# Regression

## PHYS 2601

# Regression

- Regression analysis is an attempt to estimate the relationship between an independent variable and one or more dependent variables.
- Regression analysis is often used for two possible ends:
  - Prediction - to predict a y value given an x value.
  - Casual Relationship - to understand how y is caused by x.
- In either case our analysis maybe the same, but our interpretation of what we have done may differ.  Also, what assumptions we are willing to make differ.
- Most importantly, regression analysis only uncovers correlations in the data set, extrapolating beyond that requires additional arguments.

# Regression

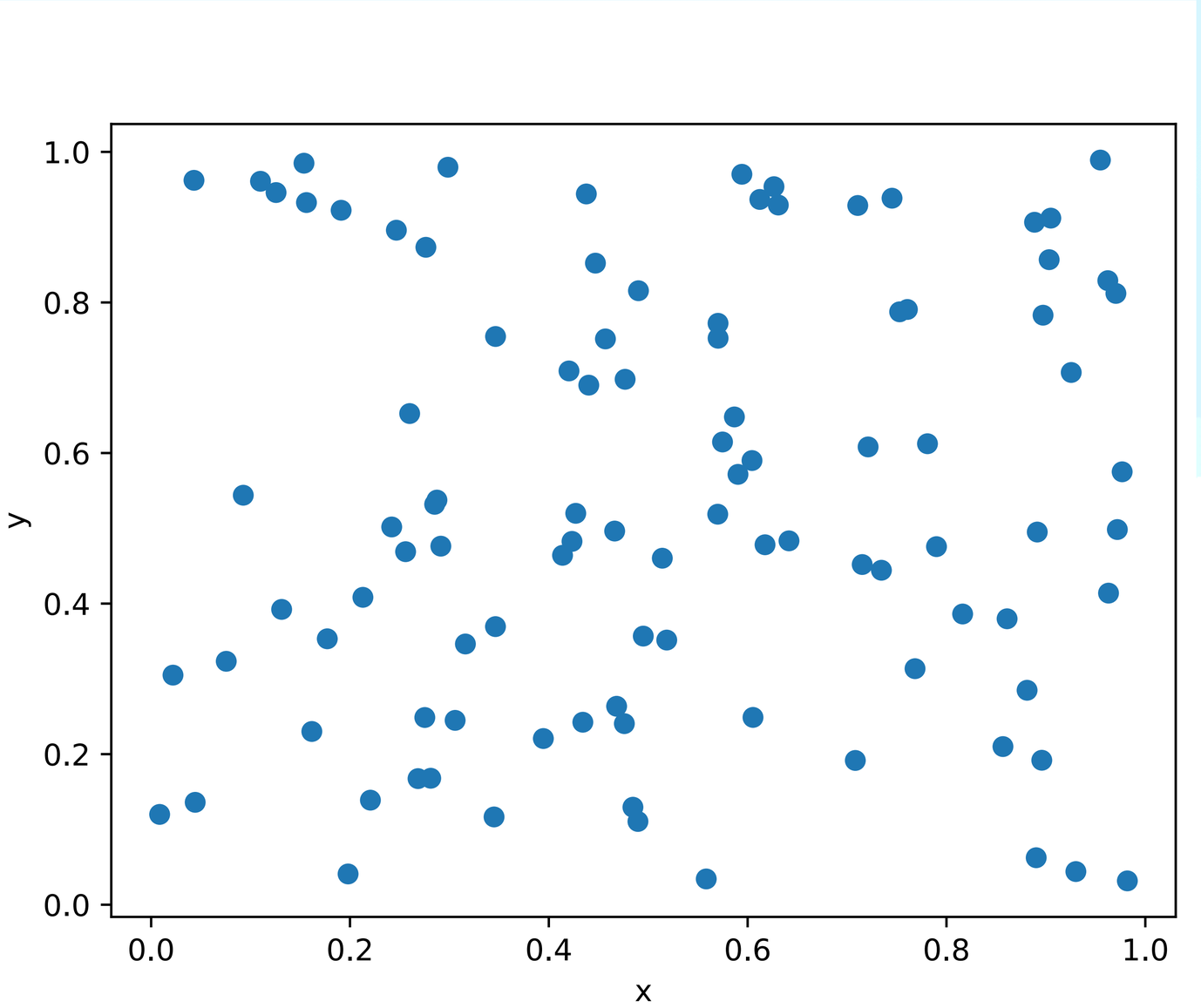
- In regression one first chooses a model (e.g. linear) and then uses a method (e.g. ordinary least squares) to find the unknown parameters of that model.
- So one has unknown parameters, **β**, an independent variable, x, and a dependent variable y (or a vector of dependent variables **y**) and some errors ε.

$$y_i = f(x_i, \beta) + \epsilon_i$$

- One is looking for an estimate of the function f that bests fits the data. This then becomes the best values of **β**.

# correlation coefficient

- How do we know if we should even try to fit some function to our data?
- We can start by estimating a correlation coefficient for our data. There are many ways to do this but the two most important are.
  - Pearson correlation coefficient
  - Spearman correlation coefficient



# Pearson correlation coefficient

- The Pearson correlation coefficient measures the linear correlation between two variables. It does this by measuring the covariance of the variables divided by their standard deviations.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

- The value of $r_{xy}$ ranges from -1 to 1 where 1 is a perfect linear relationship and -1 is also a perfect linear relationship but with a negative slope.

- One can see that by trying both y=x and y=-x in the above equation.

# Spearman correlation coefficient

- The Spearman rank correlation coefficient is a nonparametric measure of the rank correlation between two variables. It asses how monotonic the relation is between the two variables.
- To rank order, you simply take a variable with n values, put the values in order from lowest to highest and then replace their values with 1 to n.
- To calculate the Spearman rank correlation, you do this for each variable and then calculate the Pearson correlation of the ranks.
- The advantage of Spearman is that the relation between the two variables doesn’t have to be linear. As long as it is a monotonic function the the Spearman correlation will be 1 (or -1 if monotonic decreasing function).

# summary

## correlation coefficients

- The Pearson correlation, $r_{xy}$, measures how close your data is to a line. In scipy also returns a p value,

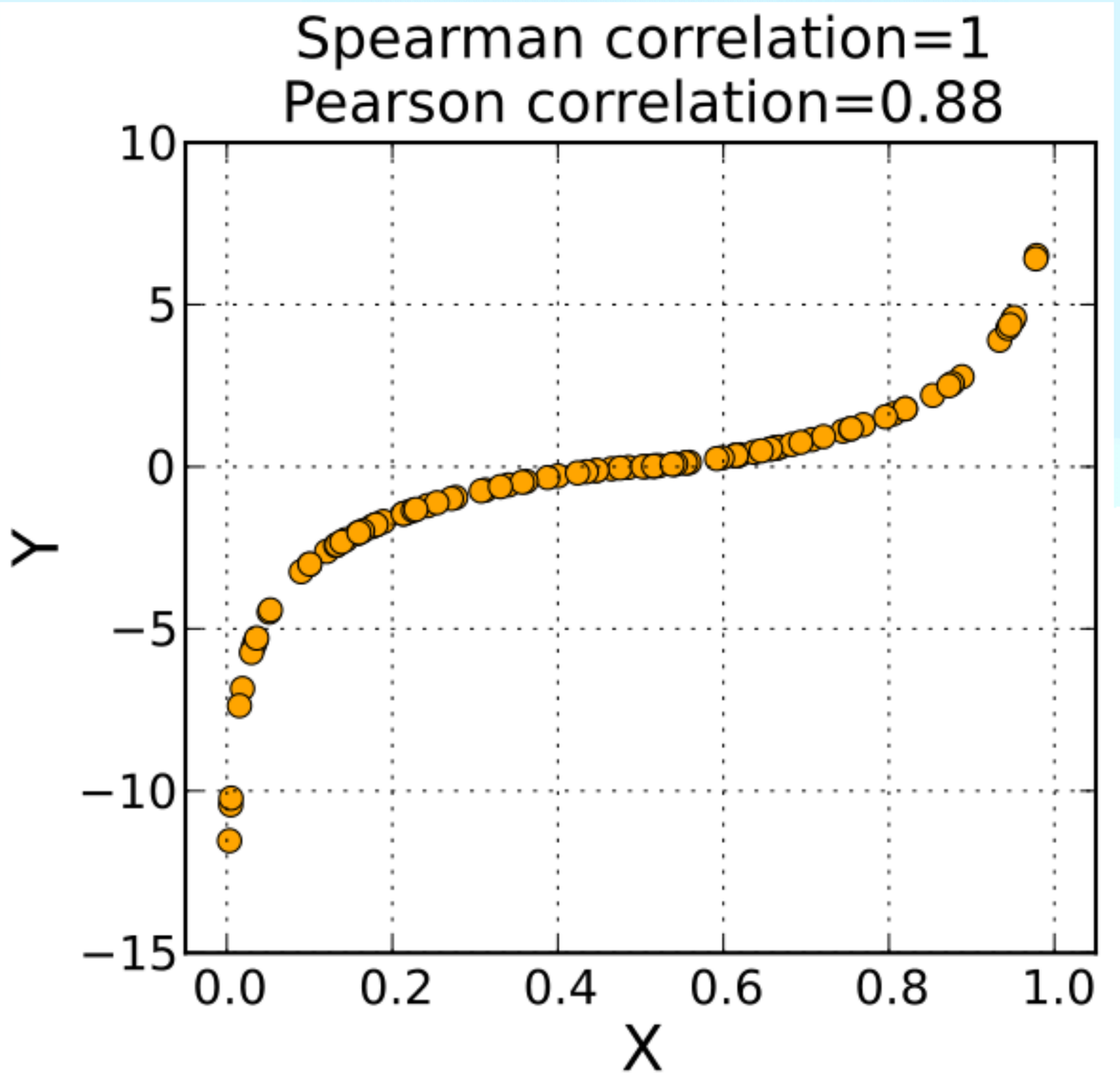
```
r,pval = scipy.stats.pearsonr(x, y)
```

- The Spearman correlation, ρ, measures how close your data is to a monotonic function. In scipy,

```
rho,pval = scipy.stats.spearmanr(x, y)
```

- Or in pandas with two series x and y

```
r = x.corr(y)
rho = x.corr(y, method='spearman')
```



# simple linear least square regression

- Now let us return to regression analysis and consider the simplest and most common example, ordinary least square regression or OLS.

$$y_i = \alpha + \beta x + \epsilon_i$$

- In this case we will fit a linear model (^) and the method we use to fit the model is called least squares.

- We can assume that to get a good model we would like to have small residuals. The method essentially is how to we define small.

- Least squares as the method means we want to have the smallest sum of the square of the residuals.

$$SS_{res} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

# simple linear least squares

$$SS_{res} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Now we simply want the values of α and β that give us the smallest residual sum squared ($SS_{res}$). Note that there are other choices for how to sum the residuals, but we do want them to all add positively. Squaring is nice mathematically and allows this formula to be solved exactly.

- However, squaring also gives a larger weight to large residuals, which may make outliers bias your results.

- The values of α and β that give the smallest $SS_{res}$ are then:

$$\hat{\alpha} = \bar{y} - (\hat{\beta}\bar{x}) \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{\sigma_y}{\sigma_x}$$

# Coefficient of Determination

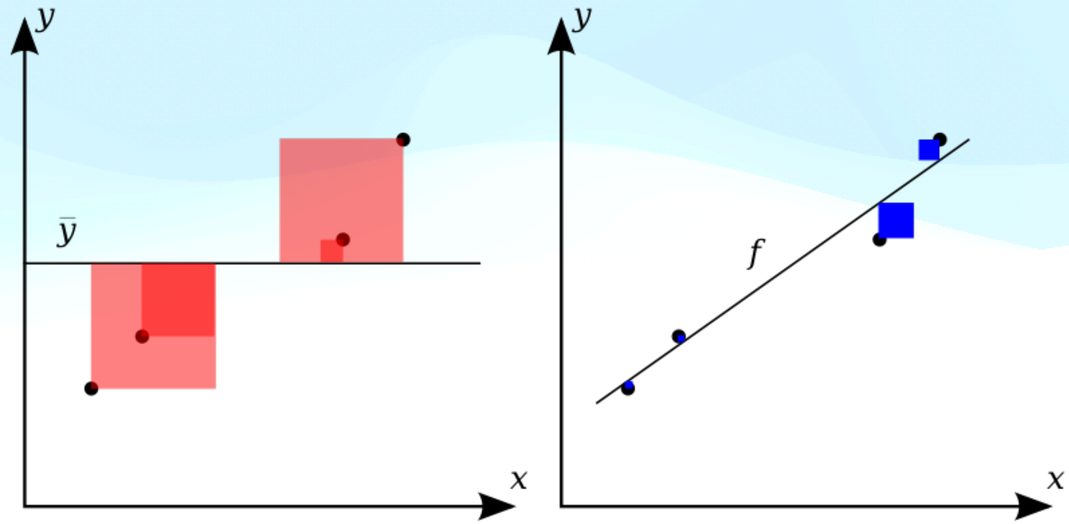
## $R^2$

- The coefficient of determination is a measure of how well the residuals from the model explain the variance in the data.

$$SS_{res} = \sum_{i=1}^{n} \epsilon_i^2$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$




- The red squares are $SS_{tot}$, the blue squares are $SS_{res}$.
- We can see that $SS_{res}$ is much smaller than $SS_{tot}$, so $R^2$ is close to 1.
- The variance is the model accounts for most of the variance in the data.

# Ordinary Least Squares

- If the method we use to find our model is minimizing $SS_{res}$, then we are using least square regression. However, we are free to choose any model we like

$$y_i = f(x_i, \beta) + \epsilon_i$$

- If it is a line then we have simple linear least squares.
- If the function only has linear terms in β, then we have linear least squares. Note that a polynomial of any degree is an example of linear least squares. These problems have exact solutions using matrix algebra.
- Function that are not linear in β give nonlinear least squares. These in general to not have exact solutions and must be solved numerically.
- Ordinary refers to no weighting of the $\varepsilon_i$. If they are weighted we would have weighted or general least squares.

# Linear Least Squares

- As long as the model function is linear in β, we have linear least squares. So if

$$f(x, \beta) = \sum_{j=1}^{n} \beta_i \phi_i(x)$$

- it is useful to express the solution in matrix notation. **y** = **Xβ**, where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 & \phi(x_1) & \dots & \\ 1 & x_2 & \phi(x_2) & \dots & \\ & \vdots & & & \\ 1 & x_n & \phi(x_n) & \dots & x_n \end{bmatrix} \quad \beta = [\beta_1 \ \beta_2 \ \dots \ \beta_k]$$

# Linear Least Squares

- So staring with $\mathbf{y} = \mathbf{X\beta}$, we get $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X\beta}$ and therefore $\boldsymbol{\beta}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, which gives us the formula for a solution in terms of linear algebra.
- Note that with N data points and k terms in $\boldsymbol{\beta}$, the problem is over determined for $N > k$. If $k > N$ then the problem is under constrained and there is no unique solution. You can't constrain a best fit model with more free parameters than data points.
- While if k=N there can be a model with no residuals, this is rarely what we are looking for. This is called overfitting in machine learning language. While the model has no residuals we don't think it will be very good and predicting the correct y for a new x, nor do we think it reveals anything about the casual relationship between or x and y.

# Nonlinear Least Squares

- If we wish to use a statistical model that is not linear in the parameters β, then we can still use the some of the squares as our method of finding the best model, but we can not do it with linear algebra.
- In general there is no way to do this in closed from and one must use some type of interactive approach.
- With computers of course this is not challenging. The scipy function curve_fit can be used to do this for us

```
def my_func(x, beta):

    return x**beta[0] + np.exp(x*beta[1])

popt,pcov = scipy.curve_fit(my_func, x,y)
```

# Weighted Least Squares

- So far we have only been discussing ordinary least squares, but we can imagine making a slight modification to get weighted least squares.
- Instead of treating all residuals the same, what if we some residuals should be larger than others. What if we have an independent estimate for the variance of each data point?
- Then we could replace our $SS_{res}$ with

$$SS_{res} = \sum_{i=1}^{n} \frac{\epsilon_i}{\sigma_i}$$

- where $\sigma_i$ are the y standard deviation errors for each data point.

# General Least Squares

- We can generalize our least squares even more. Remember we have always assumed that our y data values are independent and therefore so are the residuals, but what if this is not true.
- What if there is covariance between the residuals? Well if we have a measure of the covariance of the errors called the precision matrix Ω we can use that to weigh the least squares sum.

$$\beta = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

- Note that if our model is still linear in β then this still solved with linear algebra.

# Gaussian Processes

## fitting the precision matrix

- Under some assumption about everything being the sum of normally distributed variables it can be shown that all the information about the statistical model is in the covariance matrix of the residuals (or the inverse precision matrix).

- One can then approach regression as not only finding the parameters β of the statistical model, but also fitting the components of the residual covariance matrix $c_{ij}$.

- This is done by assuming some model of the covariance matrix, just as we choose some function for the regression. Typical choices are a Gaussian or exponential decrease away from the diagonal elements of the matrix. In any case this model will have some parameters γ.

- One then goes back to finding the value of the parameters β and γ that give the smallest sum of the residuals, $SS_{res}$.

# Selecting a Model

- We have discussed different methods (based on $SS_{res}$) for finding our best fit model (ordinary least squares, weighted least squares, general least squares and gaussian processes), but how do we select our statistical model?
- First let's consider that our model must have fewer parameters then we have data points or it will be unconstrained.
- However, increasing the number of parameters will always improve the fit of our model ($SS_{res}$) or at least not make it worse.
- Note that increasing the number of parameters past the true generating function will still improve the fit. This is called overfitting.

# Over fitting

- Let’s imagine our generating function is linear with Gaussian random noise.
- If we use regression to fit a quadratic to fit the data our $SS_{res}$ will always be equal to or smaller then what we get when fitting a line. However, since we know that our generating function was a line, the reduction in $SS_{res}$ is because we are now fitting the Gaussian errors.
- This is the meaning of over fitting at some point in any regression adding more parameters will not be fitting the relationship between x and y, but instead fitting this data’s random errors.
- One way to recognize over fitting is if one can acquire new data then the underlying relationship f(x) will be the same, but the errors will be totally different. A fit that does equally well to both data sets is fitting the relationship. A fit that changes drastically for the two data sets is fitting the errors or over fitting.

# Least is Best

- For this reason we usually want to fit our data with the fewest parameters that still provide a good fit. That is we want the least complex model as possible.
- In this way we are least likely to over fit and if we are trying to uncover a causal relationship we can start with the simplest relationship instead of a more complex one.
- However, it is important to remember that all we can say is that the data does not require a more complex model. That doesn’t mean the generating function isn’t a more complex model. In fact when the errors are large we will never be able to show that a more complex model is required.
- So if one finds a line is a good fit to the data, one’s conclusion should be a line is the simplest model that fits the data, but all higher order polynomials will also fit the data even better.