Regression/Correlation

Evaluating the model

- Intuitively, if points are grouped closely around the line of best fit we would expect the model to make very accurate predictions.
- We can measure how close points are grouped around the line using the correlation coefficient r.

Correlation Coefficient r

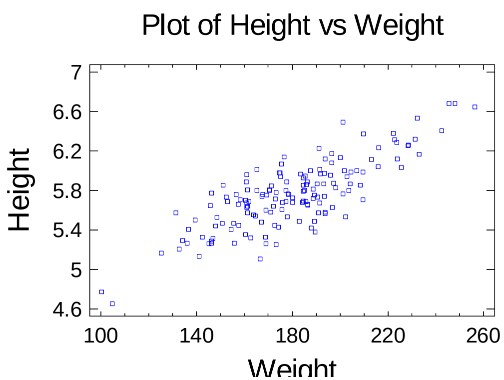
- This is a measure of how closely the values of x and y are (linearly) related .
-
- $r = (n\Sigma xy - (\Sigma x)(\Sigma y)) / [(n\Sigma x^2 – (\Sigma x)^2)]^{1/2}[(n\Sigma y^2 – (\Sigma y)^2)]^{1/2}$
-
- -1 <= r <= 1

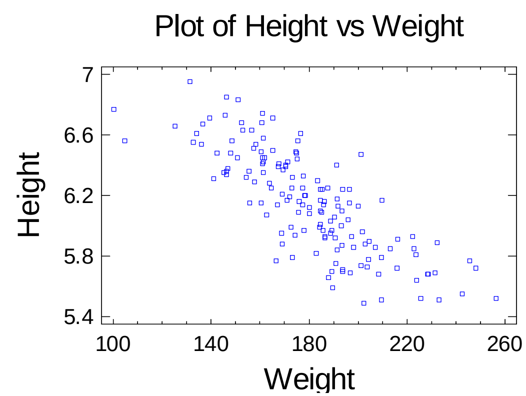
Correlation Coefficient (r)

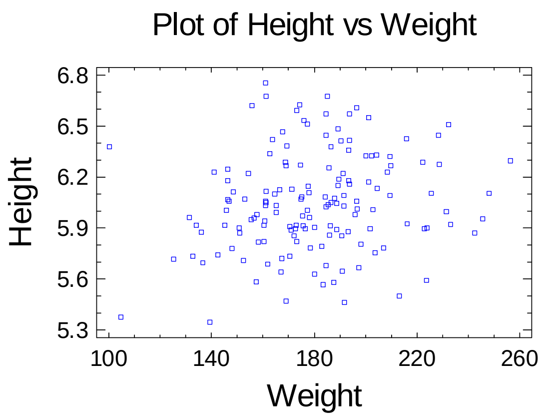
- If the correlation coefficient is close to +1 that means you have a strong positive relationship.
- If the correlation coefficient is close to -1 that means you have a strong negative relationship.
- If the correlation coefficient is close to 0 that means you have no correlation.

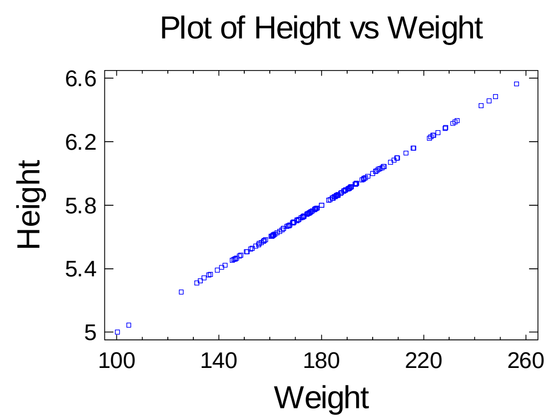
Correlation Coefficient (r)

- If we are interested in determining whether a relationship exists :-









Coefficient of Determination

- r^2 or R^2.
- Gives the percentage of variation of y explained by the variation in x.
- The rest is dues to noise/ random flutations.
- More relevant than r.

Coefficient of Determination

```
model = LinearRegression()
model.fit(x,y)

print('coefficient of determination:', model.score(x, y))
print('correlation coefficient:', math.sqrt(model.score(x, y)))
```

Coefficient of Determination

- y' = predicted value of y
- m = mean of y
- $SSE = \Sigma(y - y')^2$ (unexplained variance)
- $SST = \Sigma(y - m)^2$ (total variance)
- $R^2 = 1 - SSE/ SST$
- $R^2 = (SST - SSE) / SST$
- (explained variance / total variance)

Coefficient of Determination

```
y_hat = model.predict(x)
print('predicted response:', y_hat)

residuals = y_hat - y
print("residuals: ", residuals)



# calculate Coefficient of determination
SSres = sum(np.square(residuals))
print(SSres)

SStot = sum(np.square(y-np.mean(y)))
print(SStot)

rSquared = 1 - (SSres/SStot)
print("Calculated coefficient of determination", rSquared)

print("From model:", model.score(x,y))
```

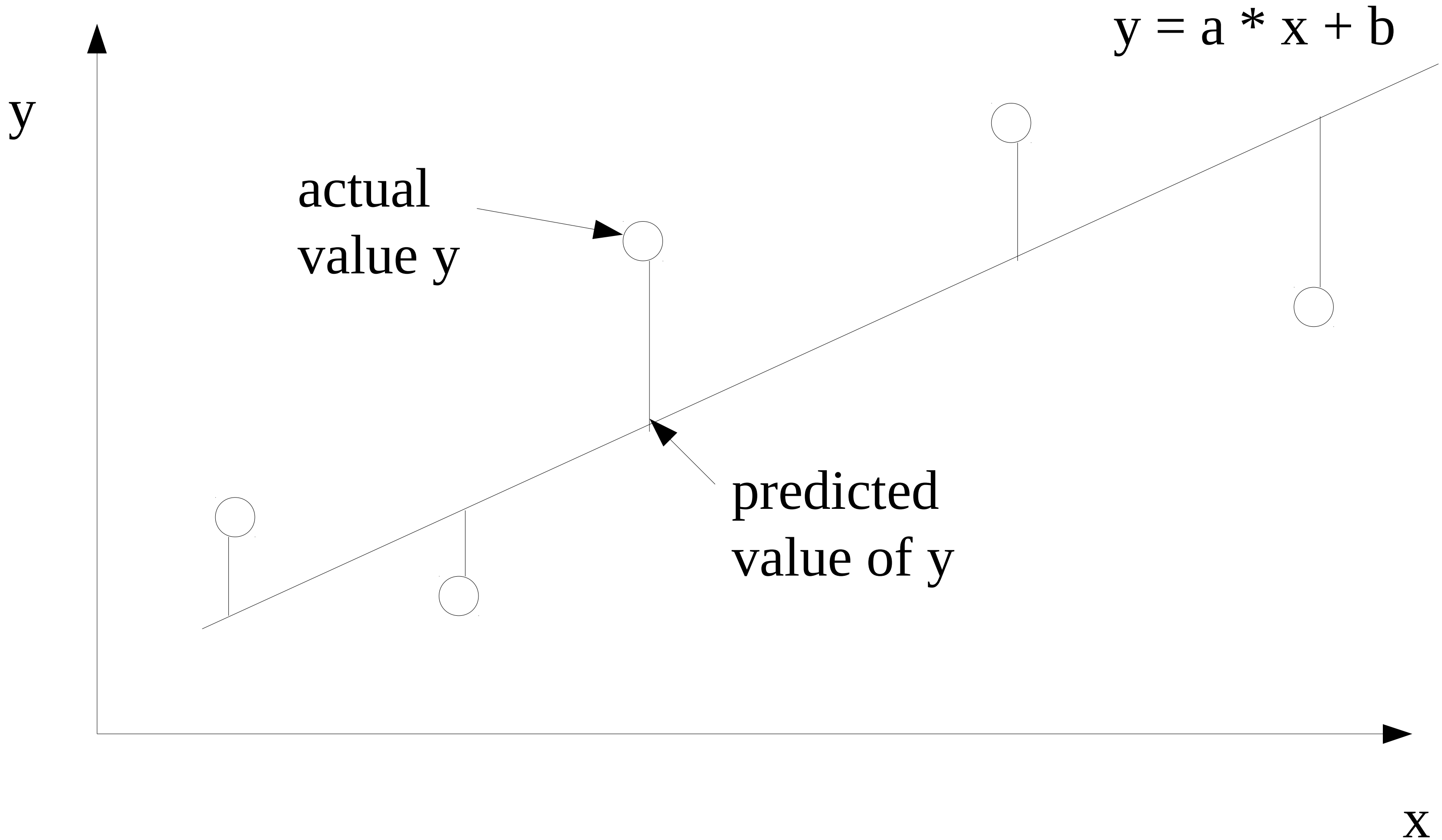
Coefficient of Determination

- The coefficient of determination is a good measure of how good the linear regression model is.
- It is the proportion of the total variation in the data that is explained by the model.
- For simple regression (one independent variable) it is equal to the square of the correlation coefficient.

Evaluating the Model – Other Measures

- Although R^2 is probably the best measure of how good the model, lets consider a number of alternatives.
- This is useful for demonstrating why R^2 is so useful.

Recap - Errors



Evaluating the Model – Other Measures

- SSE – Sum of Squares of Errors?
  - No. the more points we have the larger the value
- MSE – Mean of the Squares of the Errors?
  - Better, but still if on average the error is 5, MSE will be 25.
- RMSE – Root mean square error
  - Yes - Gives good idea about the average error of predictions.
- MAE – Mean Absolute error
  - Also good. Not as popular as RMSE

Alternatives to RMSE

- If RMSE is 5 and average value of y is 10, then thats a lot.
- If RMSE is 5 and average value of y is 100, then its not.
- We can compare the RMSE calculated using the model predictions with the RMSE calculated using the mean value of y as the prediction.
- This gives us R^2, the coefficient of determination.

R^2 – The Coefficient of Determination (again)

- $R^2 = 1 - SSE/SSEusingMean$
- SSEusingMean is the SSE when using the mean of y as the prediction of y
- If the model is doing no better than the mean, $R^2 = 0$.
- If it is perfect (RMSE = 0), $R^2 = 1$
- $R^2 * 100$ gives the percentage of the variance of y that is explained by the model.
- (With simple regression, R^2 is the same as r^2 , where r is the correlation coefficient.)