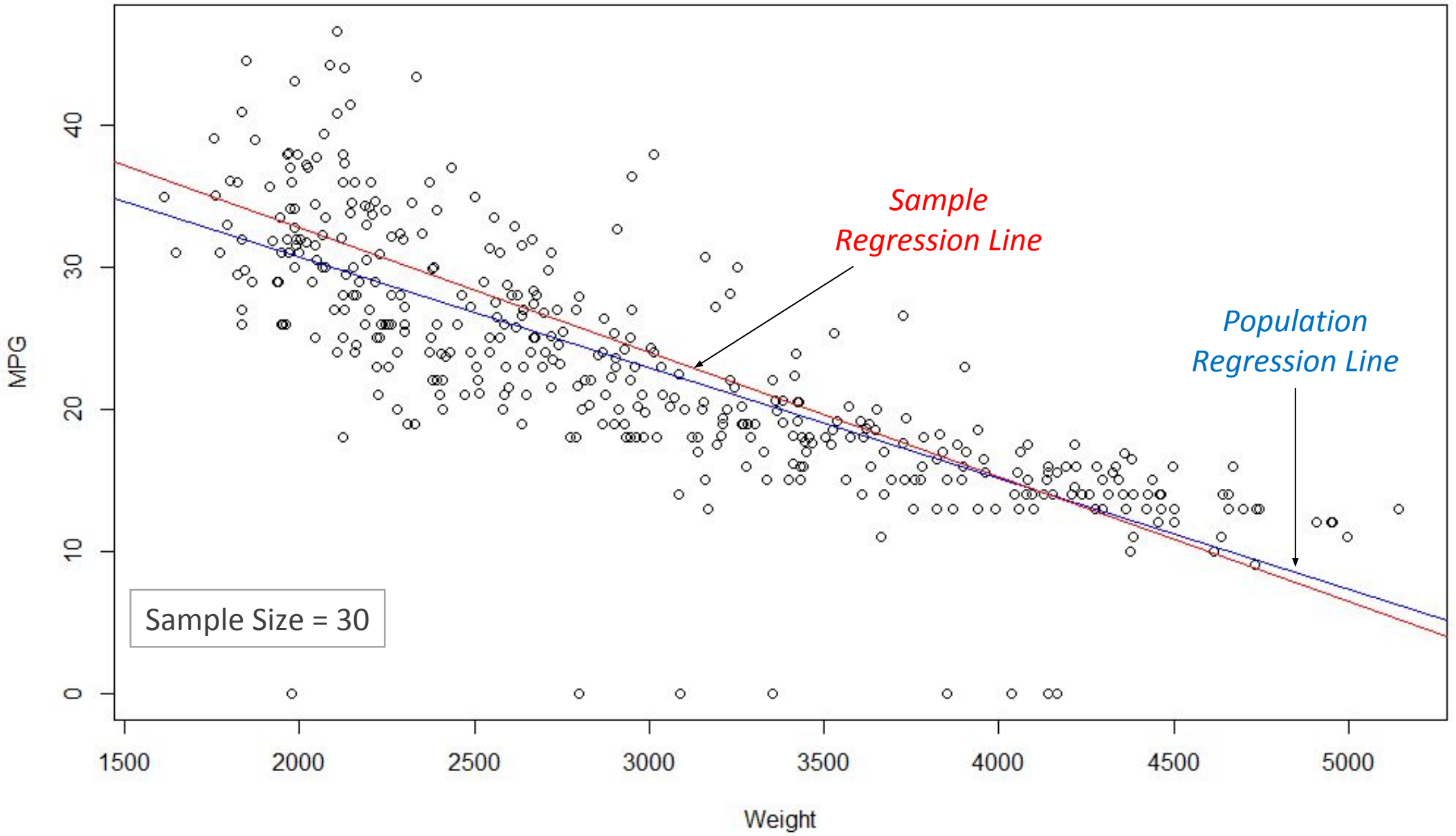
POPULATION REGRESSION LINE



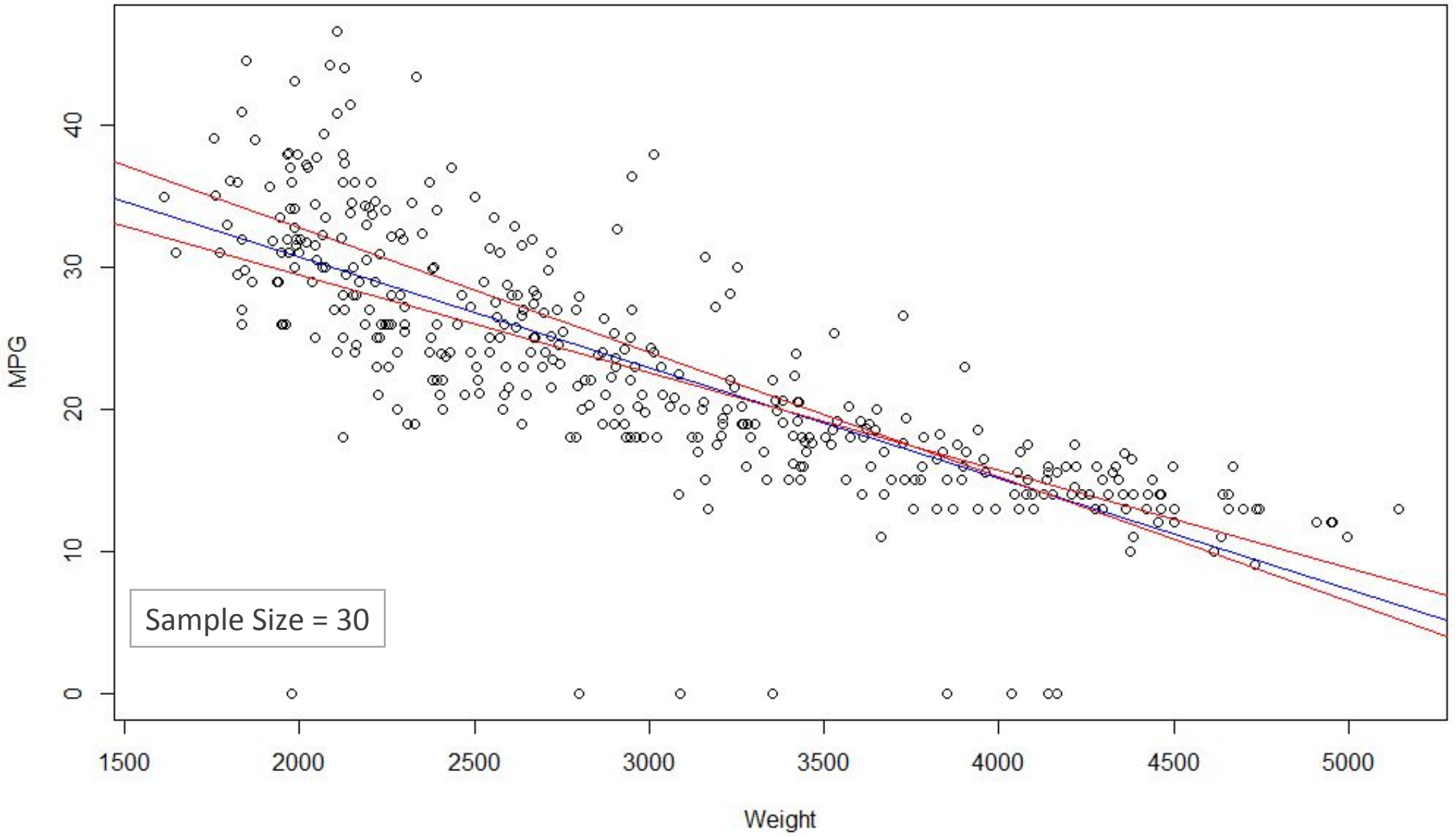
IRREDUCIBLE ERROR



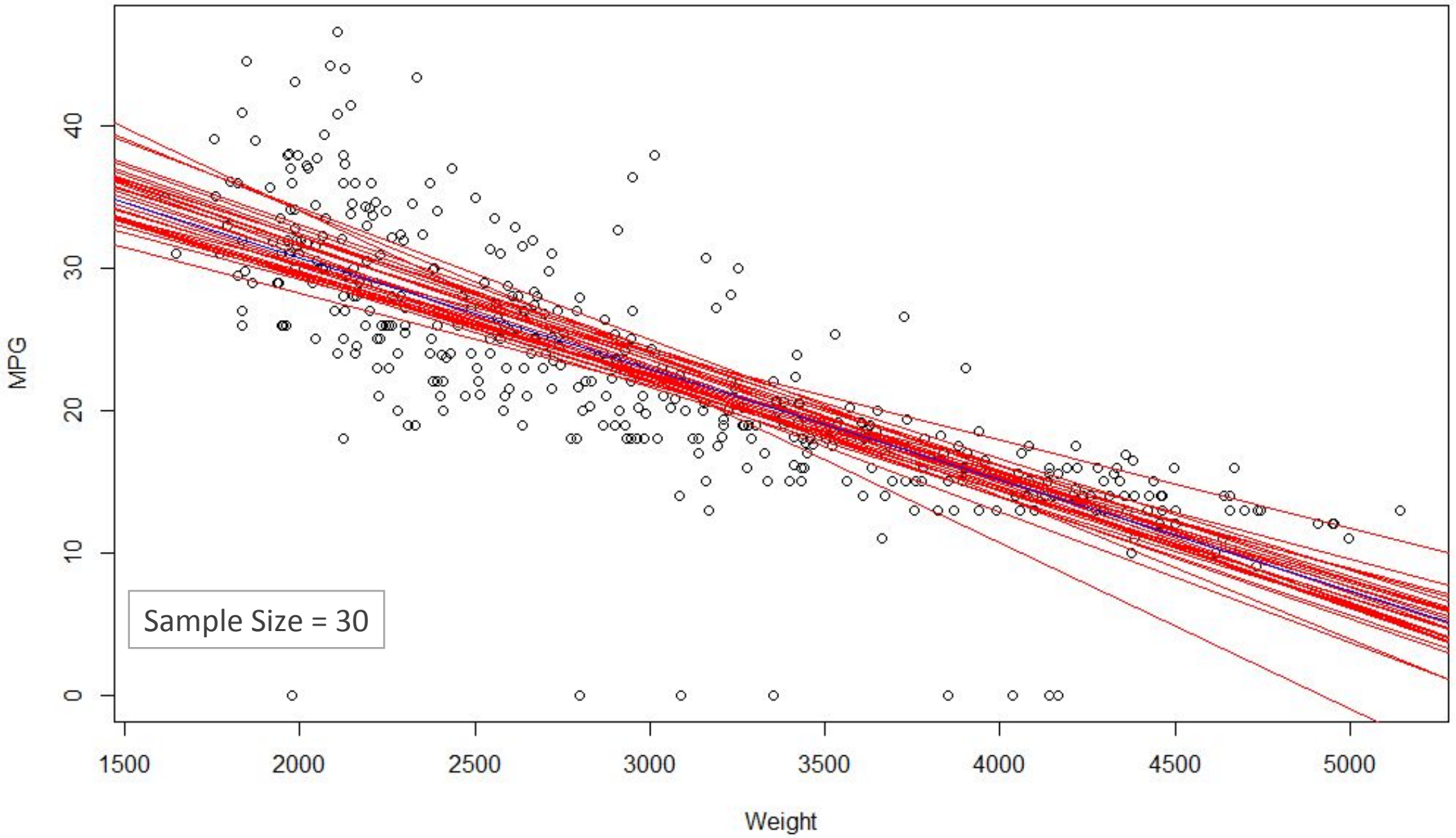
REGRESSION LINE ON 1 SAMPLE



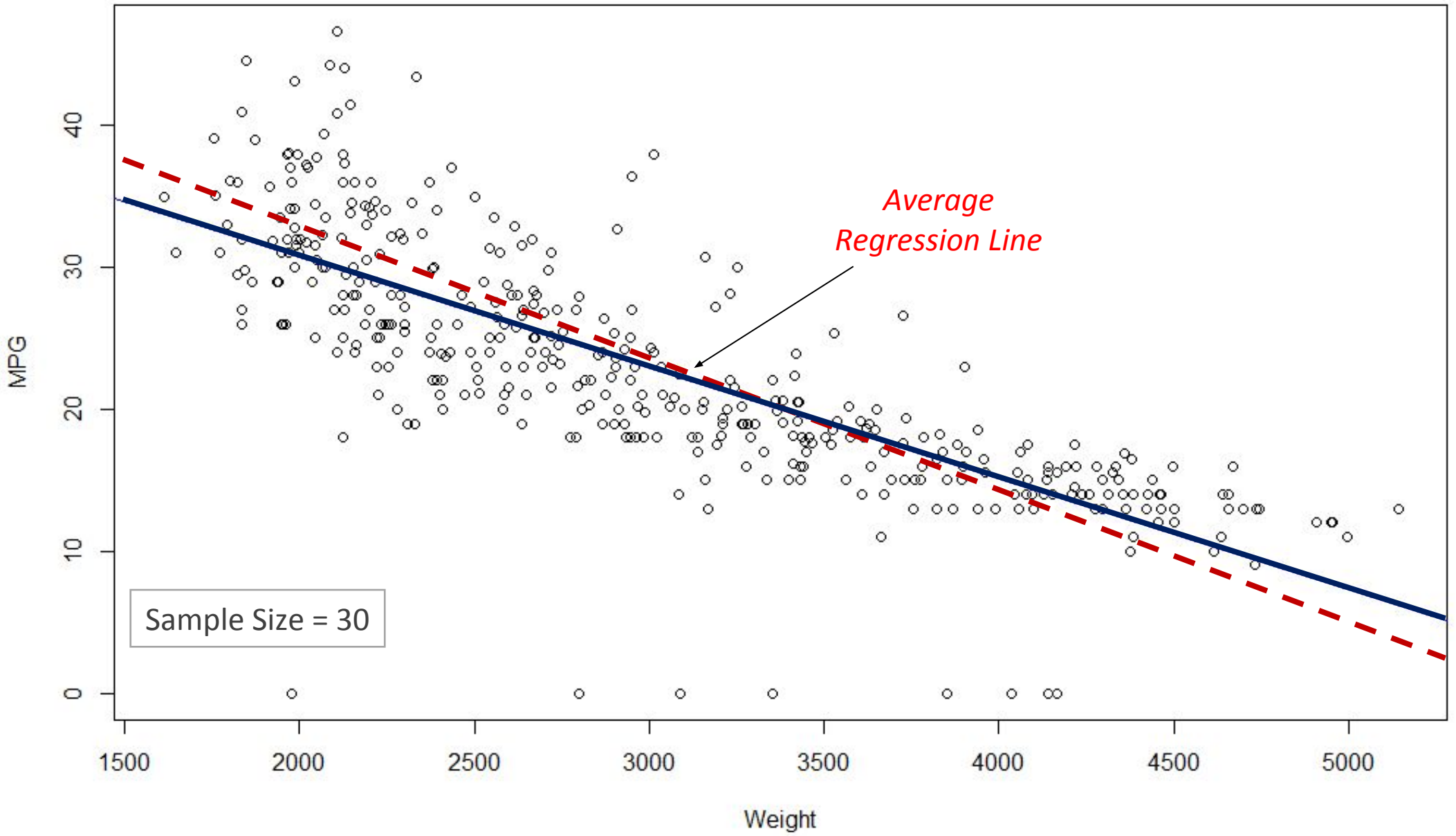
REGRESSION LINES FITTED ON 2 DIFFERENT SAMPLE



REGRESSION LINES FITTED ON 30 DIFFERENT SAMPLE



THE AVERAGE REGRESSION LINE



NOTATIONS

Let $\boldsymbol{x}$ be an observation.

$f(\boldsymbol{x}) =$ True regression line. That is, the regression line obtained by fitting on the population data.

$\widehat{f_{sample}}(\boldsymbol{x}) =$ The fitted sample regression line. That is, the regression line obtained by fitting on a sample.

$\bar{f}(\boldsymbol{x}) =$ The average of all the regression line fitted over all possible sample.

OUTPUTS OF THE FUNCTIONS

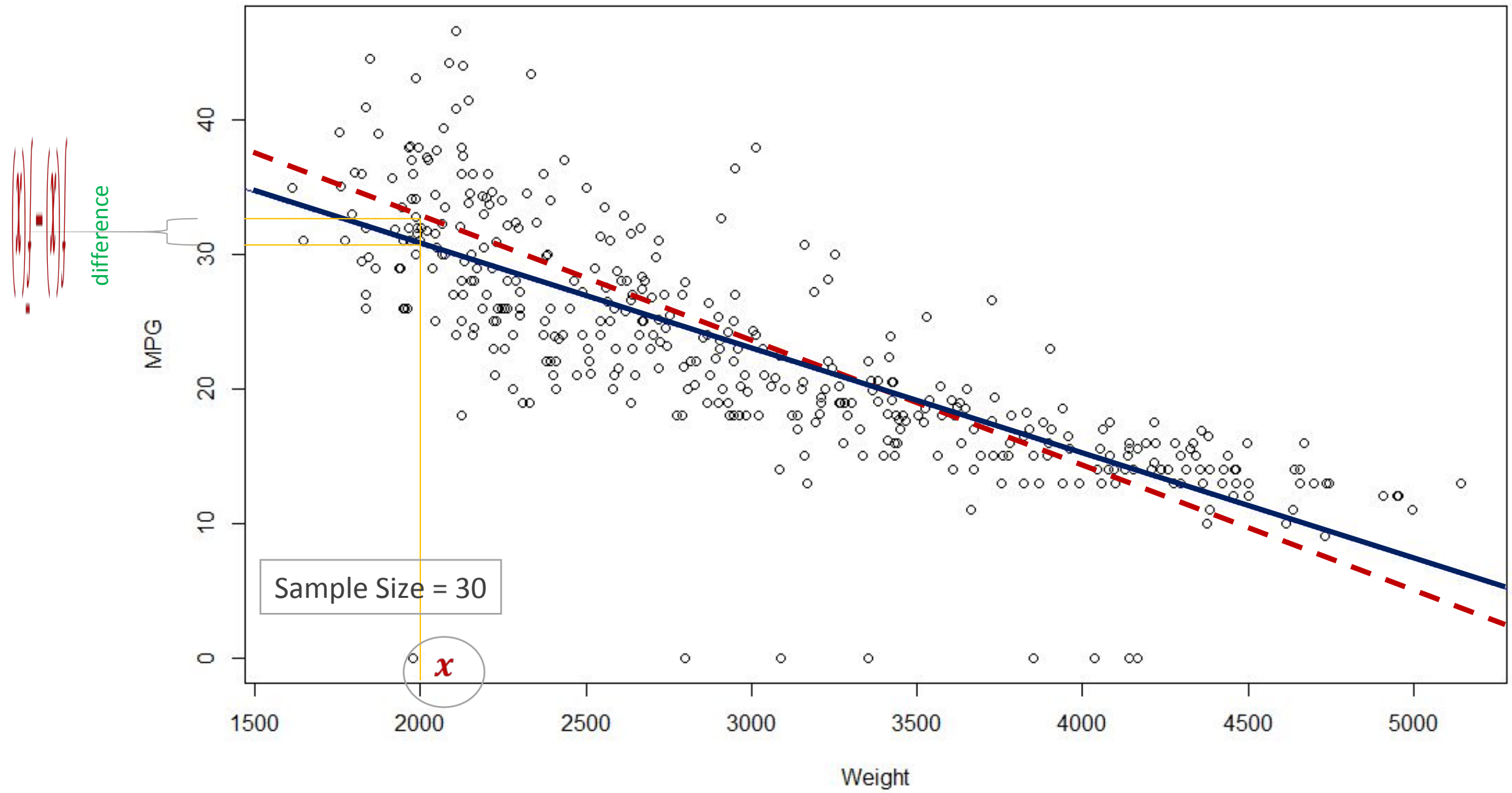
Let $\boldsymbol{x}$ curl be an observation.

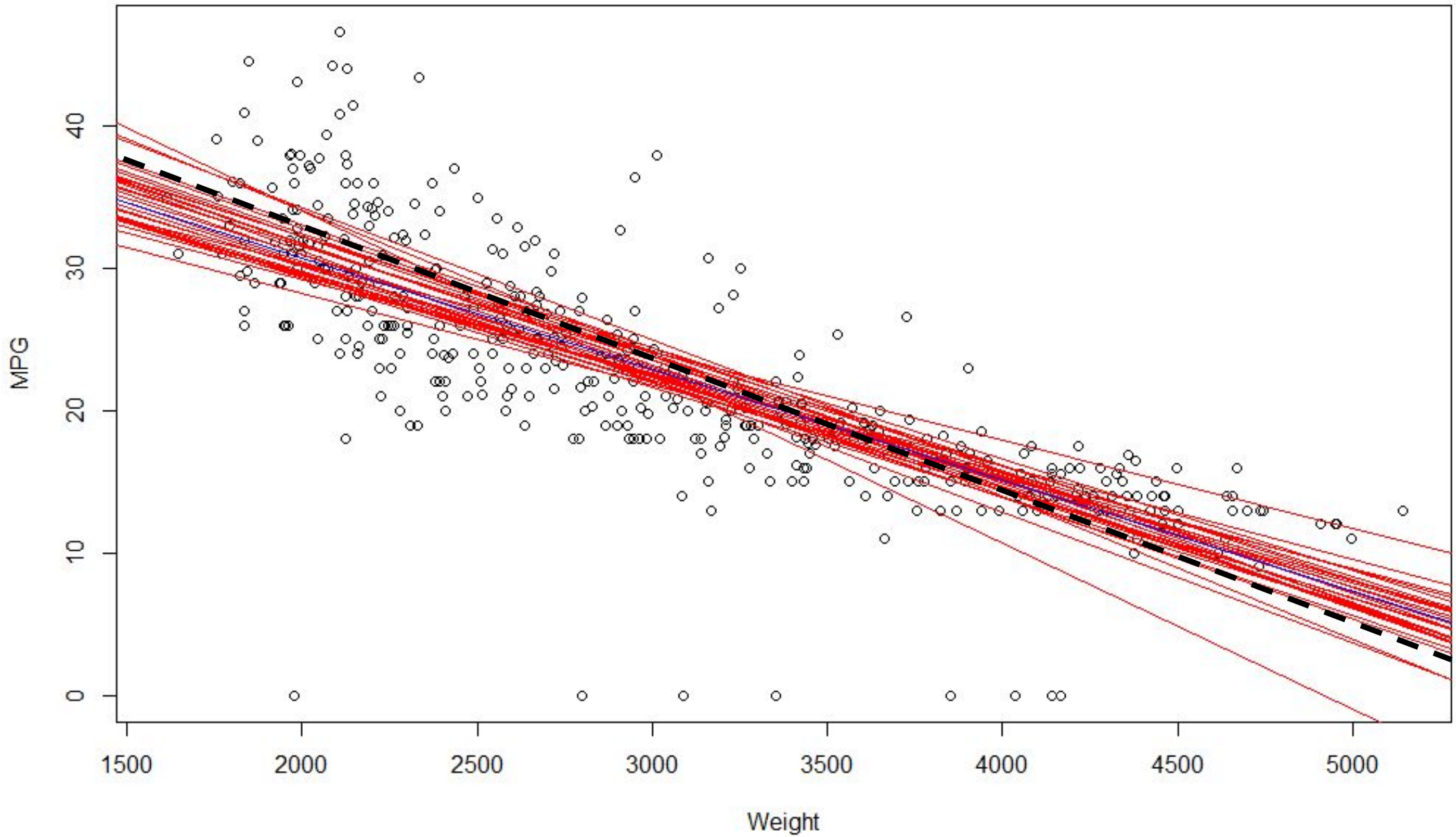
$f(\boldsymbol{x}) =$ Returns the predicted value of $\boldsymbol{x}$ using the fitted population regression model

$\widehat{f_{sample}}(\boldsymbol{x}) =$ Returns the predicted value of $\boldsymbol{x}$ using the estimated regression model based on the sample.

$\bar{f}(\boldsymbol{x}) =$ Returns the average of all the predicted value of $\boldsymbol{x}$ predicted using all possible regression models fitted over all possible samples.

THE DIFFERENCE BETWEEN ACTUAL AND PREDICTED




MPG
0
10
20
30
40
1500
2000
2500
3000
3500
4000
4500
5000
Weight

BIAS AND VARIANCE

$$Bias(\boldsymbol{x}) = f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})$$

How far the average of all the models is from the true model?

$$Variance(\boldsymbol{x}) = E_{sample}[\widehat{f_{sample}}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})]^2$$

How much does the outputs of each models (fitted over different samples of same size) varies from the average of all the models?

MSE at $\boldsymbol{x}$

$$MSE = Irreducible\ Error + bias^2(\boldsymbol{x}) + variance(\boldsymbol{x})$$

$$\bar{x} = \frac{1}{n}\sum x_i$$

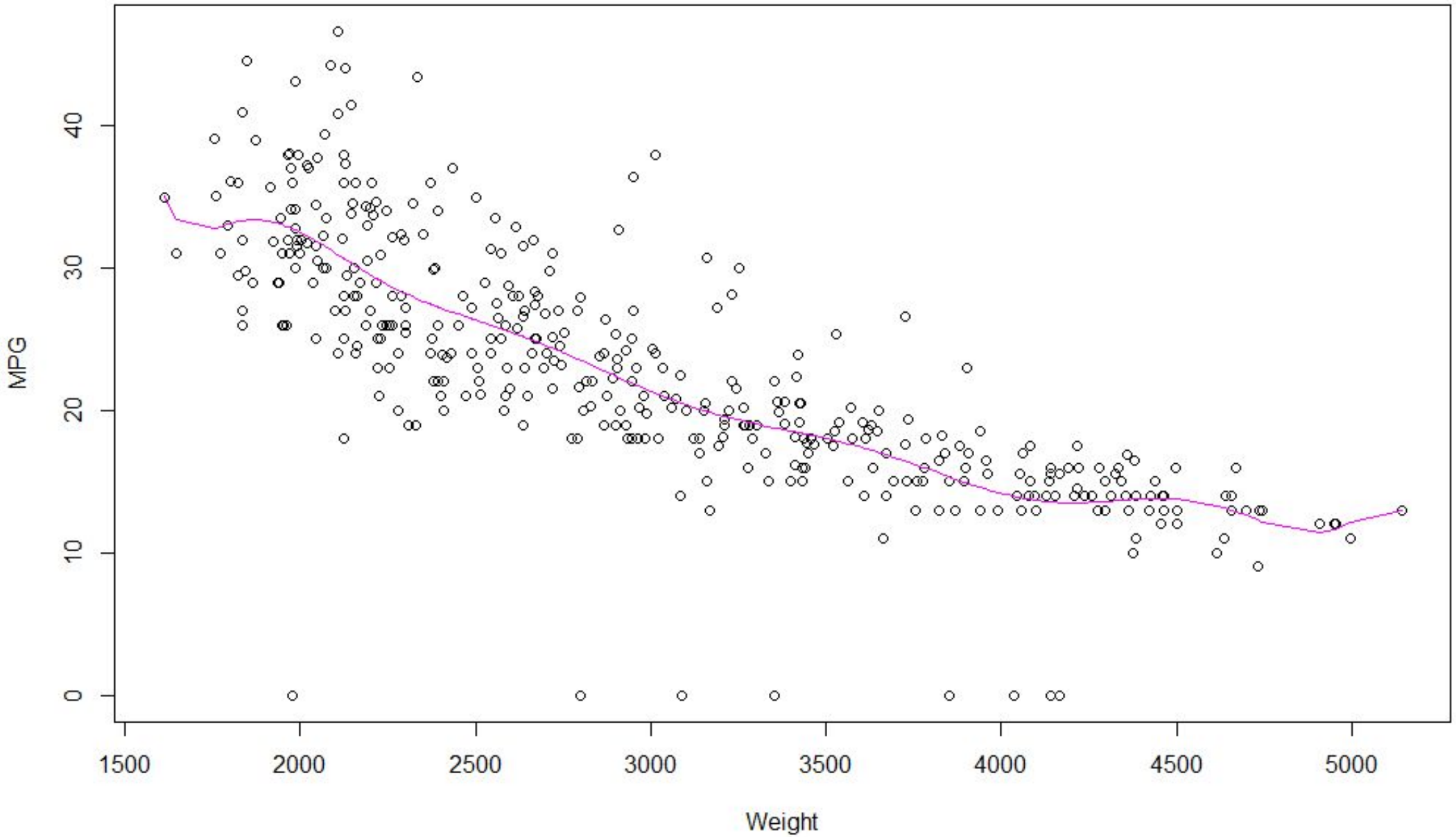
$$variance = \frac{1}{n}\sum (x_i - \bar{x})\text{^}2$$

$$E(X) = \sum x_i p_i$$

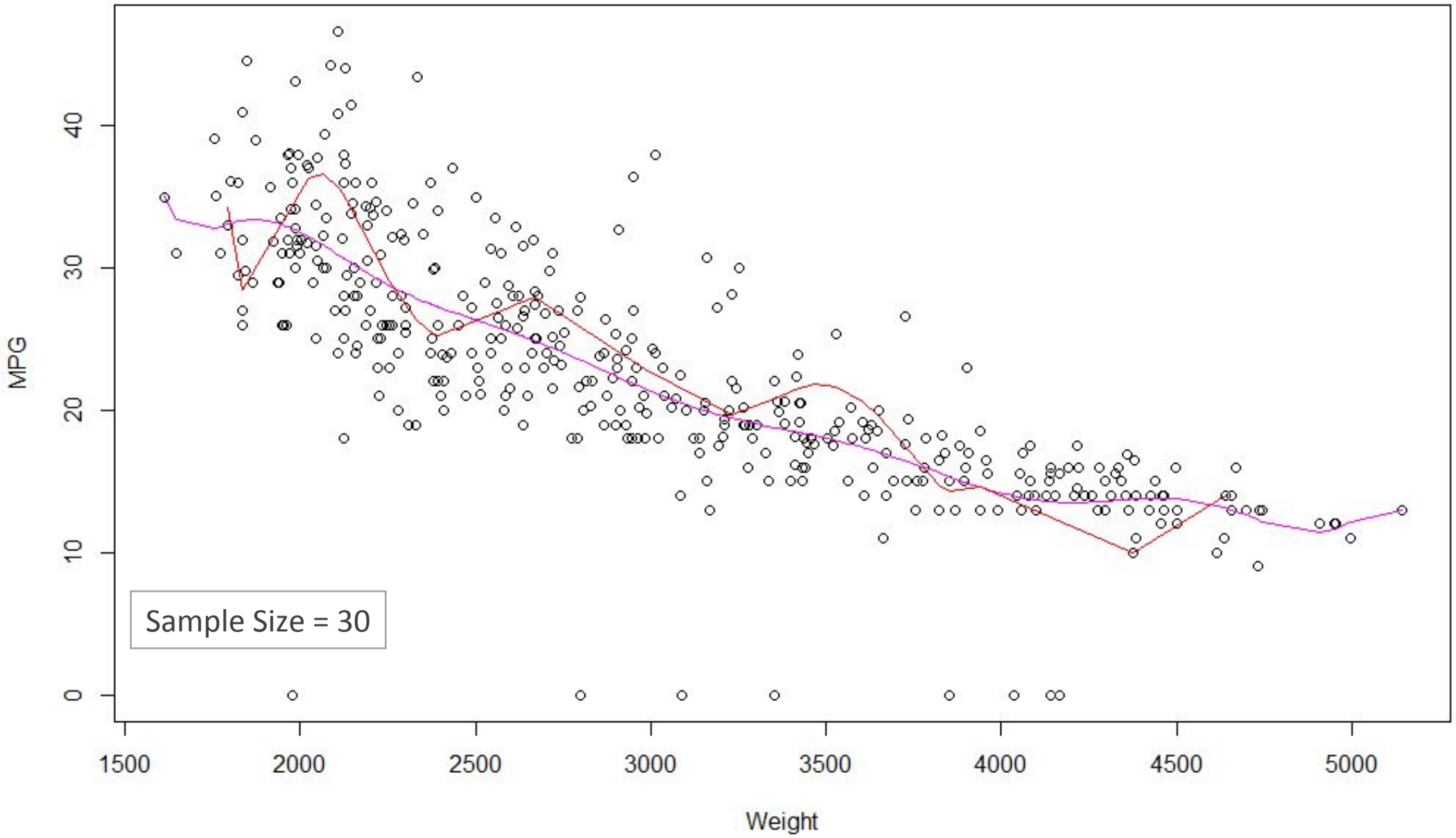
$$Var(X) = E[\left(X - E(X)\right)^2]$$

MODEL COMPLEXITY AND VARIANCE

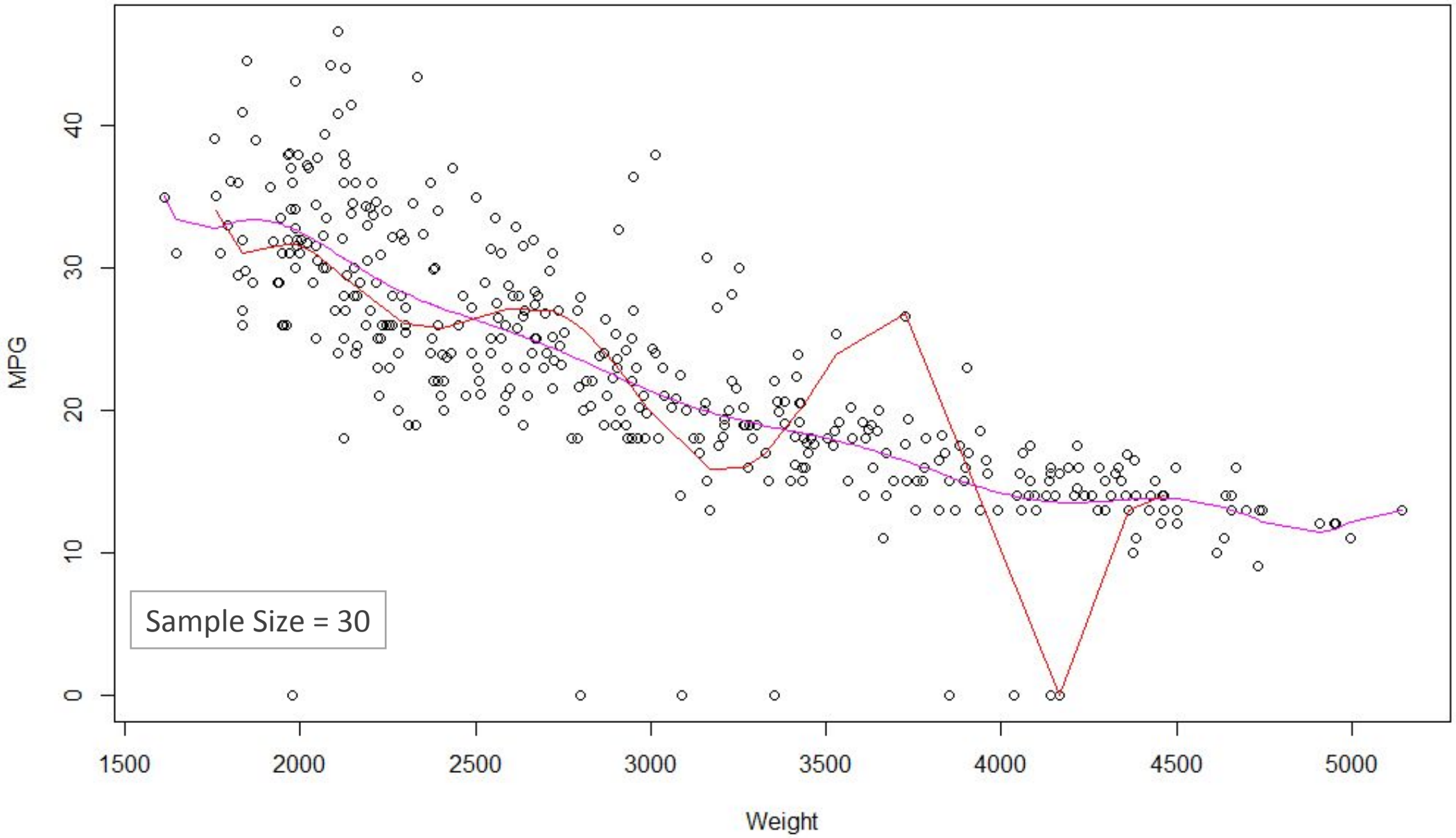
POLYNOMIAL OF ORDER 12



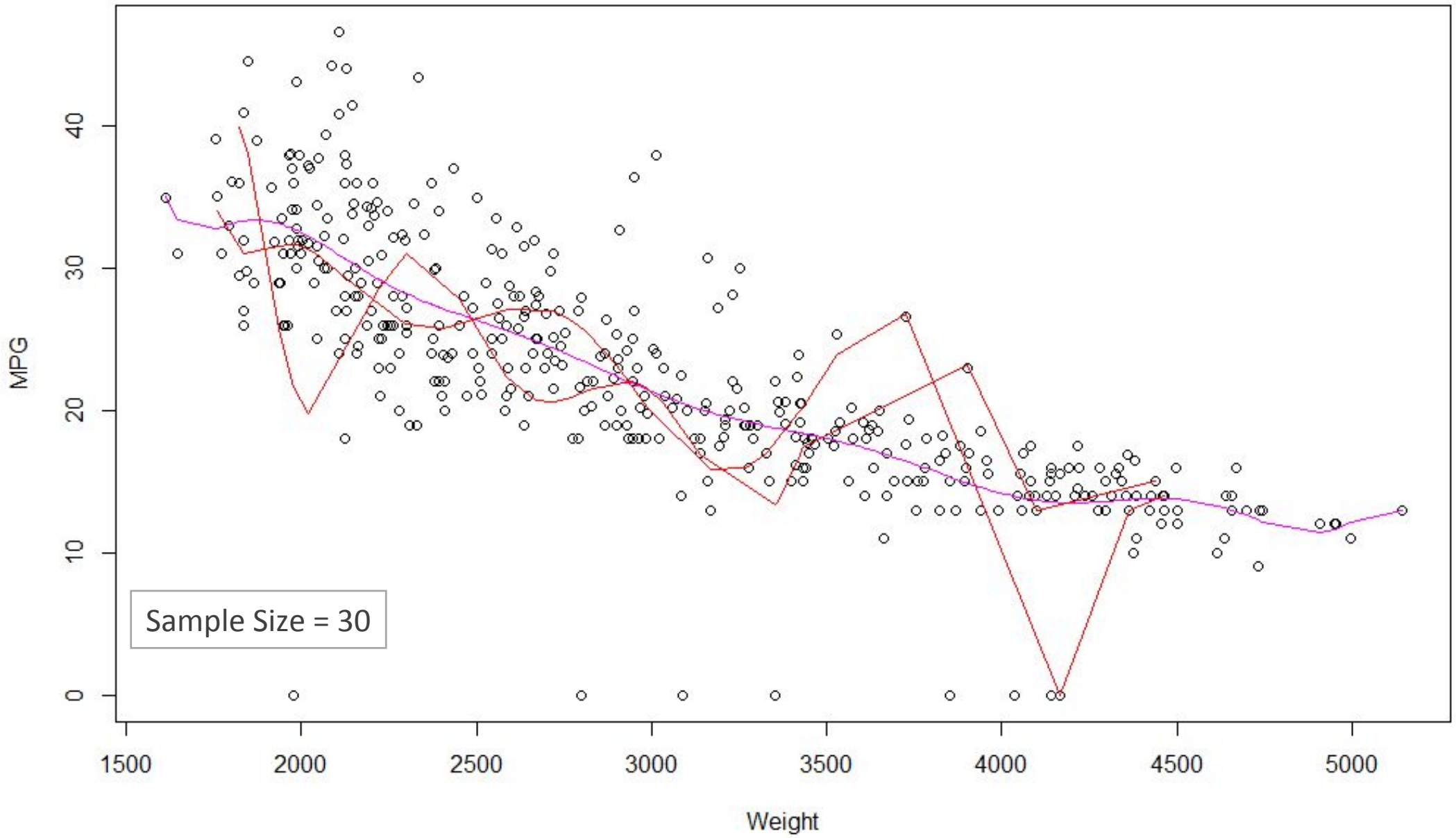
POLYNOMIAL OF ORDER 12 FITTED ON 1 SAMPLE



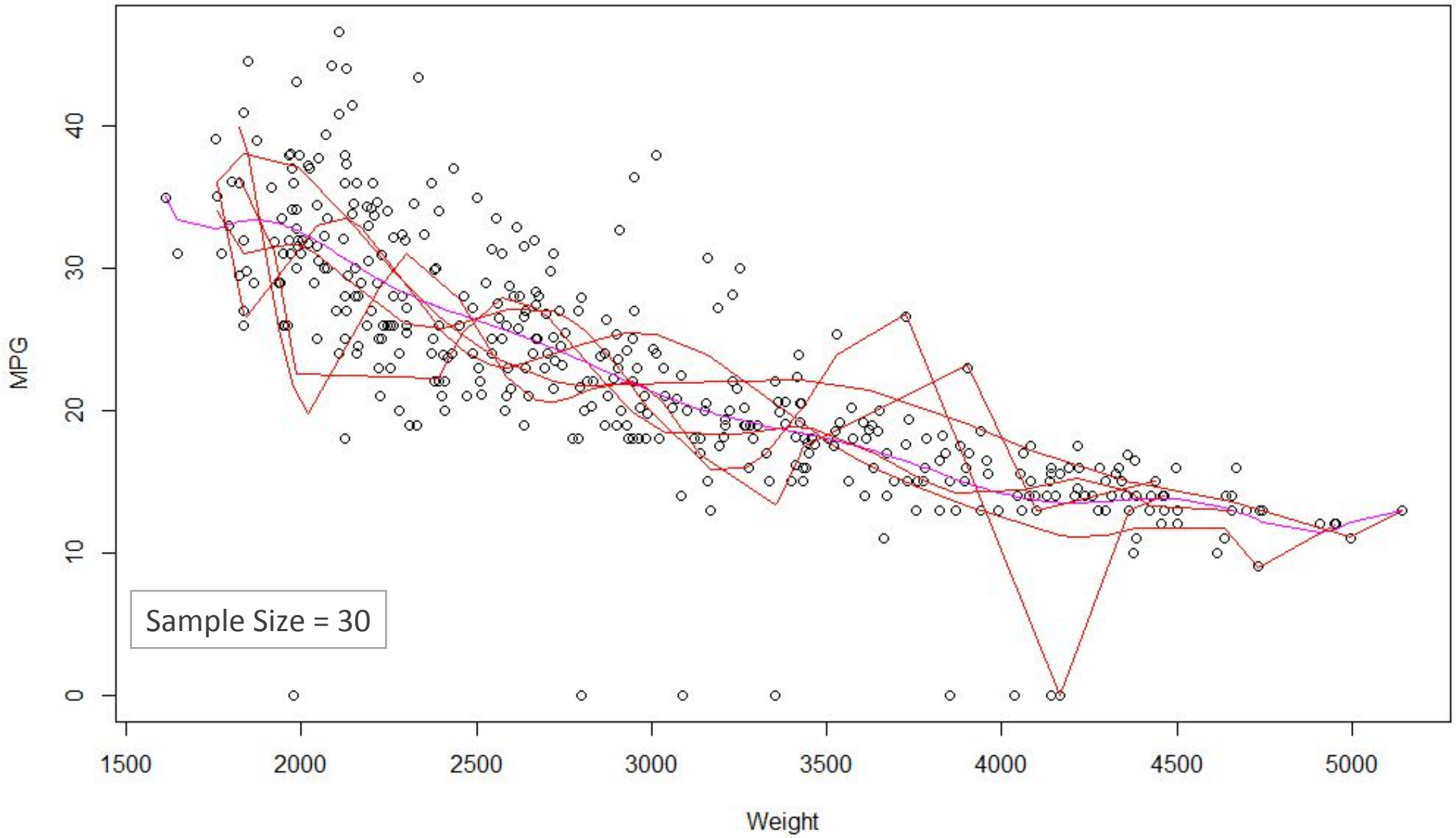
POLYNOMIAL OF ORDER 12 FITTED ON A DIFFERENT SAMPLE



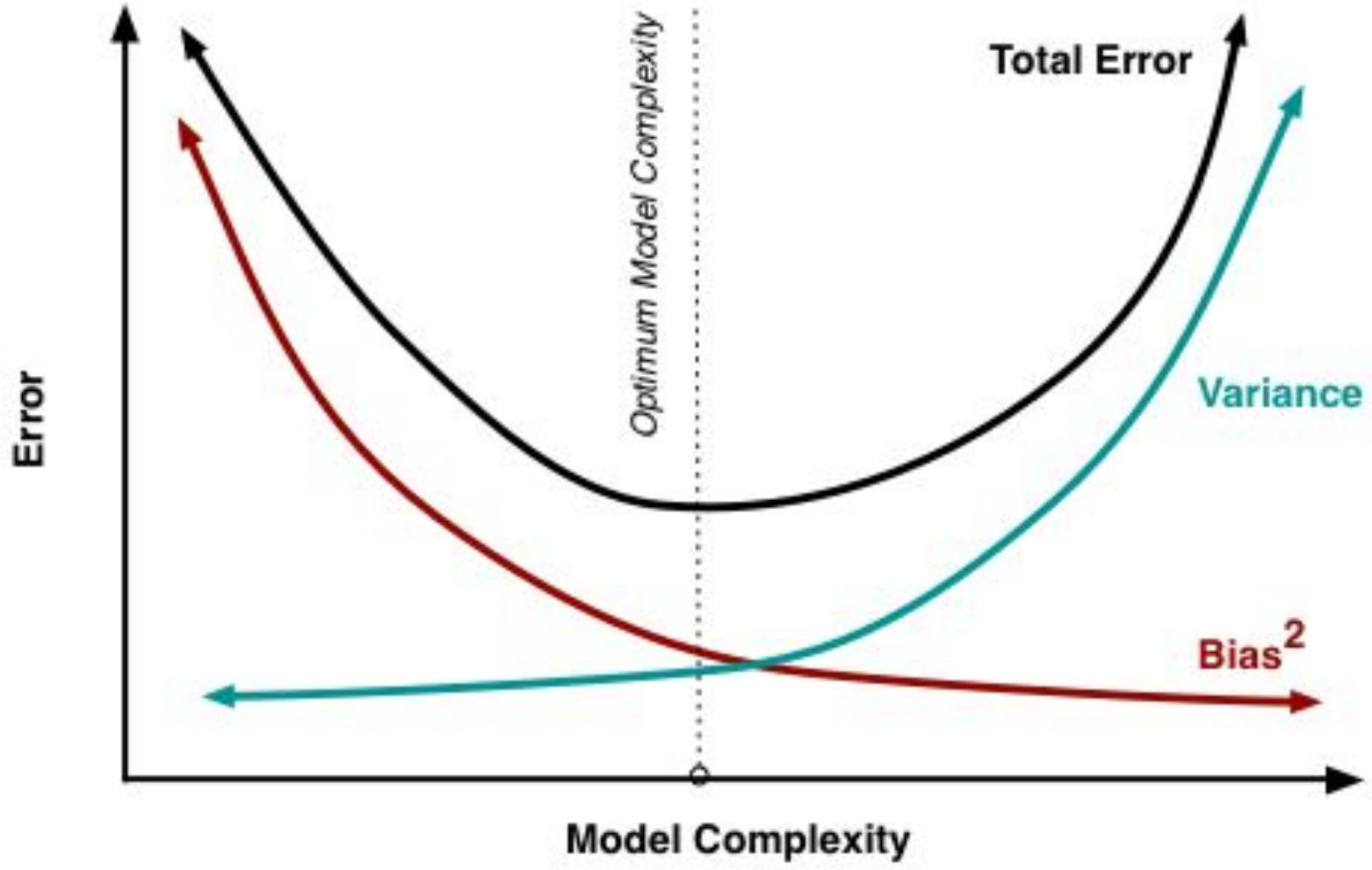
POLYNOMIAL OF ORDER 12 FITTED ON 2 DIFFERENT SAMPLES



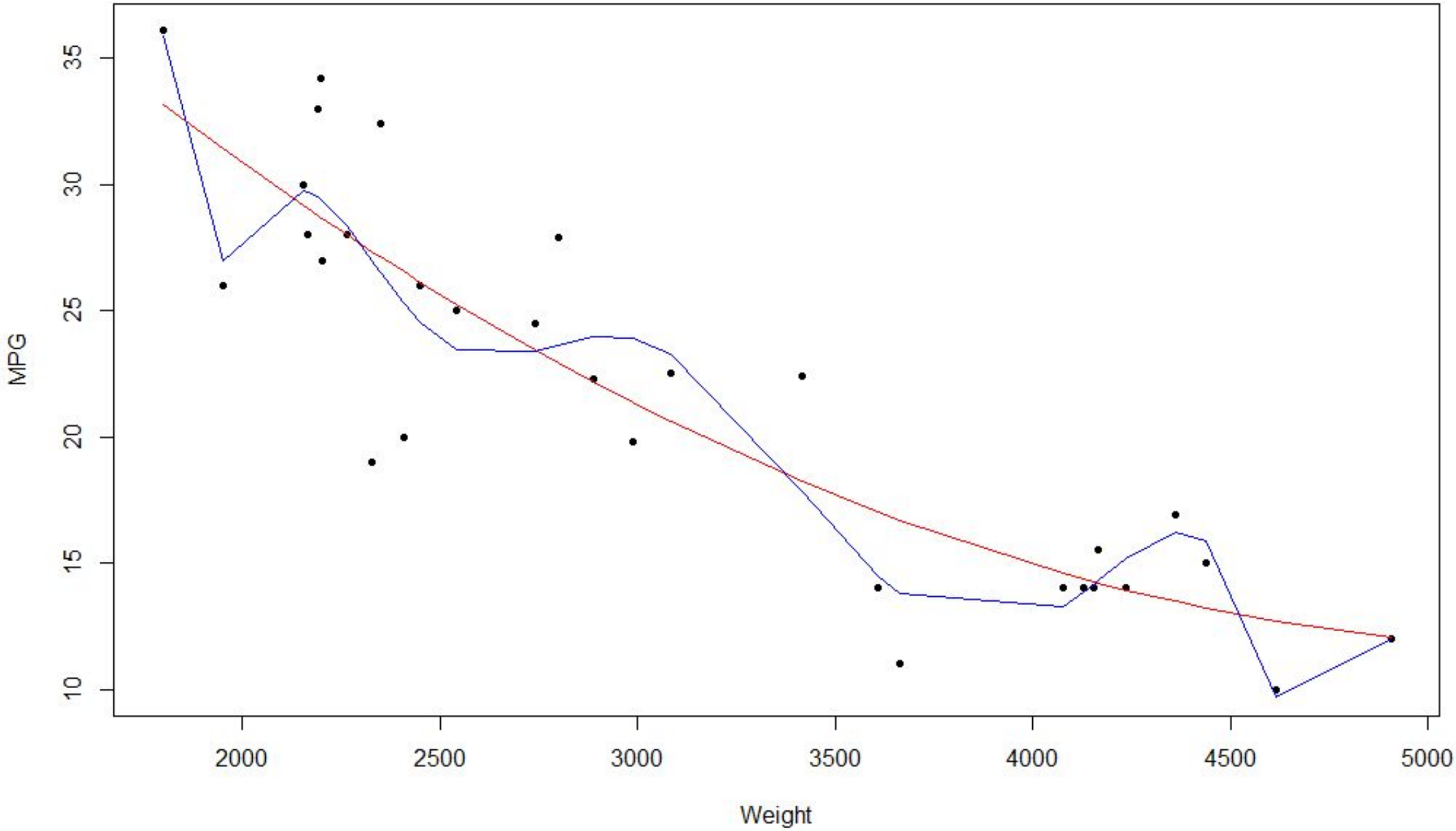
POLYNOMIAL OF ORDER 12 FITTED ON 5 DIFFERENT SAMPLES



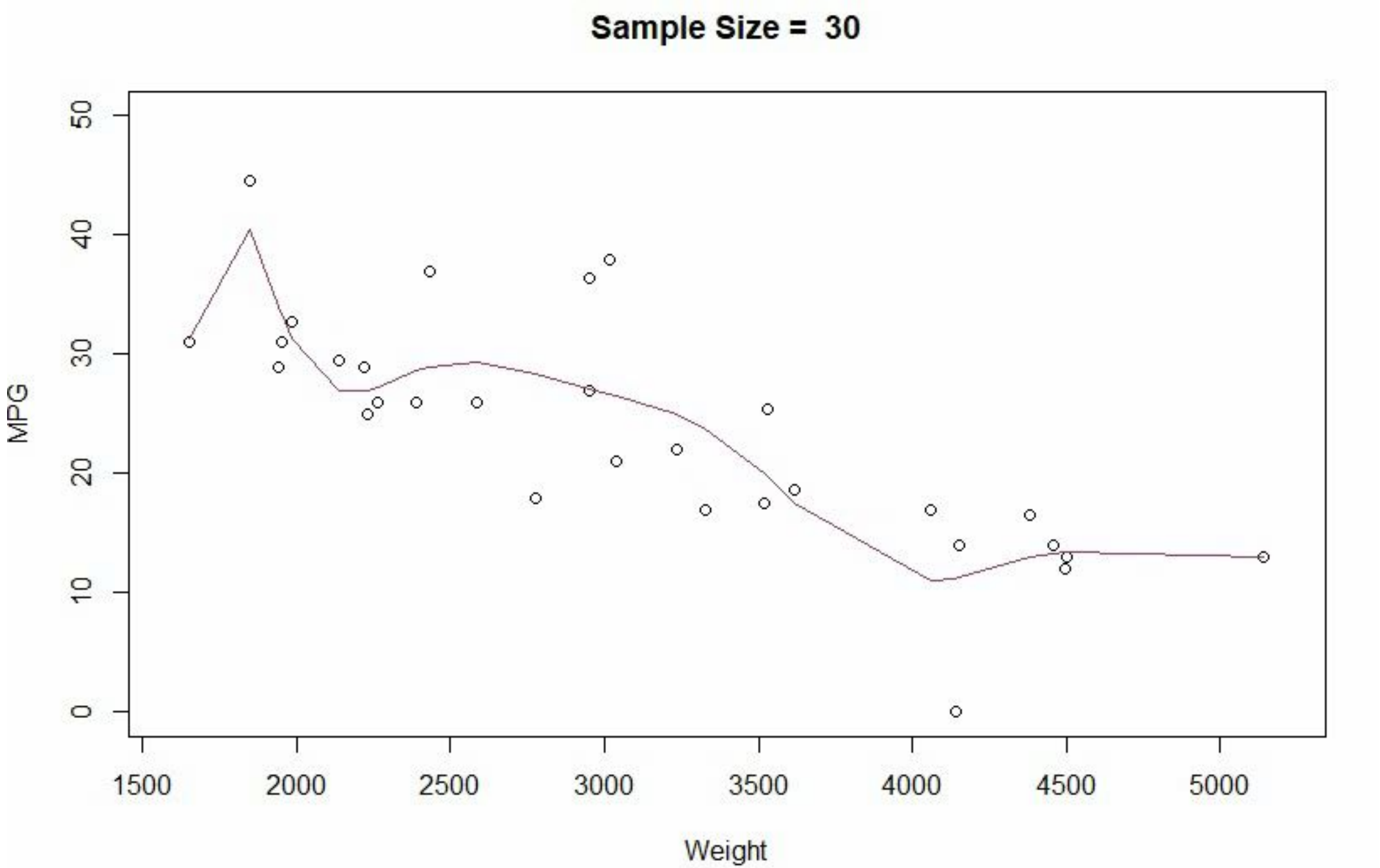
MODEL COMPLEXITY – BIAS - VARIANCE

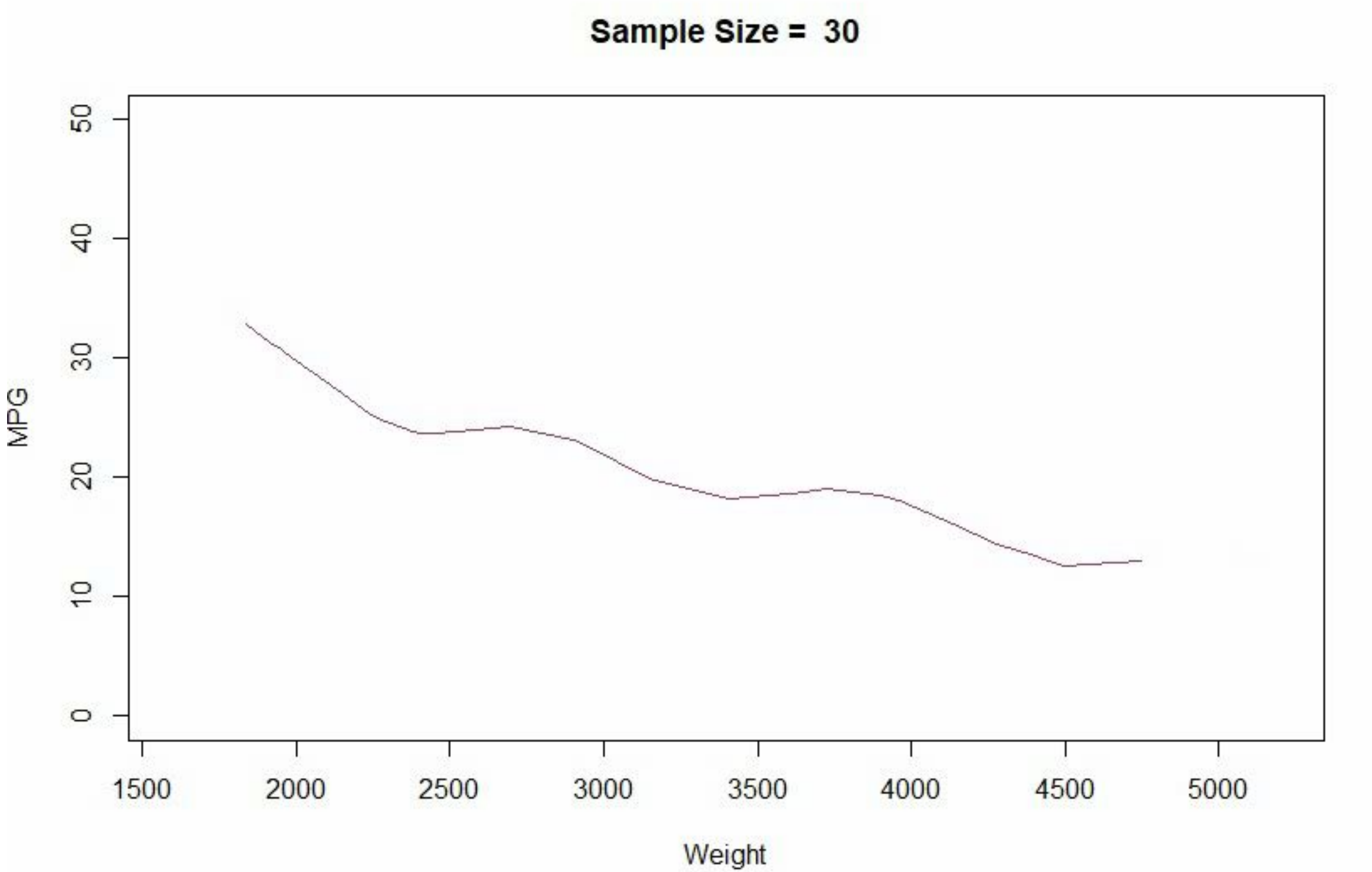


COMPLEXITY VS OVERFITTING

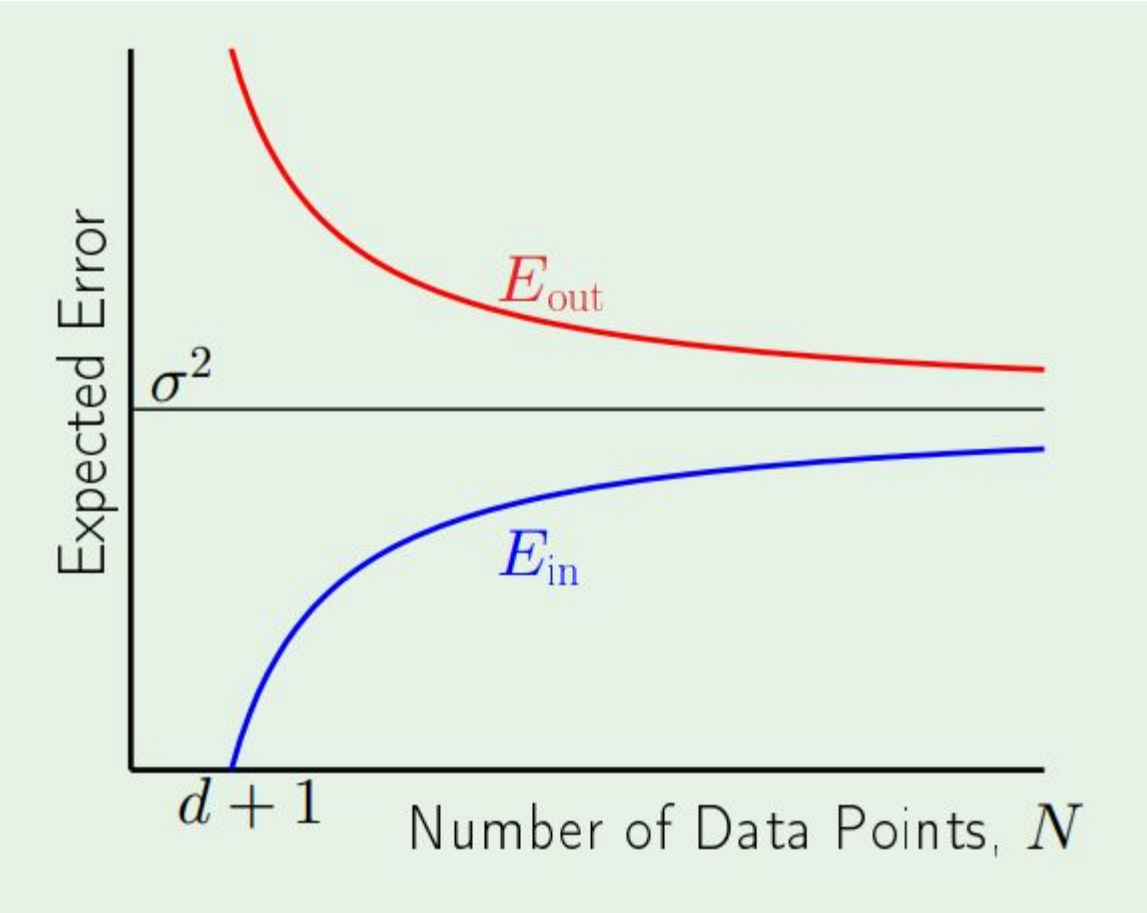


MODEL VARIANCE AGAINST SAMPLE SIZE


Sample Size = 30
MPG
0
10
20
30
40
50
1500
2000
2500
3000
3500
4000
4500
5000
Weight


Sample Size = 30
MPG
0
10
20
30
40
50
1500
2000
2500
3000
3500
4000
4500
5000
Weight

EXPECTED ERROR VS SAMPLE SIZE



http://work.caltech.edu/slides/slides08.pdf