Introduction to Regression

V. Ashley Villar
PSU

PSU Astrostats Summer School

What is regression?

Regression aims to construct a model that relates the response variable Y to the predictor variables $X_1, X_2, \ldots, X_p$.

Example 1: Predicting redshift from photometry

Y = redshift

X_1, X_2, X_3, X_4, X_5 = magnitudes in u, g, r, i, z bands

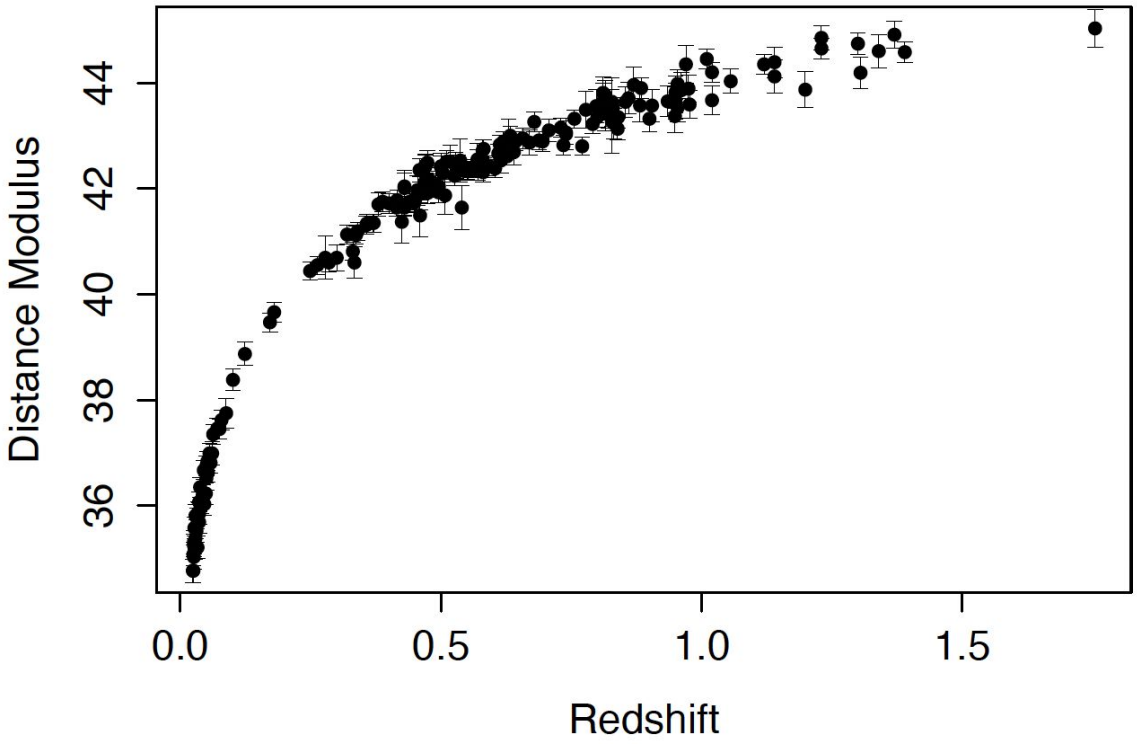
Example 2: Modelling relationship between distance modulus and redshift for Type Ia Supernovae

Y = distance modulus

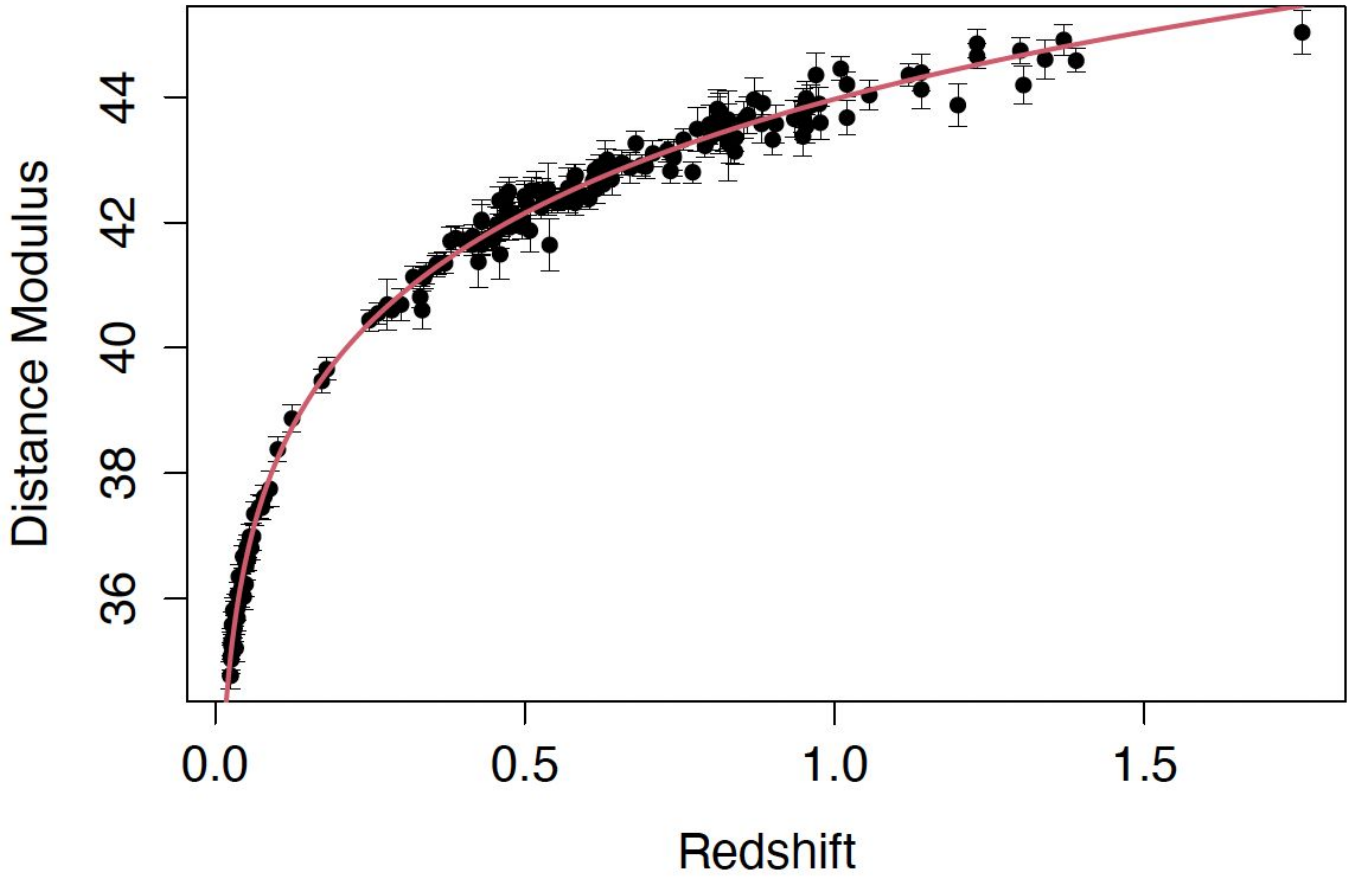
X_1 = redshift

The ΛCDM model proposes a simple relationship between redshift and the distance modulus which depends on two cosmological parameters: Ω_m and H_0.

Learning the relationship between redshift and DM therefore constrains these parameters.



From Riess et al 2007



The case where $H_0 = 72.76$ and $\Omega_m = 0.34$.

The two above examples contrast two of the typical motivations for using regression:

In **Example 1** we want to make **predictions** for new/future response values (redshifts) given the ugriz magnitudes of an object. Very important problem for LSST and other photometric surveys.

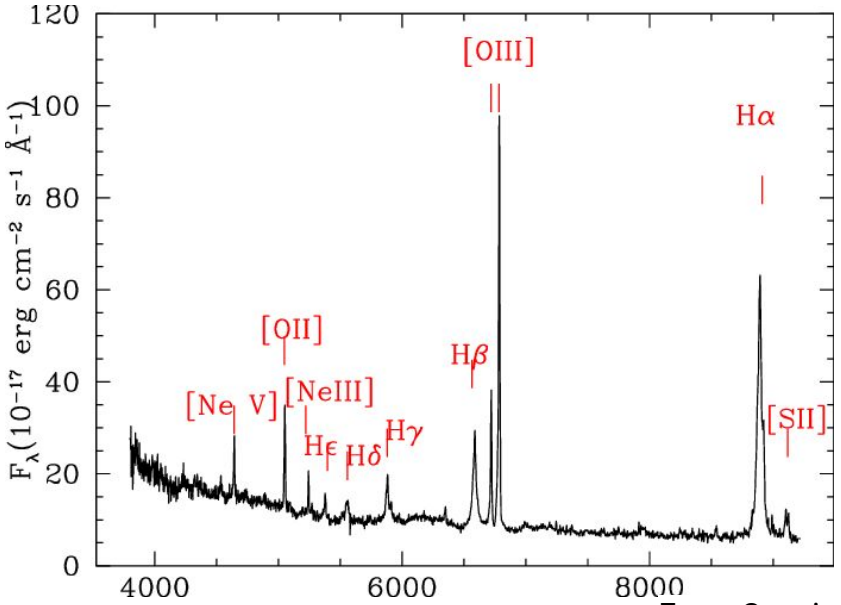
In **Example 2** we want to learn something about the **relationship** between redshift and distance modulus for Type Ia supernovae, since different physical theories make different predictions for its shape.

Which of the two scenarios we are in will dictate our level of concern for model fit and violations of other theoretical assumptions.

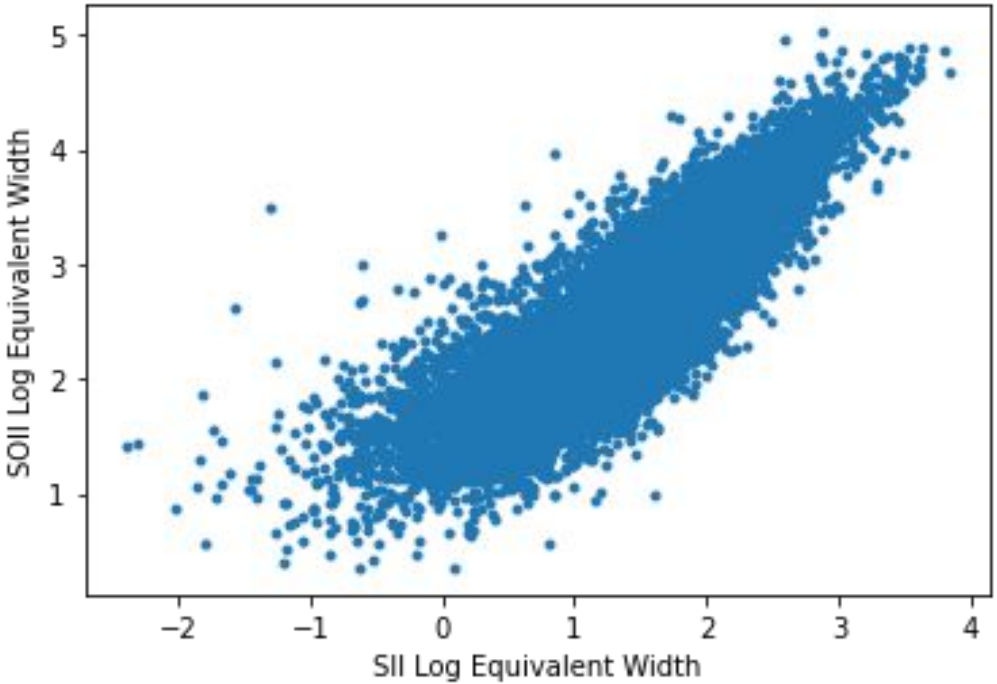
Modelling Galaxy Emission Lines

We will be working with SDSS emission line galaxy data, understanding the relationships between various emission line widths.

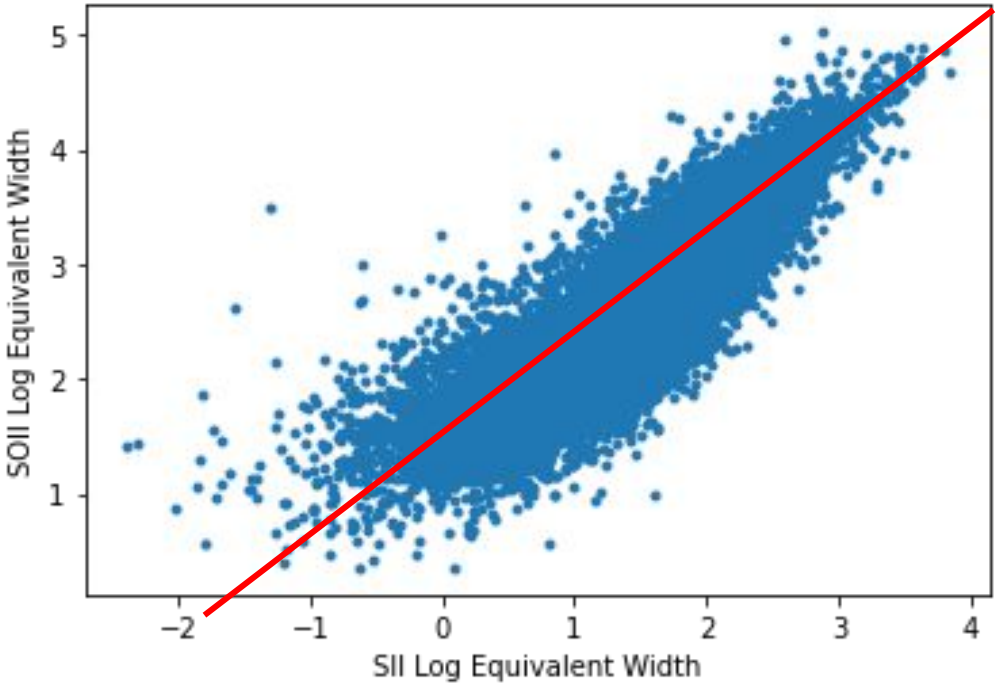
Colab link.



From Caccianiga et al.


SOII Log Equivalent Width
SII Log Equivalent Width
5
4
3
2
1
−2
−1
0
1
2
3
4

Is this a good model?



Our assessment of this line would depend, in part, on our motivation for fitting the regression model.

If our objective was solely to make good predictions of the OII log equivalent width using the SII line, then it is not doing a bad job.

On the other hand, we would be very reluctant to claim that the linear model is revealing some “truth” regarding the relationships between these emission lines.

In what follows, we will consider a range of increasingly-complex models.

The starting point will be linear regression.

Regression using MOO

1. **Model** to fit the data (e.g. physics).
2. **Objective Function** (or ‘loss/cost function’) which is a metric that you will choose to quantify how well the model fits the data (e.g. chi-squared).
3. **Optimization Method** which you will use to find the best model (e.g. gradient descent).

Linear Regression

Linear regression models predict the response variable as a linear combination of the predictors, plus random scatter

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

If we want to stress that there is a sample of n observations being used, we can write the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

The available sample used for fitting the model is often referred to as the **training sample**.

Linear Regression can use transforms of the original features

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

A note on that ε....

ε is often called an **irreducible error** – it is ever present and cannot be ignored! We do **not** seek to fit models that eliminate truly random scatter, as this would lead to **overfitting**.

A note on that ε....

ε_i is often called an **irreducible error** – it is ever present and cannot be ignored! We do **not** seek to fit models that eliminate truly random scatter, as this would lead to **overfitting**

It is almost always assumed that $E(\varepsilon_i) = 0$. If $Var(\varepsilon_i) = \sigma^2$ for all i, then the errors are **homoskedastic**. In contrast, **heteroskedastic** errors describing changing error properties. Often irreducible errors are assumed to be **uncorrelated** and “independent and identically distributed” (iid) Gaussian.

Finally, let's fit a line (with MOO)!

Model ("simple" linear regression):

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

Where we assume that ε_i are iid Gaussian variables with variance σ^2

So what are the **three** free parameters of our model?

Finally, let's fit a line (with MOO)!

Model ("simple" linear regression):

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

Where we assume that ε_i are iid Gaussian variables with variance σ^2

So what are the **three** free parameters of our model?

β_0, β_1, σ

What is our Objective Function?

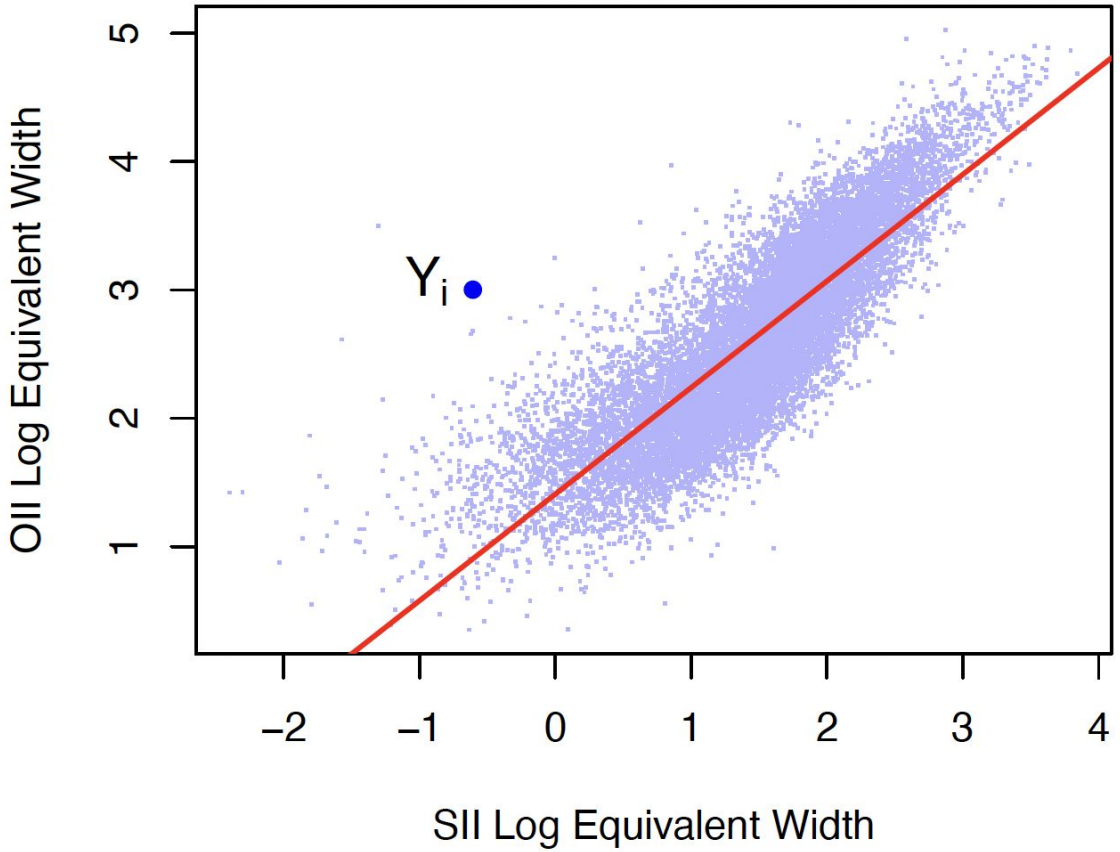
The standard approach to estimating β_0, β_1 is to use **least squares** regression.

The associated objective function is:

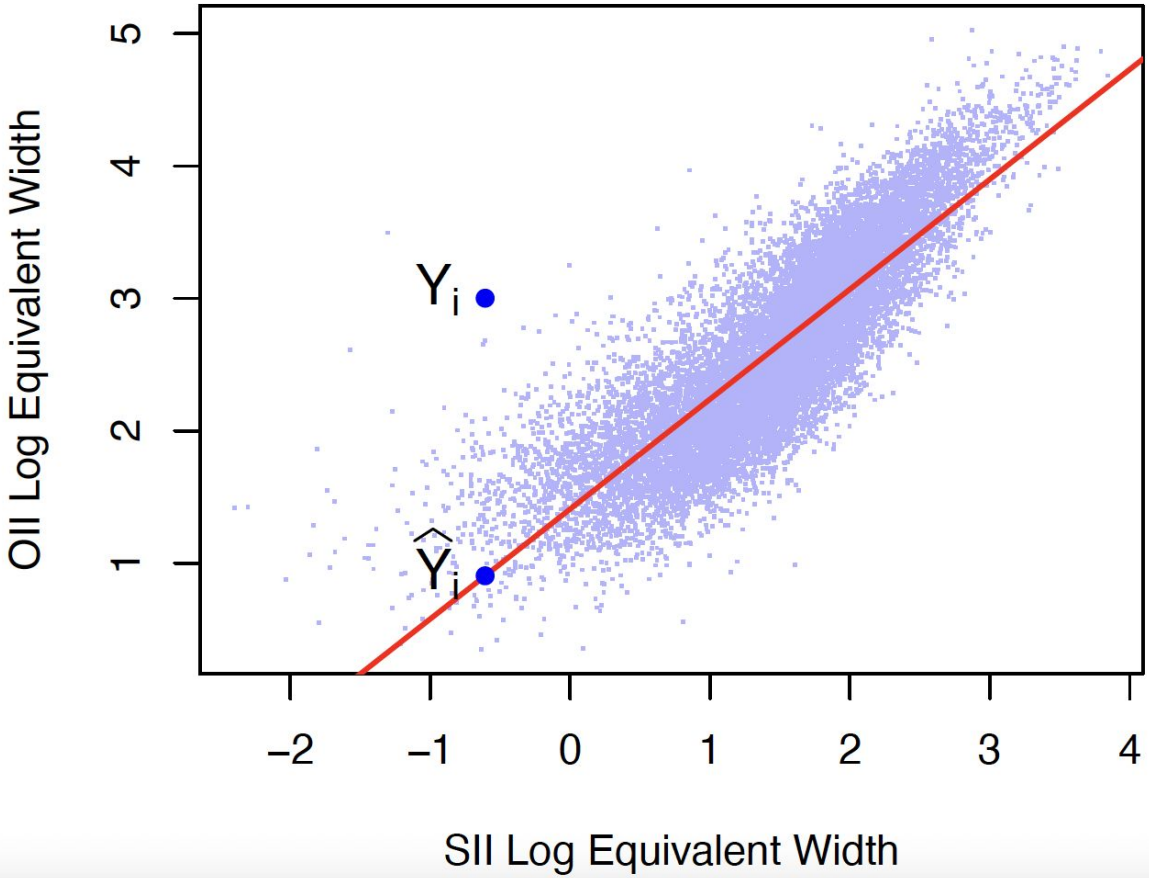
$$\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Will this objective function be **minimized** or **maximized**?

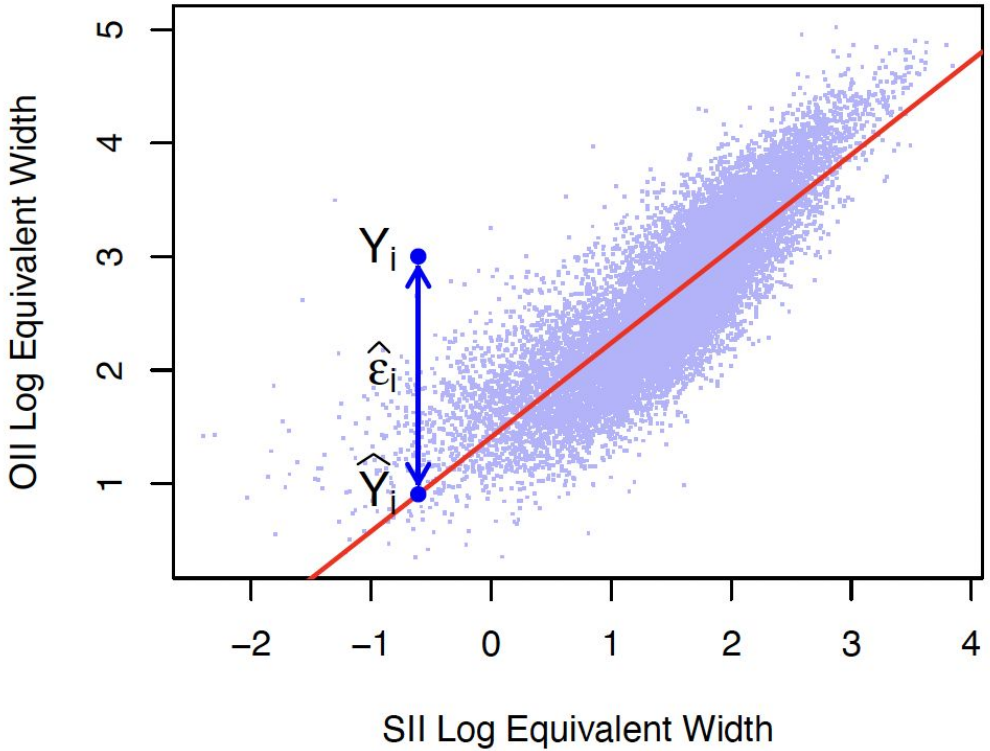
What does our objective function look like geometrically?



What does our objective function look like geometrically?



The difference is the **residual**, $\hat{\epsilon}_i = Y_i - \widehat{Y}_i$:



The least squares regression line minimizes $\sum_i \hat{\epsilon}_i^2$.

How do we Optimize this Objective Function?

There are many valid ways of optimizing the objective function. Ideally, we want one which is computationally **efficient** and **accurate**. We could:

1. Guess and check randomly, or on a grid
2. Systematically guess (following some algorithm)
3. Analytically solve

How do we Optimize this Objective Function?

We can actually find a global minimum analytically! (An exercise left for the student ;))

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

Correlation Coefficients

The **Pearson correlation coefficient** provides a measure of linear association between two variables.

The (sample) **covariance** between two variables is:

$$\mathrm{Cov}(X,Y) = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \Big/ (n-1)$$

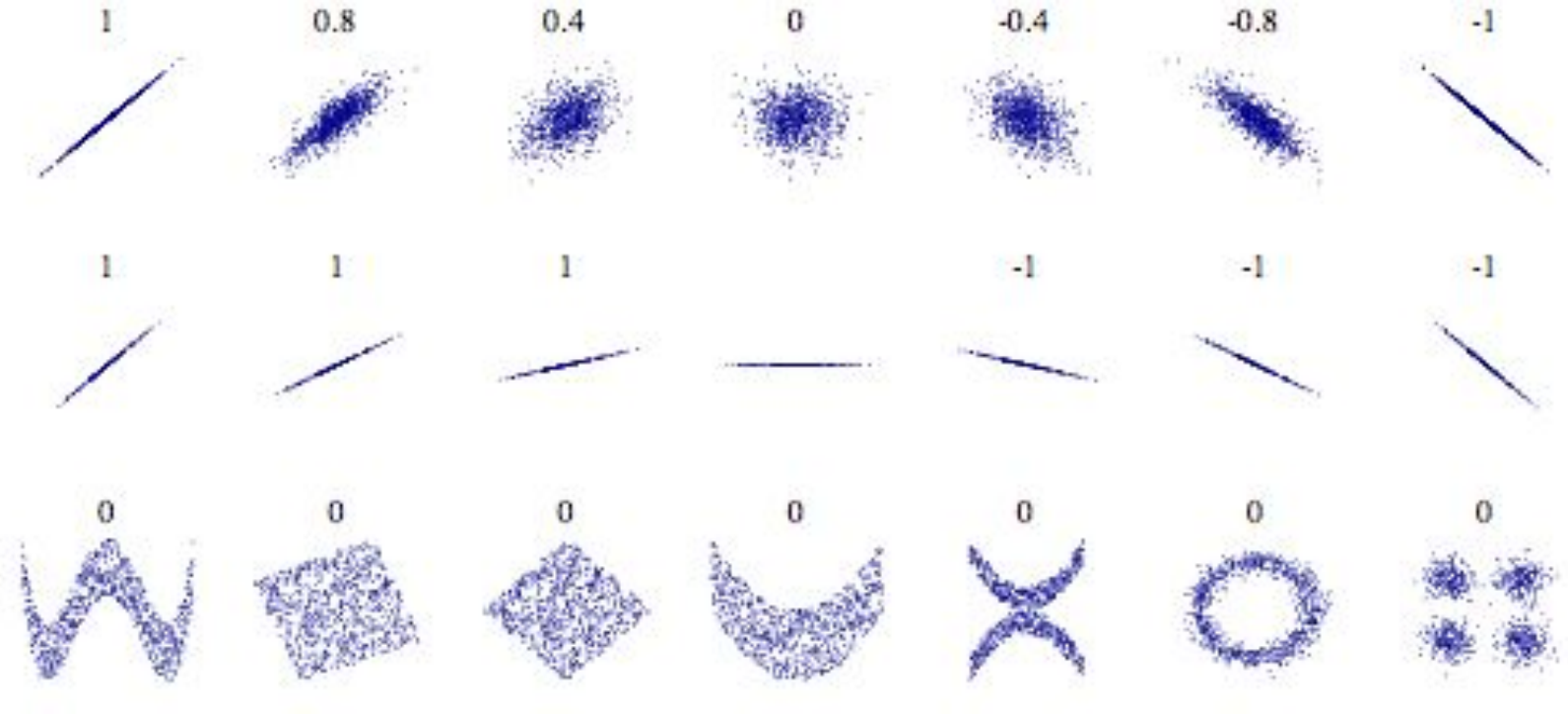
And the correlation coefficient is

$$r_{xy} = \mathrm{Cov}(X,Y) \Big/ s_x s_y$$

Where s denotes the **sample standard deviation**

Note that -1 ≤ r ≤ 1, and 1 represents a perfect linear relationship.

Here are some examples of how correlation coefficient can vary with the data. Any interesting observations?



The Pearson correlation looks a lot like our simple regression!

$$\widehat{\beta}_1 = \frac{\mathrm{Cov}(X,Y)}{s_x^2} = r_{xy}\left(\frac{s_y}{s_x}\right).$$

It's worth noting that if X and Y have been rescaled to have mean of zero and standard deviation of one, then $\widehat{\beta}_1 = r_{xy}$, and $\widehat{\beta}_0 = 0$.

Linear regression in python...

See Colab.

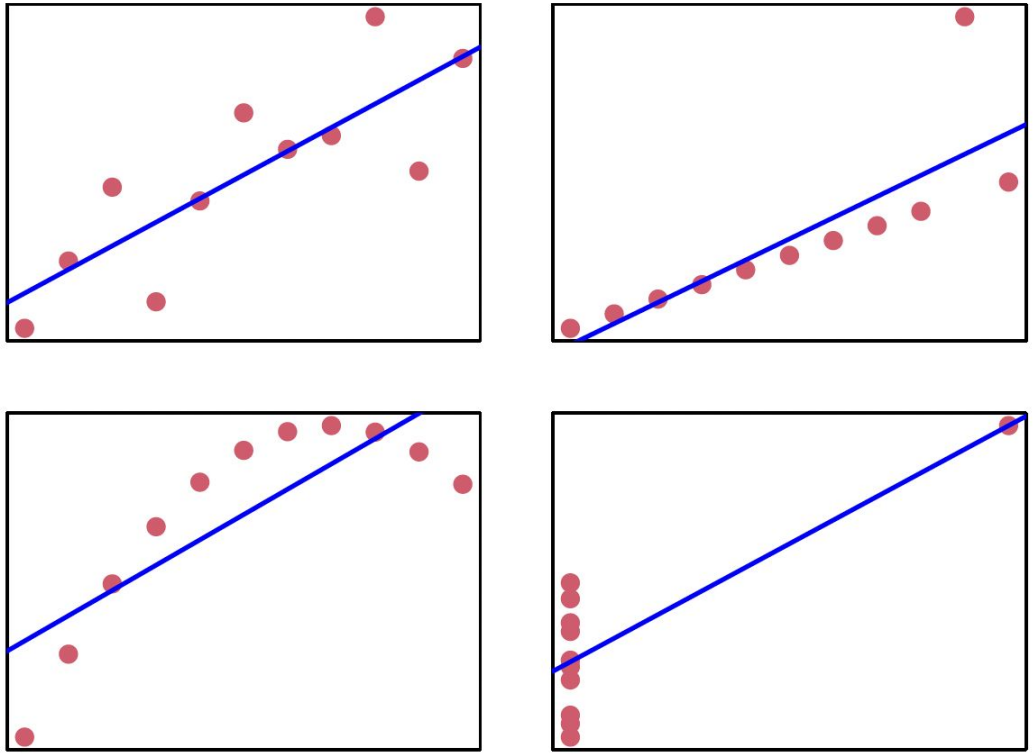
Coefficient of Determination

This can be interpreted as “the proportion of the variance in the response explained by the model”.

Unfortunately, R^2 is often used as a proxy for the quality of a model fit, but it is **not** designed to assess this.

$$R^2 = 1 - \frac{\sum_i \left(Y_i - \widehat{Y_i}\right)^2}{\sum_i \left(Y_i - \overline{Y}\right)^2}$$

Consider the four scatter plots below, Anscombe's Quartet (1973).



In all four cases, the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$, r_{xy}, R^2, and so forth, are equal.

Residuals as Diagnostic Tools

Clearly we can't use our correlation coefficients to judge the quality of our model fits.

However, one way we can judge model fit is to examine the residuals. If our model is a reasonable fit, the remaining scatter (the residuals) should be a reasonable estimate of ε, the irreducible scatter:

$$
\begin{aligned}
\widehat{\epsilon}_i &= Y_i - \widehat{Y}_i \\
&= Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \cdots + \widehat{\beta}_p X_{pi}) \\
&\approx Y_i - (\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}) \\
&= \epsilon_i
\end{aligned}
$$

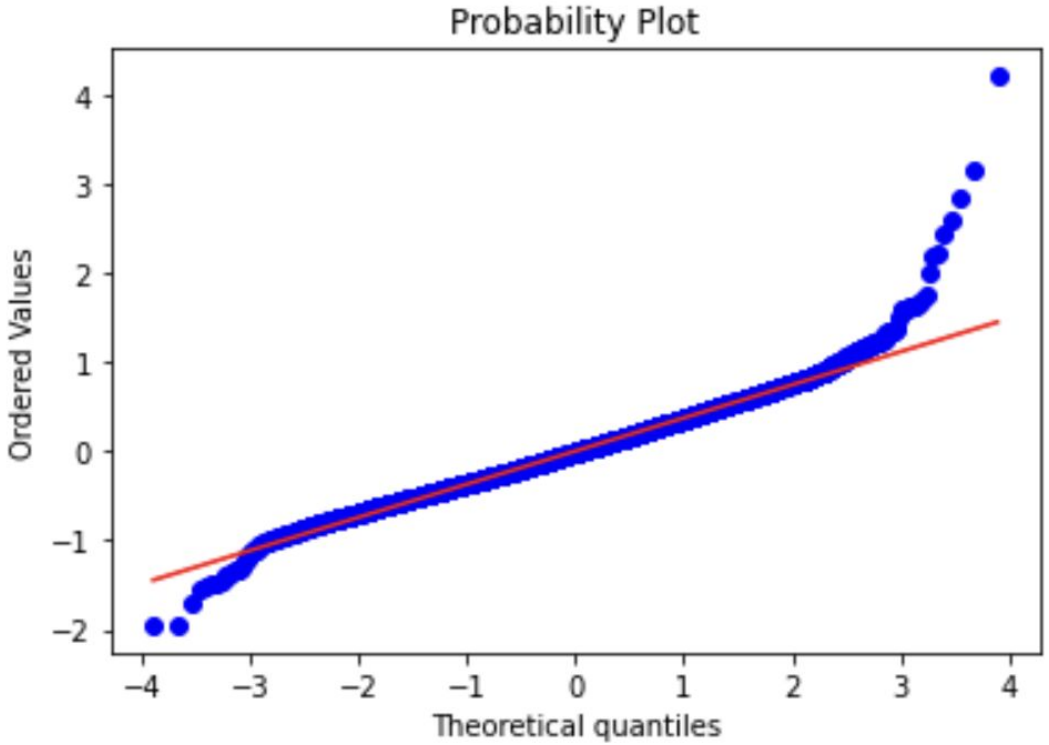
Residuals as Diagnostic Tools

Therefore, if our model is a **good** fit, our residuals should be random scatter, lack any relationship with predictors, etc (be iid and Gaussian).

So let’s plot them!

Looking at the normal probability plot for residuals

While normality of the residuals is not necessarily true, probability plots still help to highlight extreme residuals



An aside on logistic regression

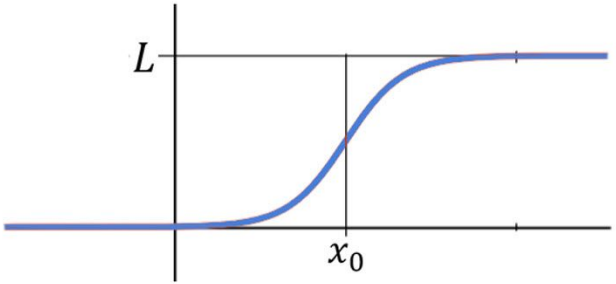
What if I want to **model** a discrete outcome, such as one which predicts if an image is of a galaxy or star?

Logistic regression can model these discrete outputs

Logistic Function

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

x_0 = x value of midpoint
L = maximum value
k = growth rate



An example logistic regression Model

$$P(Y=1|X)=\frac{1}{1+\exp\left(w_0+\sum_{i=1}^{n} w_i X_i\right)}$$

and

$$P(Y=0|X)=\frac{\exp\left(w_0+\sum_{i=1}^{n} w_i X_i\right)}{1+\exp\left(w_0+\sum_{i=1}^{n} w_i X_i\right)}$$

From Kumar et al. 2016

What is an appropriate objective function for classification?

$$H(P^* \mid P) = -\sum_i \underbrace{P^*(i)}_{\text{TRUE CLASS DISTIRBUTION}} \log \underbrace{P(i)}_{\text{PREDICTED CLASS DISTIRBUTION}}$$

From Adian Liusie

And what Optimization method should I use?

1. Analytical formula?
    a. Not closed form!
2. Random guess and check?
    a. Not ridiculous! Harder to get an understanding of uncertainty
3. Clever guess and check (via an algorithm)
    a. Gradient descent, MCMC, etc.

Model Selection

The process of **model selection** involves deciding which of the available predictors should be utilized in the model.

Model Selection

The process of **model selection** involves deciding which of the available predictors should be utilized in the model.

Including too many predictors leads to **overfitting**, a situation in which a model fits better to the training sample than it would to external observations not used in the fitting. This is a big problem.

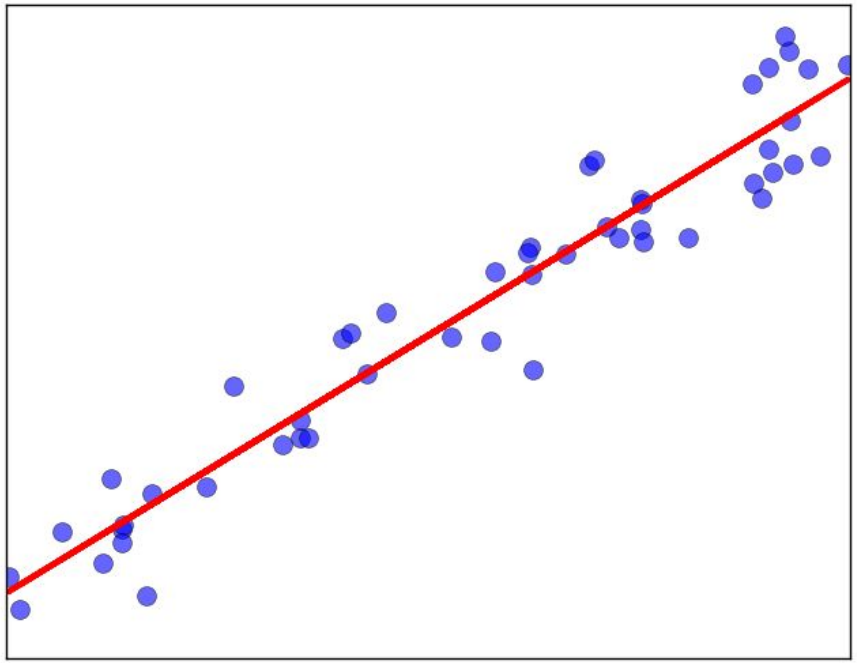
Model Selection

The process of **model selection** involves deciding which of the available predictors should be utilized in the model.

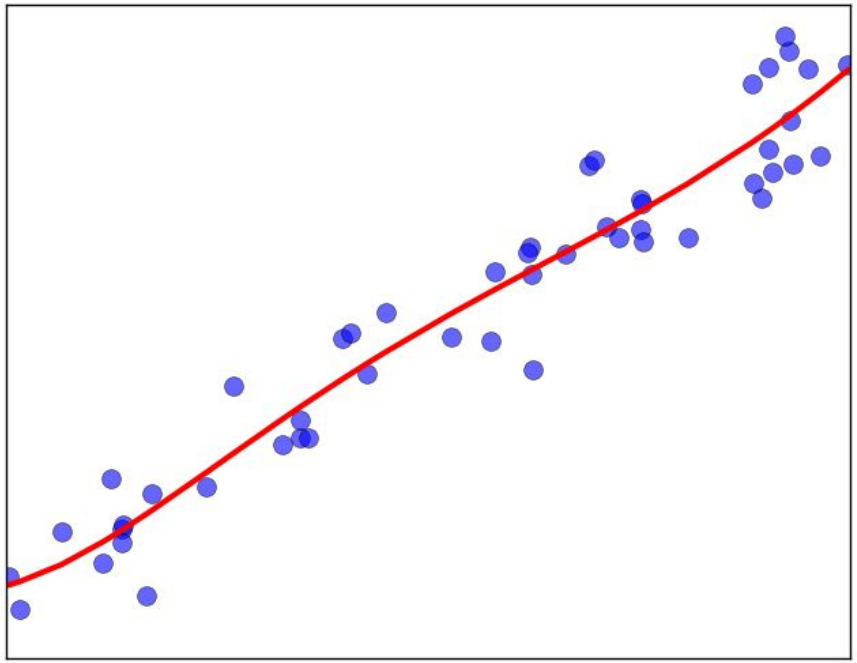
Including too many predictors leads to **overfitting**, a situation in which a model fits better to the training sample than it would to external observations not used in the fitting. This is a big problem.

Important: Increasing the number of predictors will necessarily decrease the sum of the squared residuals, and necessarily increase the value of R^2. These **metrics are not useful** for model selection.

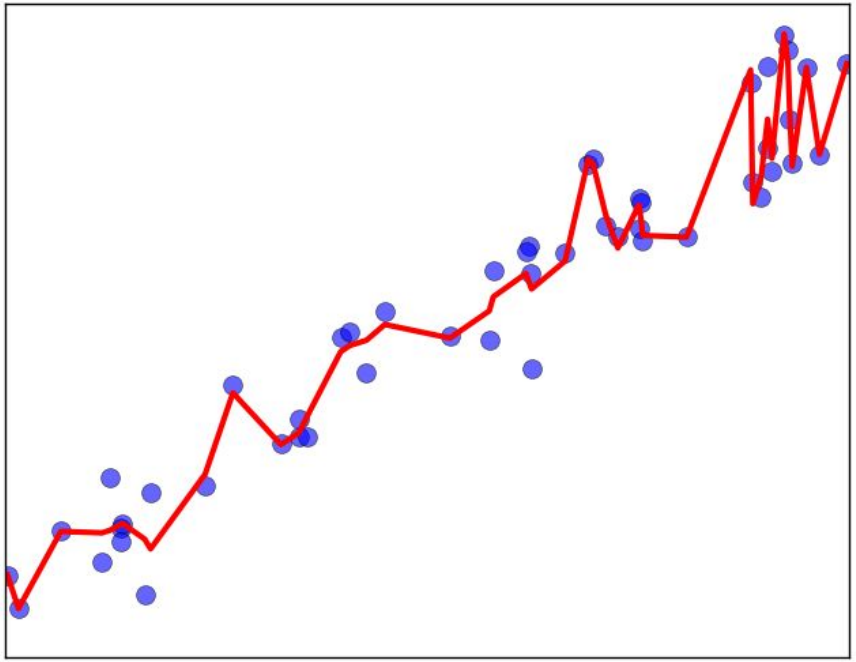
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... = \vec{\beta}\vec{X}$$



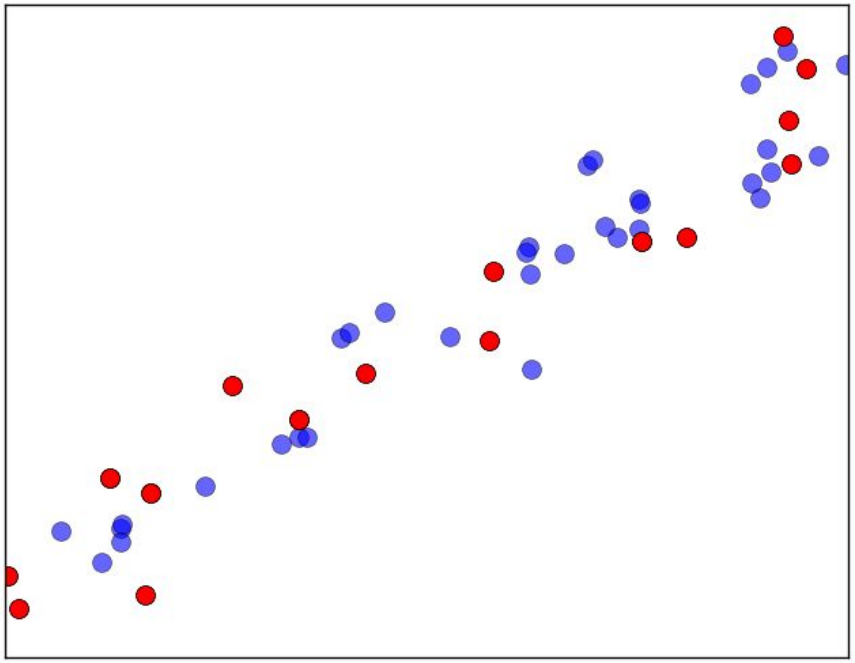
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... = \vec{\beta}\vec{X}$$



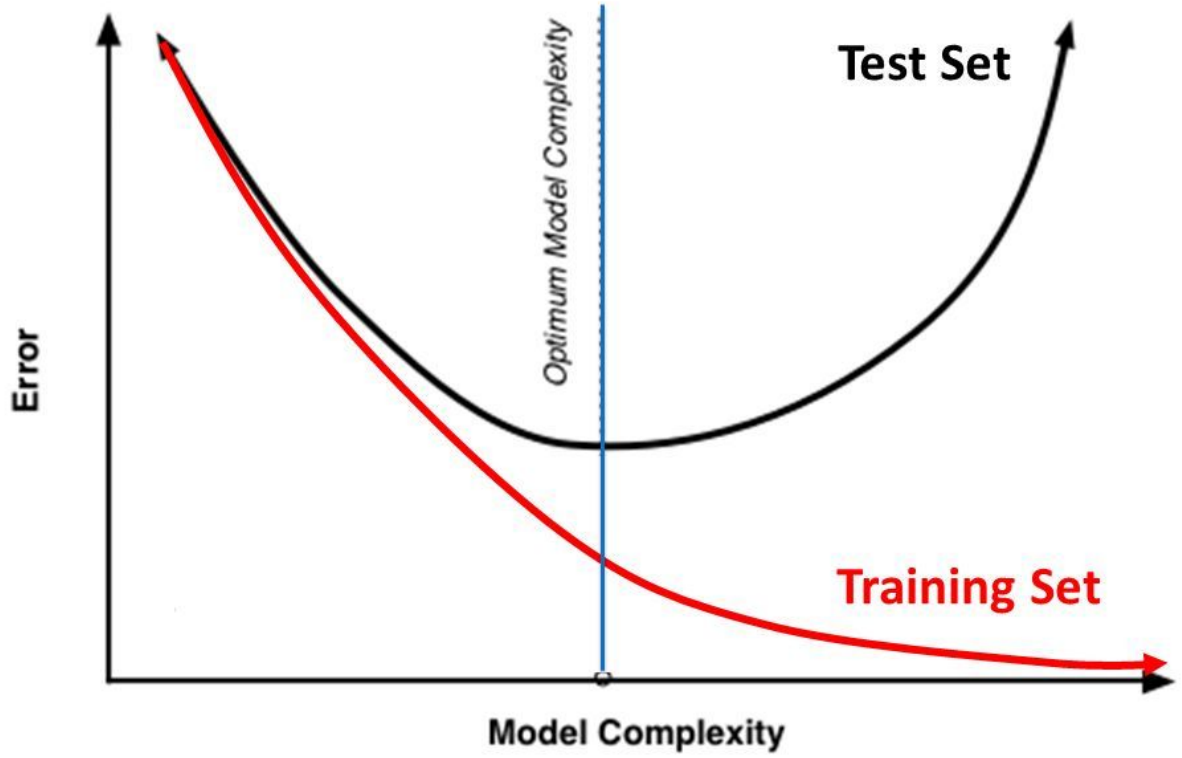
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... = \vec{\beta}\vec{X}$$



Split data into training (,validation) and test sets



Training Vs. Test Set Error



Heath 2015

Leave-one-out cross-validation

LOOCV is an important tool for model selection.

For each observation i, imagine fitting a model which “leaves out” that observation. The model is fit on this remaining training set.

The fit model is then used to predict the response for observation i. We can track how well our model predicts the observation which is left out. We can accumulate these errors into a score:

$$\text{LOOCV Score} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_{(-i)}\right)^2$$

Confidence and Prediction Intervals

As described above, these models are ultimately used for inference, i.e., drawing conclusions and making predictions. Any inference should be accompanied by an **uncertainty measure**. Let's contrast two inference objectives:

1. When **making predictions** of the response for "new" X values, then uncertainty is best described by a **prediction interval.** This interval gives a range of values that we can claim, with some confidence, the true value of Y falls into.

2. When **characterizing the relationship** between the response and the X values, a **confidence interval** on the regression function is often of interest.

Confidence and Prediction Intervals

1. When **making predictions** of the response for “new” X values, then uncertainty is best described by a **prediction interval.** This interval gives a range of values that we can claim, with some confidence, the true value of Y falls into.

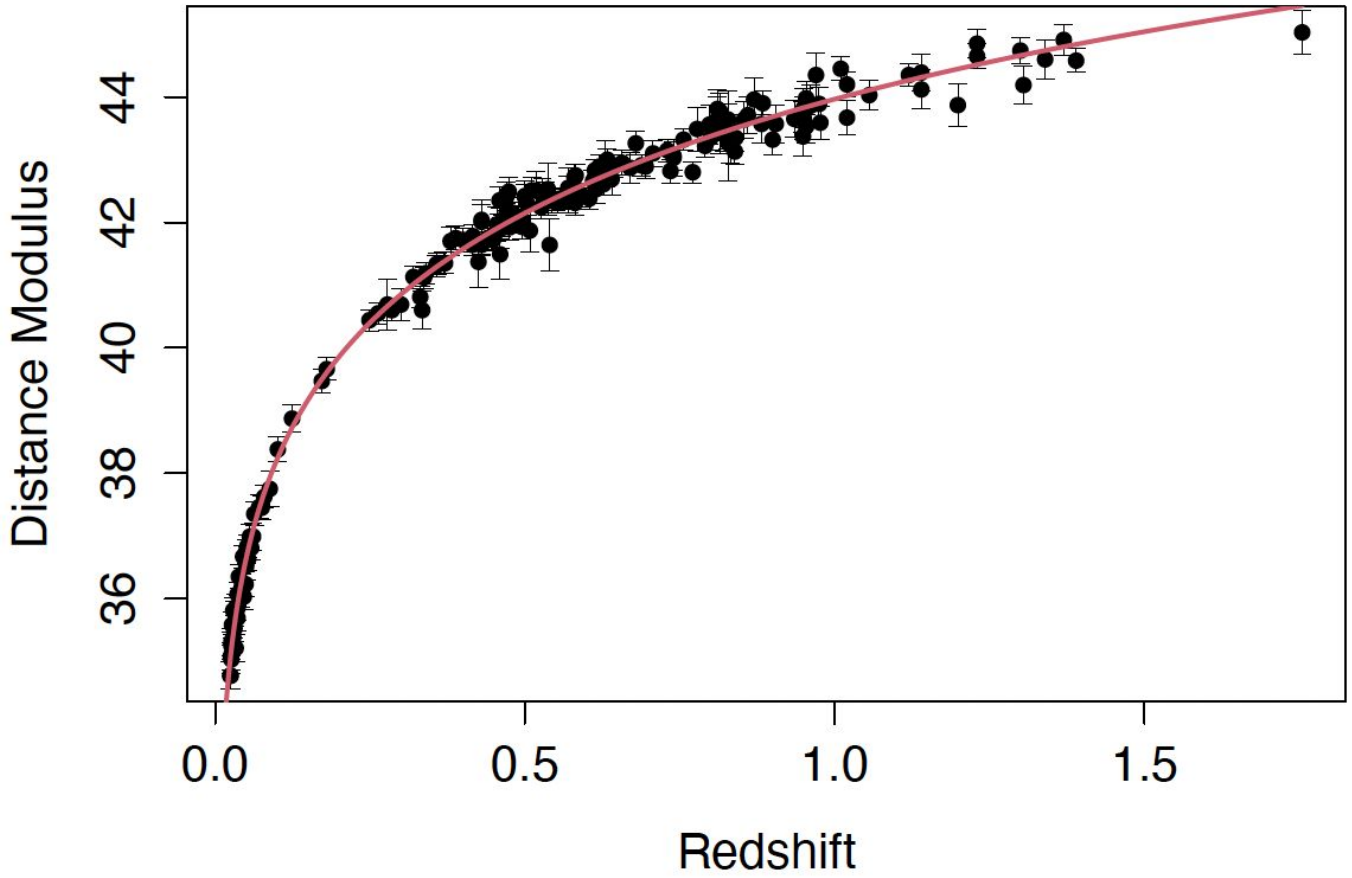
2. When **characterizing the relationship** between the response and the X values, a **confidence interval** on the regression function is often of interest.

The **prediction** interval tends to be wider, as it accounts for the intrinsic scatter. For large n, 95% prediction intervals have a half-width which is approximately 2σ.

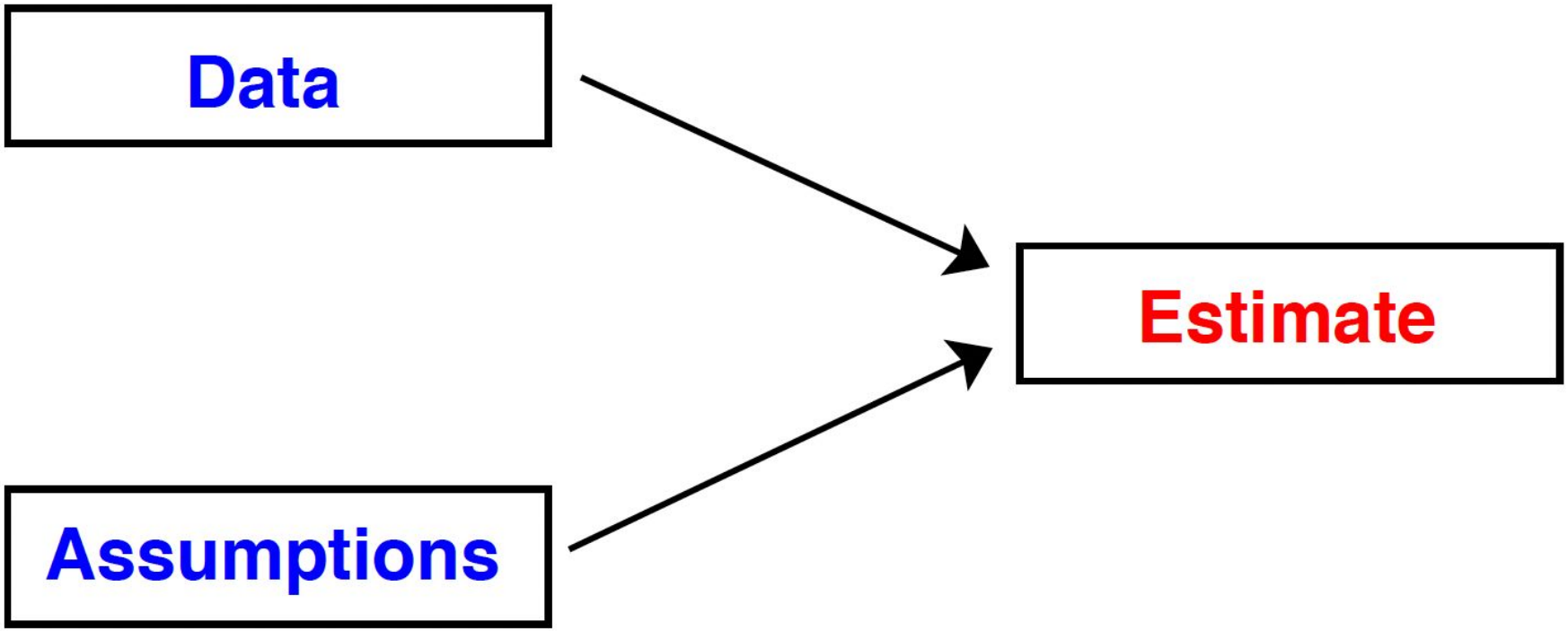
Nonparametric Regression

The linear models considered thus far are **parametric** models, meaning that there is a fixed form for the relationship between the response and predictor(s), the only unknowns are real-valued parameters.

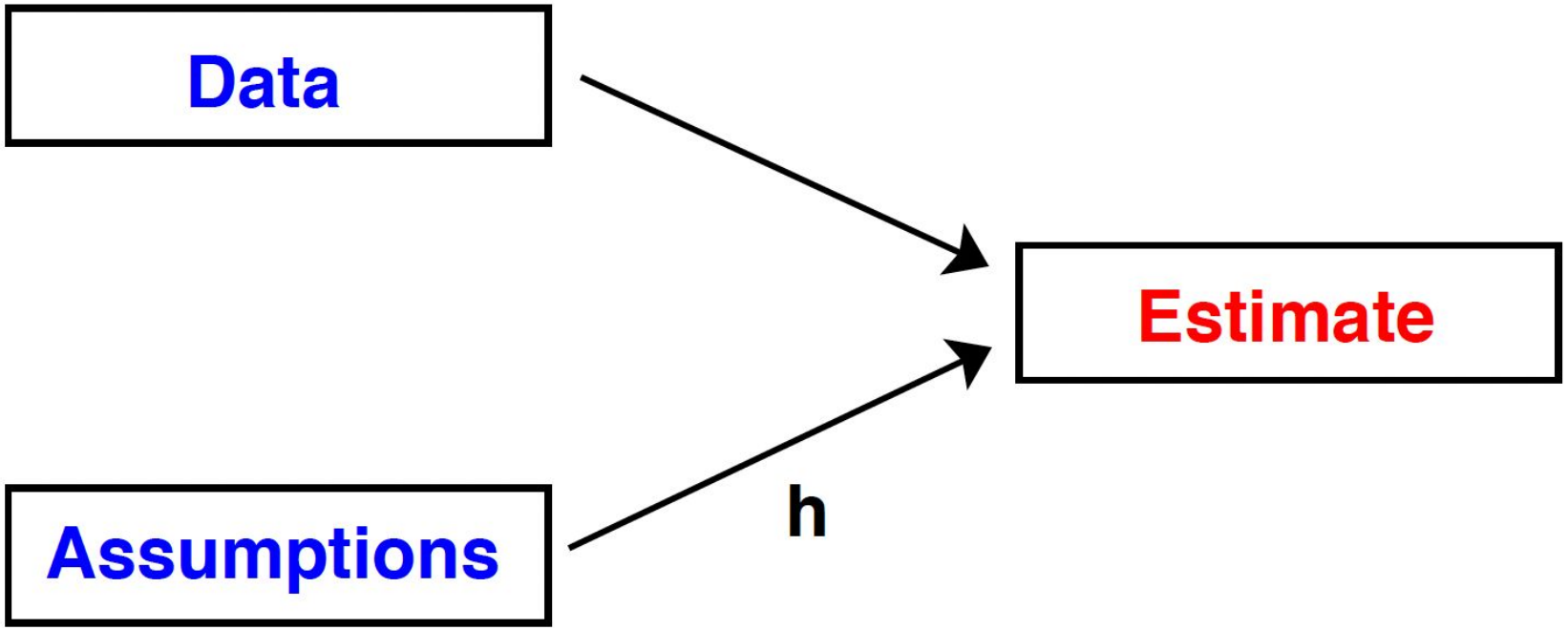
We saw at the start of the lecture that parametric models come in nonlinear forms as well (e.g., the redshift vs. distance modulus).



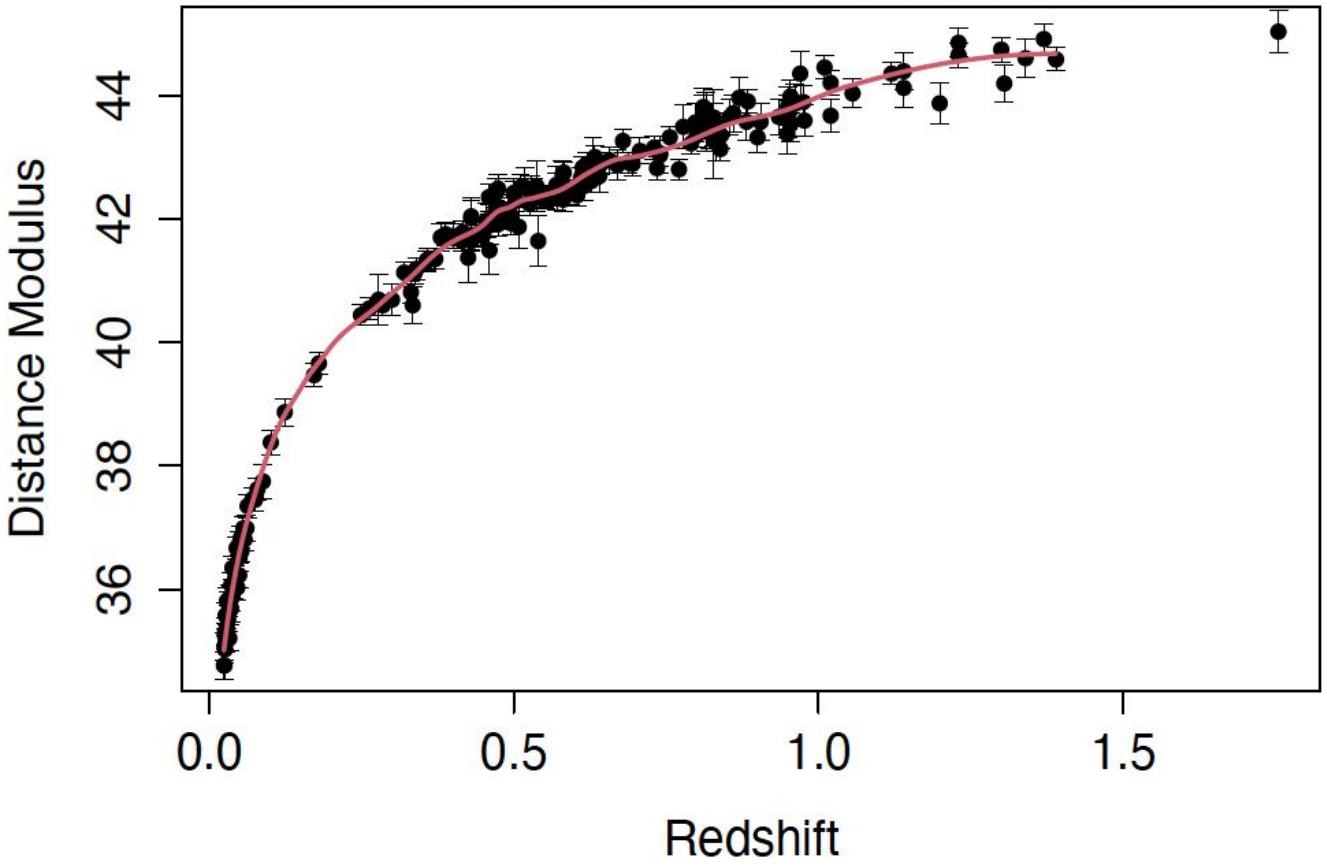
The case where $H_0 = 72.76$ and $\Omega_m = 0.34$.



In the **parametric case**, the contribution to the estimate from the assumptions is fixed.



In the **nonparametric case**, the influence of the assumptions is controlled by a **smoothing parameter** h. The value of h can be chosen smaller with larger sample sizes.



A nonparametric estimate of the relationship.

Local Linear Regression

The figure on the previous slide shows a **nonparametric regression** or **nonparametric smooth** of the relationship.

There are many variants of nonparametric regression, but here we focus on **local linear regression**, a version of **local polynomial regression**.

Local Linear Regression

The figure on the previous slide shows a **nonparametric regression** or **nonparametric smooth** of the relationship.

There are many variants of nonparametric regression, but here we focus on **local linear regression**, a version of **local polynomial regression**.

Briefly stated, this procedure works by fitting a sequence of linear models: Each is fit not to the entire data set, but to only data within a **neighborhood** of a target point. The size of this neighborhood is the smoothing parameter: Large neighborhoods yield a large degree of smoothing, while small neighborhoods result in minimal smoothing.

Local Linear Regression: Model

Our model here is that we observe (x_i, Y_i) for i=1,2,...,n and that:

$$Y_i = f(x_i) + \epsilon_i$$

Where ε are iid with mean zero and variance σ^2

Local Linear Regression: Model

Our model here is that we observe (x_i, Y_i) for i=1,2,...,n and that:

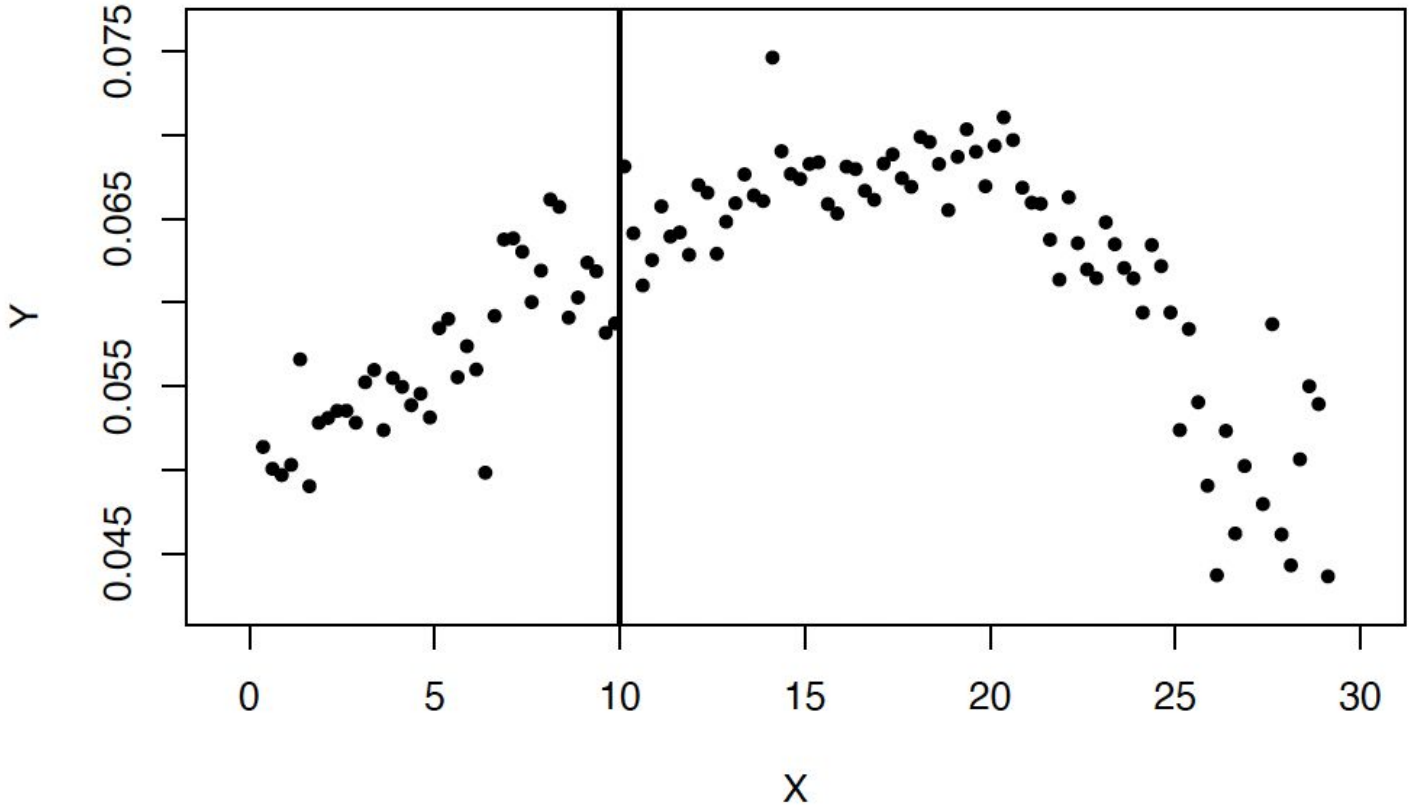
$$Y_i = f(x_i) + \epsilon_i$$

Where ε are iid with mean zero and variance σ^2

In order to construct the local linear regression estimate of f(x), we will consider a sequence of steps for each fixed x_0 at which $f(x_0)$ will be estimated.

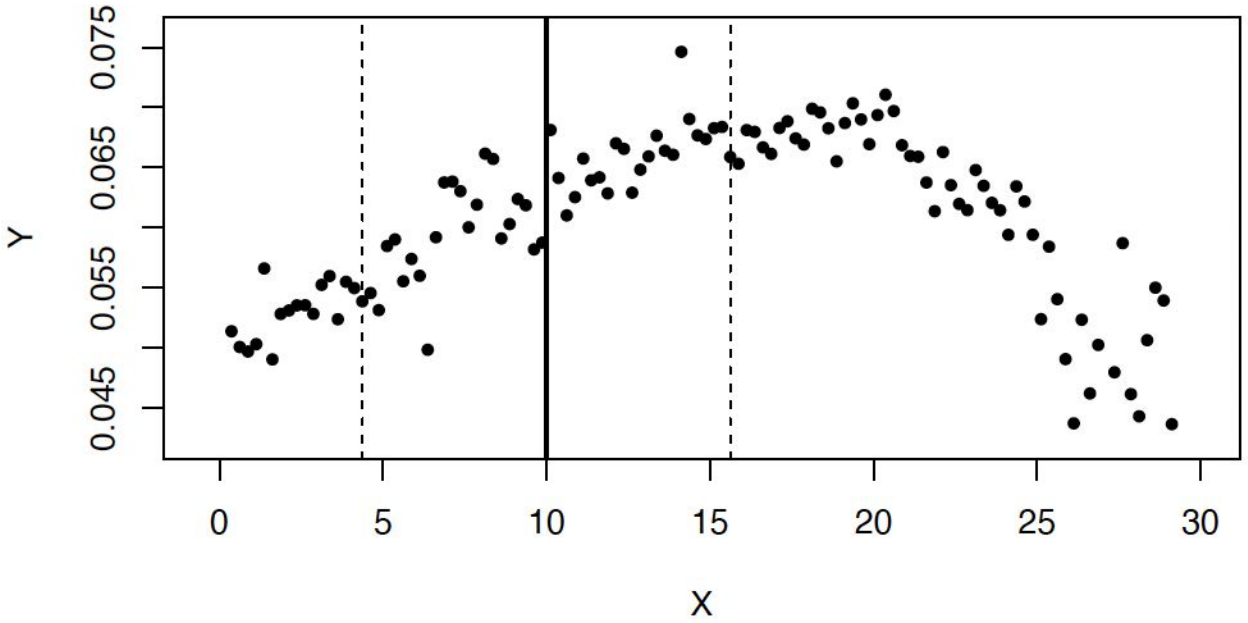
Step One: Fix the target point x_0.

Our objective is to estimate the regression function at x_0.



Step Two: Create the neighborhood around x_0.

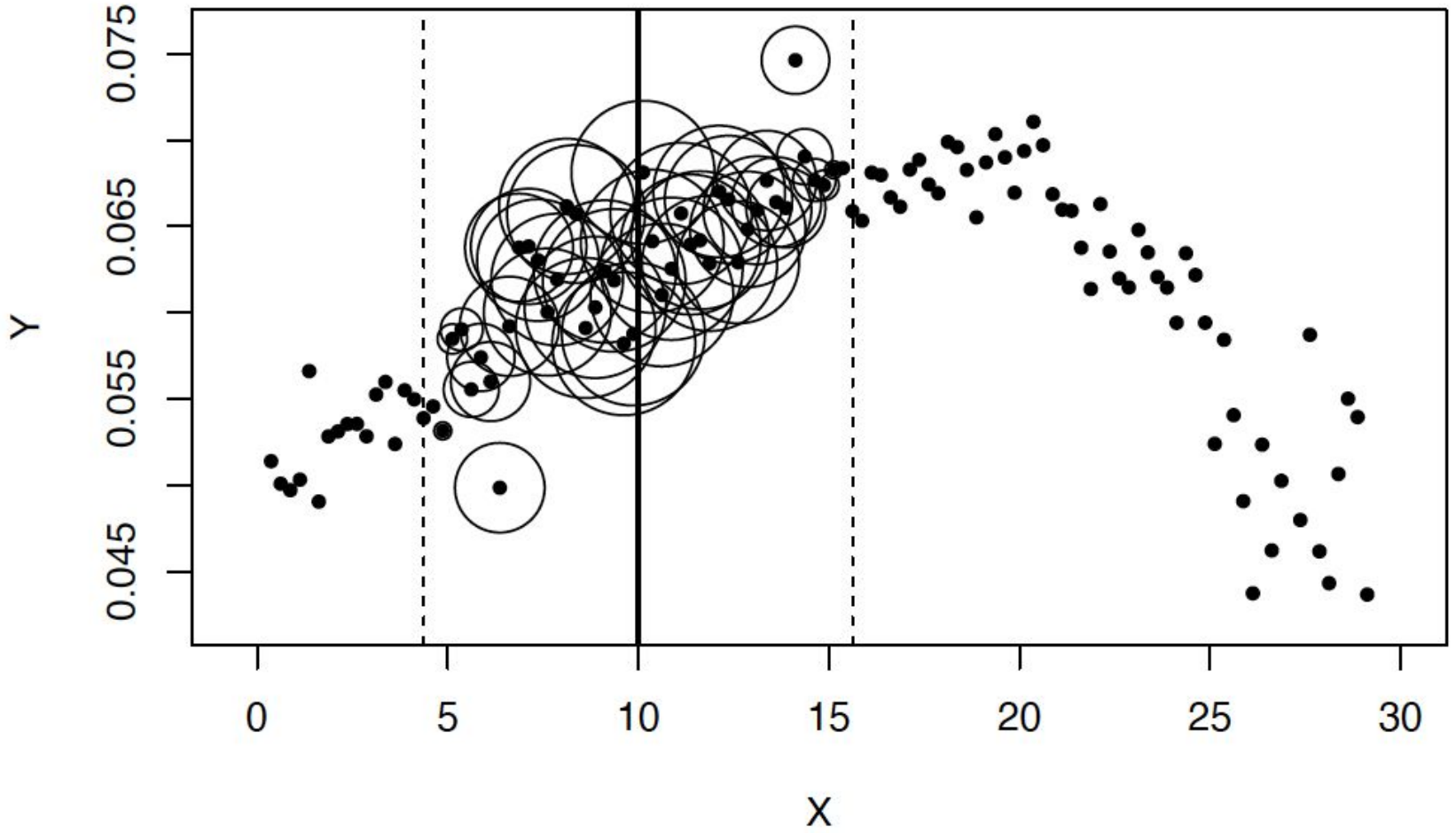
A common way to choose the neighborhood size is to choose is large enough to capture proportion α of the data. This parameter α is often called the **span**. A typical choice is $\alpha \approx 0.5$.



Step Three: Weight the data in the neighborhood.

Values of x which are close x_0 will receive a larger weight than those far from x_0. Denote by w_i the weight placed on observation i. The default choice is the **tri-cube weight function**:

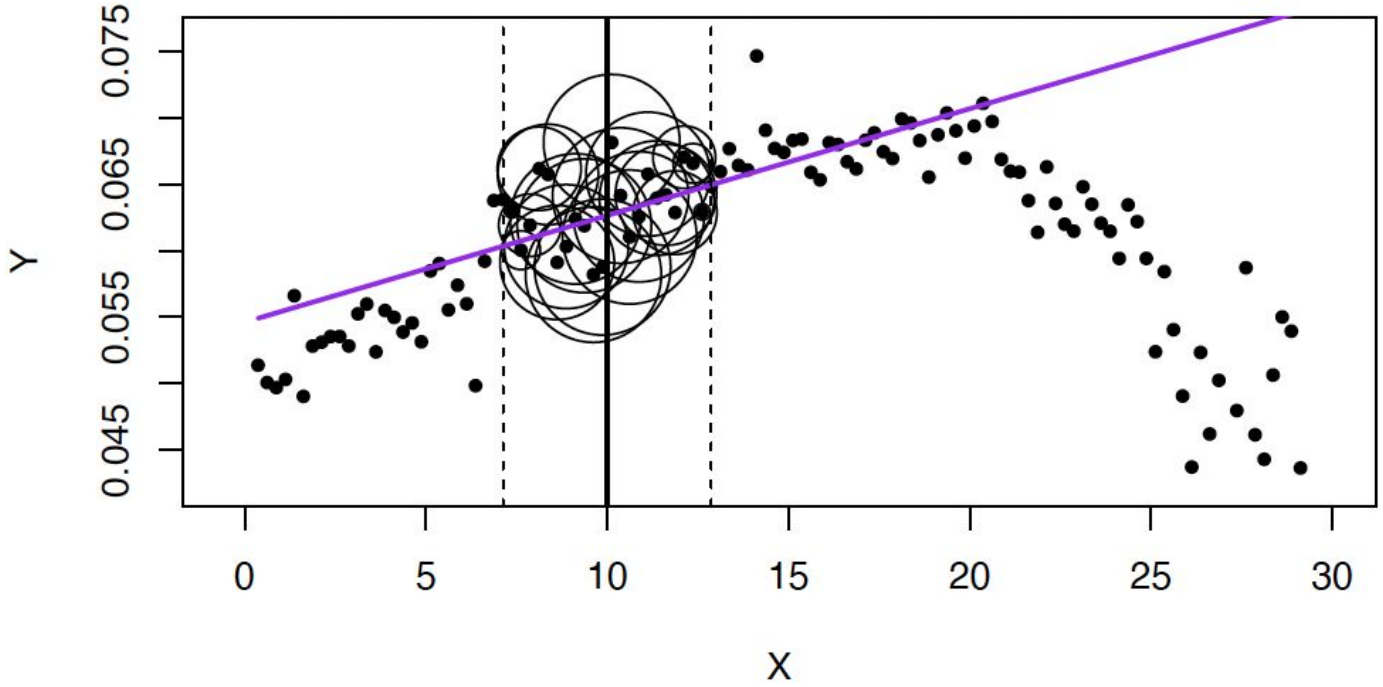
$$w_i = \begin{cases} \left(1 - \left|\frac{x_i - x_0}{\text{max dist}}\right|^3\right)^3, & \text{if } x_i \text{ in the neighborhood of } x_0 \\ 0, & \text{if } x_i \text{ is not in neighborhood of } x_0 \end{cases}$$


Y
0.045
0.055
0.065
0.075
0
5
10
15
20
25
30
X

Step Four: Fit the local regression line.

This is done by finding β_0 and β_1 to minimize the weighted sum of squares

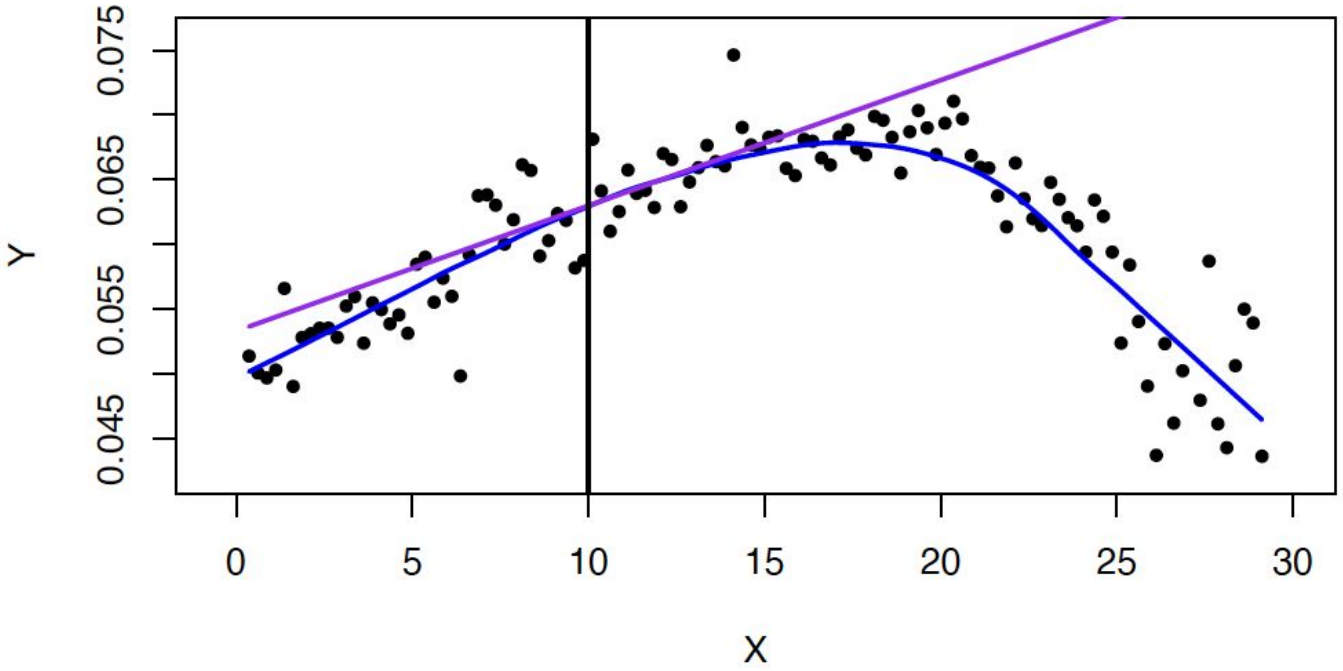
$$\sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$



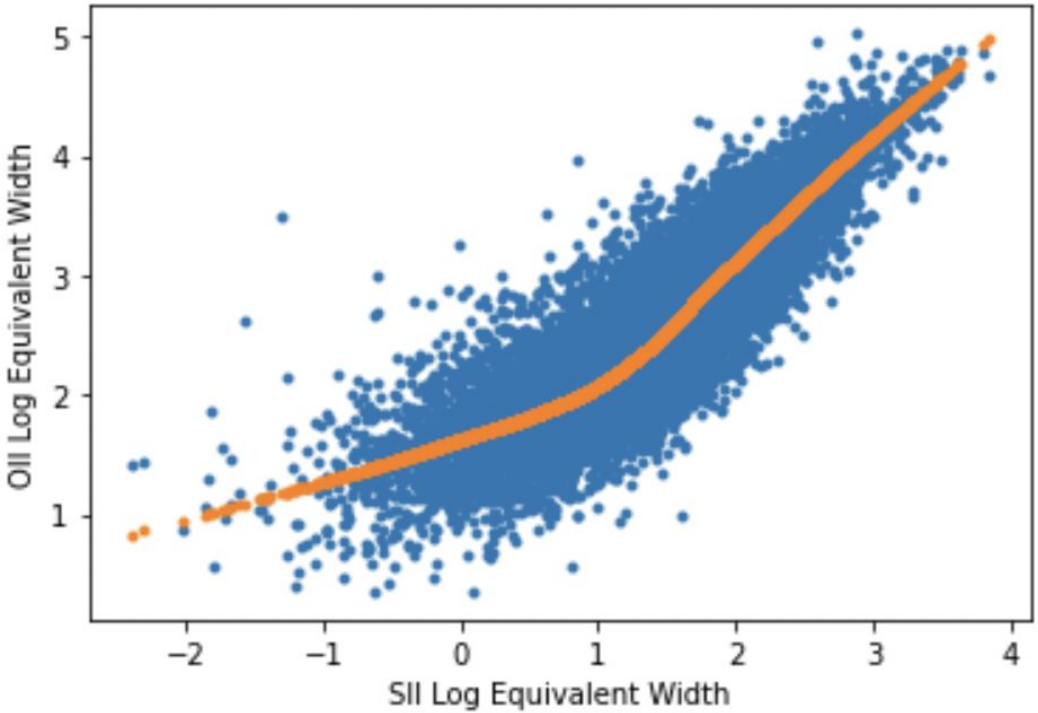
Step Five: Estimate $f(x_0)$.

This is done using the fitted regression line to estimate the regression function at x_0:

$$\widehat{f}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$



To the Colab!


OII Log Equivalent Width
SII Log Equivalent Width
5
4
3
2
1
−2
−1
0
1
2
3
4

Smoothing Parameter Selection

The smoothing parameter in nonparametric regression controls how “smooth” the resulting estimate is: **A smaller value leads to a “rougher” estimate, a larger value gives a “smoother” estimate.**

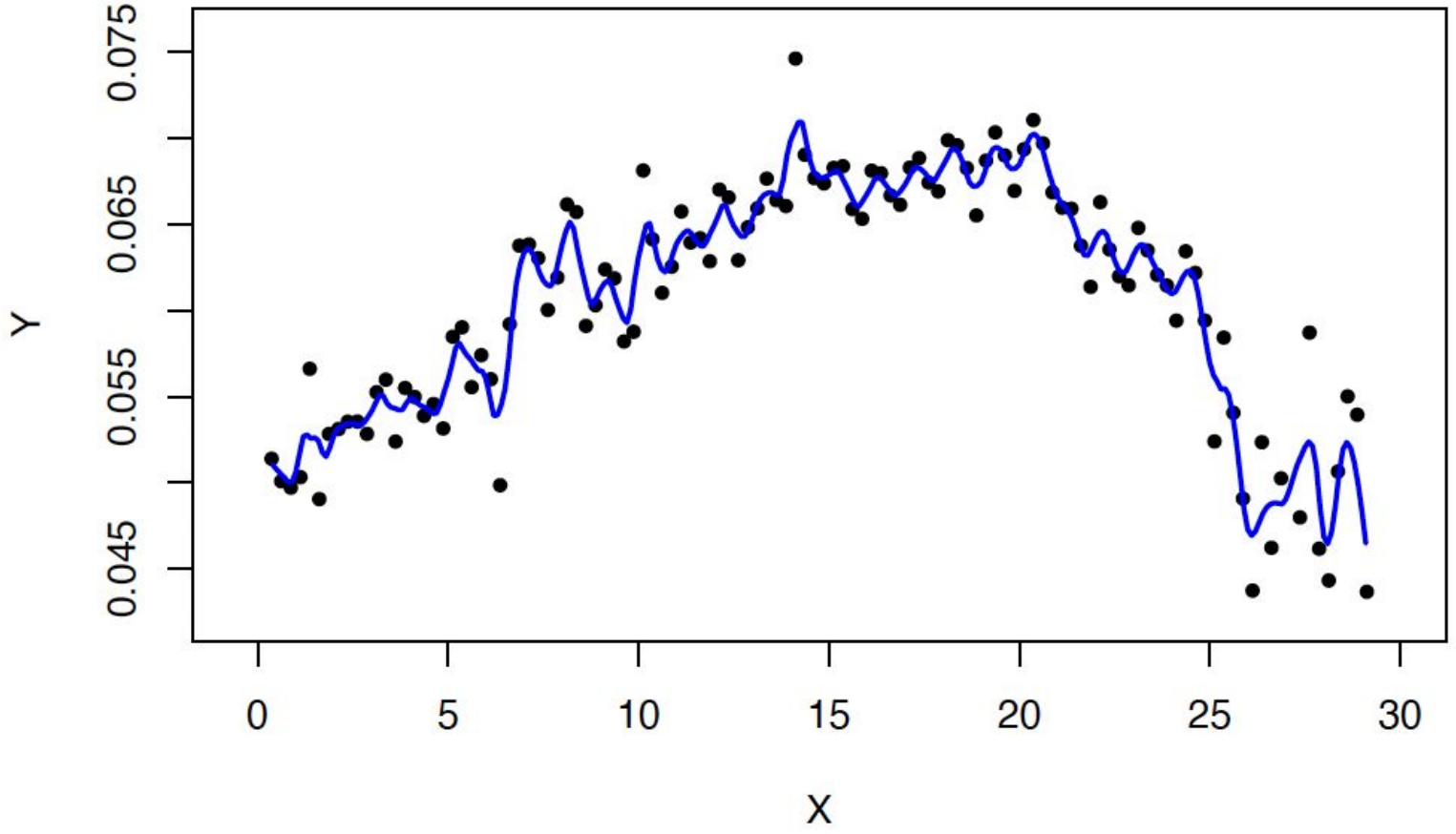
The argument span to loess() controls the span, as defined above. Reasonable values are ~0.2-0.8.

Smoothing Parameter Selection

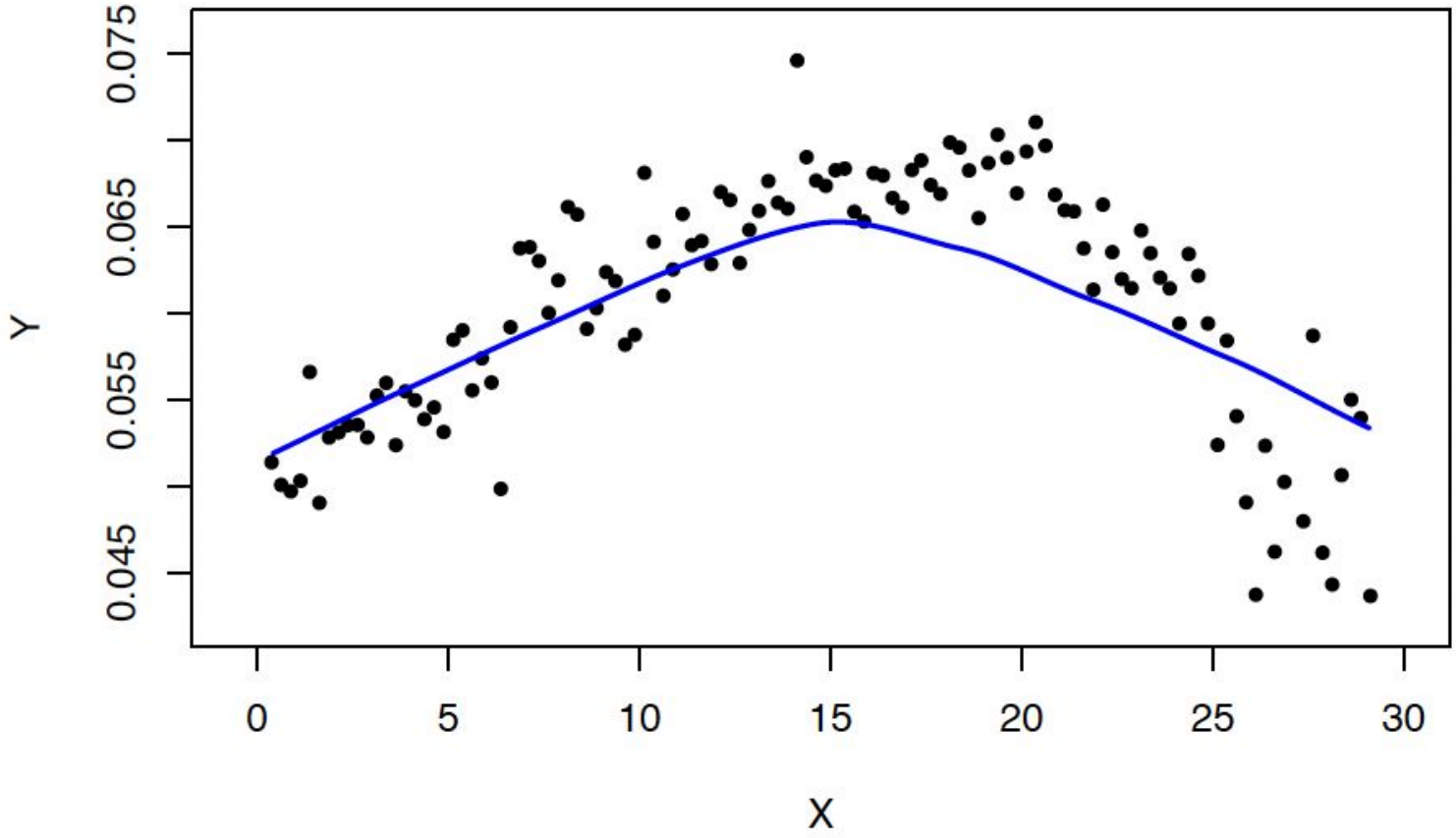
The smoothing parameter in nonparametric regression controls how “smooth” the resulting estimate is: **A smaller value leads to a “rougher” estimate, a larger value gives a “smoother” estimate.**

The argument span to loess() controls the span, as defined above. Reasonable values are ~0.2-0.8.

The smoothing parameter **cannot** be chosen by minimizing the sum of squared residuals (“least squares”). Doing so would lead to overfitting,since for small enough choice, the residuals could all be made close to zero.

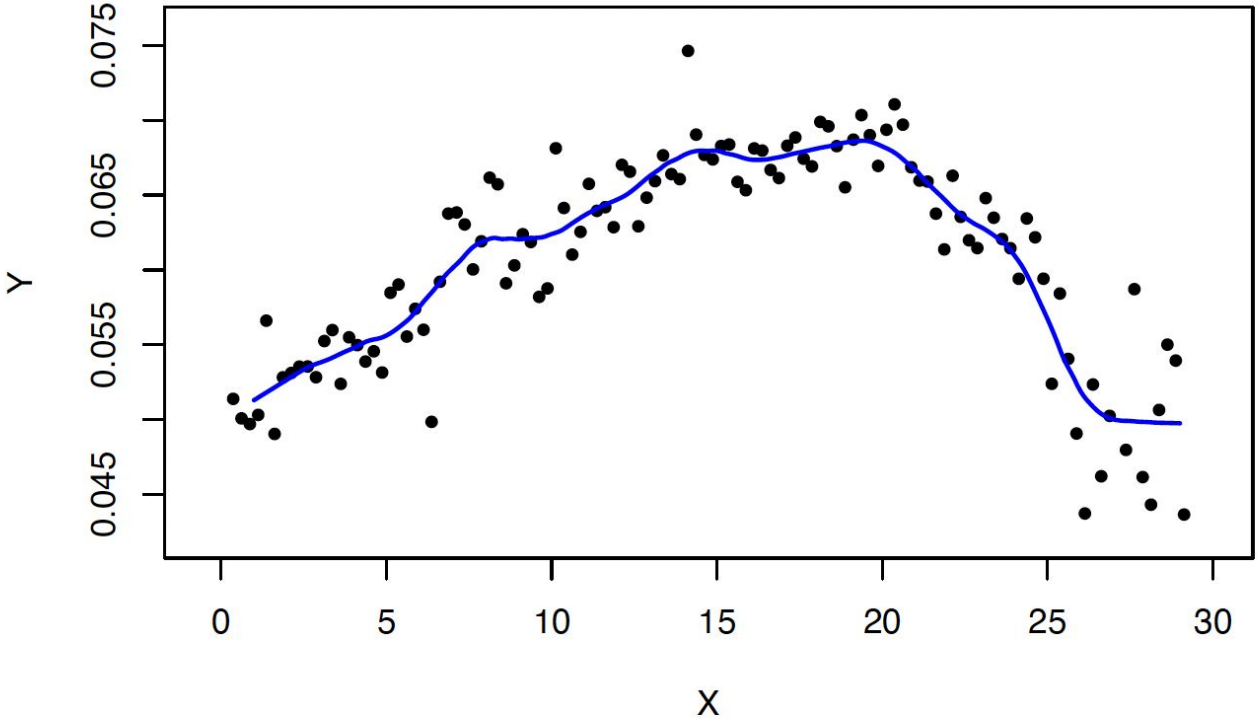


Using span = 0.05. Clearly not enough smoothing.



Using span = 0.9. Too much smoothing, missing important features.

We can optimize span with cross validation!



When the sample size is small, cross validation can be unreliable. Reconsider our example from above, with span = 0.16 chosen using GCV.

Smoothing Splines

Another nonparametric approach to regression is the **penalized spline** or **smoothing spline.**

This approach starts with an optimization problem: Find the twice differentiable function f(x) such that:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b \left[f^{(2)}(x) \right]^2 dx$$

where $f^{(2)}(x)$ indicates the second derivative of f evaluated at x and $\lambda > 0$ is the smoothing parameter.

Typically, a = min{xi} and b = max{xi}.

Note that the **penalty term**

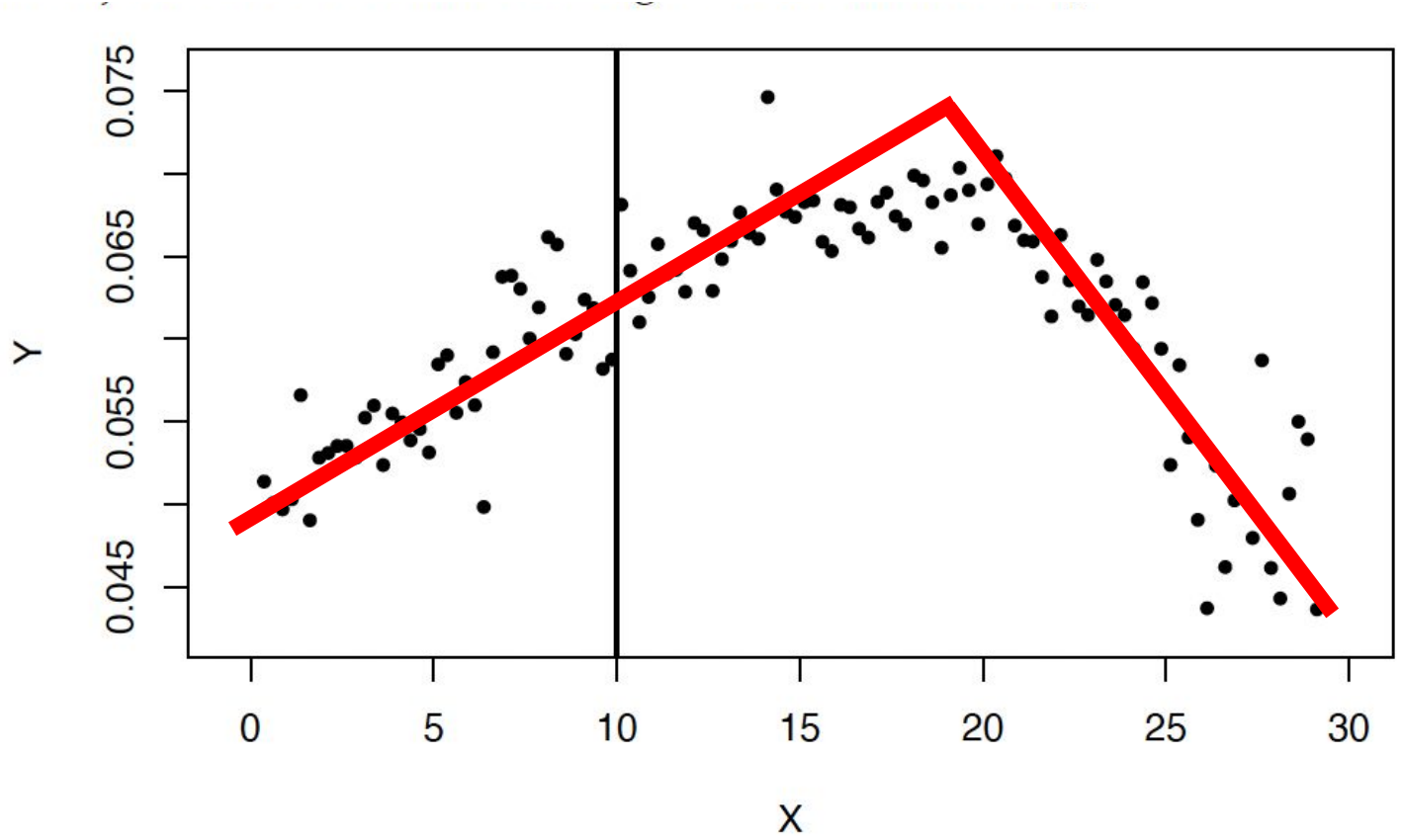
$$\int_a^b \left[f^{(2)}(x)\right]^2$$

will be large if the function is "wiggly.'' Of course, if $f(x) = a + bx$, then this penalty equals zero.

Large values of λ lead to smooth functions $f()$.

As before, λ can be chosen via cross-validation.

Broken Stick Regression



At each **knot** a slope can change.

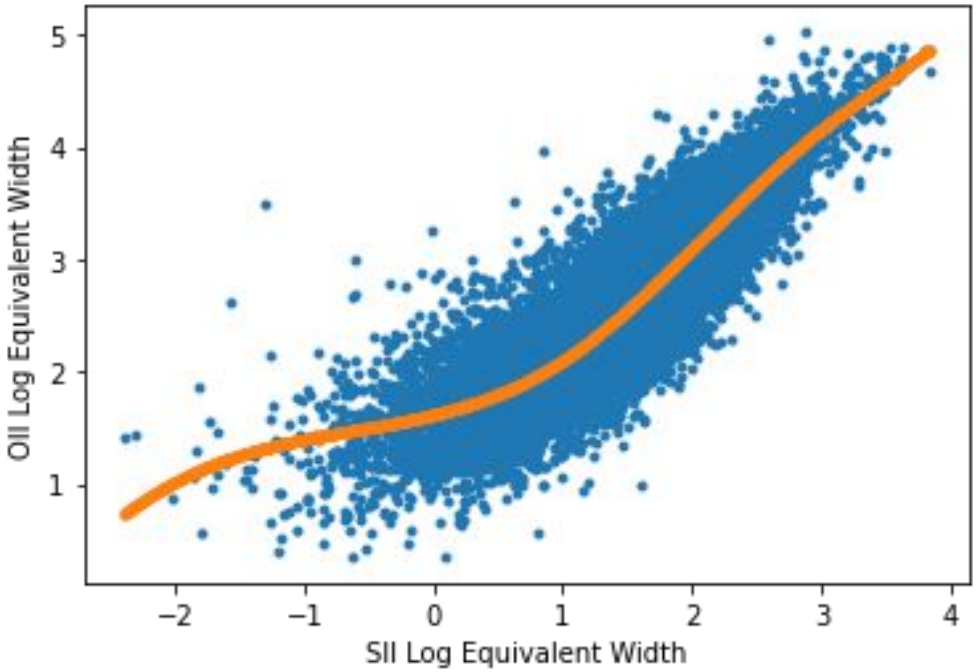
For higher order (order p) polynomials, the basis vectors are used at each knot:

$$1, x_i, \ldots, x_i^p, (x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p$$

Only the highest order coefficients change at each knot.

The number of knots it typically << the number of data points

To the Colab!


OII Log Equivalent Width
SII Log Equivalent Width
5
4
3
2
1
−2
−1
0
1
2
3
4

Comment: The Use of Nonparametrics in Astronomy

A common objection raised to using nonparametric regression in astronomy is the failure to obtain a **functional form** for the relationship between the variables.

Comment: The Use of Nonparametrics in Astronomy

A common objection raised to using nonparametric regression in astronomy is the failure to obtain a **functional form** for the relationship between the variables.

Although there are situations where this is a drawback, there are many other situations in which a researcher wants a model for either (1) making predictions (e.g., photometric redshifts), or (2) summarizing data.

“Letting the data speak for themselves” more important as samples grow.

Conclusions

Linear regression is a powerful tool for data-driven modelling of astronomical data.

MOO (Model, Objective Function, Optimization Method) is a generic method for solving inference problems, including regression.

Parametric and non-parametric methods alike can be used to fully interpret our data.

Let’s take a break, then take questions!