# Simple Linear Regression for Data Science

## Linear Regression using `statsmodels`

Suppose we have a dataset named measurements with columns height and weight . If we want to fit a model that can predict height based on weight, we would use the formula 'height ~ weight' as shown in the example code.

```
import statsmodels.api as sm
model = sm.OLS.from_formula('height ~ weight', data = measurements)
results = model.fit()
print(results.summary())
```

## Prediction using a Simple Linear Model

In order to use a simple linear regression model to make a prediction, we need to plug in the slope and intercept to the equation for a line (y=mx+b). For example, suppose we fit a linear model to predict weight based on height and calculate an intercept of -200 and slope of 5. The equation is:

```
weight = 5*height - 200
```

Therefore, a person who is 60 inches tall would be expected to weigh 100 pounds:

```
weight = 5*60-200 = 100
```

## Interpreting Regression Parameters

In simple linear regression the intercept is the expected (or average) value of the outcome variable when the predictor variable is equal to zero. The slope is the expected difference in the outcome variable for a one unit increase in the predictor. For example, if we fit a linear regression with weight as the outcome variable and height as the predictor, then:

- The intercept is the expected weight of a person with height = 0 (even though this is impossible).
- The slope is the expected difference in weight for each additional unit of height.

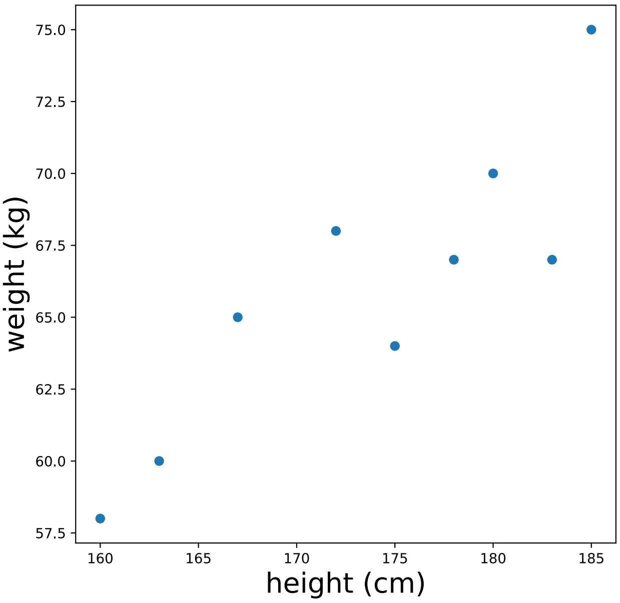
## Linear Regression Assumptions

The assumptions of simple linear regression are:

- linear functional form: the relationship between the outcome and predictor variable must be linear (not curved)
- normality: the residuals should be approximately normally distributed
- homoscedasticity: the variance of the residuals should be equal for all values of the predictor

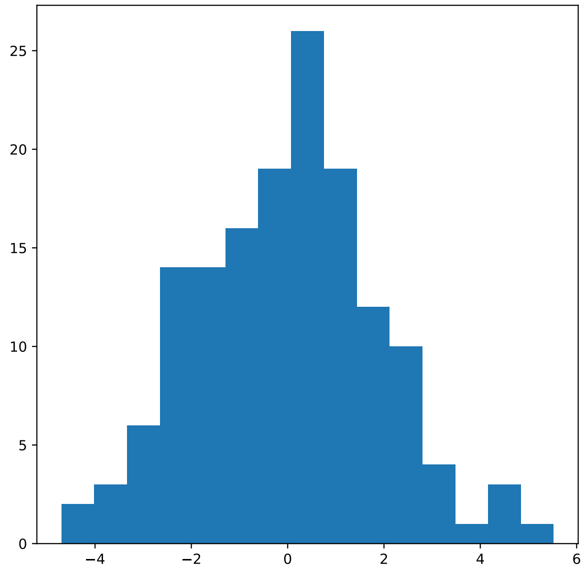
## Linear Functional Form Assumption

In order to check the linear functional form assumption for simple linear regression, we can plot a scatter plot of the outcome variable and predictor variable, then check whether the relationship is linear (can be represented by a straight line). For example, the provided plot of weight vs. height shows a linear relationship.



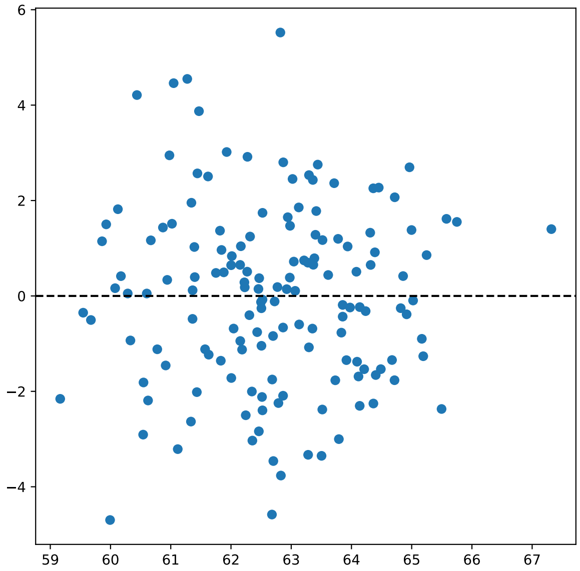
## Normality Assumption

In order to check the normality assumption for simple linear regression, we can plot a histogram of the residuals and check whether they appear approximately normal (no skew or multimodality). For example, the provided histogram of residuals would meet the normality assumption.



## Homoscedasticity Assumption

In order to check the homoscedasticity assumption for linear regression, we can plot the residuals against the fitted values. If the assumption is met, the residuals should be symmetrically scattered around 0, with no funneling or other patterns. The provided plot demonstrates what this should look like if the assumption is met.



## Fitted Values

The fitted values for a linear regression model are the predicted values of the outcome variable for the data that is used to fit the model. For a statsmodels model object named `model` that was fit using a dataframe named `data`, the provided code shows how we could calculate the fitted values.

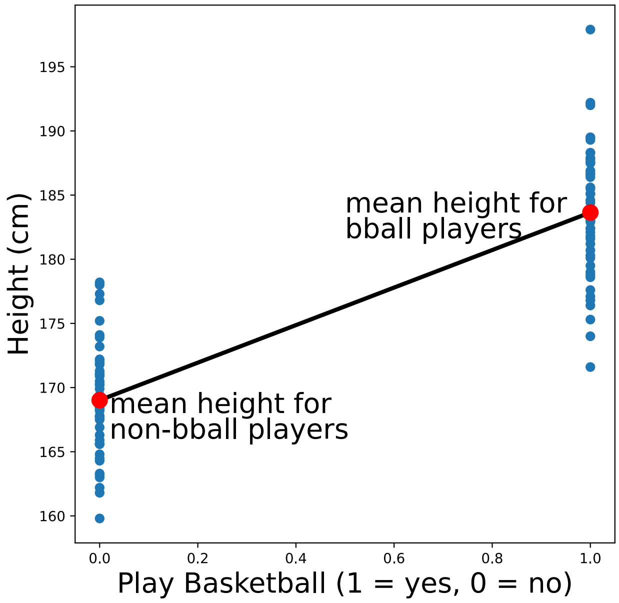
```
fitted_values = model.predict(data)
```

## Residuals

In linear regression, the residuals are the differences between each of the fitted values and true values of the outcome variable. They can be calculated by subtracting the fitted values from the true values.

## Simple Linear Regression with a Categorical Predictor

We can fit a simple linear model with a categorical predictor. When we fit a regression with a binary categorical predictor, one category will be coded as $0$ and the other as $1$. The intercept will be the average value of the outcome variable for the category coded as $0$ and the slope will be the difference in average value of the outcome variable for the two groups. The provided image shows this graphically.


mean height for
bball players
mean height for
non-bball players
Height (cm)
Play Basketball (1 = yes, 0 = no)