POLITECNICO
MILANO 1863

Data Analysis for Smart Agriculture

- Regression -

Prof. Matteo Matteucci – matteo.matteucci@polimi.it

Regression Problem

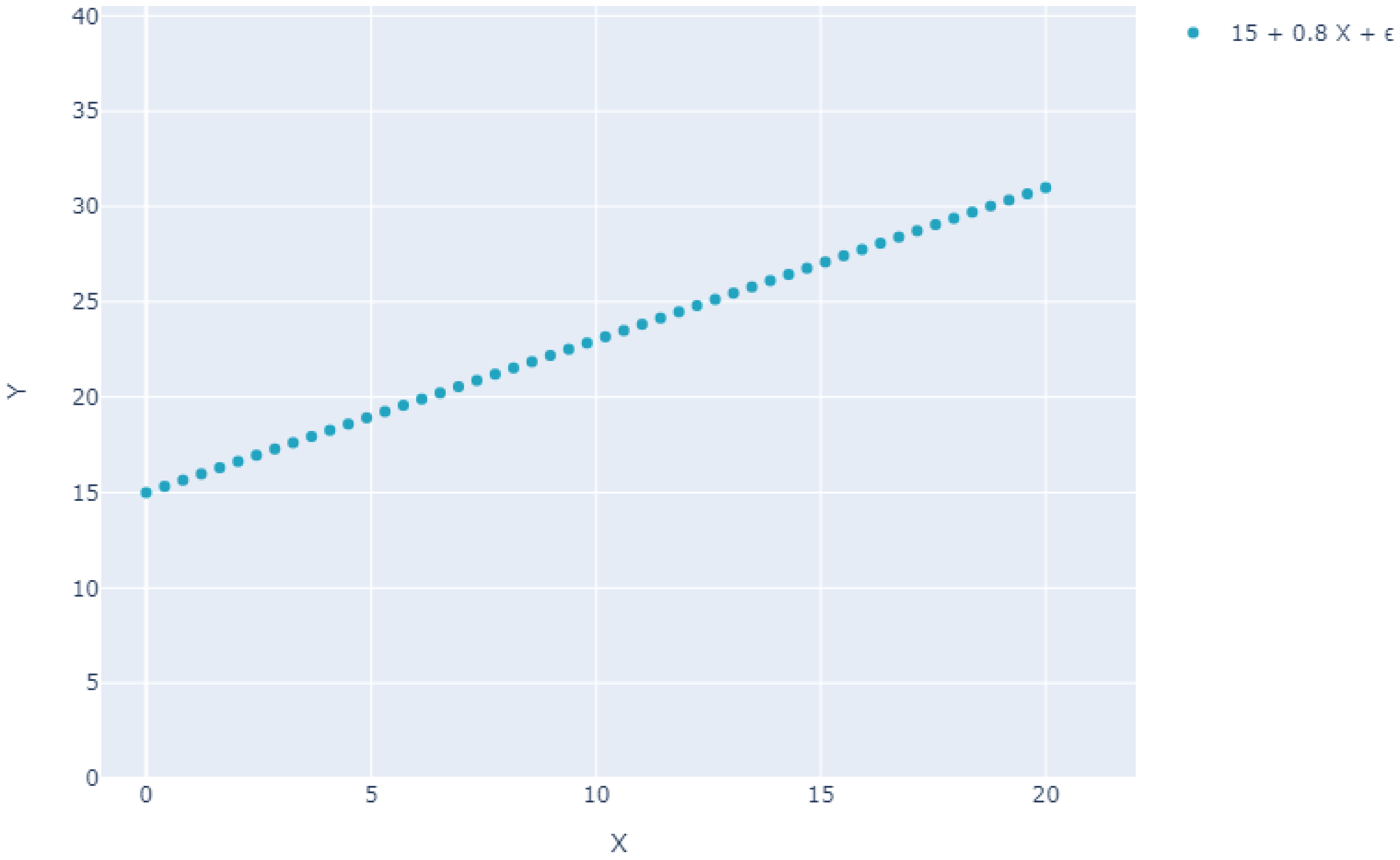
```
Boston house prices dataset
---------------------------

- CRIM     per capita crime rate by town
- ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS    proportion of non-retail business acres per town
- CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX      nitric oxides concentration (parts per 10 million)
- RM       average number of rooms per dwelling
- AGE      proportion of owner-occupied units built prior to 1940
- DIS      weighted distances to five Boston employment centres
- RAD      index of accessibility to radial highways
- TAX      full-value property-tax rate per $10,000
- PTRATIO  pupil-teacher ratio by town
- B        1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
- LSTAT    % lower status of the population
- MEDV     Median value of owner-occupied homes in $1000's
```

Can we learn to predict a target variable (e.g., MEDV) by observing input features (e.g., CRIM, ZN, INDUS, …)?

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

Regression Problem

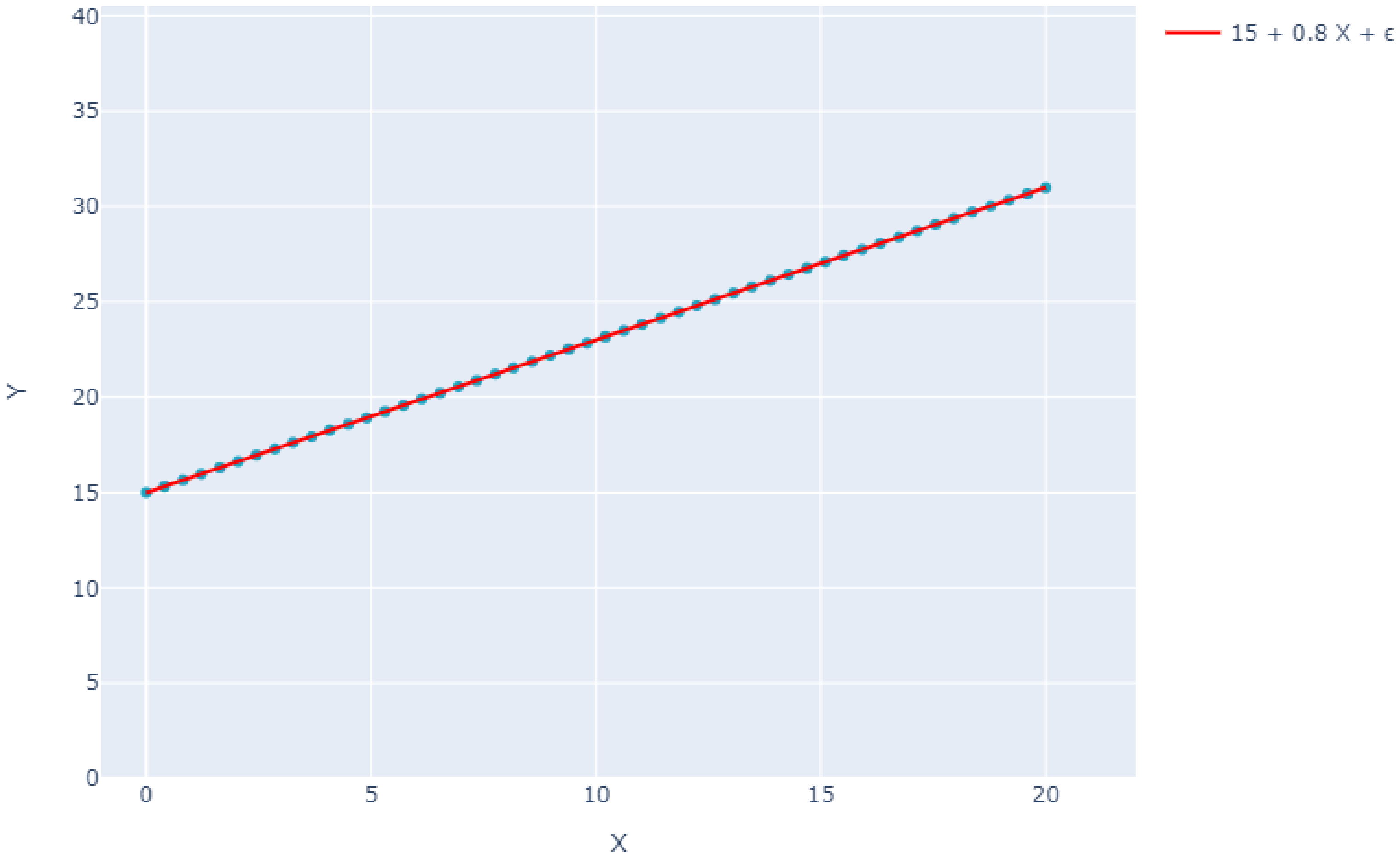
Let's start with a simple example … can we predict Y from X?



Regression Problem

Let's start with a simple example … can we predict Y from X?

$$y = \underbrace{w_0}_{\text{Intercept}} + \underbrace{w_1}_{\text{Slope}} x$$

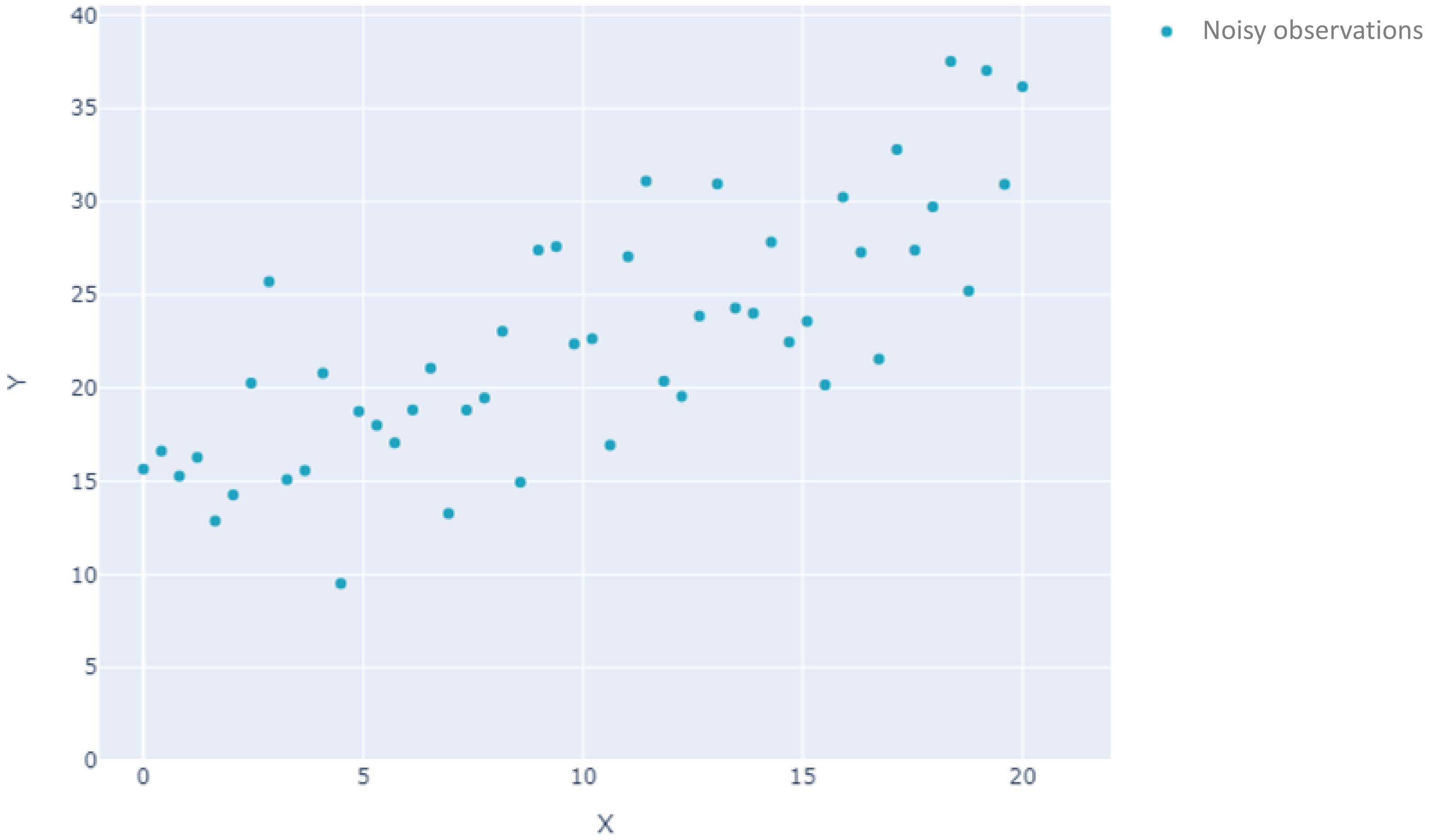


In the example

- $w_0 = 15$
- $w_1 = 0.8$

Regression Problem

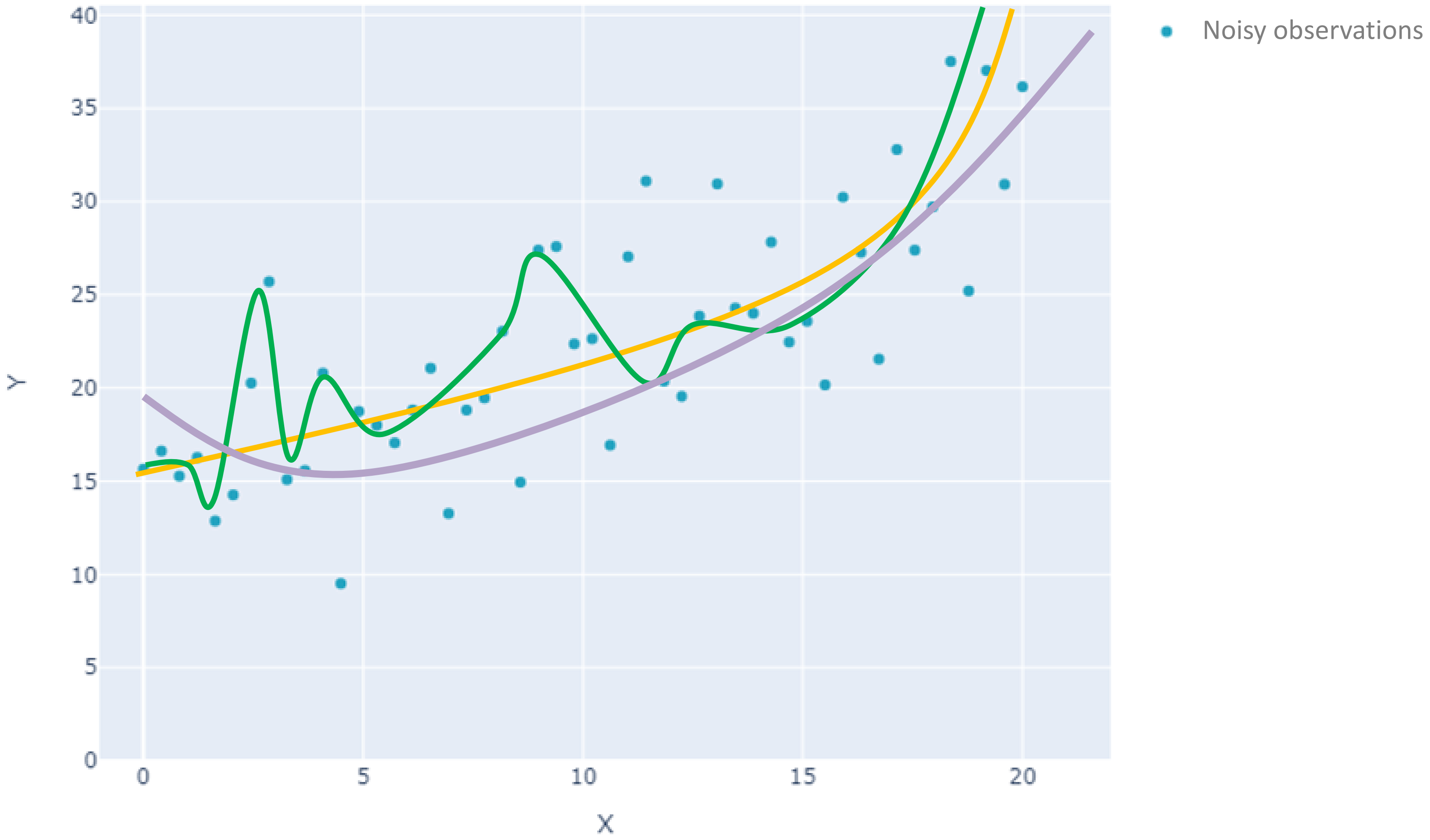
Let's add noise … can we still predict Y from X?



Regression Problem

Let's add noise … can we still predict Y from X?

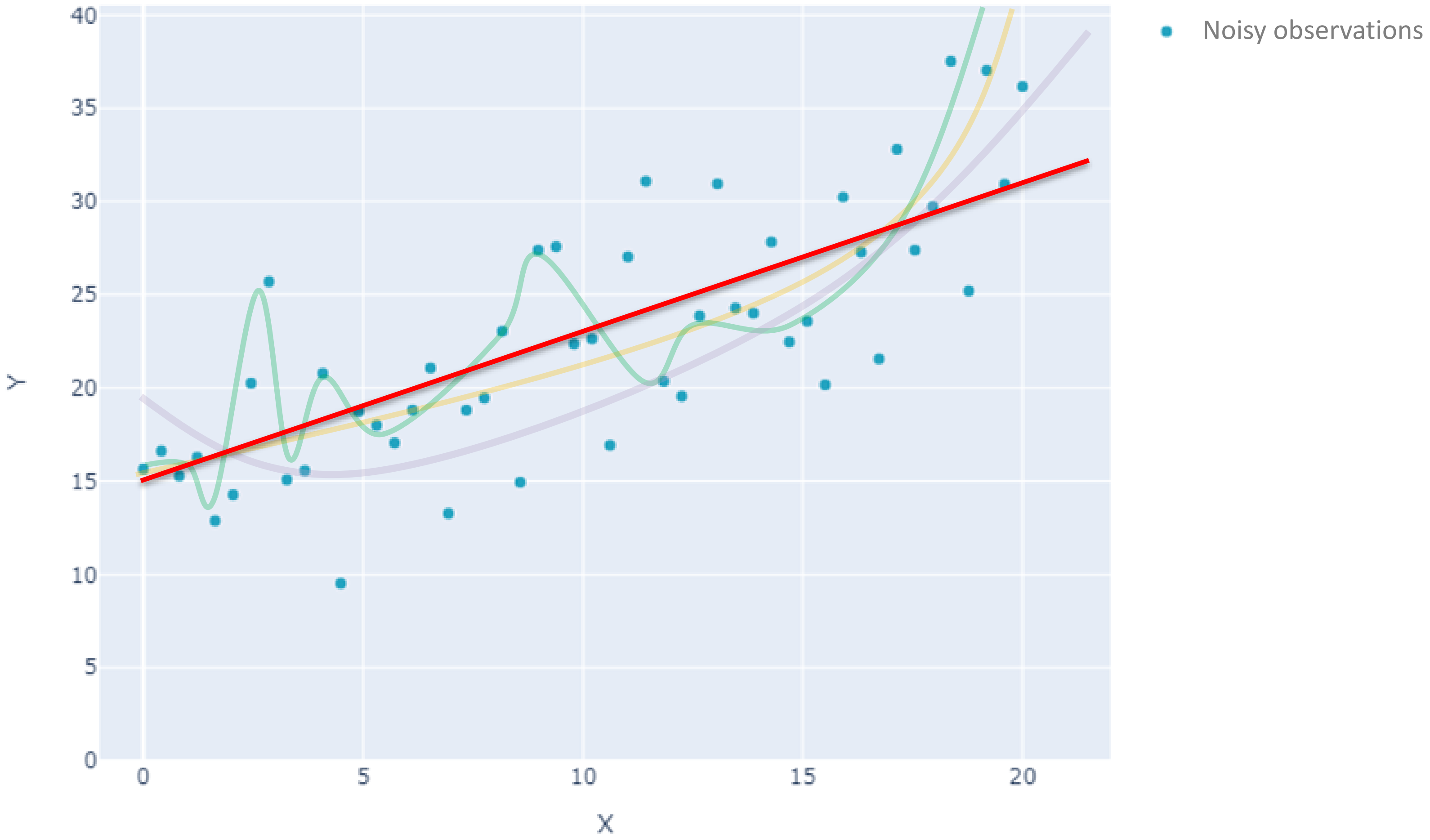
- Many alternative solutions



Regression Problem

Let's add noise … can we still predict Y from X?

- Many alternative solutions
- Linear modeling is **an assumption**

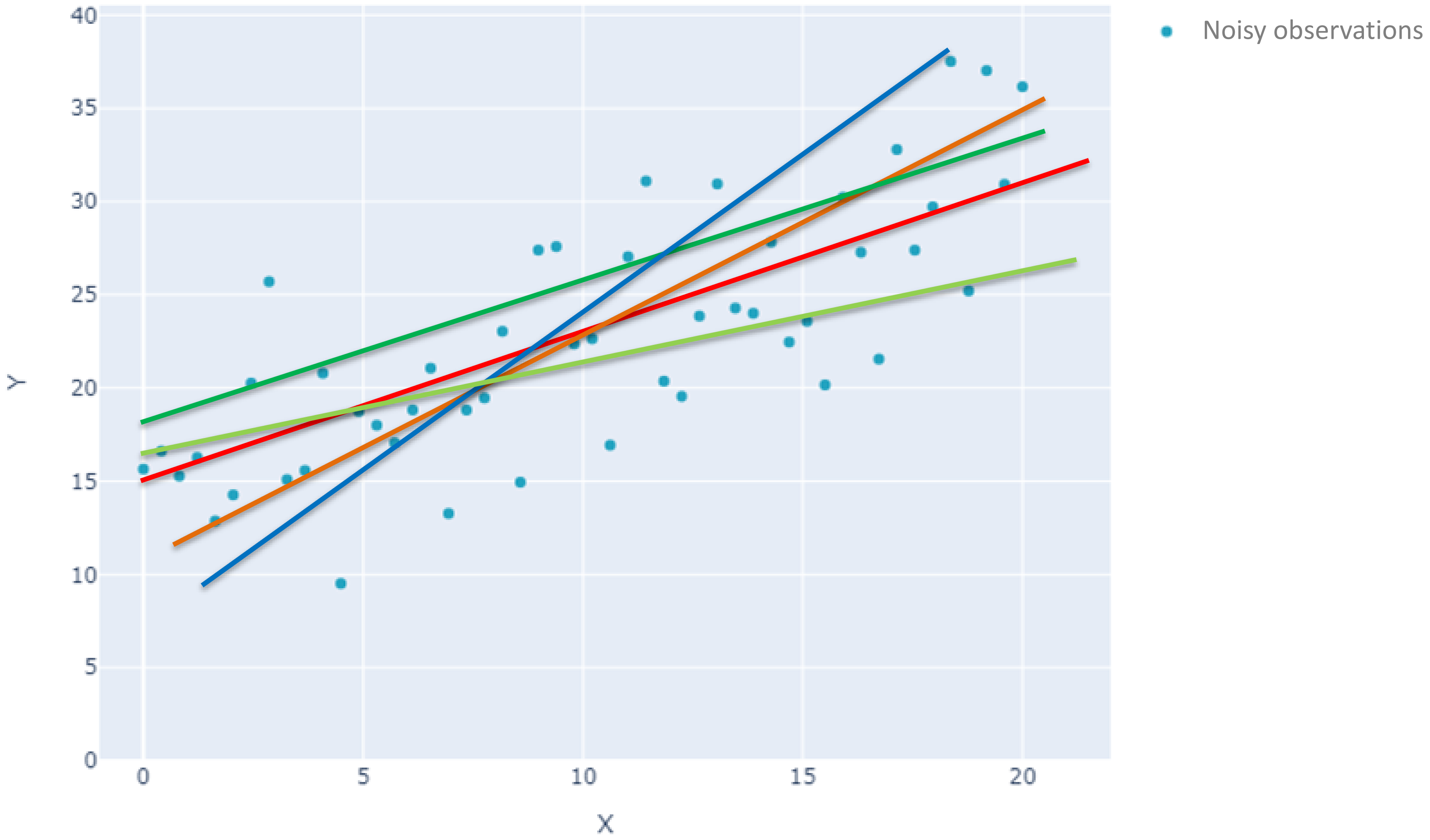


Regression Problem

Let's add noise … can we still predict Y from X?

- Many alternative solutions
- Linear modeling is **an assumption**

Which one is the best?

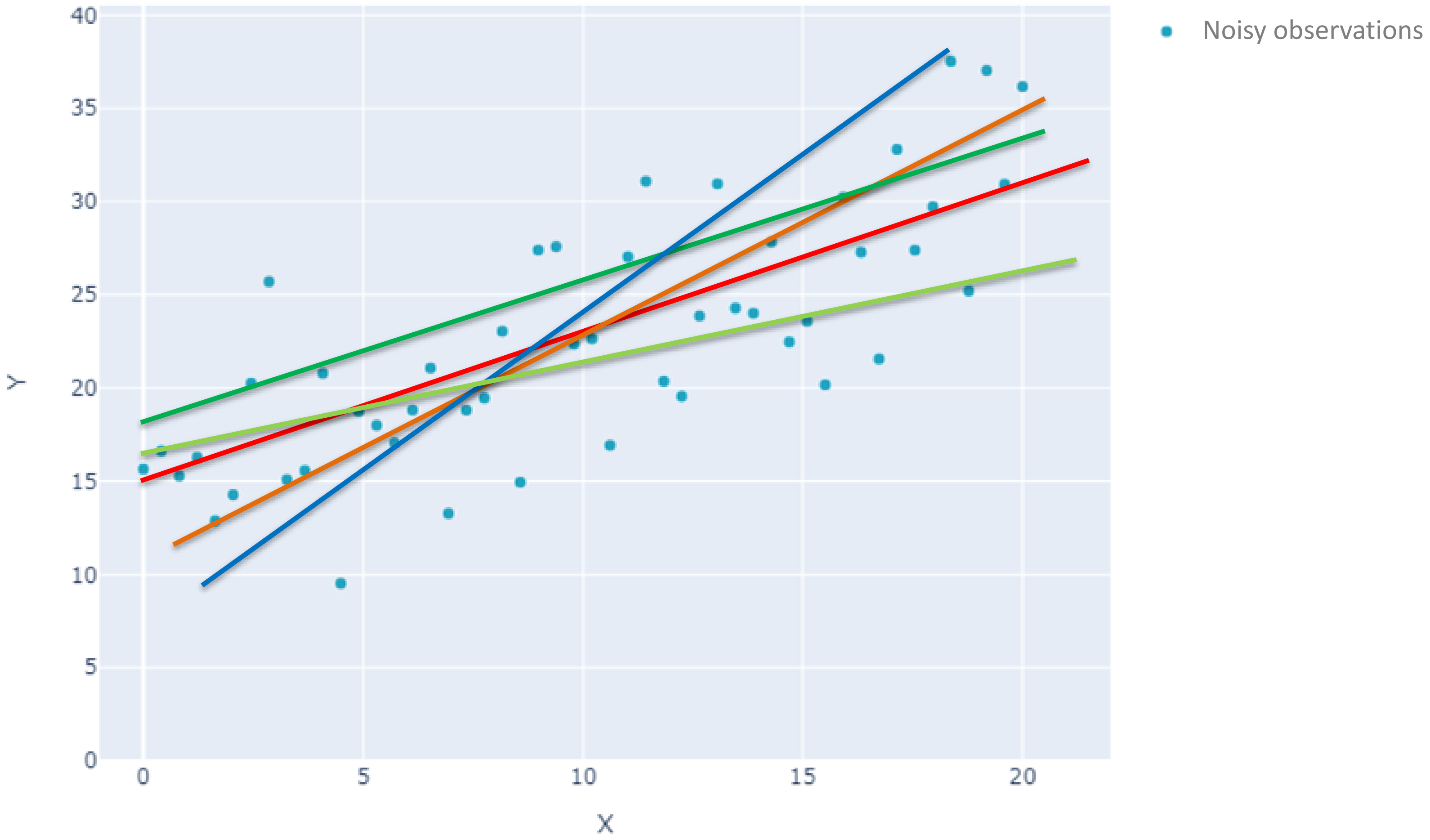


Linear Regression

Given observed pairs $< x_i, y_i >, \forall i \in N$, find the best linear model

$$y = w_0 + w_1\, x$$

- Infinite solutions exist
- Need to define an optimality criterion



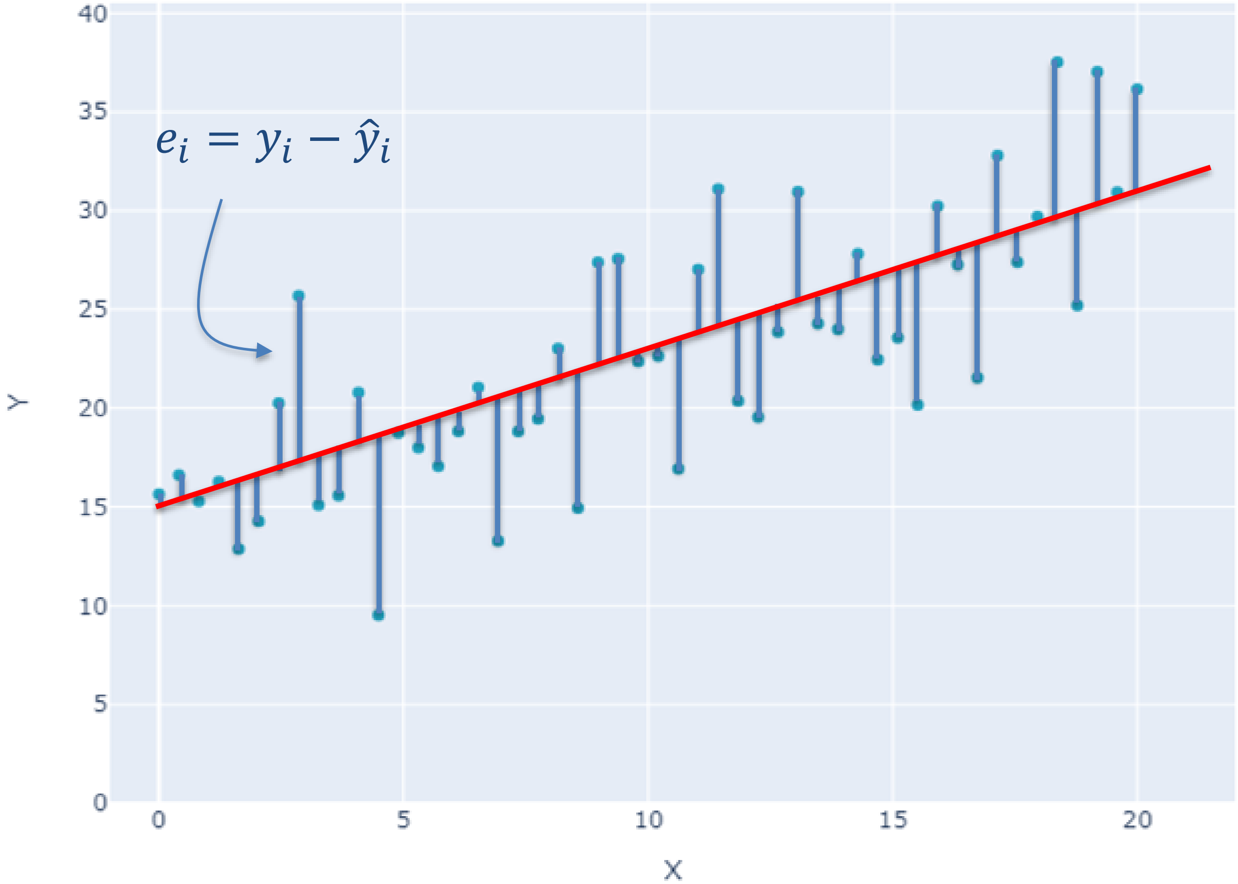
Least Squares Regression

The Residual Sum of Squares (RSS) is used to evaluate the model. It is defined as the sum of the squared residues $\boldsymbol{e_i = y_i - \hat{y}_i}$, i.e.,

$$RSS = e_1^2 + e_2^2 + \cdots + e_N^2$$

Rewriting as a function of parameter $\boldsymbol{w_0}$ and $\boldsymbol{w_1}$, we obtain

$$RSS(w_0, w_1) = \sum_{i=1}^{N} \left(y_i - (w_0 + w_1 x_i)\right)^2$$



Least Squares Regression

Then, the goal is to find the value of the weights / parameters w_0, w_1 (sometimes named as β_0, β_1) which minimize the RSS

$$\mathrm{w}_0, w_1 = \underset{w_0,w_1}{argmin}\, RSS(w_0, w_1) =$$
$$= \underset{w_0,w_1}{argmin} \sum_{i=1}^{N} \left(y_i - (w_0 + w_1 x_i)\right)^2$$

This is a function minimization problem we can solve by

- Computing gradients $\nabla RSS(w)$ w.r.t. weights $w = [w_0, w_1]$
- solving $\nabla RSS(w) = 0$.

Least Squares Regression

$$\mathrm{w}_0, w_1 = \underset{w_0, w_1}{argmin}\, RSS(w_0, w_1) =$$

$$= \underset{w_0, w_1}{argmin} \sum_{i=1}^{N} \left(y_i - (w_0 + w_1 x_i)\right)^2$$

Putting the gradient $\nabla RSS(w_0, w_1) = \mathbf{0}$, we obtain

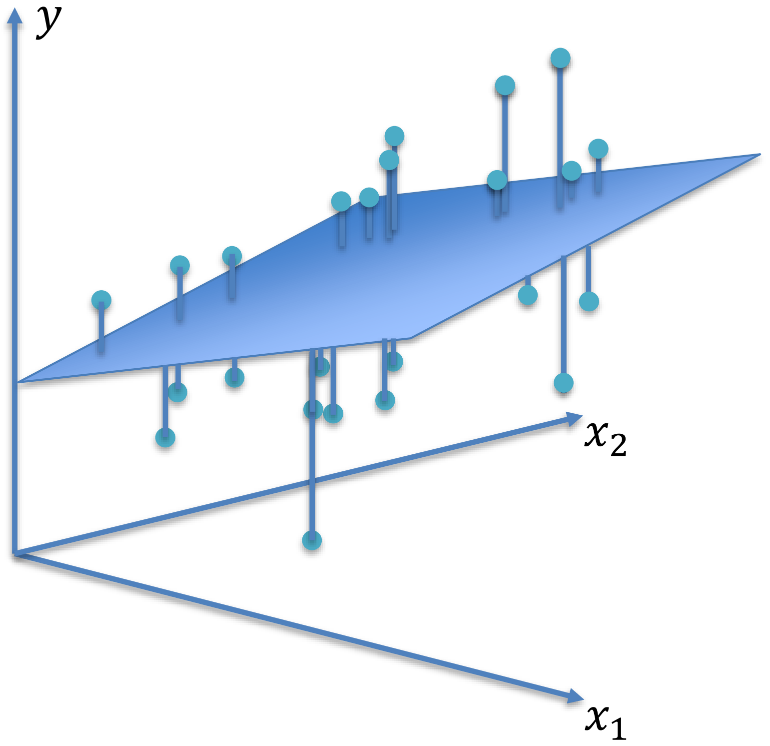
$$\frac{\partial RSS}{\partial w_0} = -2 \sum_{i=1}^{N} \left(y_i - (w_0 + w_1 x_i)\right) = 0$$

$$\frac{\partial RSS}{\partial w_1} = -2 \sum_{i=1}^{N} \left(y_i - (w_0 + w_1 x_i)\right) x_i = 0$$

Multivariate Linear Regression

Suppose to have M features, then the multivariate regression problem is

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M + \epsilon$$

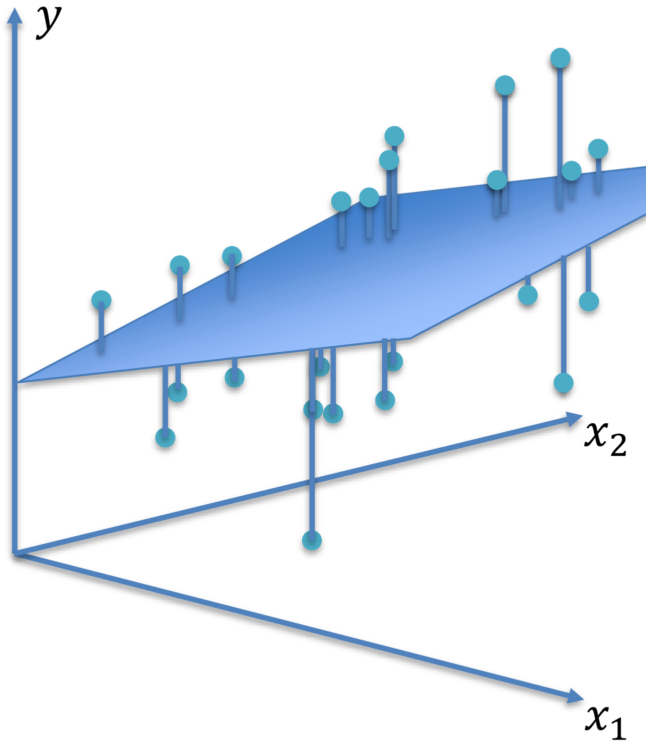


Example:
A regression with 2 features $[x_1, x_2]$ *and 1 target variable* y*. The least squares solution is a plane chosen by minimizing the distances between the observations and the plane.*

Multivariate Linear Regression

Suppose to have M features, then the multivariate regression problem is

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M + \epsilon$$



$$RSS = \sum_{i=1}^{N} \left(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_M x_{iM})\right)^2$$

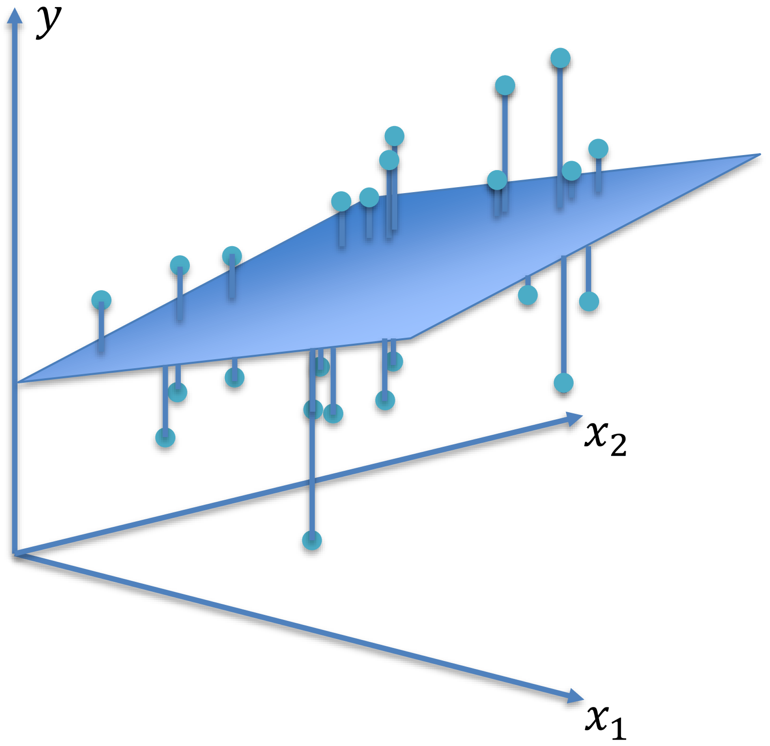
$$= \sum_{i=1}^{N} \left(y_i - \left(w_0 + \sum_{m=1}^{M} w_m x_{im} \right) \right)^2$$

$$RSS(w) = (y - \mathrm{X}w)^T (y - \mathrm{X}w)$$

Multivariate Linear Regression

Suppose to have M features, then the multivariate regression problem is

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M + \epsilon$$



$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \quad \begin{bmatrix} \boxed{\begin{matrix} 1 \\ 1 \\ \cdots \\ 1 \end{matrix}} & \begin{matrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & x_{NM} \end{matrix} \end{bmatrix} \quad \begin{bmatrix} \boxed{w_0} \\ w_1 \\ w_2 \\ \cdots \\ w_M \end{bmatrix}$$

- y = N x 1 vector of target values
- X is a N x (M+1) data matrix
- w is a (M+1) x 1 vector of weights

$$RSS(w) = (y - \mathrm{X}w)^T (y - \mathrm{X}w)$$

Multivariate Linear Regression

$$RSS(w) = (y - \mathrm{X}w)^T(y - \mathrm{X}w)$$

Quadratic function, thus convex, thus unique minimum …

Let's compute the derivatives

$$\frac{\partial RSS}{\partial w} = -2\mathrm{X}^T(y - \mathrm{X}w)^T$$

Putting the derivatives equal to zero and solving for $\boldsymbol{w}$ we obtain

$$w = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T y$$

which are known as the normal equations for the least squares problem.

Quite expensive with many samples …

Moore-Penrose pseudo-inverse of X

Non linearity of data

$$y = w_0 + \underbrace{w_1 x_1} + \underbrace{w_2 x_2} + \cdots + \underbrace{w_M x_M} + \epsilon$$

Multivariate linear regression assumes

- Linear relationship between features and target variables
- Additive relationship between features and target variables

Linear models could not be sufficient to fit observed data as a linear relationship between features and target may not hold.

Generalized Linear Regression

Given a set of input variables x, a set of N examples $< x_i, y_i >$ and a set of D features h_j computed from the input variables x_i, we get the model

$$y = w_0 + w_1 f(x)_1 + w_2 f(x)_2 + \cdots + w_D f(x)_D + \epsilon$$

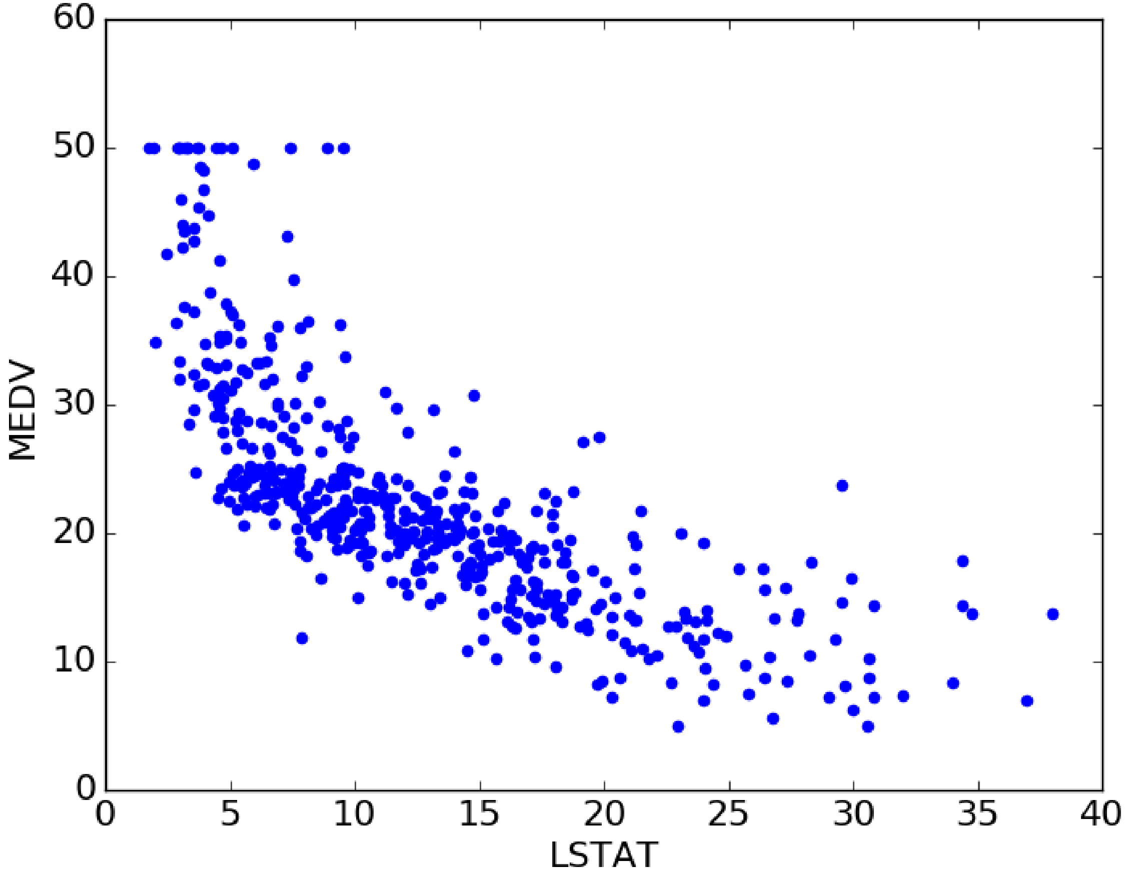
- $f_j(.)$ identify variables derived from the original inputs
- $f_j(.)$ could be derived from an existing variable, e.g., the squared value, a trigonometric function, the age given the date of birth, etc.

<u>Notes</u>: the solution can still be computed via matrix pseudo inverse (but you get a non linear model)

$$W = \left(f(X)^T f(X)\right)^{-1} f(X) Y$$

Polinomial Regression

Input: LSTAT - % lower status of the population
Output: MEDV - Median value of owner-occupied homes in $1000's

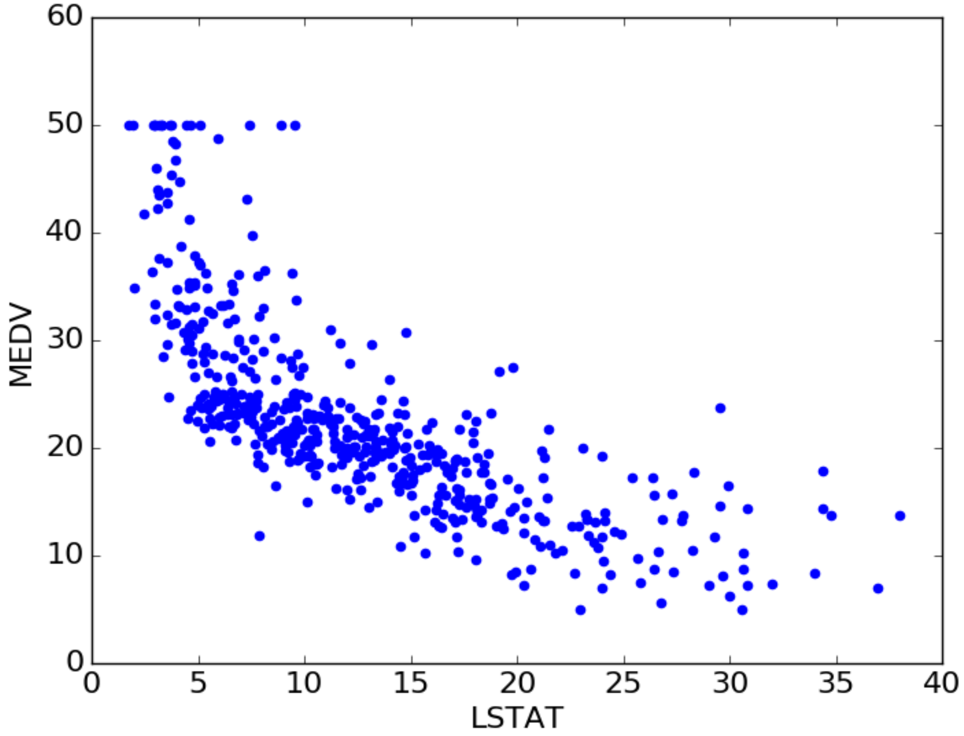


Polinomial Regression

Given a set of examples associating $LSTAT_i$ values to $MEDV_i$ values, nonlinear regression finds a function $f(.)$ such that

$$MEDV_i = f(LSTAT_i) + \varepsilon_i$$

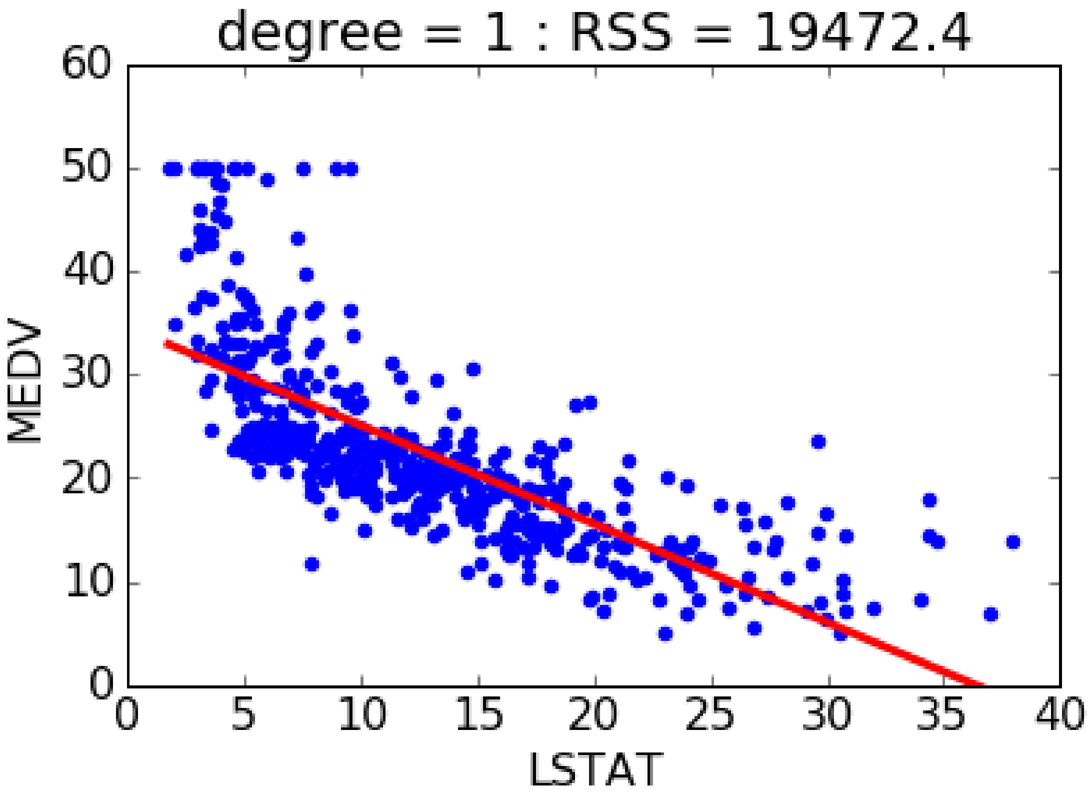
where $\boldsymbol{\varepsilon}_i$ is the error to be minimized

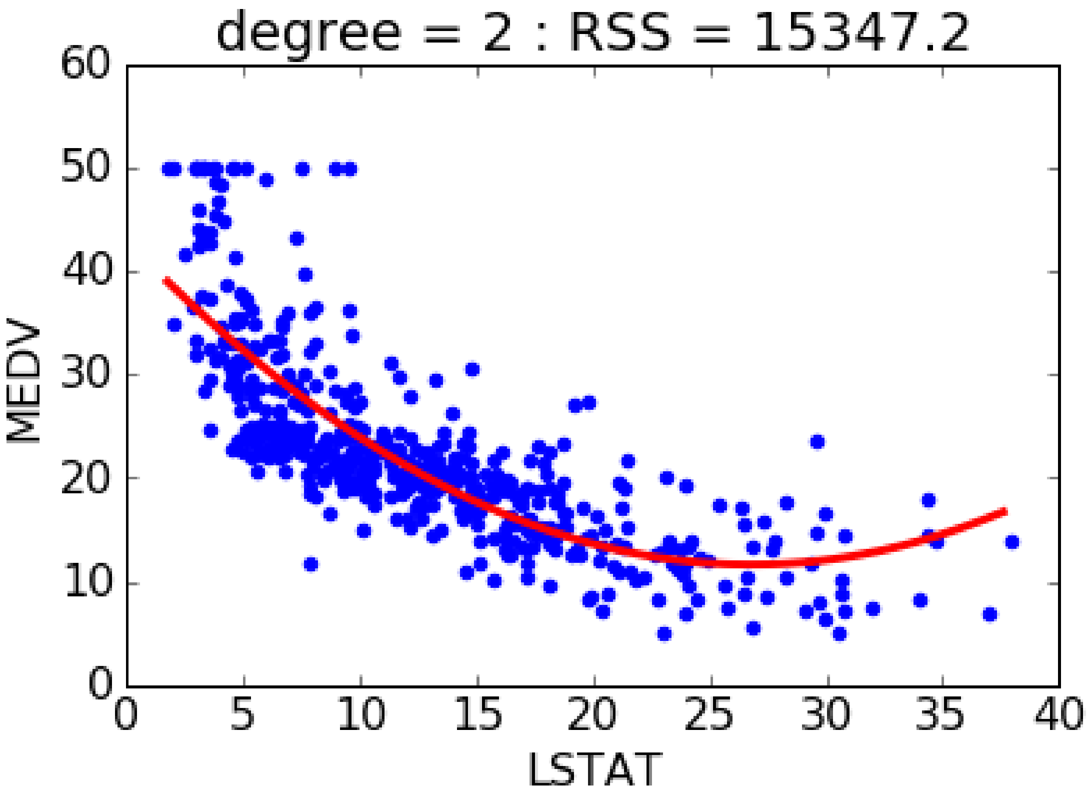


A polynomial model would fit the data points with a function,

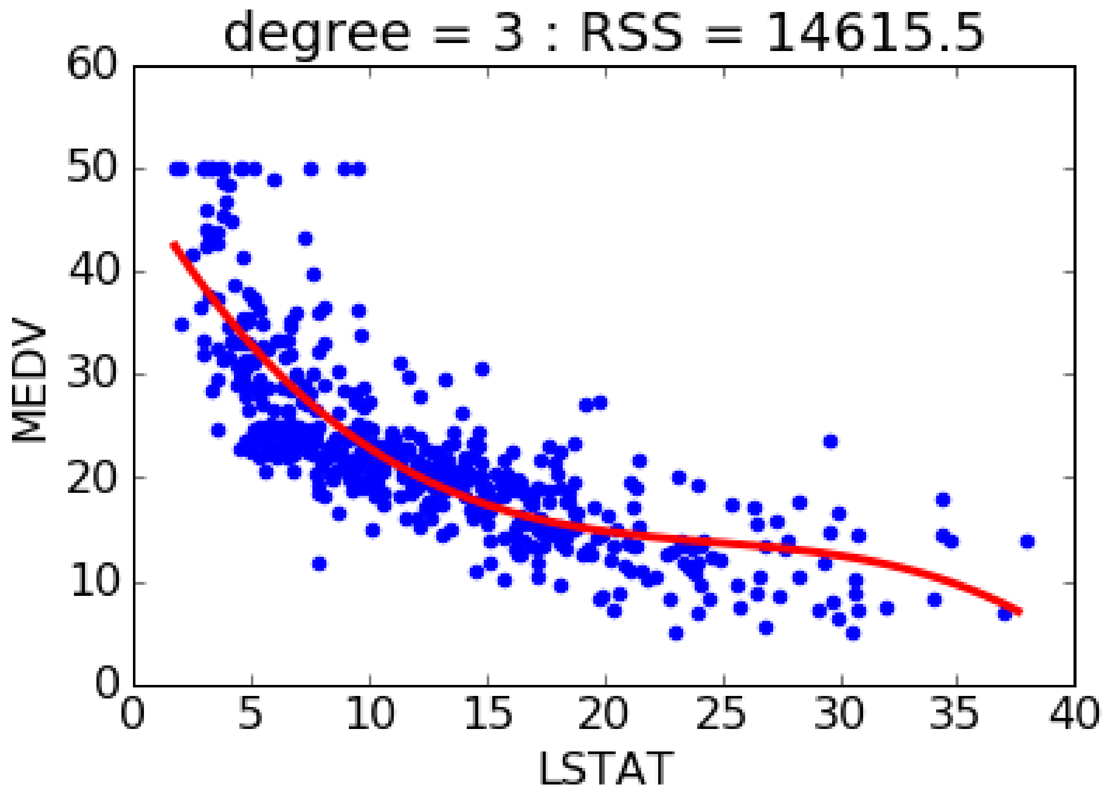
$$f(LSTAT_i) = w_0 + \sum_{j=1}^{D} w_j \times LSTAT_i{}^j$$

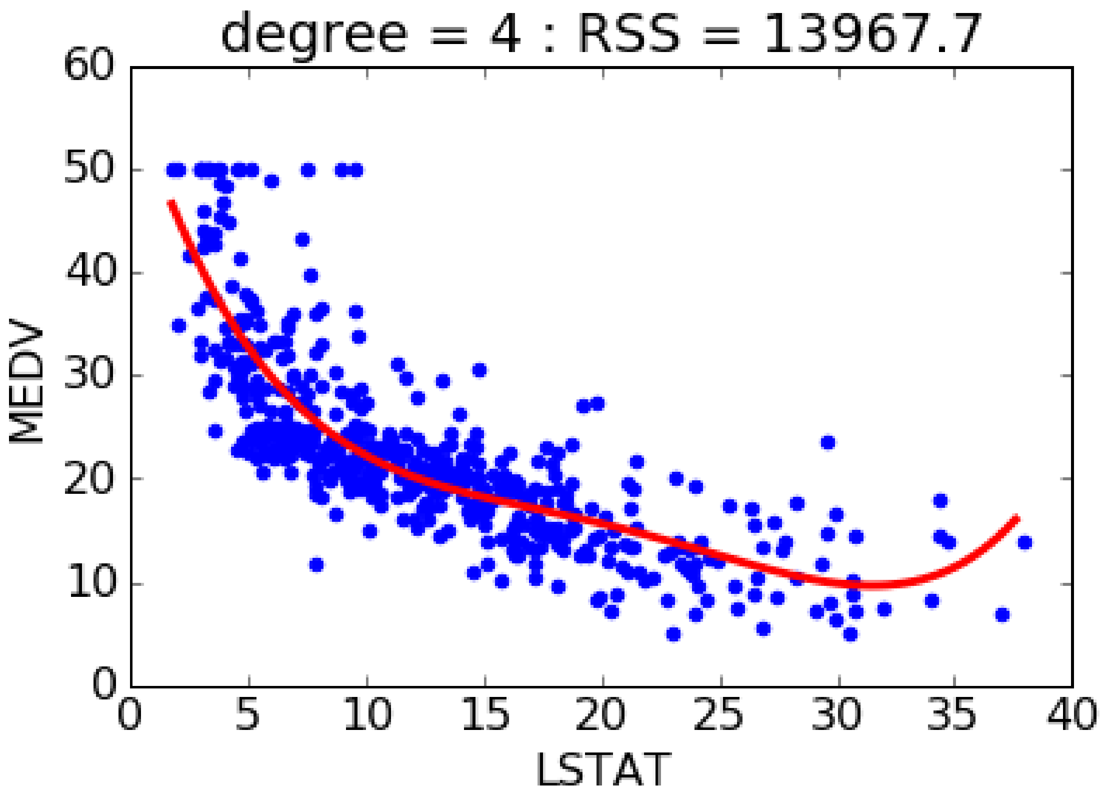
Polinomial Regression



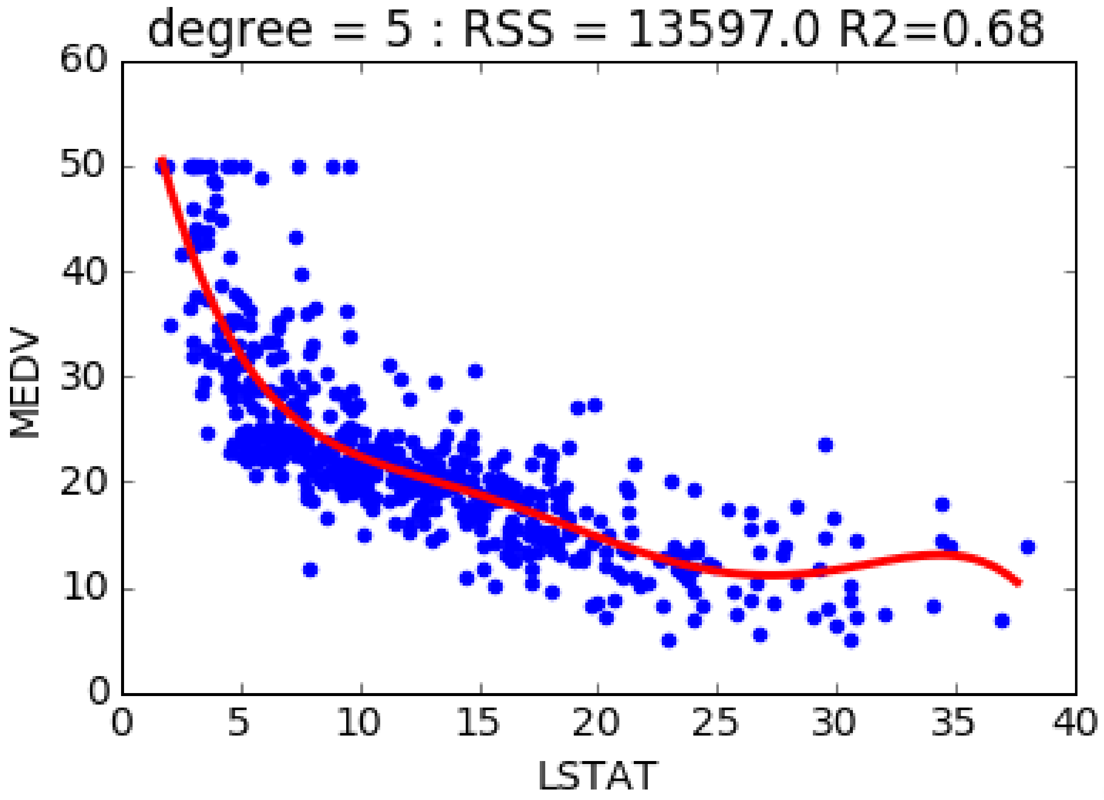


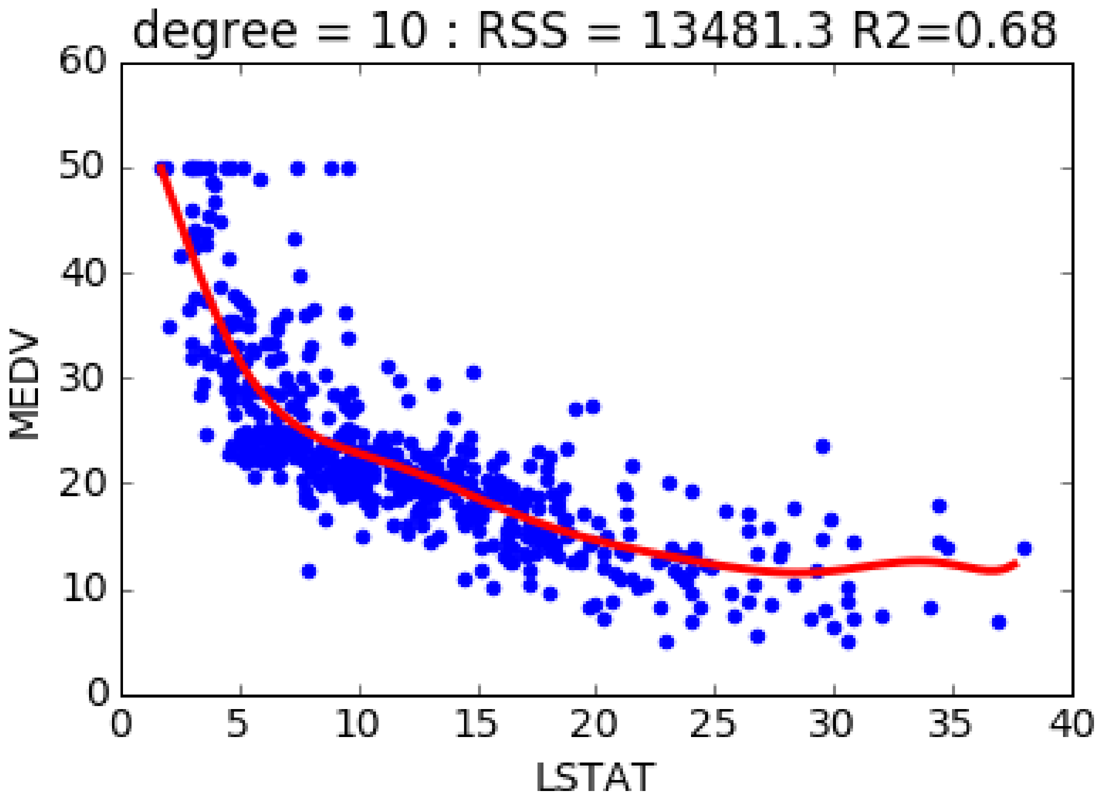
Polinomial Regression





Polinomial Regression





Which one do you prefer?

Coefficient of Determination R^2

Total sum of squares

$$TSS = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

Coefficient of determination

$$R^2 = 1 - \frac{RSS}{TSS}$$

What about new data?

R^2 measures of how well the regression line approximates the real data points. When R^2 is 1, the regression line perfectly fits the data.

Model Evaluation

Would be feasible to evaluate students using exactly the same problems solved in class?

Overfitting: perfect output on the training data,
terrible outcome on data which has never seen before ☹

Models should be evaluated using data that have not been used to build the model itself:

- Training data will be used to build the model
- Test data will be used to evaluate the model performance

Hold Out Evaluation

Reserves a certain amount for testing and uses the remainder for training

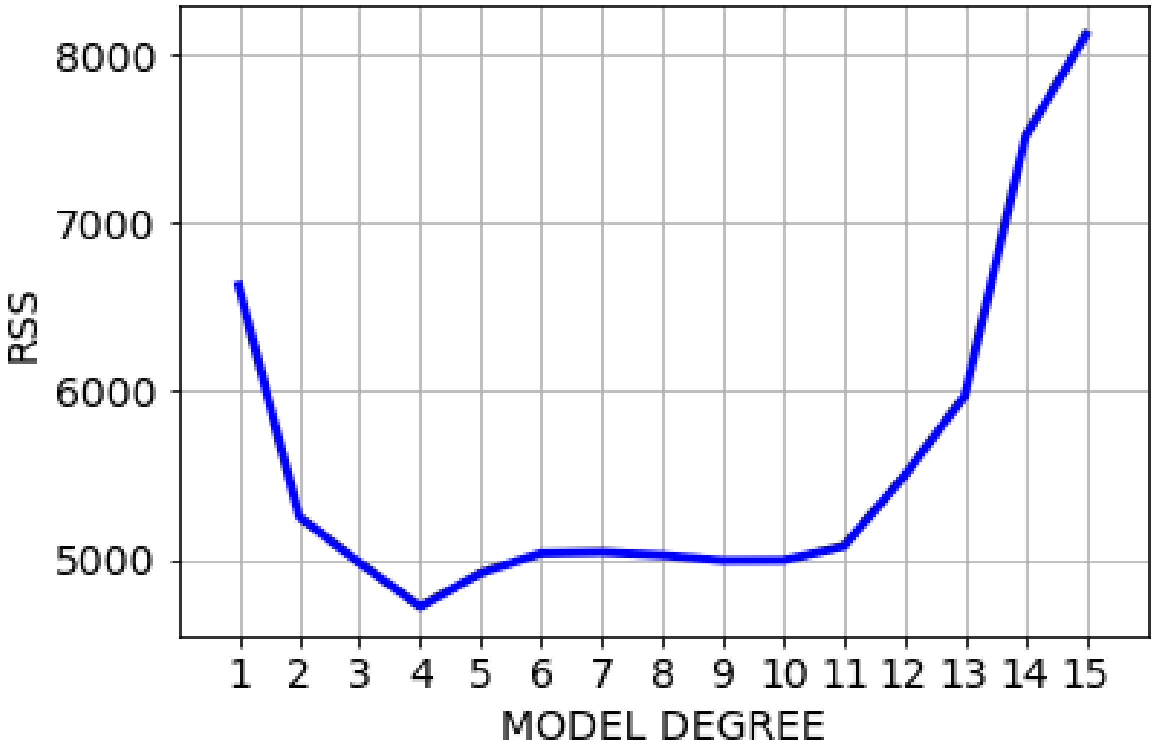
- Too small training sets might result in poor weight estimation
- Too small test sets might result in a poor estimation of future performance

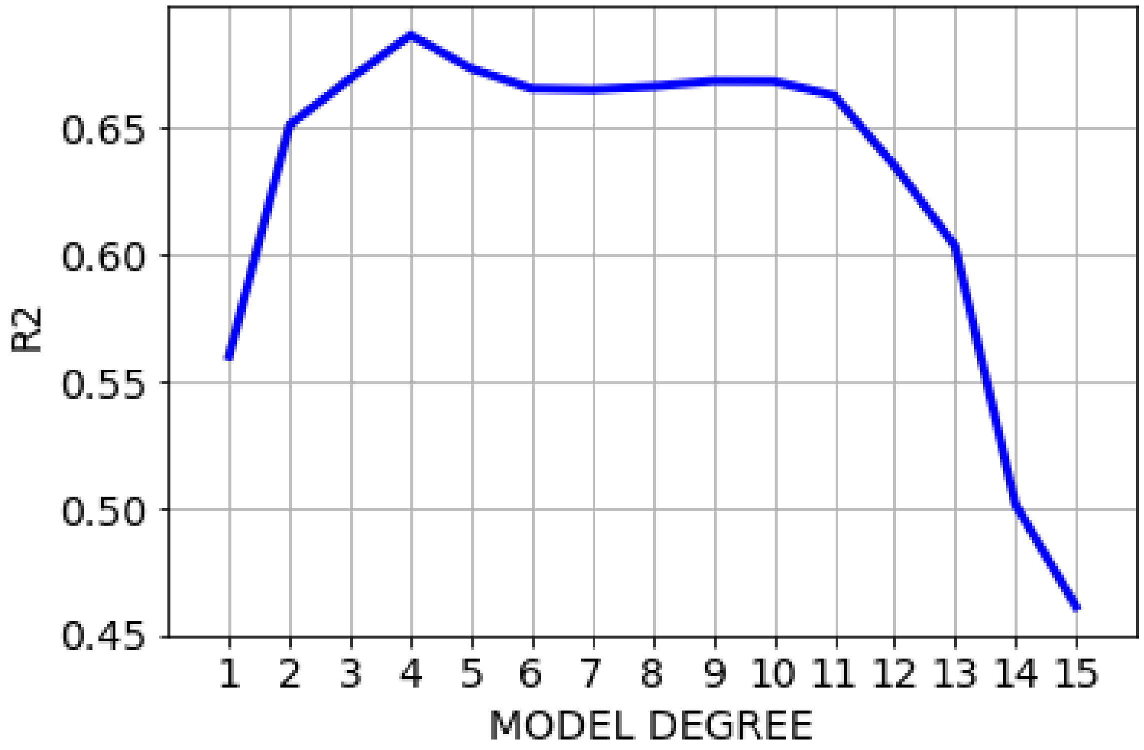
Typically,

- Reserve 2/3 for training and 1/3 for testing (but percentage may vary)

Hold Out Evaluation on Housing Data

Given the original dataset, split the data into 2/3 train and 1/3 test and then apply linear regression using polynomials of increasing degree.





Cross-Validation

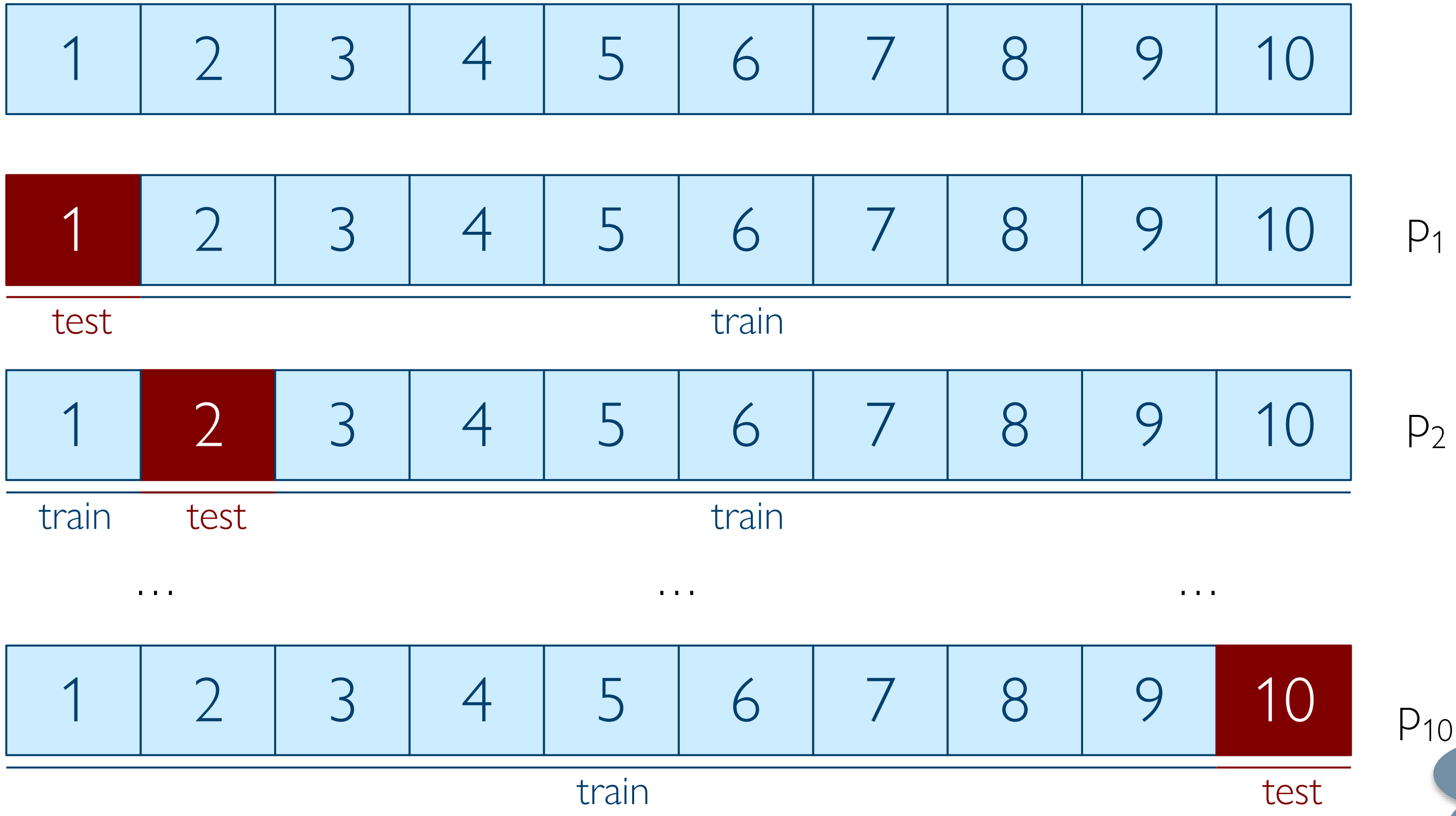
For small or “unbalanced” datasets, Hold Out samples might not be representative, thus k-fold Cross-Validation can be used:

- Data is split into k subsets of equal size
- Each subset in turn is used for testing and the remainder for training
- The error estimates are averaged to yield an overall error estimate

Cross-Validation

For small or "unbalanced" datasets, Hold Out samples
representative, thus k-fold Cross-Validation can be use

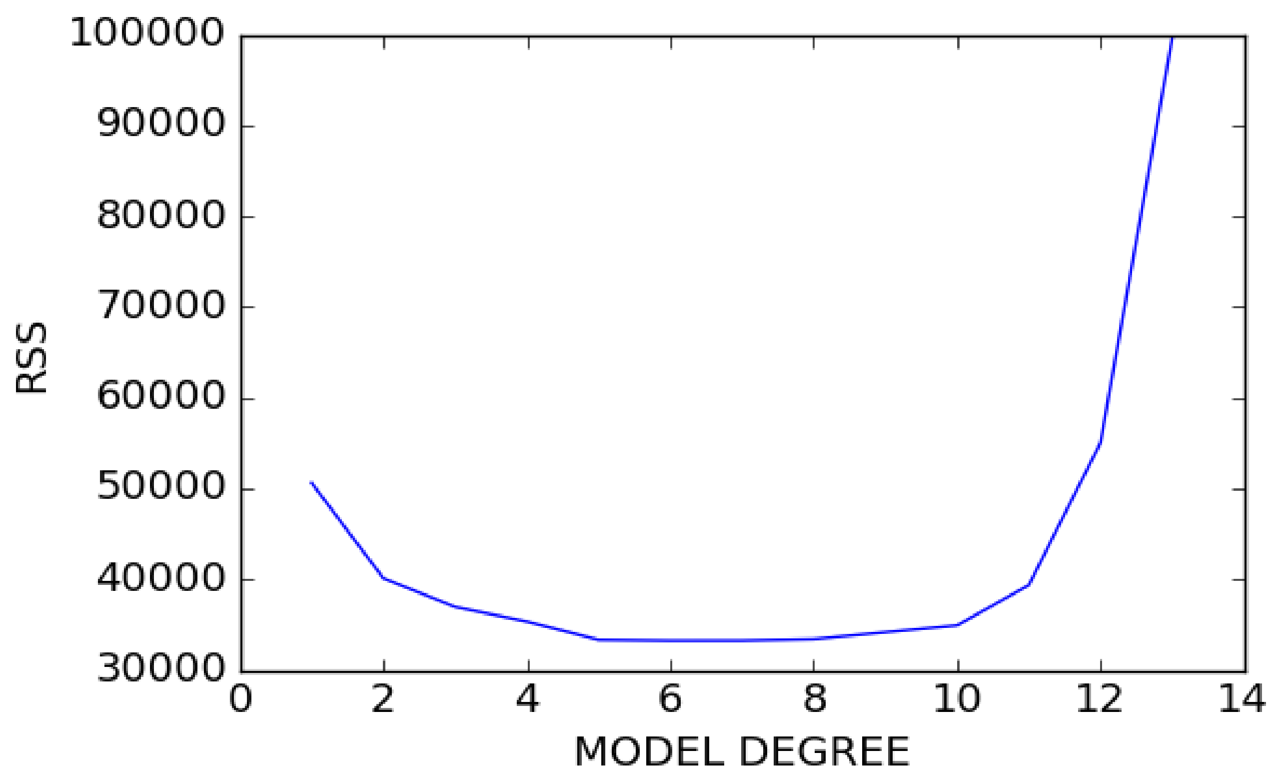
K=10 gets accurate estimantes

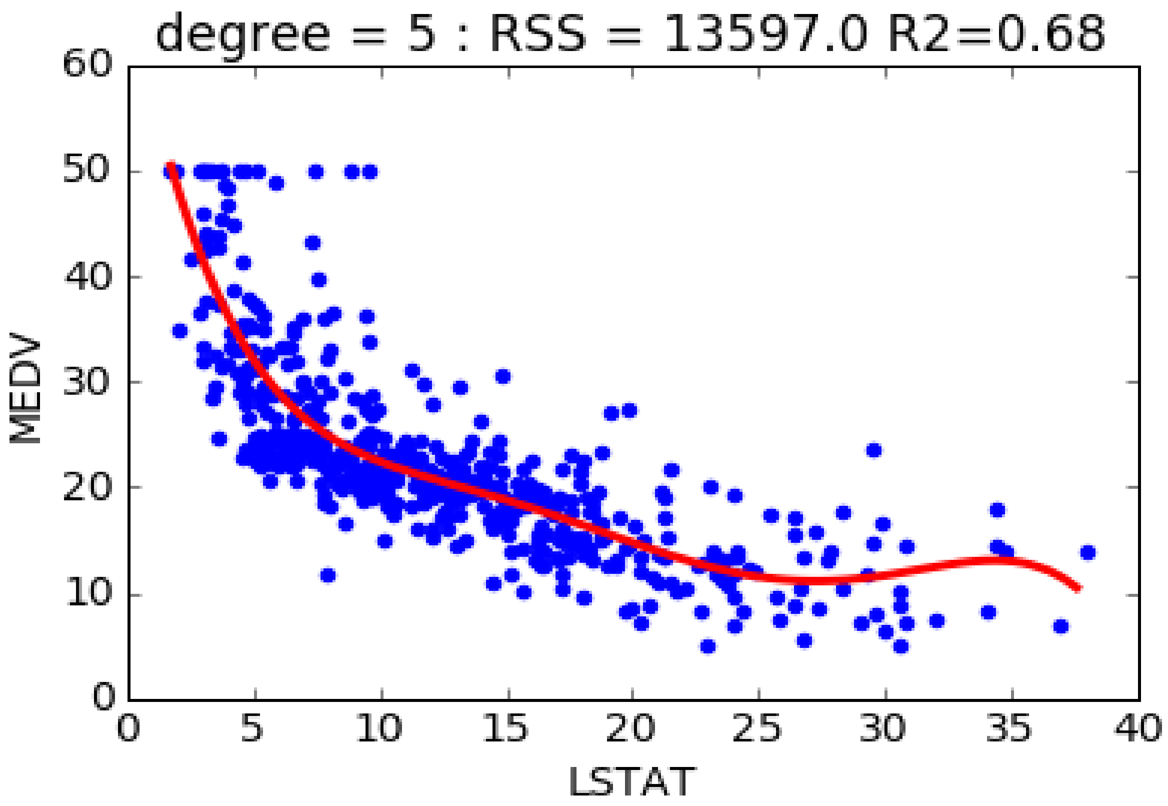


Sometimes repeated multiple times (e.g., 10)

K-fold Cross-Validation on Housing Data

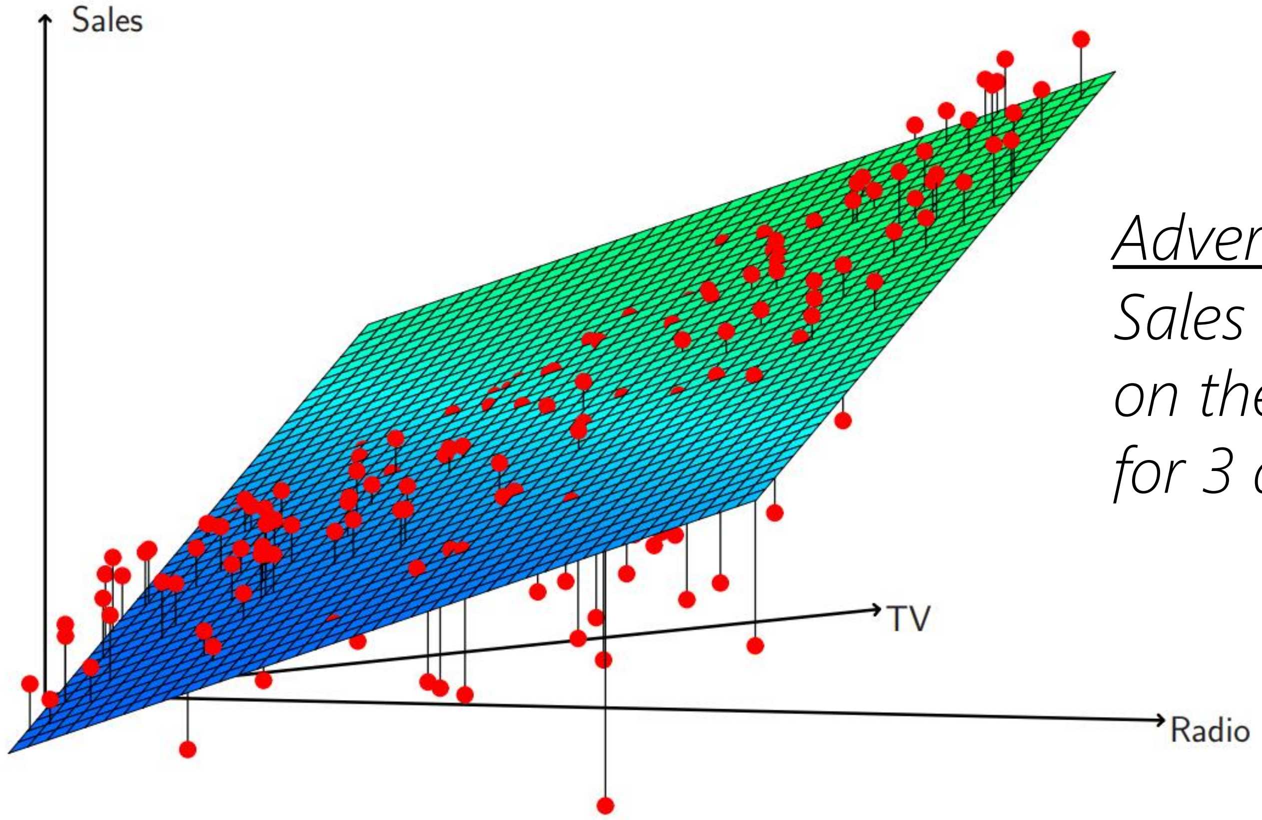
Given the original dataset, perform k-fold Cross-Validation on linear regression using polynomials of increasing degree.





Other non linearities: Sinergy

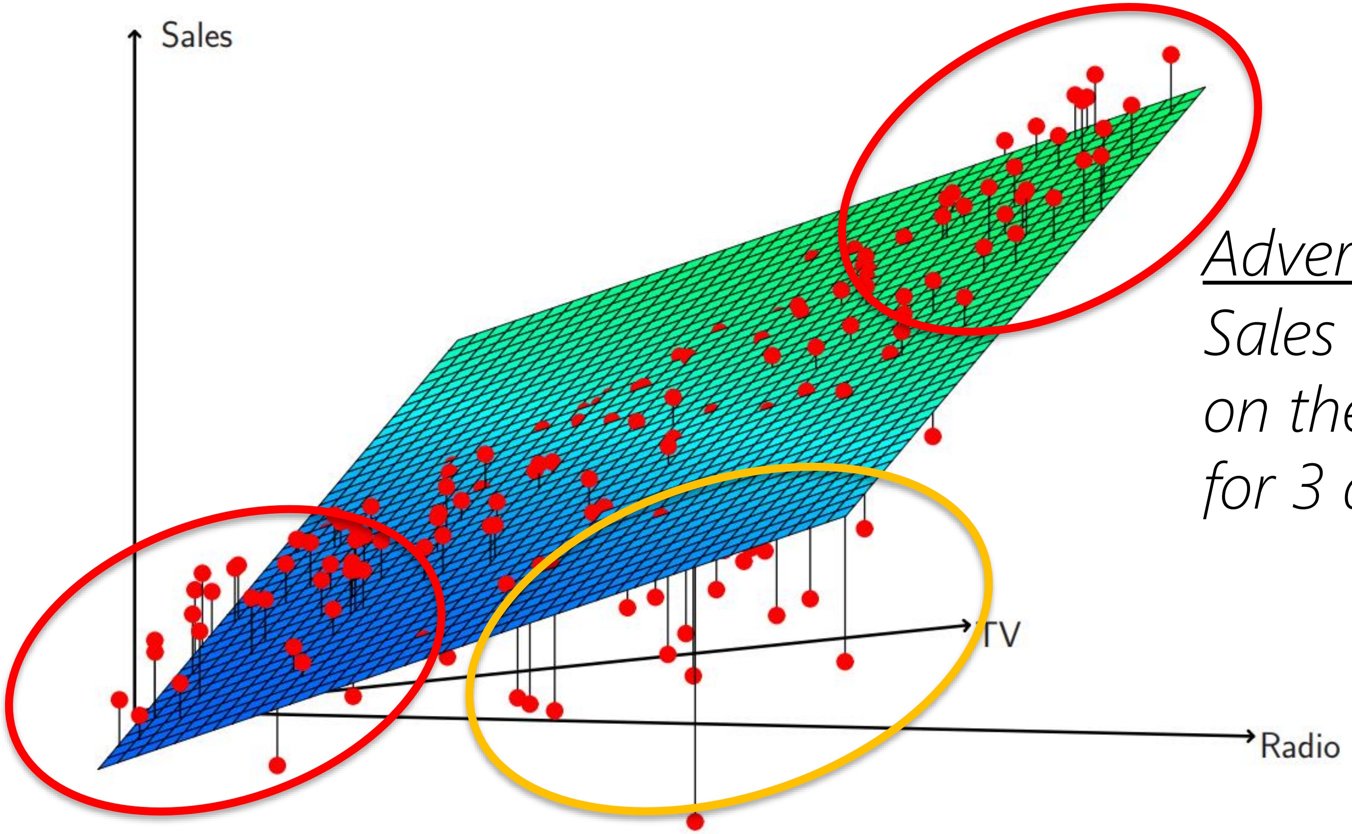
We could have interactions between variables (or sinergies)



<u>Advertising dataset example</u>
Sales of a product in 200 markets, based on the advertising budgets for the product for 3 different media: TV, radio, newspaper.

Other non linearities: Sinergy

We could have interactions between variables (or sinergies)



<u>Advertising dataset example</u>

Sales of a product in 200 markets, based on the advertising budgets for the product for 3 different media: TV, radio, newspaper.

This effect in marketing is called *sinergy,* i.e., acting on one variable modifies the other variables

Linear model underestimates red regions and overestimates yellow ones

Modeling Sinergy

Let consider the classical Linear Regression model with 2 variables

$$y = w_0 + w_1 x_1 + w_2 x_2 + \epsilon$$

- An increase in $\boldsymbol{x_1}$ of 1 unit increases $\boldsymbol{y}$ on average by $\boldsymbol{w_1}$ units
- Presence or absence of other variables does not affect this

We can extend the linear model with an interaction term

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + \epsilon$$

- Non-linearity w.r.t. the $\boldsymbol{x}$ variables
- Linear w.r.t. the parameters $\boldsymbol{w}$

Modeling Sinergy

We can extend the linear model with an interaction term

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + \epsilon$$

- Non-linearity w.r.t. the x variables
- Linear w.r.t. the parameters w

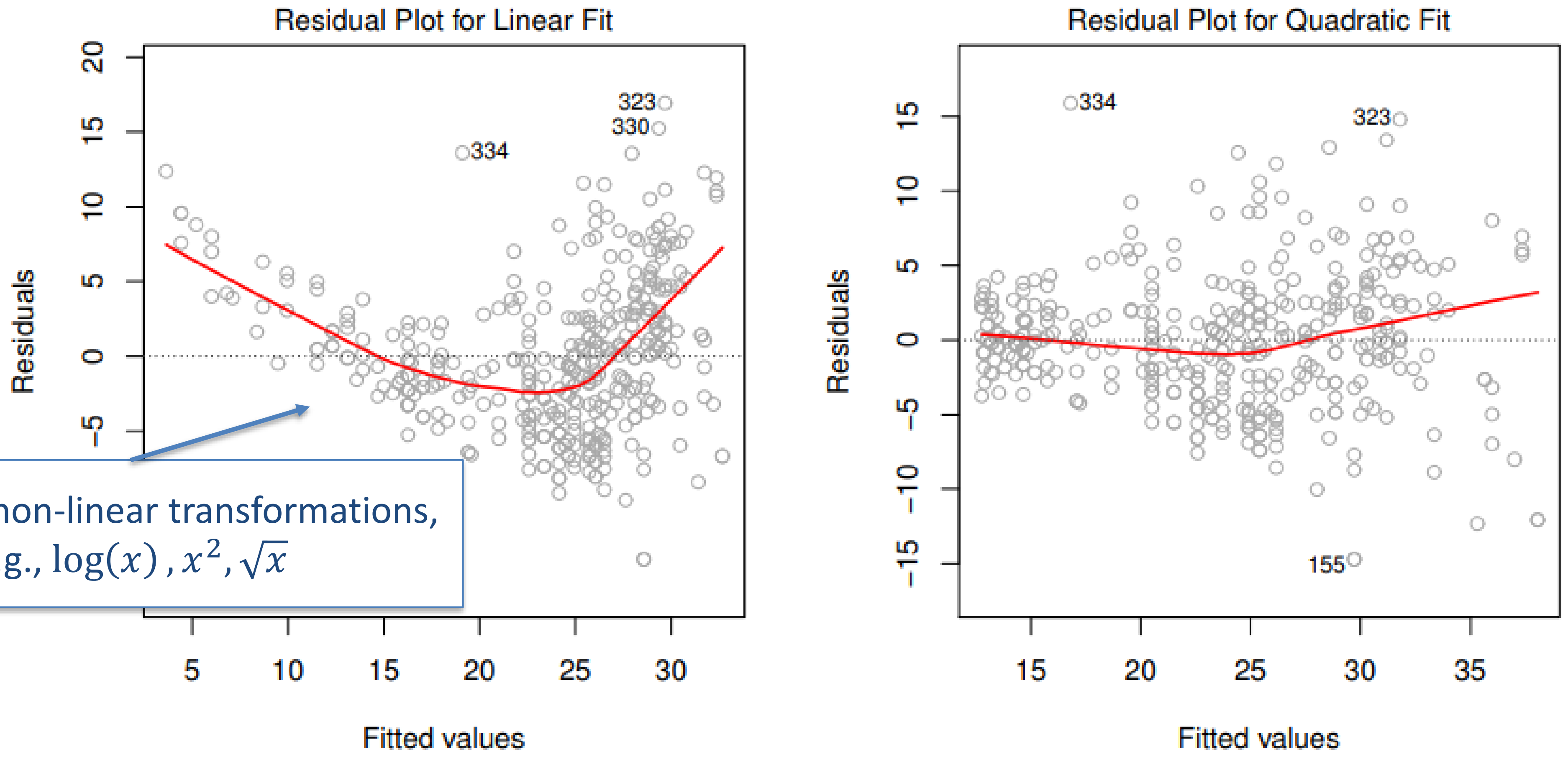
This translates in a linear model by adding a new variable $x_3 = x_1 x_2$

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \epsilon$$

We can use the same training algorithms for linear regression!

Checking for non linearities

Use residuals plot to check if the linearity assumption does not hold



Try to use non-linear transformations, e.g., $\log(x)$, x^2, $\sqrt{x}$

Non-Constant Variance of Error Term

Linear Regression assumes no heteroscedasticity in the noise

$$y = w_0 + w_1 x_1 + \ \dots + w_N x_N + \epsilon \qquad \text{with } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

constant

This might not be true, and the variance is a function σ^2 (x) of the data

- this effect is called heteroscedasticity
- if we have a constant σ^2 we have instead homoscedasticity

Non-Constant Variance of Error Term

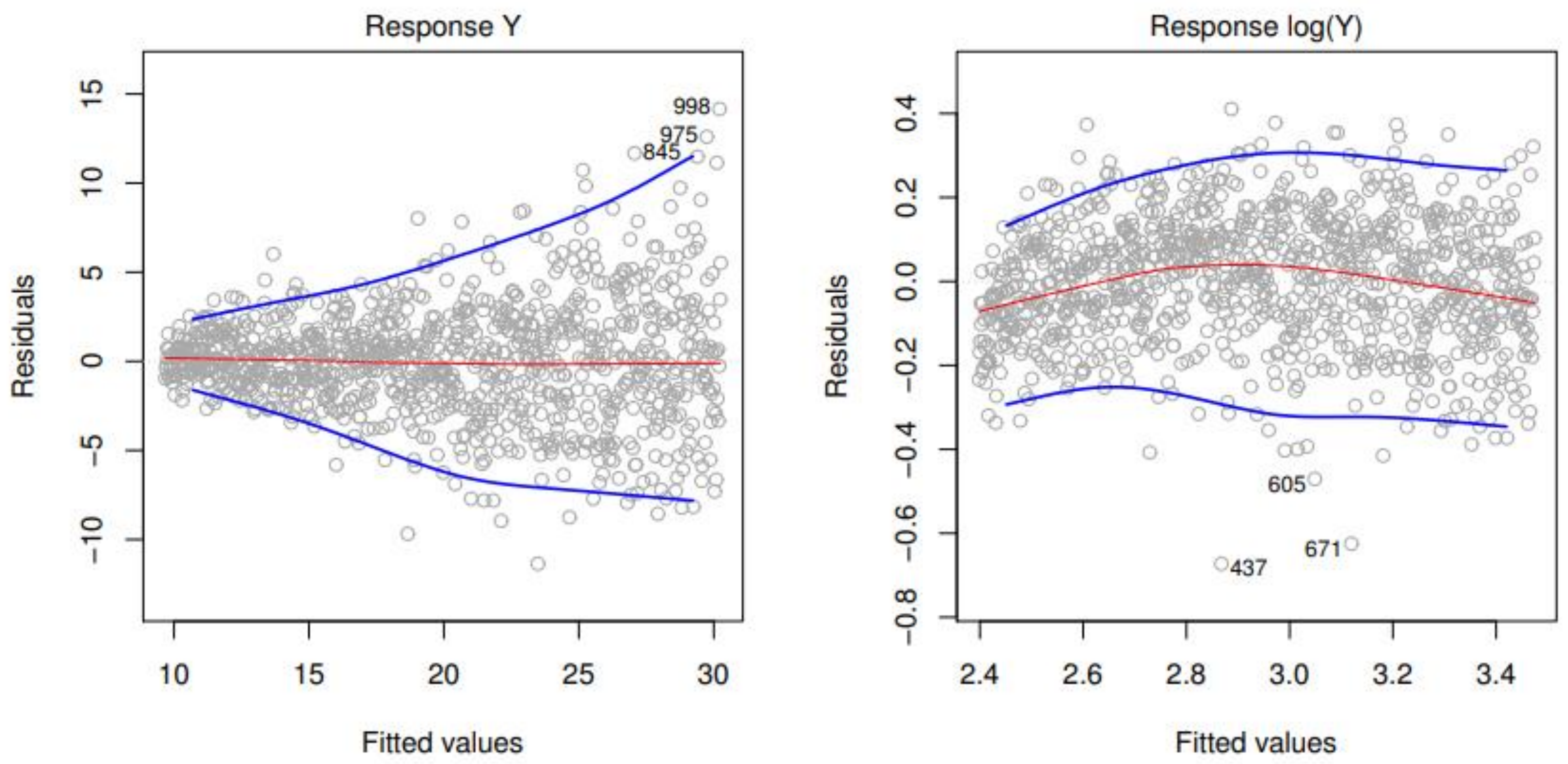


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

Outliers, residual plot, and studentized residuals

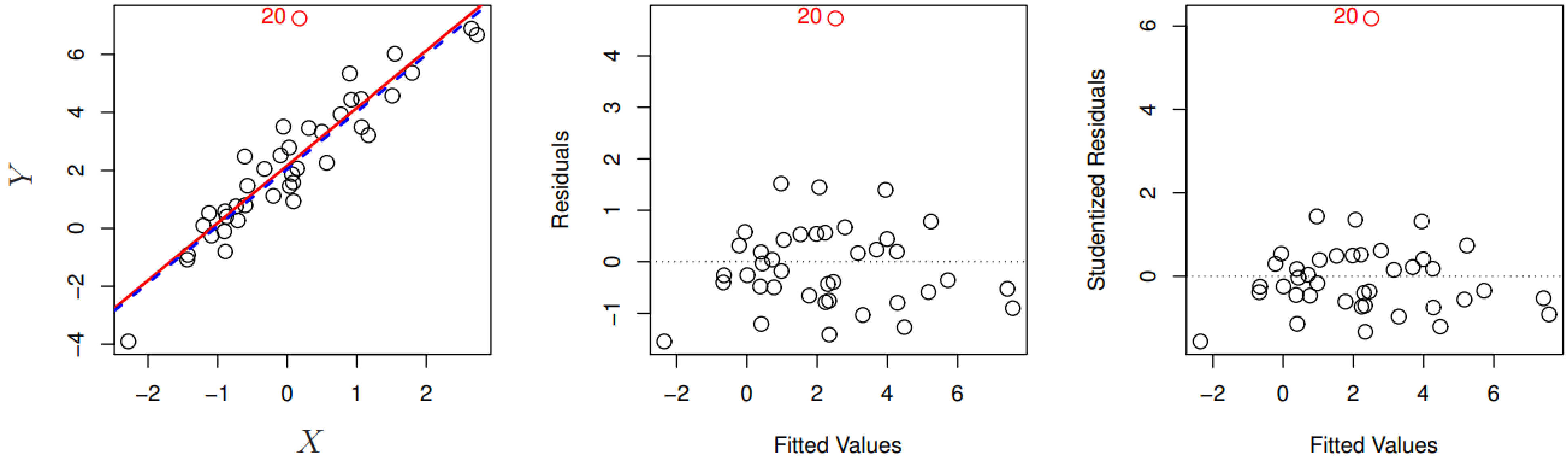
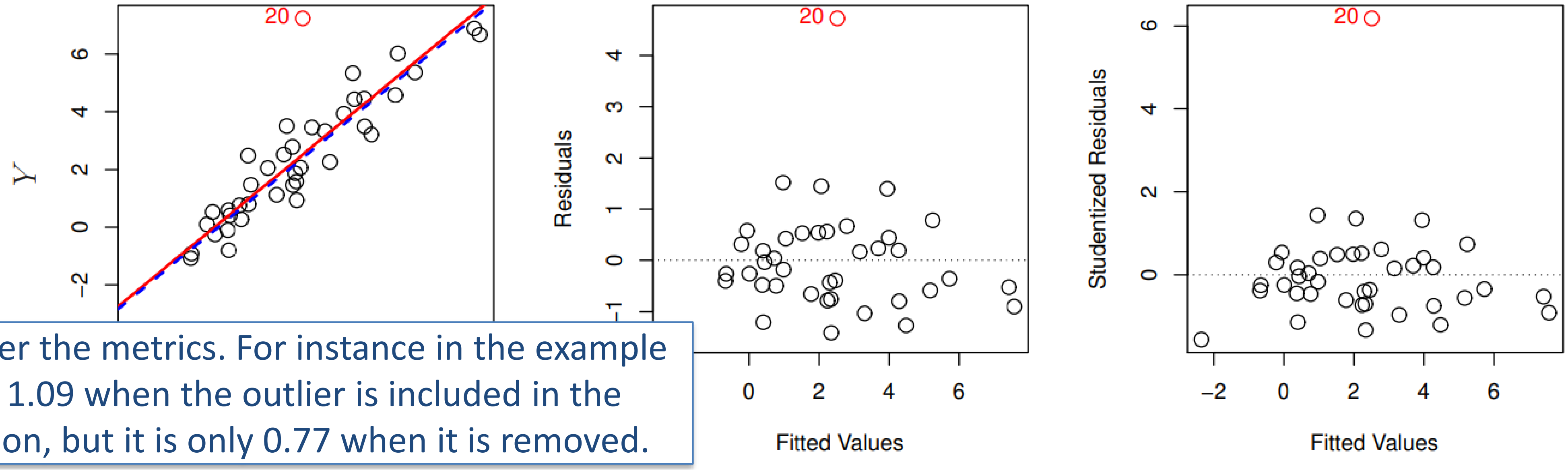


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between* -3 *and* 3.

Outliers, residual plot, and studentized residuals



It can alter the metrics. For instance in the example RSE is 1.09 when the outlier is included in the regression, but it is only 0.77 when it is removed.

FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between* -3 *and* 3*.*

High Leverage Points (unusual x)

Easy to observe with only one feature

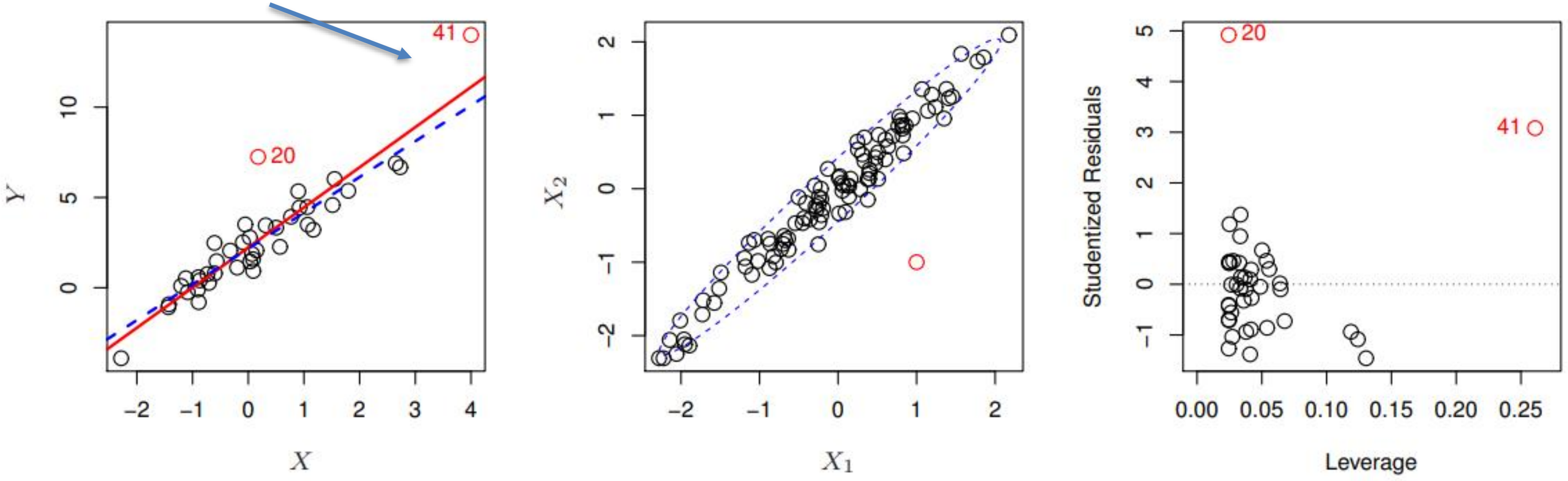


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

High Leverage Points (unusual x)

Easy to observe with only one feature

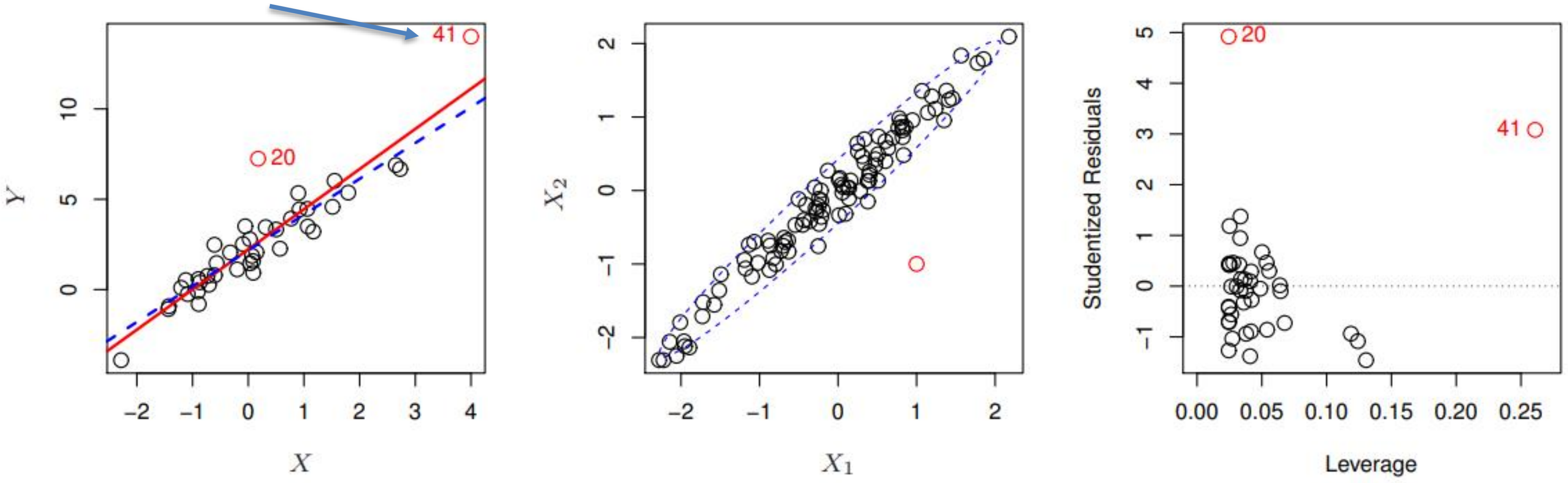


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

High Leverage Points (unusual x)

The value of the high leverage point is in the range of each individual feature's values, but it is an high leverage point

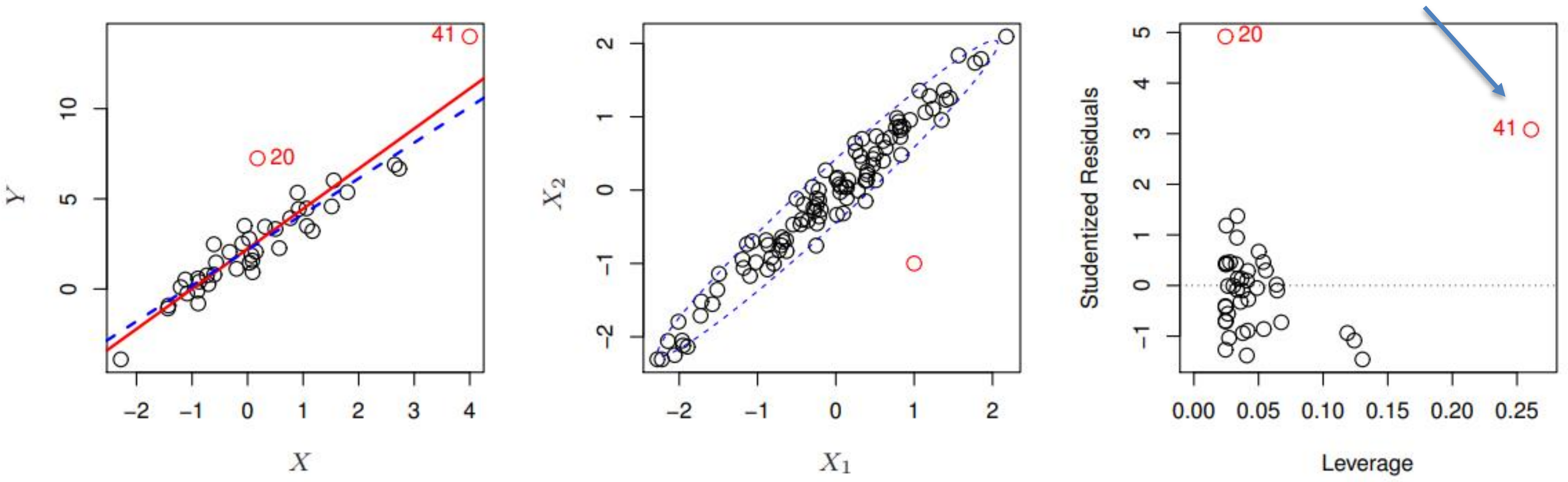
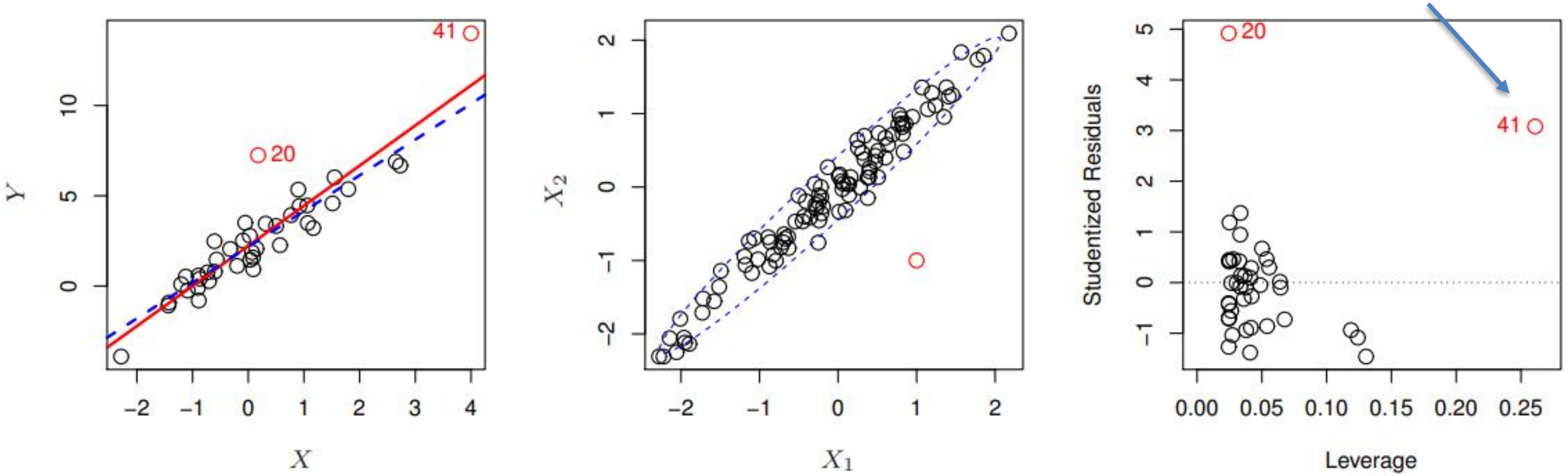


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

High Leverage Points (unusual x)

The value of the high leverage point is in the range of each individual feature's values, but it is an high leverage point



The leverage of an observation is computed via the *leverage statistic*

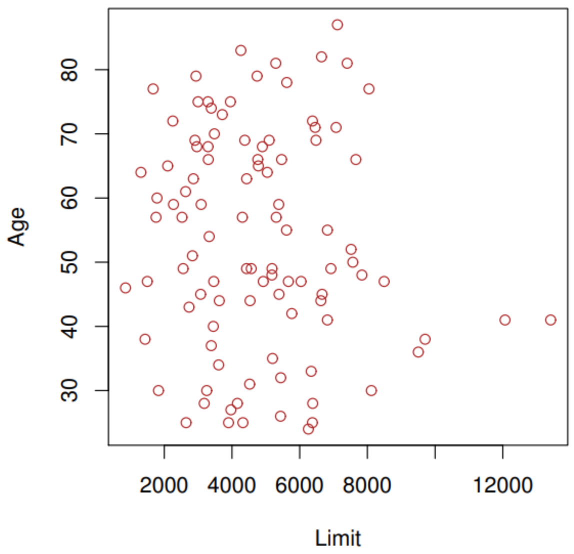
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

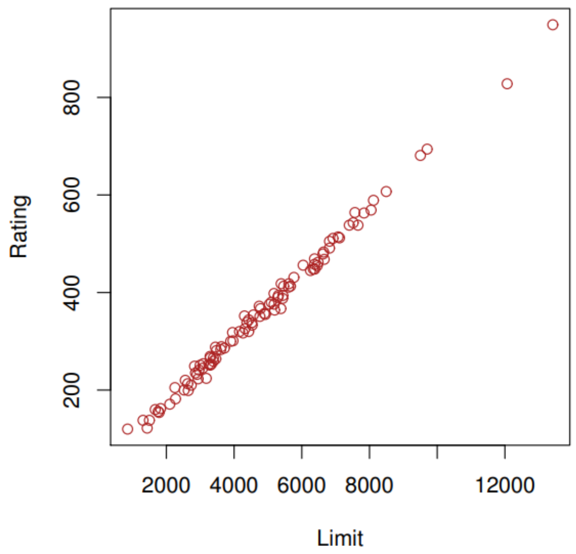
which is always between $\frac{1}{n}$ and 1

FIG ... *, while 20 is not.*
The ... *with observation*
41 re ... *ns of its* X_1 *value*
or its X_2 *value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation 41 has a high leverage and a high residual.*

Colinearity

Colinearity happens when two variables are highly correlated





Credit dataset example

- Limit and Age variables do not show a correlation
- Limit and Credit Rating variables instead show a marked correlation

$$y = w_0 + w_1 x_1 + \ldots + w_N x_N + \epsilon$$

$$y = w_0 + w_1 x_1 + \ldots + w_N x_N + \epsilon$$

Check for correlation and remove variables or use PCA …

If two variables are correlated, it can be difficult to estimate the relationship between each variable separately with the response

Improved Linear Regression

We can devise alternative procedures to least squares

- Improve prediction accuracy: if number of data is limited (or p is big) we might have “low bias” but too “high variance” (overfitting) and a poor prediction
- Improve model interpretability: irrelevant variables, beside impacting on accuracy, make models unnecessary complex and difficult to interpret

Several alternatives to remove unnecessary features (predictors)

- Subset Selection: selection of the input variables
- Shrinkage (or regularization): reduction of model variance
- Dimension reduction: projection on an input subspace

Shrinkage Methods: Ridge Regression

Ordinary Least Squares (OLS) minimizes

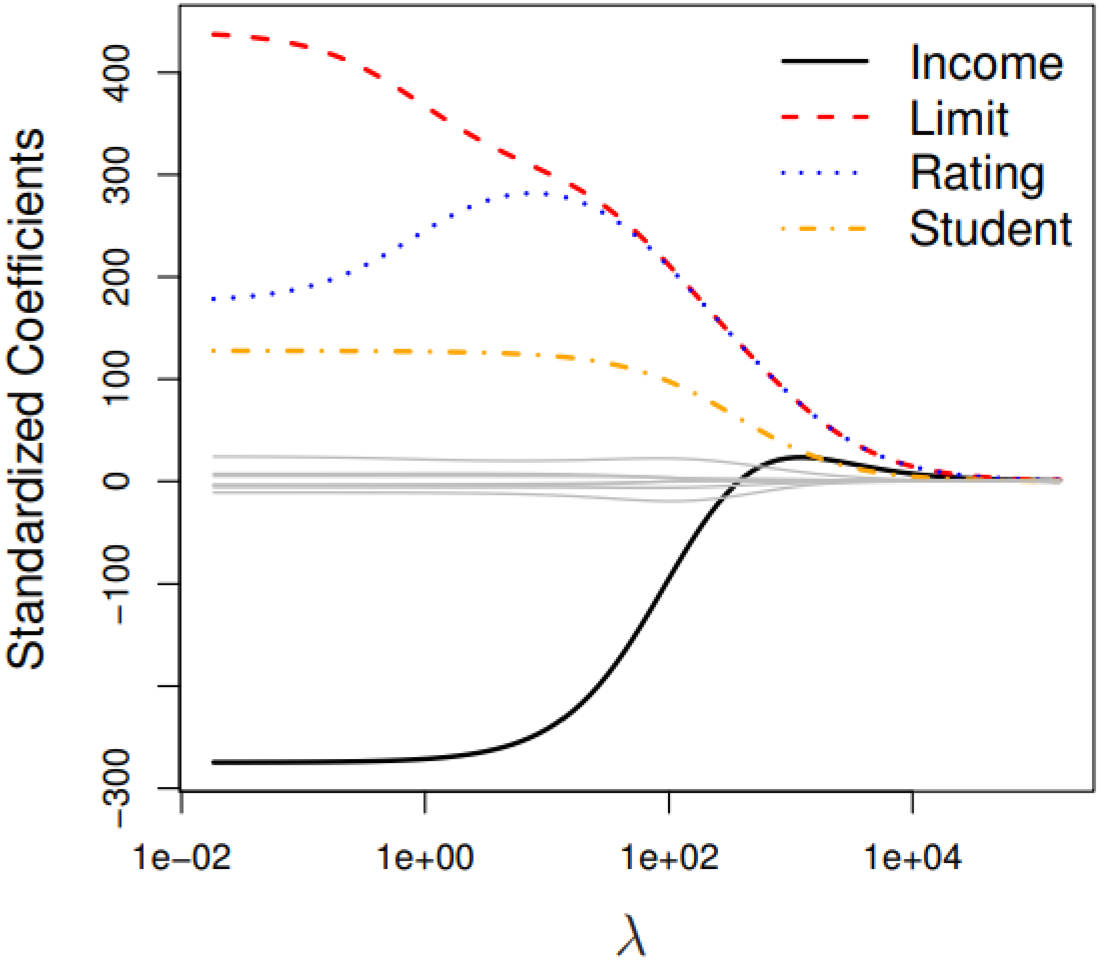
$$RSS = \sum_{i=1}^{N}\left(y_i - \left(w_0 + \sum_{m=1}^{M} w_m x_{im}\right)\right)^2$$

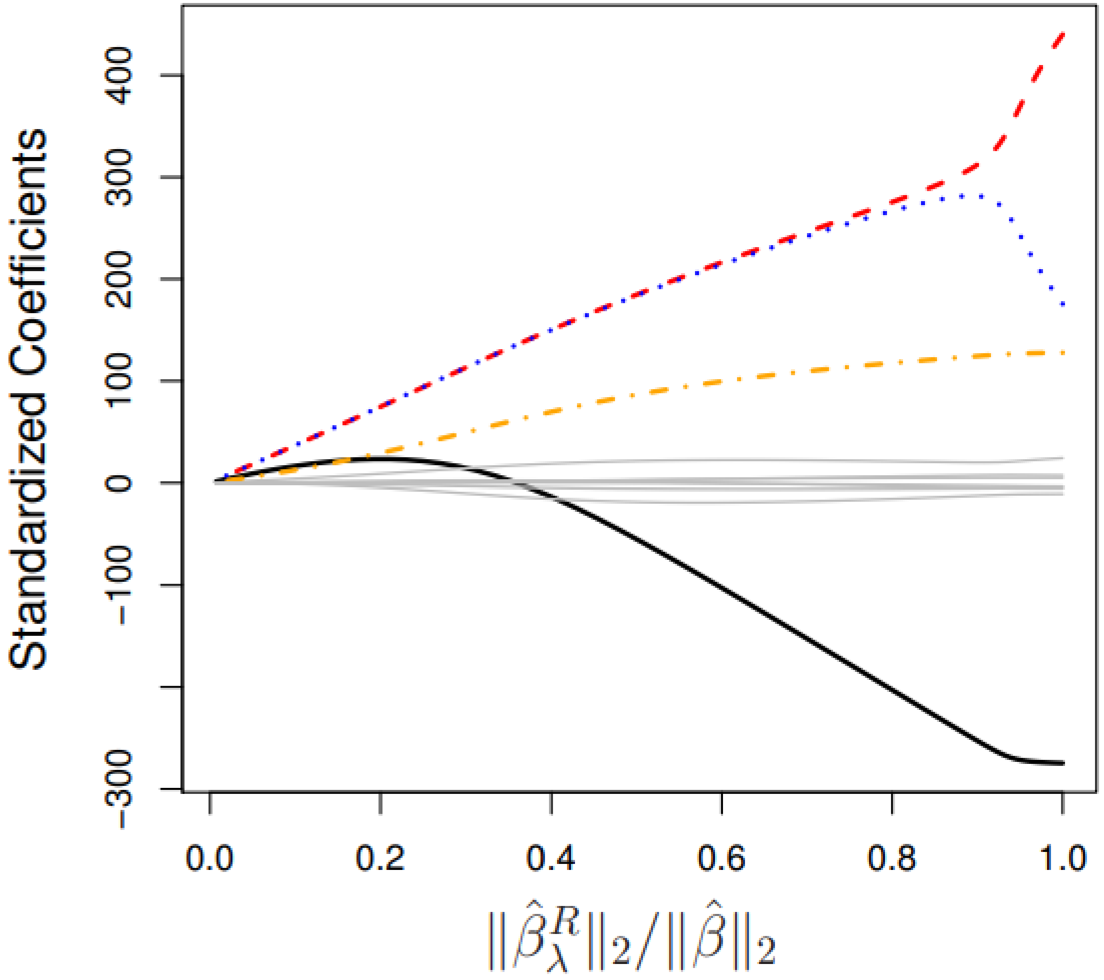
Ridge Regression minimizes a slightly different function

$$\sum_{i=1}^{N}\left(y_i - \left(w_0 + \sum_{m=1}^{M} w_m x_{im}\right)\right)^2 + \lambda \sum_{m=1}^{M} w_m^2 = RSS + \lambda \sum_{m=1}^{M} w_m^2$$

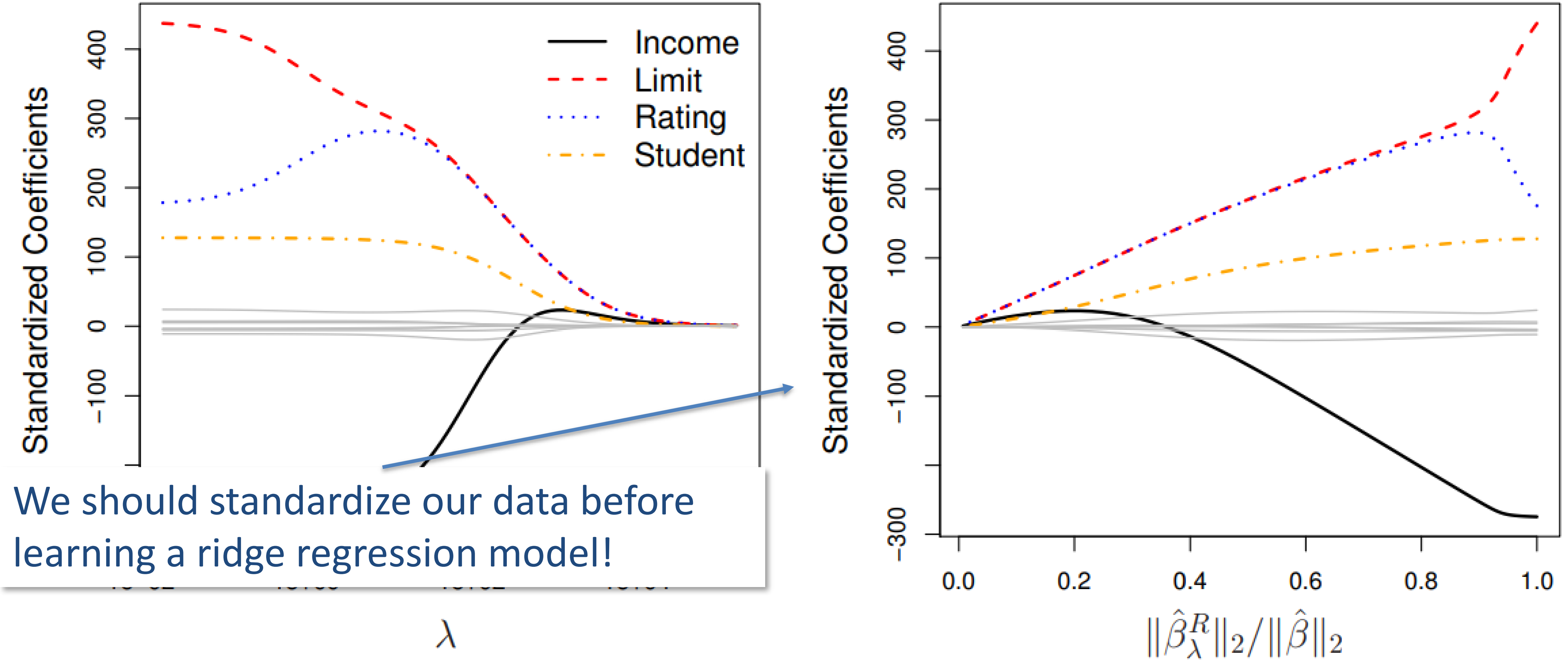
- $\lambda \geq 0$ is a tuning parameter to be estimated experimentally
- $\lambda \sum_{m=1}^{M} w_m^2$ is called shrinkage penalty
- as $\lambda \to \infty$ parameters shrink to zero

Shrinkage Methods: Ridge Regression





Shrinkage Methods: Ridge Regression



We should standardize our data before learning a ridge regression model!

Shrinkage Methods: The Lasso

It is an alternative to the ridge regression

$$\sum_{i=1}^{N}\left(y_i - \left(w_0 + \sum_{m=1}^{M} w_m x_{im}\right)\right)^2 + \lambda \sum_{m=1}^{M} |w_m| = RSS + \lambda \sum_{m=1}^{M} |w_m|$$

- $\lambda \geq 0$ is a tuning parameter to be estimated experimentally
- $\lambda \sum_{m=1}^{M} |w_m|$ is called lasso penalty
- as $\lambda \to \infty$ parameters shrink to zero

It forces some of the coefficients to be exactly zero, it performs variable selection …

Shrinkage Methods: Lasso