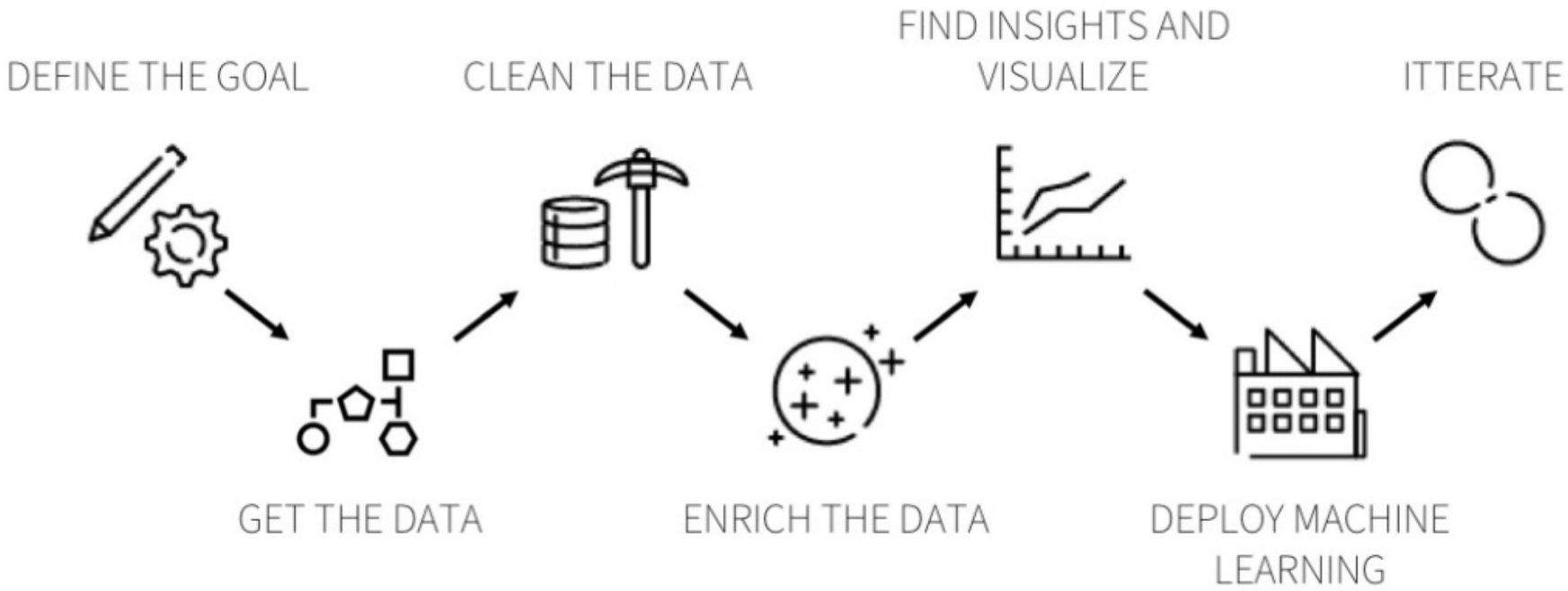
Exploring Data with `Statistics`

The Data Science Process



Agenda

- What is Regression?
- From scatterplots to regression lines using highschool math.
- Residuals
- Method of Least Squares
- R^2 and goodness-of-fit

Linear Regression

This is where we transition from the realm of traditional statistics and venture into the world of machine learning.

> Linear Regression (and Generalized Linear Models) is among the foundational tools for Data Science, employed for prediction or to evaluate whether there is a linear relationship between two numerical variables..

Specifically, we use regression to answer questions such as

```
* How does sales volume change with changes in price. How is this affected by changes in the weather?
* How does the amount of a drug absorbed vary with dosage and with body weight of patient?  Does it depend on blood p
* How are the conversions on an ecommerce website affected by two different page titles in an A/B comparison?
* How does the energy released by an earthquake vary with the depth of it's epicenter?
```

by creating a **model** which is essentially a probabilistic formula that defines the relationship between a dependent variable and the independent variables. For **simple linear regression**, this model takes the form of an equation such as

$$y = \beta_0 + \beta_1 x + \epsilon_i$$

When there are multiple x_i, we use the more general form of the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon_i$$

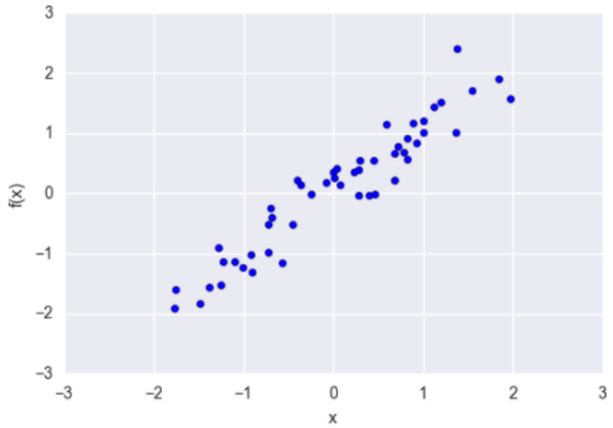
where the β_i are known as the **coefficients**, the x_i are called the **predictors** and the y is the dependent variable.

Visualize

We have established so far that in order to explore the relationship between two numeric variables, we use

- scatterplots to visualize the relationship, using `df.plot.scatter(x='X', y='Y')`
- correlation coefficients to quantify the strength and direction of the relationship, using `df[['X', 'Y']].corr()`

We know that a scatterplot is essentially a bunch of point plotted on an X-Y plane.

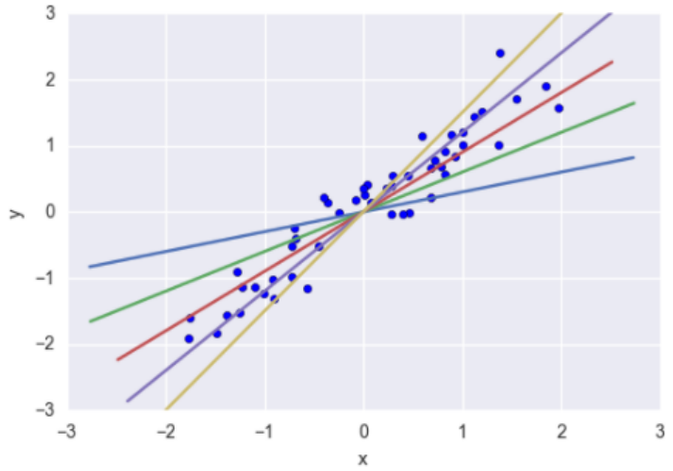


Simply put, Regression is a class of techniques for **fitting** a straight line to a set of data points where we try to find the line that best fits our data so that we can use it to predict y given a new x.

Straight Lines

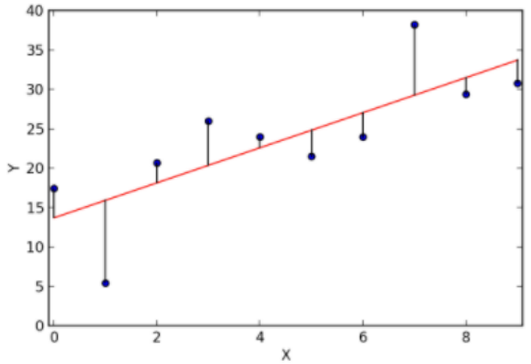
If you studied co-ordinate geometry in highschool, you would remember that the equation of a straight line is: $y = mx + c$, where m is the *slope* of the line, and c is the y-intercept (the point where the line cuts the y-axis.) This equation can be computed using any two points $(x_1, y_1), (x_2, y_2)$ in the X-Y plane, where $m = \frac{y_2 - y_1}{x_2 - x_1}$

So, if we were to find the line that **fits** our data, we could potentially draw hundreds of them on our cloud of points, and then figure out a way to determine the **line of best fit.**



Residuals

Mathematically, the residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$. This $\epsilon_i = y_i - \hat{y}_i$ is sometimes referred to as **process uncertainty**. We would like to select β_0, β_1 so that the $\sum \epsilon_i = 0$, but this is not usually possible.



Instead, we choose a reasonable criterion: ***the smallest sum of the squared differences between*** $\hat{y}$ ***and*** y. Squaring serves two purposes:

- to prevent positive and negative values from cancelling each other out, and
- to strongly penalize large deviations.

> In other words, we will select the parameters that minimize the squared error of the model.

Least Squares Regression

The least squares method minimizes the vertical distances between our data points (the observed values) and our line (the predicted values.) Mathematically, we find the values for β and β_0 in simple regression using:

$$\beta = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - \bar{X})^2}$$

The regression line will always pass through the point $(\bar{X}, \bar{Y})$ so we can plug this point into our equation $(y = \beta_0 + \beta_1 x)$ to get β_0, which is the point where the line passes through the y-axis.

$Note$

- Minimizing the sum of squares is not the only criterion we can use; it is just a very popular (and successful) one.
  - For example, we can try to minimize the sum of absolute differences:
- We are not restricted to a straight-line regression model
  - We can represent a curved relationship between our variables by introducing **polynomial** terms. For example, a cubic model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

R^2 for goodness-of-fit

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line SS_M, relative to how much variation there was to explain in the first place SS_T. Mathematically it is expressed as

$$R^2 = \frac{SS_M}{SS_T}$$

where, SS_M is the sum of squares of the residuals, and SS_T is the sum of squared deviations of y_i from the mean $\bar{y}$. In simple regression we can take the square root of this value to obtain Pearsonâ€™s correlation coefficient.

> The closer the R^2 to 1, the better our model is in explaining the variance in the dependent variable.

Linear Regression in Python

The `statsmodels` package implements least squares models that allow for model fitting in a single line

```
import statsmodels.api as sm

lr_0 = sm.OLS(y, sm.add_constant(x)).fit()
lr_0.summary()
```

and produces an output as shown on the next slide.

The package `scikit-learn` also implements a regression model in the following way

```
from sklearn.linear_model import LinearRegression

# Instantiate and fit the model
lr_1 = LinearRegression()
lr_1.fit(X, y)

# Inspect coefficients, intercept
print lr_1.coef_, lr_1.intercept_

# Make predictions
lr_1.predict(X_new)
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.891
Model:	OLS	**Adj. R-squared:**	0.864
Method:	Least Squares	**F-statistic:**	32.67
Date:	Fri, 24 Feb 2017	**Prob (F-statistic):**	0.00463
Time:	22:04:26	**Log-Likelihood:**	-12.325
No. Observations:	6	**AIC:**	28.65
Df Residuals:	4	**BIC:**	28.23
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-4.3500	2.937	-1.481	0.213	-12.505 3.805
x1	3.0000	0.525	5.716	0.005	1.543 4.457

Omnibus:	nan	**Durbin-Watson:**	2.387
Prob(Omnibus):	nan	**Jarque-Bera (JB):**	0.570
Skew:	0.359	**Prob(JB):**	0.752
Kurtosis:	1.671	**Cond. No.**	17.9

A word of caution

Regression can be misleading when

- there are outliers or a non-linear relationship
- residuals are not normally distributed, not independent
- residulas have unequal variance
- **multicollinearity** (having predictor variables that are highly correlated)
    - leads to instability, high variances in estimates, worse interpretability

Quiz

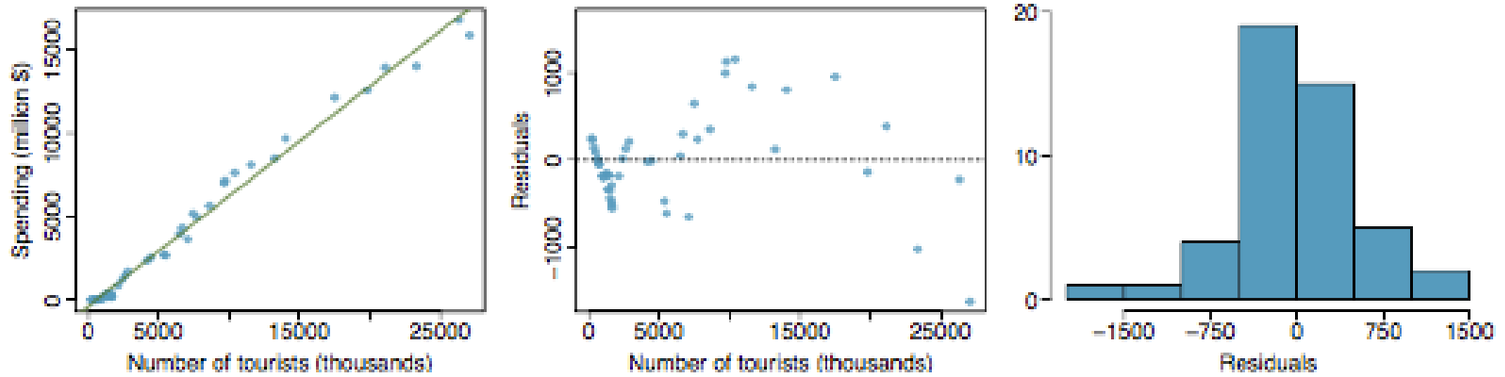
Determine if I or II is higher or if they are equal. Explain your reasoning.

> For a regression line, the uncertainty associated with the slope estimate, b1, is higher when

I. there is a lot of scatter around the regression line or
II. there is very little scatter around the regression line

Quiz

7.23 Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.[20] Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.



(a) Describe the relationship between number of tourists and spending.

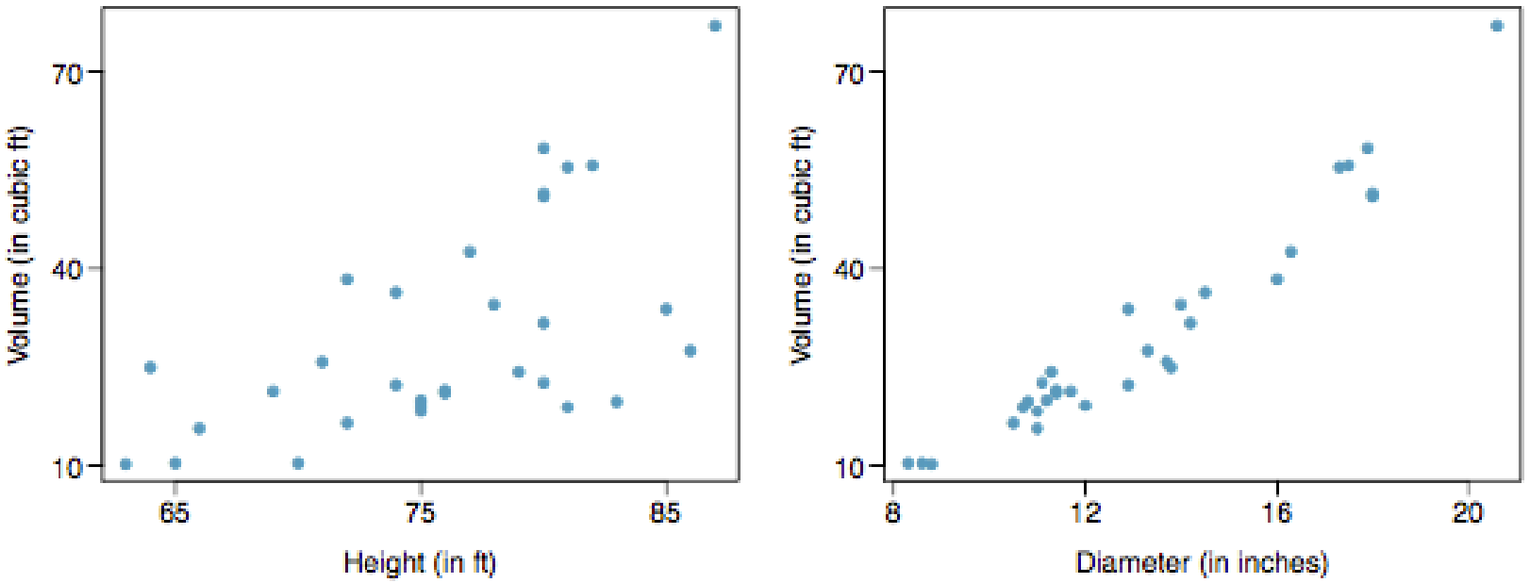
(b) What are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

Assignments

7.12 Trees. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.[17]



(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

Assignments

7.30 Cats, Part I. The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$s = 1.452$ $R^2 = 64.66\%$ $R^2_{adj} = 64.41\%$

(a) Write out the linear model.

(b) Interpret the intercept.

(c) Interpret the slope.

(d) Interpret R^2.

(e) Calculate the correlation coefficient.