Simple Linear Regression

MSc Statistics

Introduction

- Correlation and regression – for quantitative/numeric variables

  - Correlation: assessing the association between quantitative variables

  - Simple linear regression: description and prediction of one quantitative variable from another

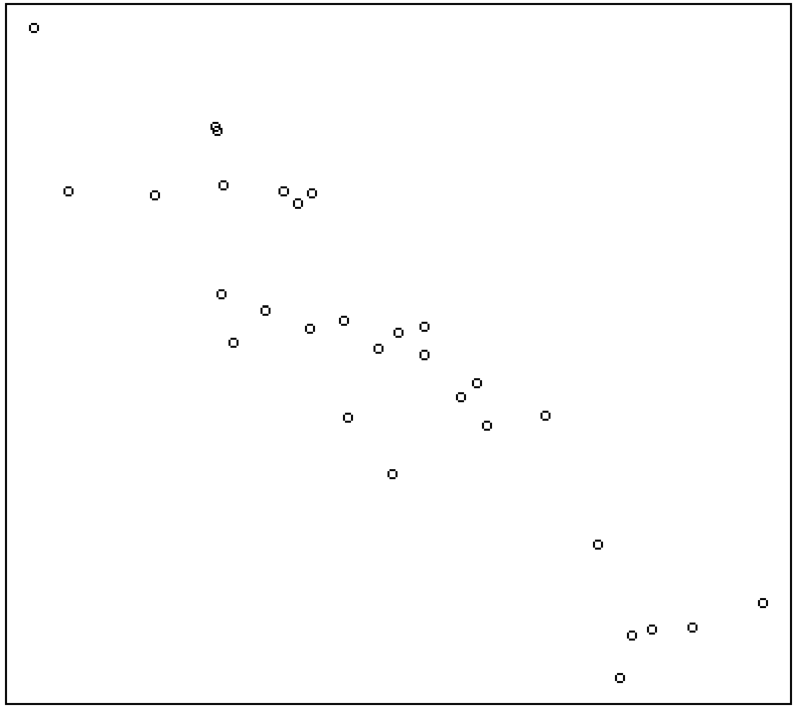
Introduction

- Simple linear regression: only considering linear (straight-line) relationships

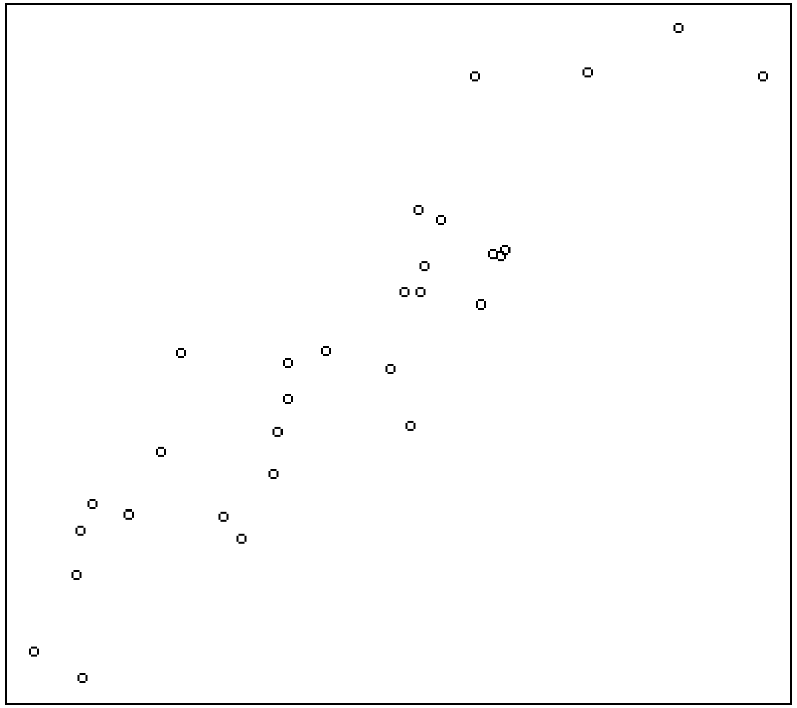
- When considering correlation or carrying out a regression analysis between two variables always plot the data on a scatter plot first

Scatter Plots

Linear

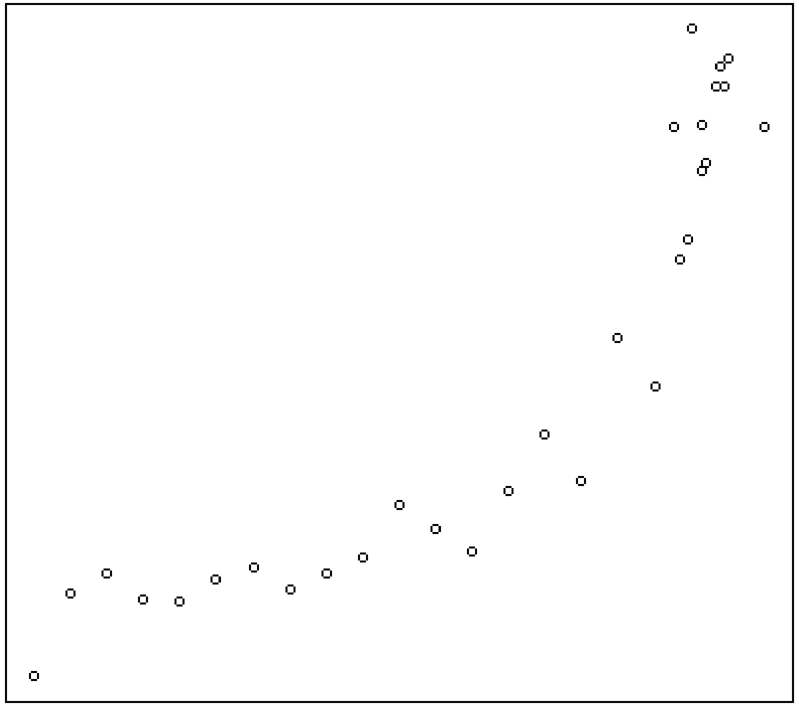


Linear

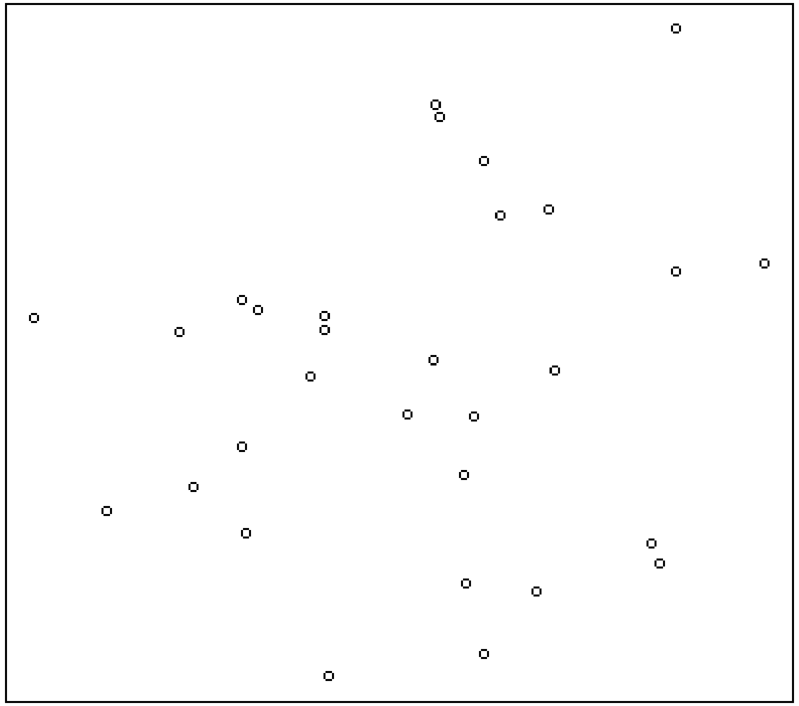


Scatter Plots

Non-Linear



No Relationship



Simple Linear Regression

- Data on two numerical variables
- Aim is to *describe* the relationship between the two variables and/or to *predict* the value of one variable when we only know the other variable
- Interested in a linear relationship between the two variables X and Y

Y	Predicted Variable	Dependent Variable	Response Variable	Outcome Variable
X	Predictor Variable	Independent Variable	Carrier Variable	Input Variable

Simple Linear Regression

- Simple linear regression - when there is only one predictor variable, which we will consider here
- Multiple or multivariate regression - when there is more than one predictor variable

Simple Linear Regression

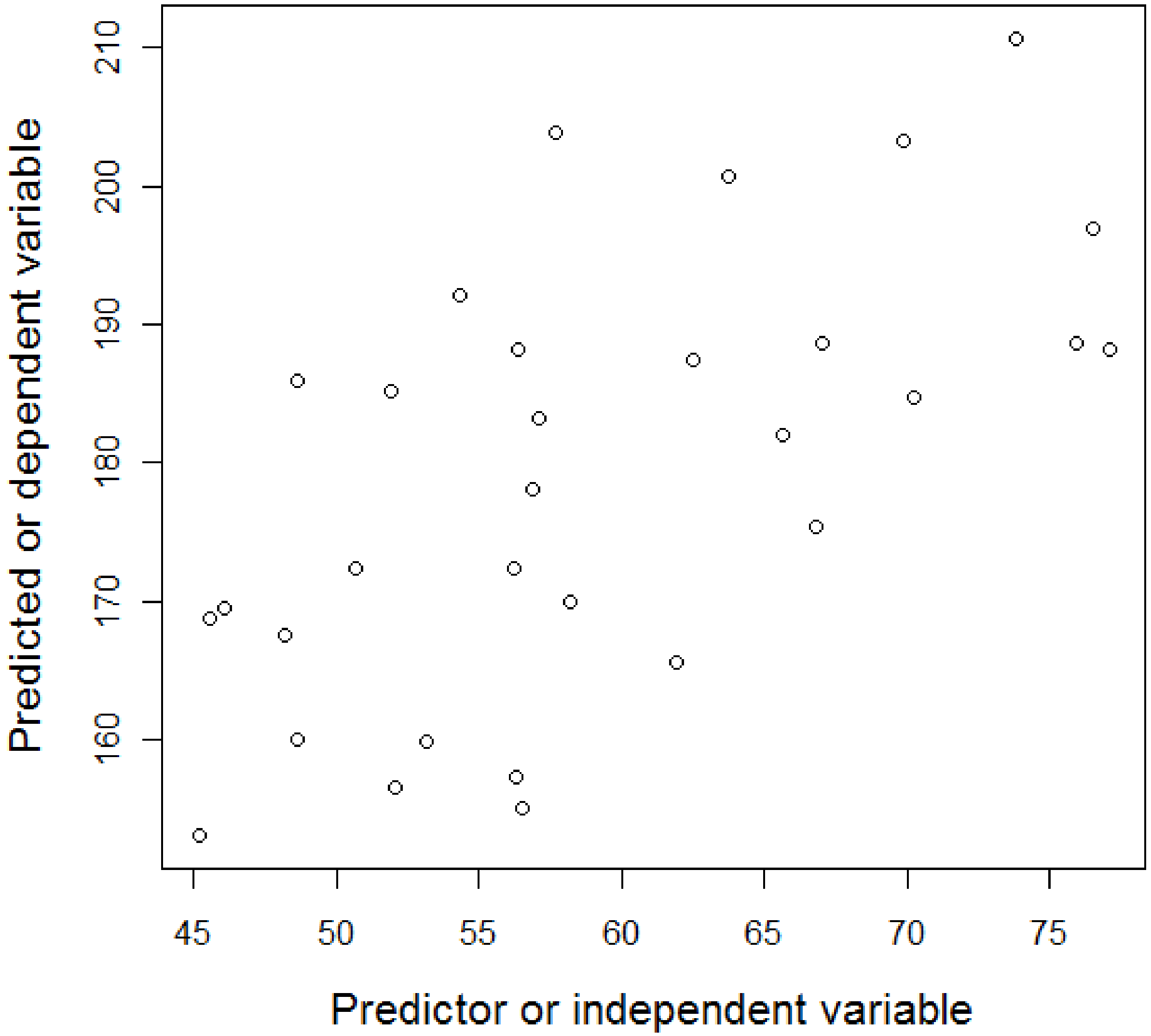
- The aim is to fit a straight line to the data that predicts the mean value of the dependent variable (*Y*) for a given value of the independent variable (*X*)

- Intuitively this will be a line that minimizes the distance between the data and the fitted line

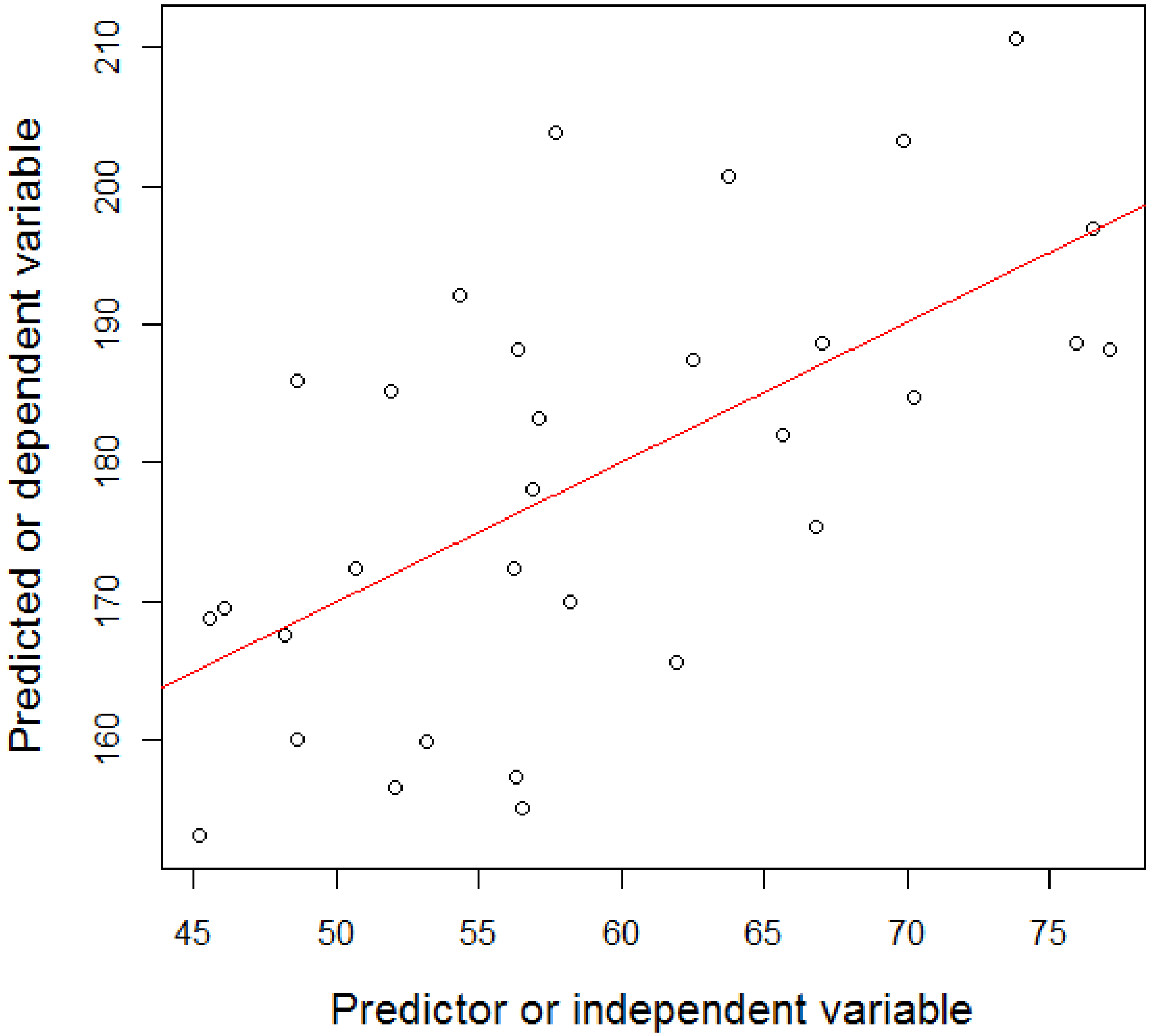
- Standard method is *least squares regression*

- Notation: n pairs of data points, (x_i, y_i)

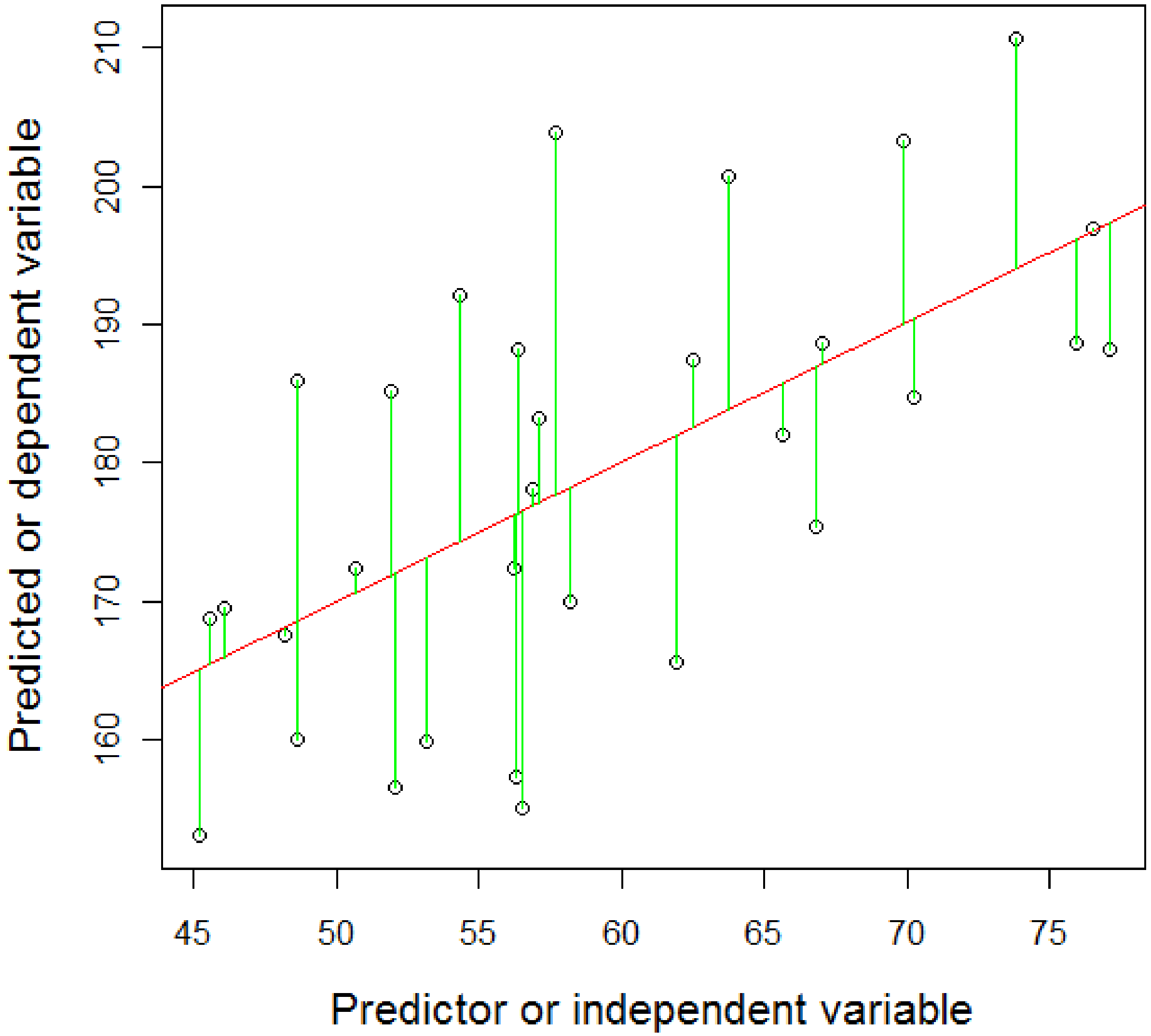
Two Quantitative Variables



Two Quantitative Variables, Regression Line



Two Quantitative Variables, Regression Line

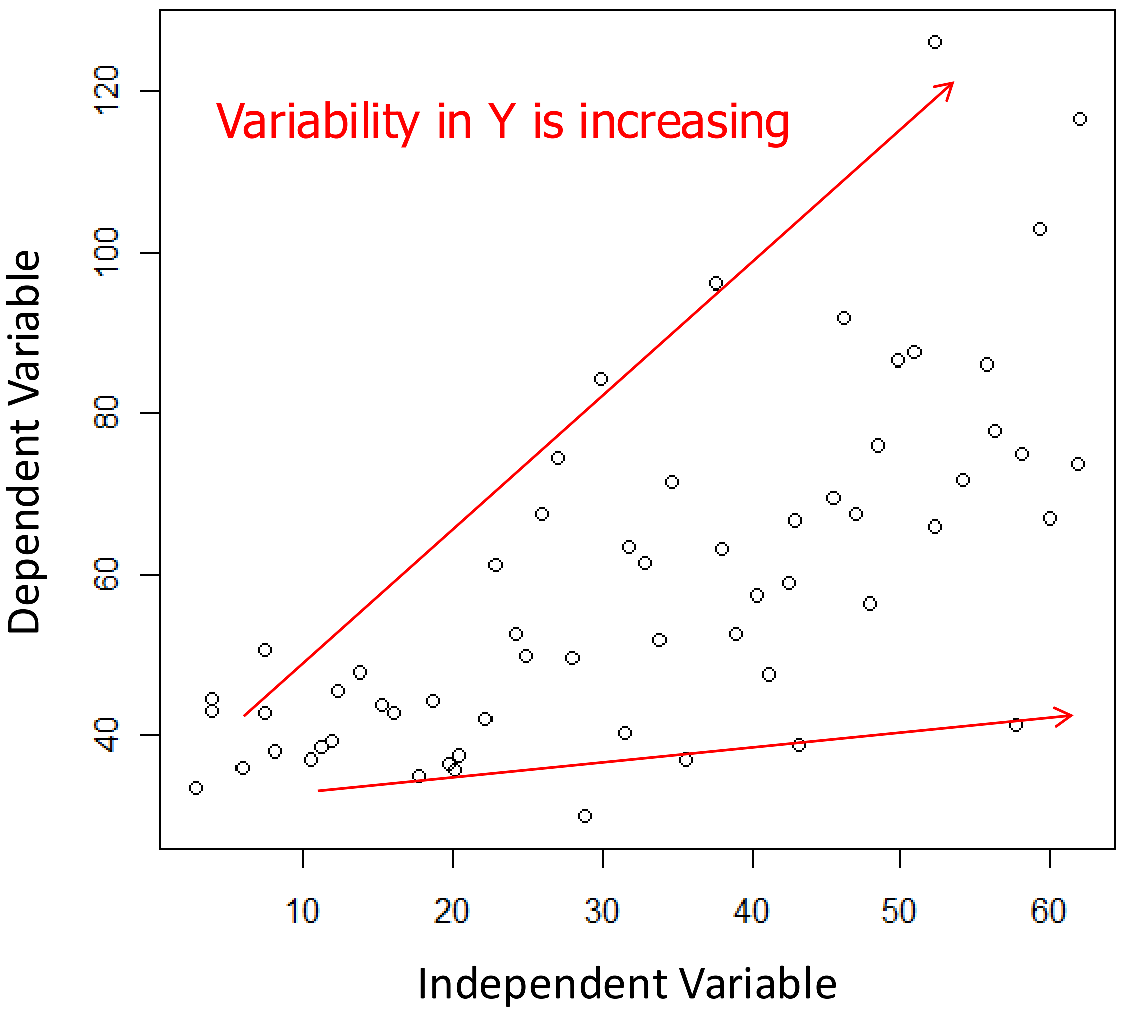


Linear Regression Assumptions

- The values of the dependent variable Y should be Normally distributed for each value of the independent variable X (needed for hypothesis testing and confidence intervals)

- The variability of Y should be the same for each value of X (homoscedasticity)

Linear Regression Assumptions



Linear Regression Assumptions

Other points to note:

- The relationship between the two variables should be linear

- The observations should be independent

- Values of *X* do not have to be random

- Values of *X* don't have to be Normally distributed

Linear Regression Assumptions

- It is easier to check many of these assumptions after the regression has been carried out
- Use residuals to do this and we will return to these later

Linear Regression Assumptions

- The straight line or linear relationship is described by the equation for a straight line

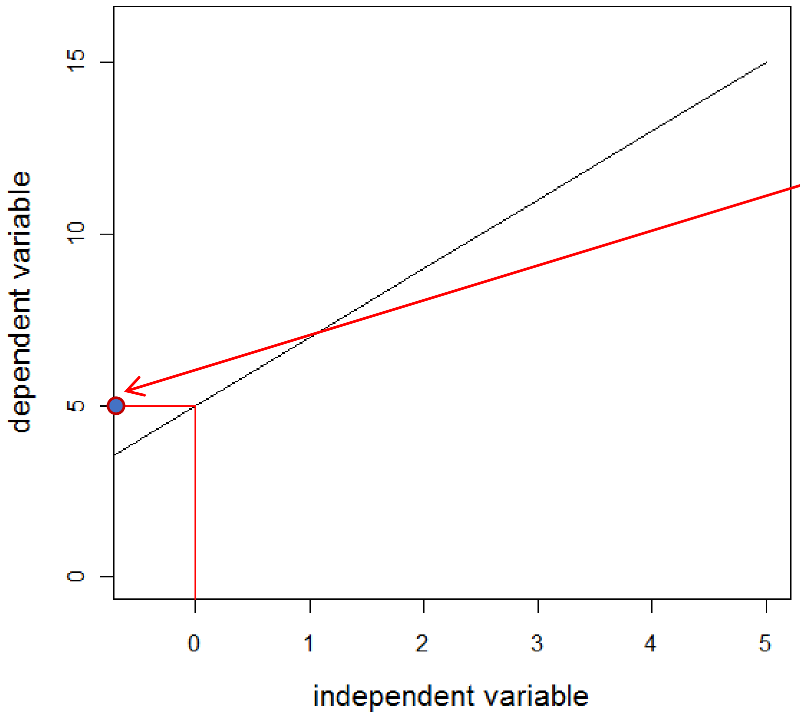
$$y = a + bx$$

Dependent variable

Intercept

Slope

Independent variable

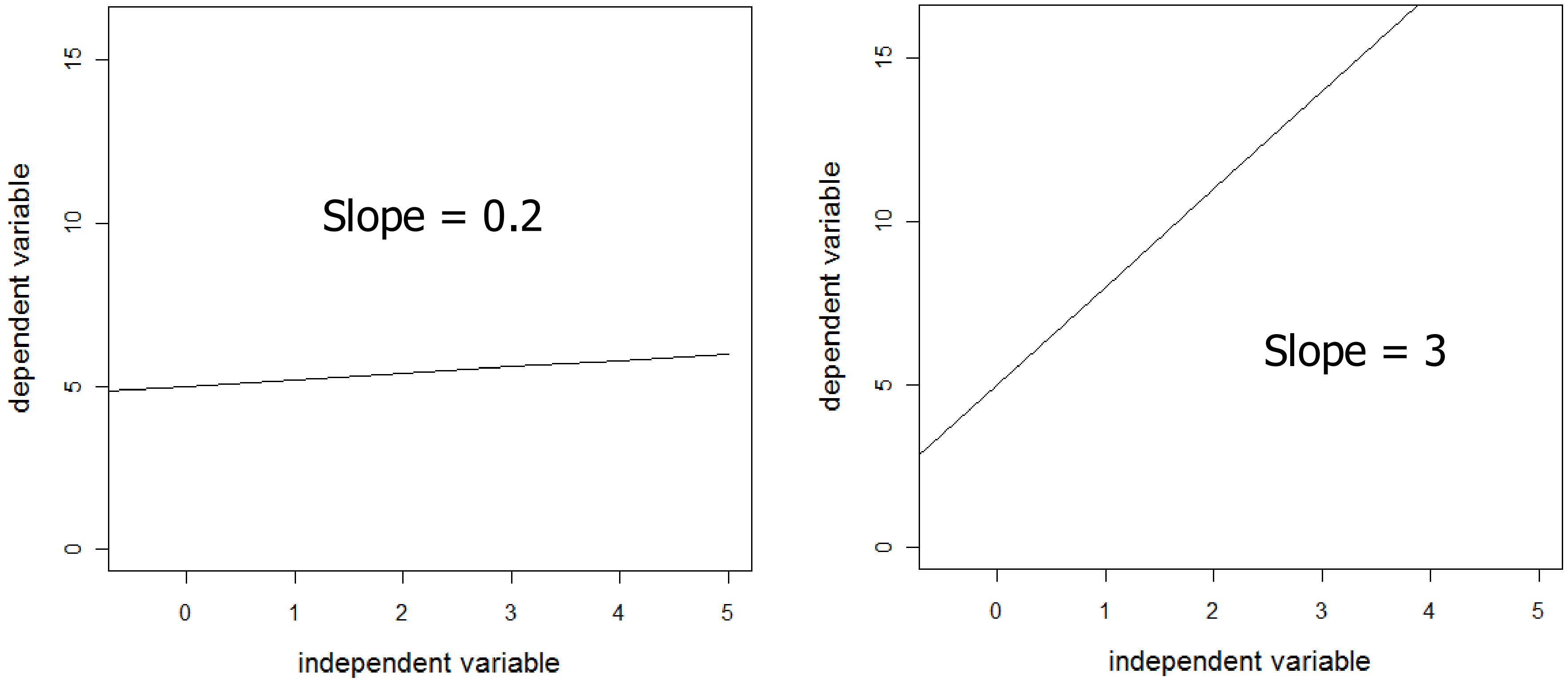


Intercept, value of y when x = 0,
a = 5, y = 5 + b*0

Slope of the line, b = 2 here

$y = 5 + 2x$

Slopes



Same intercept = 5

Least Squares Regression

- No line could pass through all the data points in our example

- We want the best “average” equation (*regression equation*) that would represent a line through the middle of the data, this is the *regression line:*

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

- The constants a, the intercept and b, the slope or regression coefficient are computed using the method of least squares

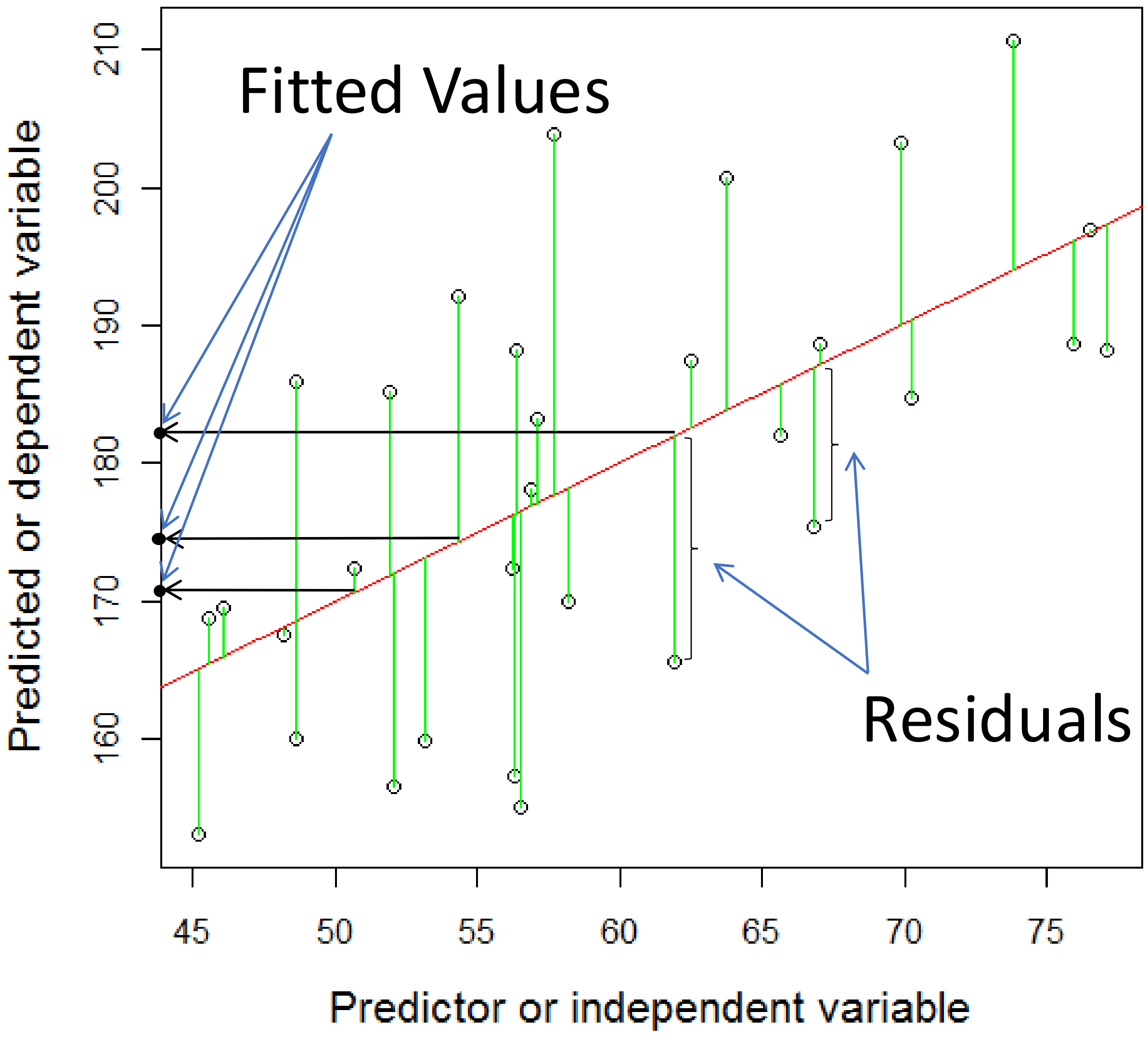
Least Squares Regression

- Fitted value = value of Y given by the line for any value of the variable X

- Residual = difference between the observed value of Y and the fitted value

- Least squares aim: to minimize the sum of squares of the residuals

Fitted Values and Residuals



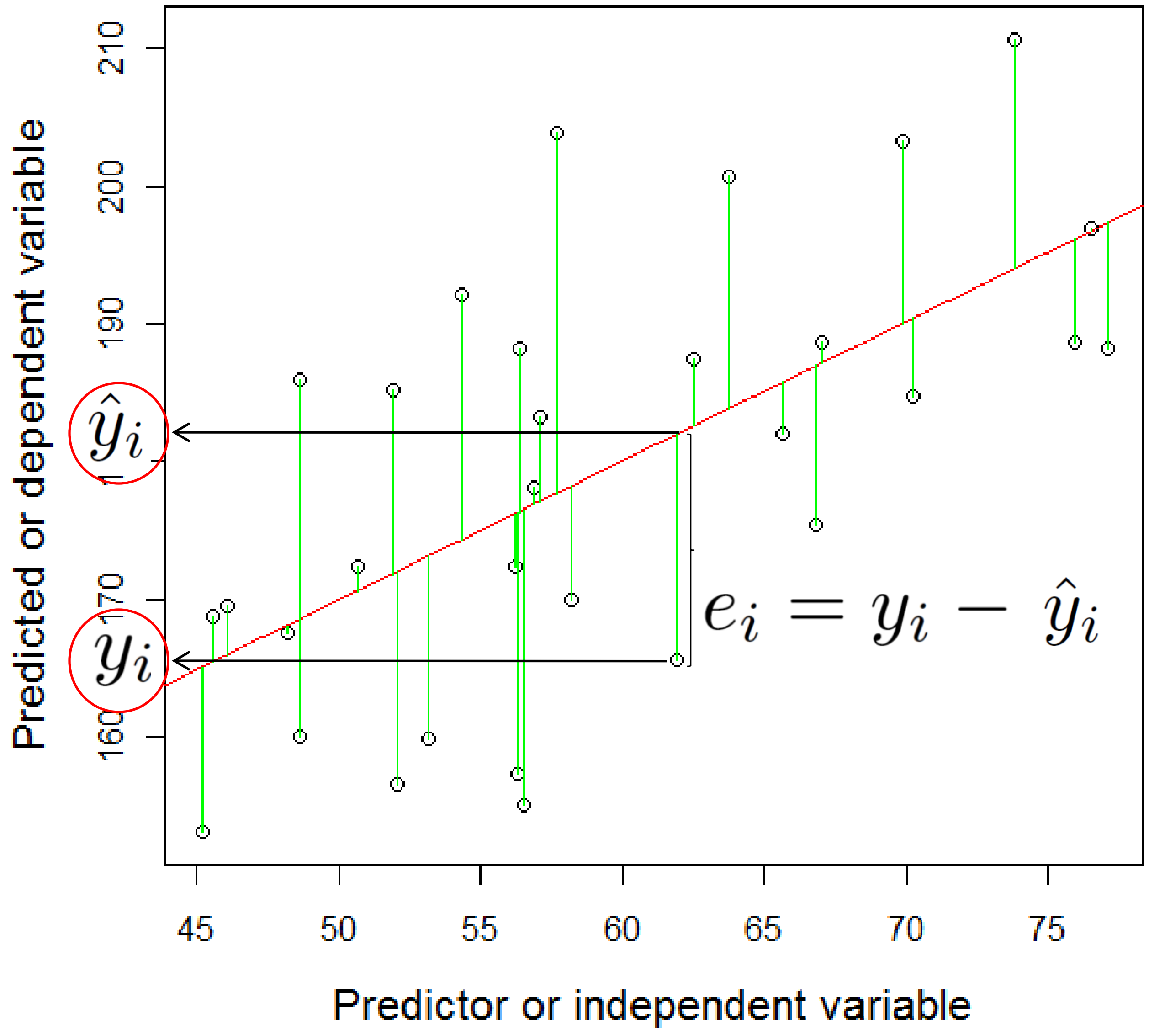
Least Squares Regression

- At any point x_i, the corresponding point on the line is given by: $a + bx_i$

Regression equation: $\hat{y}_i = \hat{a} + \hat{b}x_i$

Residuals (errors): $e_i = y_i - \hat{y}_i$

Fitted Values and Residuals



Least Squares Regression

- Linear model:

$$y_i = \hat{a} + \hat{b}x_i + e_i, \quad e_i \sim \mathrm{Normal}(0, \sigma^2)$$

- Note: if the errors/residuals are correlated or have unequal variances then least squares is not the best way to estimate the regression coefficient

Least Squares Regression

- Minimize the sum of squares *(S)* of the vertical distances of the observations from the fitted line (residuals)

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

- In order to find the intercept and regression coefficient that minimize S the mathematical technique of differentiation is employed

Least Squares Regression

- The solution for these two equations results in the following two formulae for the estimates of the intercept and regression coefficients respectively:

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n(\bar{x})^2}$$

- For the systolic blood pressure and age data in the previous plots:

$\hat{a}$ = 118.7

$\hat{b}$ = 1.0

x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
49	186	168.7	17.3
46	169	165.7	3.3
58	170	177.9	-7.9
53	160	172.8	-12.8
:	:	:	:

Residuals

- Checking assumptions:

✓ Linearity: A linear relationship between the dependent variable and the independent variables.

(scatterplot of x and y)

✓ Normal Distribution: Residuals are normally distributed with mean zero.

(histogram or qqplot of residuals)

✓ Constant Variance: The variance of the residuals are similar across the values of the independent variables.

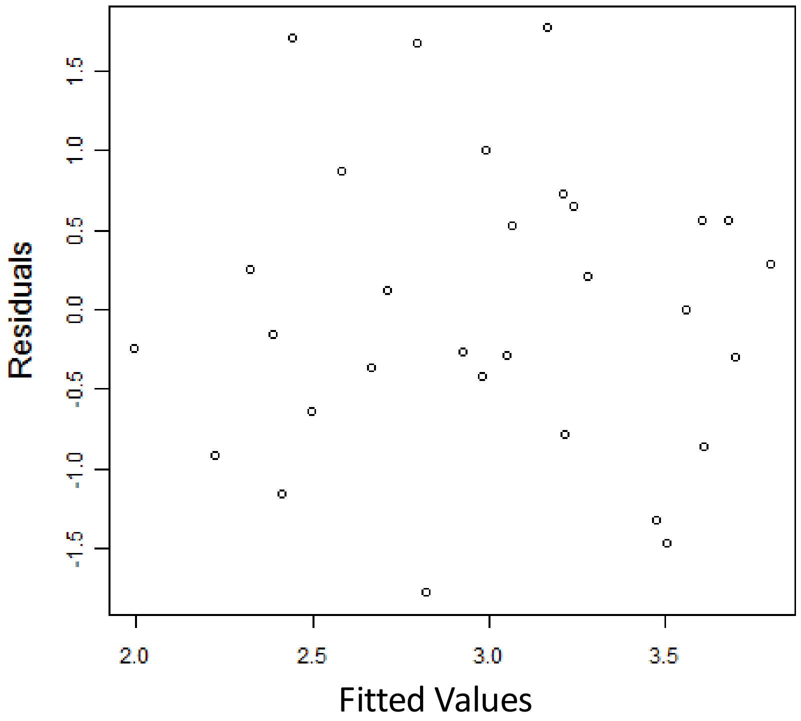
(scatterplot of residuals vs fitted values: constant spread)

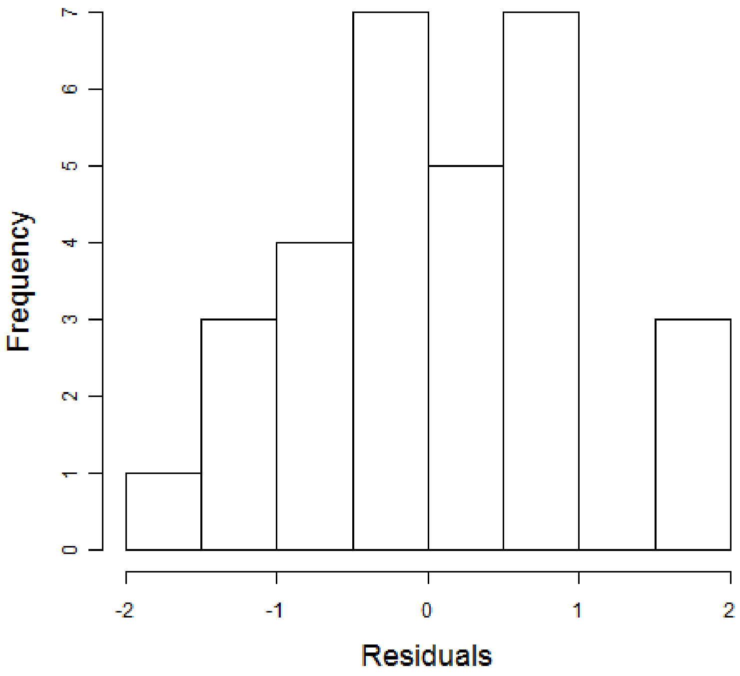
✓ i.i.d: Residuals are independently and identically distributed – random scatter.

(scatterplot of residuals vs fitted values: random scatter)

Residuals

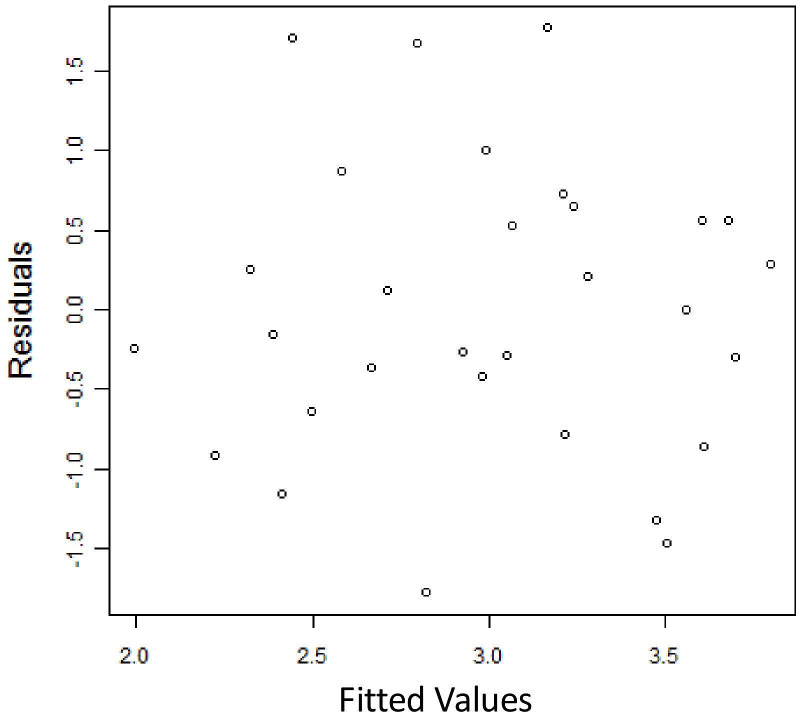
- These residuals appear reasonable

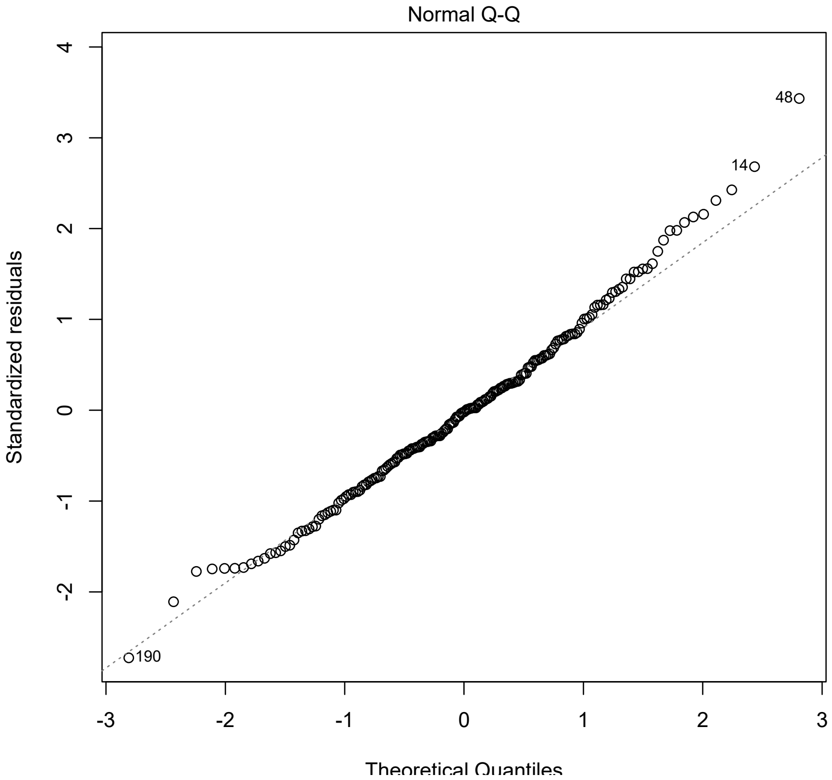




Residuals

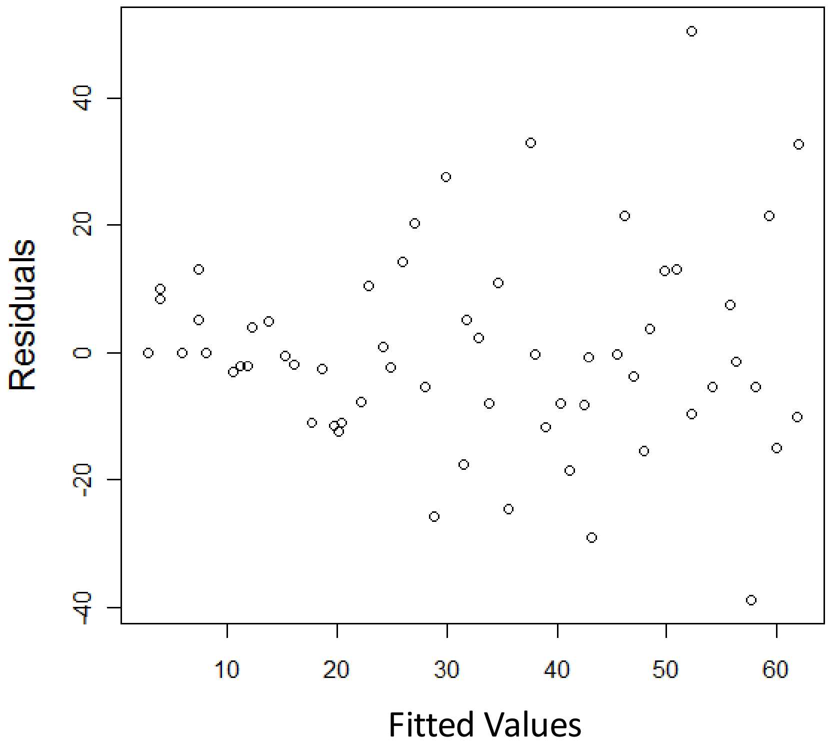
- These residuals appear reasonable

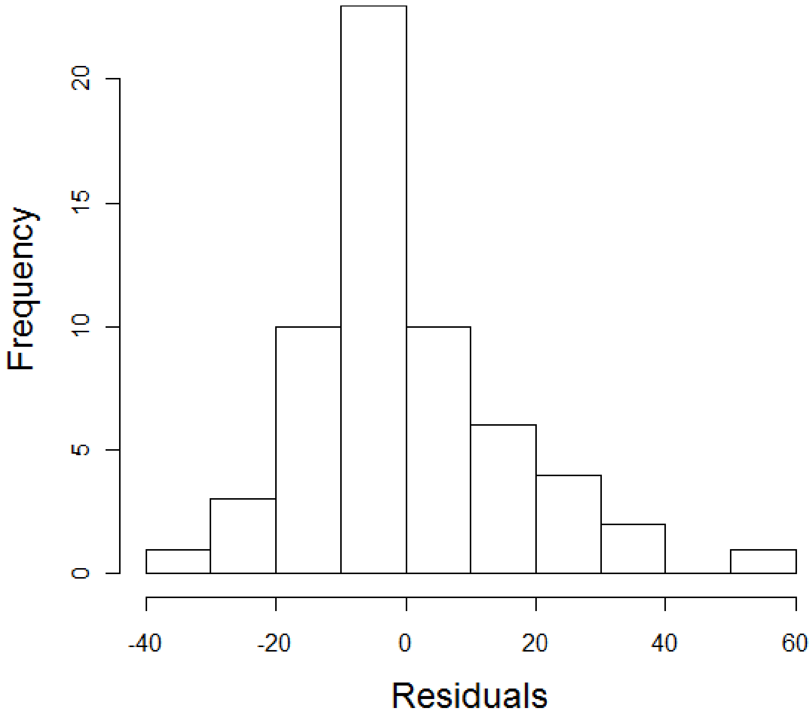




Residuals

- These residuals show increasing variability





Regression Coefficient b

- Regression coefficient:
    - this is the slope of the regression line
    - indicates the strength of the relationship between the two variables
    - interpreted as the expected change in y for a one-unit change in x

Regression Coefficient b

- Regression coefficient:
  - can calculate a standard error for the regression coefficient
  - can calculate a confidence interval for the coefficient
  - can test the hypothesis that $b = 0$, i.e., that there is no relationship between the two variables

Regression Coefficient b

- To test the hypothesis that $b = 0$, testing the hypothesis that there is no relationship between the *X* and *Y* variables, the test statistic is given by:

$$t = \frac{b}{se(b)}$$

  comparing this ratio with a *t* distribution with n-2 degrees of freedom

- Can also calculate a confidence interval for b:

$$b \pm t_{0.975} se(b)$$

Intercept a

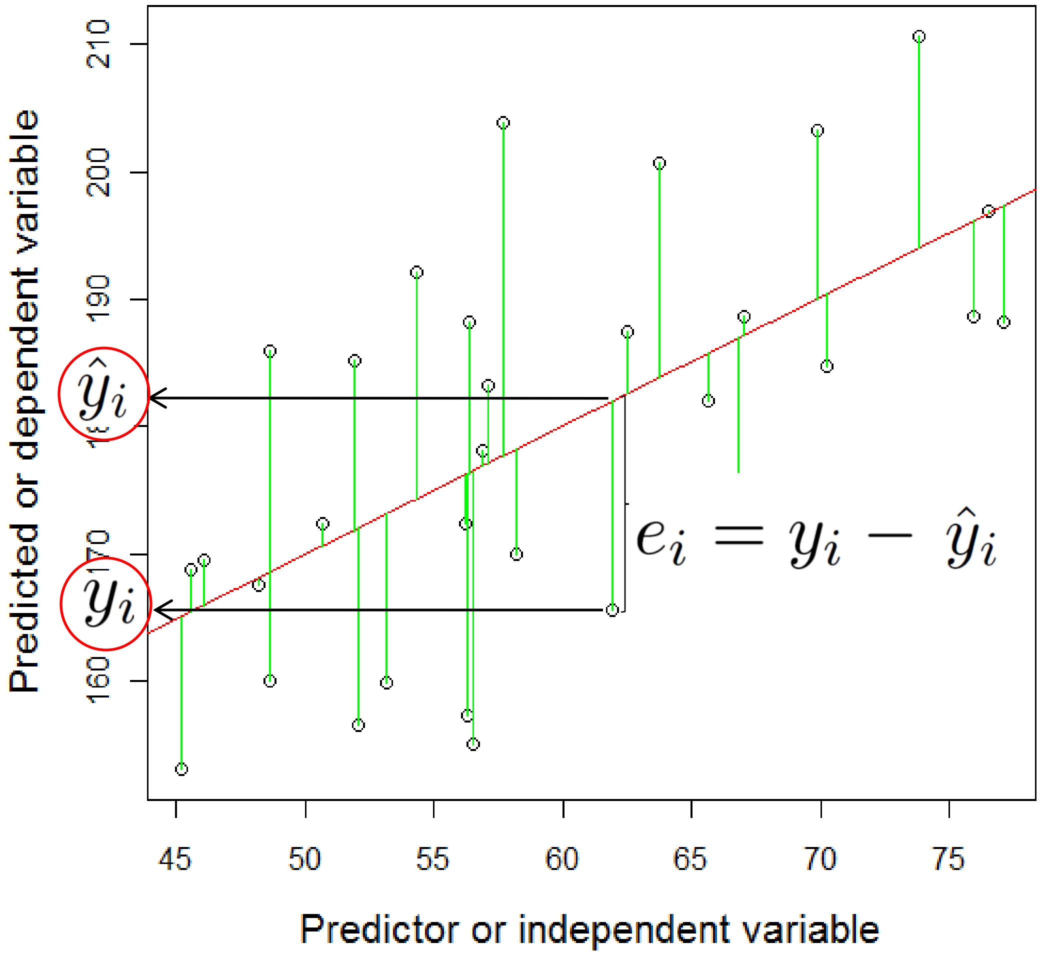
- Intercept:
  - the estimated intercept a gives the value of y that is expected when $x = 0$
  - often not very useful as in many situations it may not be realistic or relevant to consider $x = 0$
  - it is possible to get a confidence interval and to test the null hypothesis that the intercept is zero and most statistical packages will report these

Coefficient of Determination, R-Squared

- The coefficient of determination or R-squared is the amount of variability in the data set that is explained by the statistical model
- Used as a measure of how good predictions from the model will be
- In linear regression R-squared is the square of the correlation coefficient

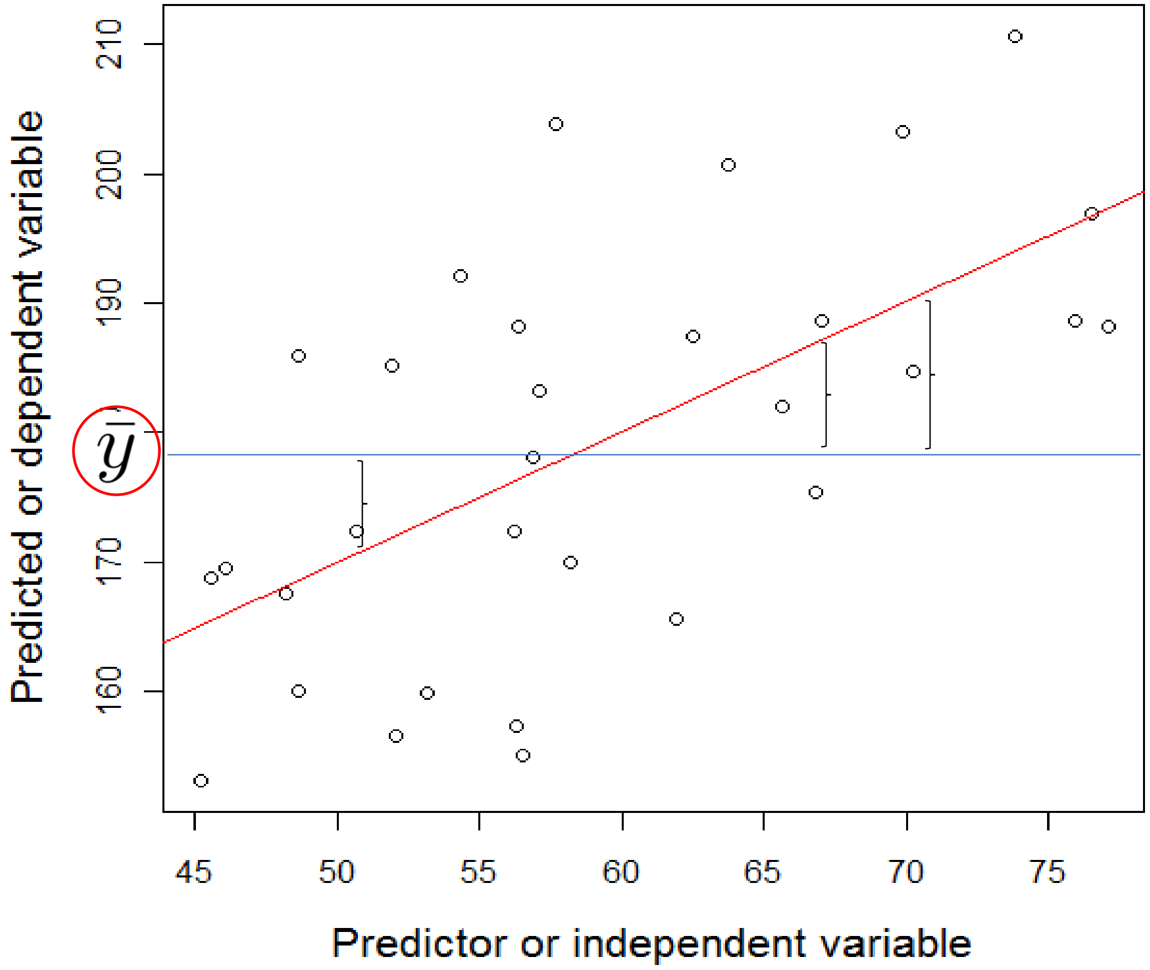
Residual Sum of Squares

$$\sum_i (y_i - \hat{y}_i)^2$$



Regression Sum of Squares

$$\sum_i (\hat{y}_i - \bar{y})^2$$

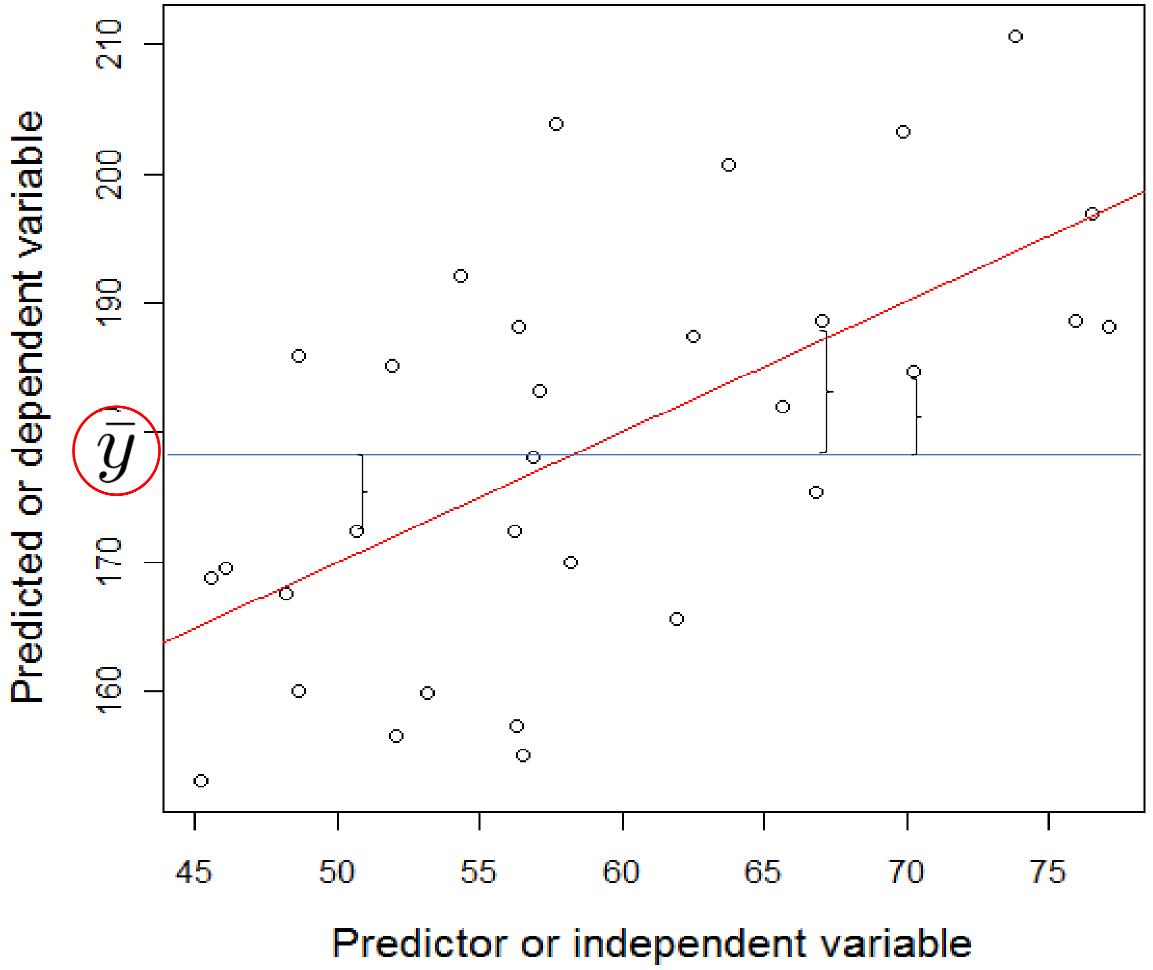


Here $\bar{y}$ = 179

Total Sum of Squares

$$\sum_i (y_i - \bar{y})^2$$

Here $\bar{y}$ = 179



Sum of Squares

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

Total sum of squares	=	Residual sum of squares	+	Regression sum of squares
Total Variation	=	Unexplained Variation	+	Explained Variation

Coefficient of Determination

Total sum of squares	=	Residual sum of squares	+	Regression sum of squares
Total Variation	=	Unexplained Variation	+	Explained Variation

$$\text{Coefficient of Determination} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$= \frac{\text{Regression SS}}{\text{Total SS}}$$

Coefficient of Determination, R-Squared

- Coefficient of determination

  = R-Squared

  = R^2

  = R-Sq

Coefficient of Determination, R-Squared

- R-Sq must lie between 0 and 1

- If it is equal to one then all the observed points must lie exactly on a straight line – no residual variability

- Often expressed as a percentage

- High R-squared says that the majority of the variability in the data is explained by the model (good!)

Adjusted R-Squared

- Sometimes an adjusted R-squared will be presented in the output as well as the R-squared

- Adjusted R-squared is a modification to the R-squared adjusting for the sample size and for the number of explanatory or predictor variables in the model (more relevant when considering multiple regression)

- The adjusted R-squared will only increase if the addition of the new predictor improves the model more than would be expected by chance

Mean Squared Error

- The **Mean Squared Error** is the mean of the squares of the errors:

$$\text{MSE} = \frac{1}{n-2}\sum_i (yi \ - \ \hat{y}_i)^2 = \frac{\text{Residual Sum of Squares}}{\text{DF}_{\text{Error}}}$$

- Degrees of Freedom in Simple Linear regression are n-2
- Lose one degree of freedom for each parameter estimated

Residual Standard Deviation

- Remember the linear model formulation:

$$y_i = \hat{a} + \hat{b}x_i + e_i, \quad e_i \sim \text{Normal}(0, \sigma^2)$$

- The <u>residual standard deviation</u> is an estimate of the standard deviation of the residuals
- The **residual standard error** is the positive square root of the mean square error

$$\text{Residual Standard Error} = \sqrt{\frac{1}{DF_{Error}} \sum_i (yi - \hat{y}_i)^2}$$

- Measures the spread of the y values about the regression line

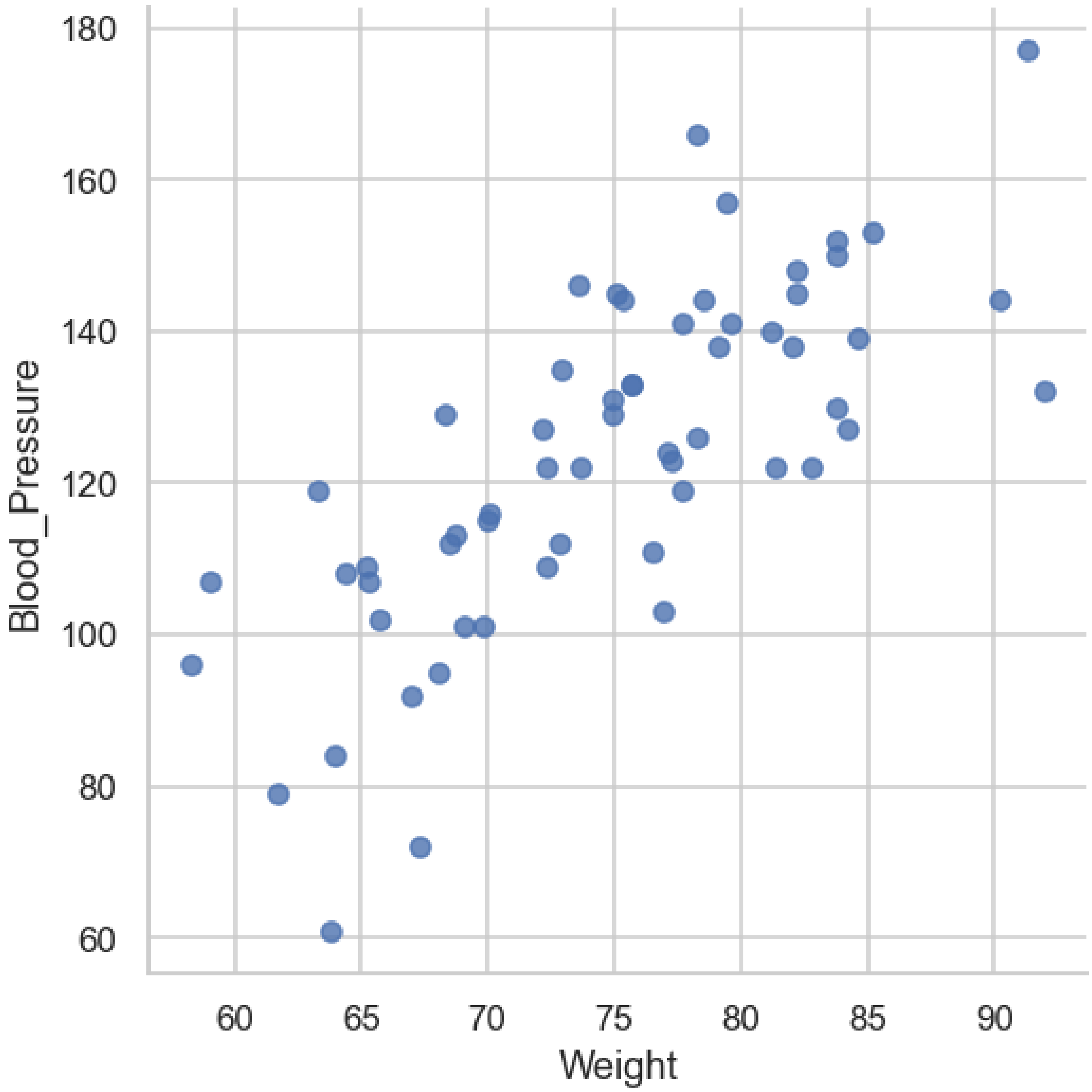
Residual Standard Deviation

- The residual standard deviation is a goodness-of-fit measure
- The smaller the residual standard deviation the closer the fit to the data

Regression: Python Example

- Data has been collected on individuals weights and blood pressure, to see if there is relationship between the two and can one be used to predict the other

- Which variable is the response and which is the predictor?



180
160
140
120
100
80
60
Blood_Pressure
60
65
70
75
80
85
90
Weight

Regression: Python Example

- Data has been collected on individuals weights and blood pressure, to see if there is relationship between the two and can one be used to predict the other

- Which variable is the response and which is the predictor?

```
X = df["Weight"]
X = sm.add_constant(X)
y = df["Blood_Pressure"]
model = sm.OLS(y, X).fit()
model.summary()
```

OLS Regression Results

```
==============================================================================
Dep. Variable:                   Blood_Pressure   R-squared:                 0.560
Model:                                      OLS   Adj. R-squared:            0.552
Method:                           Least Squares   F-statistic:               73.71
Date:                          Wed, 01 Jan 2020   Prob (F-statistic):     6.44e-12
Time:                                  12:59:00   Log-Likelihood:        -247.03
No. Observations:                            60   AIC:                       498.1
Df Residuals:                                58   BIC:                       502.2
Df Model:                                     1
Covariance Type:                      nonrobust
==============================================================================
              coef    std err        t       P>|t|     [0.025     0.975]
------------------------------------------------------------------------------
const     -37.1113     18.824    -1.971      0.053    -74.792      0.569
Weight      2.1499      0.250     8.586      0.000      1.649      2.651
==============================================================================
Omnibus:                          1.448   Durbin-Watson:             2.317
Prob(Omnibus):                    0.485   Jarque-Bera (JB):          1.095
Skew:                            -0.331   Prob(JB):                  0.578
Kurtosis:                         3.014   Cond. No.                   726.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
"""
```

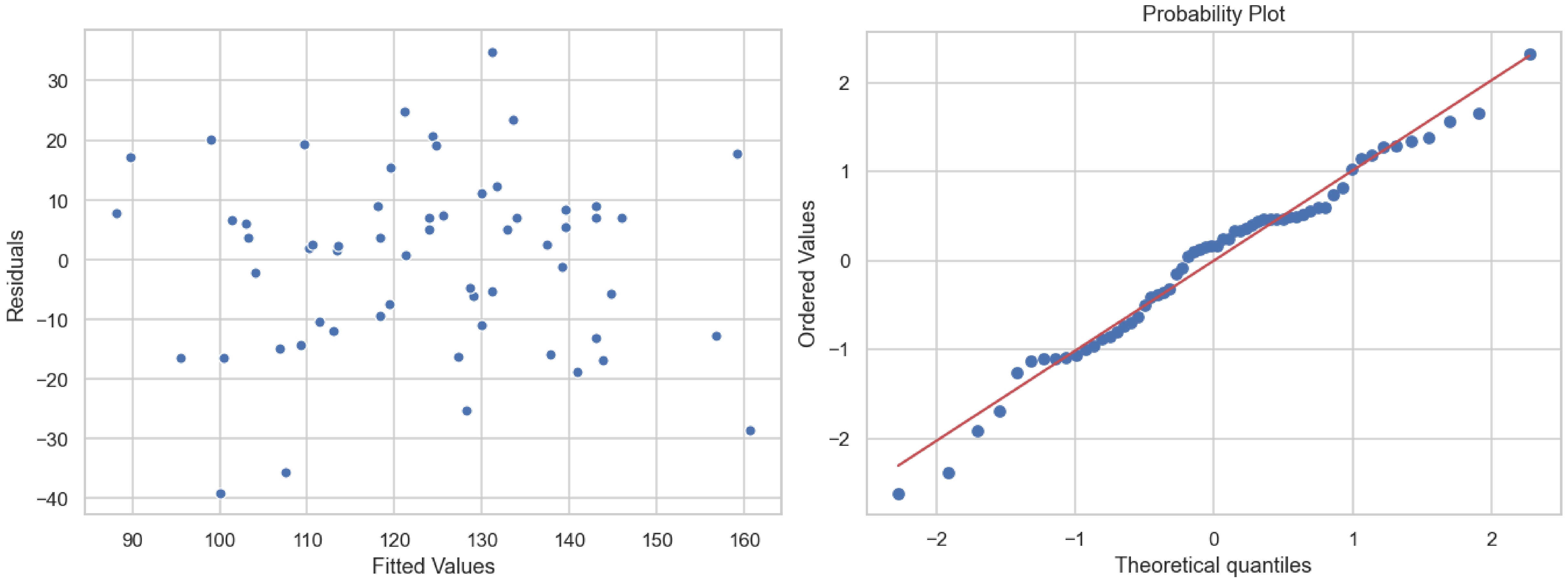
Residual plots for Checking Assumptions

```
#get the fitted values from the model output
fitted_values=model.fittedvalues
#get the residual (error) values from the fitted values and actual observations (y)
residual = y - fitted_values

#plot scatterplot to see residuals vs fits to check iid and equal variance assumptions
sns.scatterplot(x=fitted_values,y=residual)
plt.xlabel("Fitted Values")
plt.ylabel("Residuals")

#needed for qqplot to check normality of standardised residuals
import statistics
import scipy.stats as stats
#create standardised residuals
sd_red=(residual-statistics.mean(residual))/statistics.stdev(residual)
#create qqplot
stats.probplot(sd_red, plot=sns.mpl.pyplot)
```

Assumption Checking Output



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Blood_Pressure   R-squared:                       0.560
Model:                            OLS   Adj. R-squared:                  0.552
Method:                 Least Squares   F-statistic:                     73.71
Date:                Wed, 01 Jan 2020   Prob (F-statistic):           6.44e-12
Time:                        12:59:00   Log-Likelihood:                -247.03
No. Observations:                  60   AIC:                             498.1
Df Residuals:                      58   BIC:                             502.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -37.1113     18.824     -1.971      0.053     -74.792       0.569
Weight         2.1499      0.250      8.586      0.000       1.649       2.651
==============================================================================
Omnibus:                        1.448   Durbin-Watson:                   2.317
Prob(Omnibus):                  0.485   Jarque-Bera (JB):                1.095
Skew:                          -0.331   Prob(JB):                        0.578
Kurtosis:                       3.014   Cond. No.                         726.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
"""
```

R-Sq says that 56% of the variability in the data is explained by the model

Estimate of the intercept, *a*

Estimate of the regression coefficient, *b*

P-value of regression coefficient, *b*

Residual Standard Deviation in Python

- summary does not give residual standard error which can be used as a measure of fit when comparing models, but can get it from the output from OLS

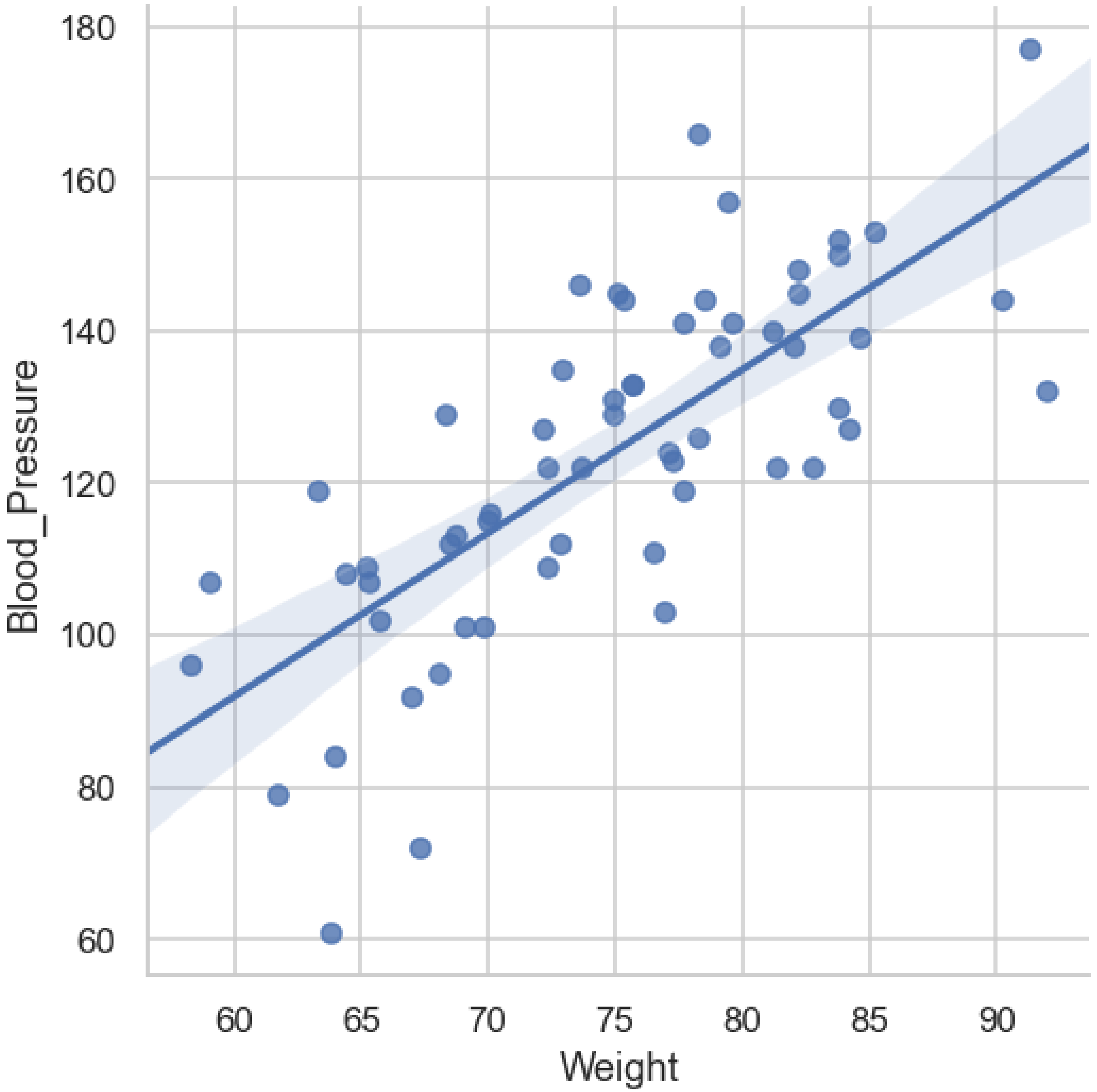
```
#mean square error using scale
In [830]: model.scale
Out[830]: 228.19489603432334

#Take square root of mean square error to get residual standard error
In [831]: np.sqrt(model.scale)
Out[831]: 15.10612114456664
```

Confidence Interval on Fitted Values

- Can calculate a confidence interval on the fitted value: $\hat{y}_i$
- This is a confidence interval for the mean value of y, given a value of x
- The width of the confidence interval depends on the value of x_i and will be a minimum at $x_i = \bar{x}$ and will widen as $|x_i - \bar{x}|$ increases

95% Confidence Interval



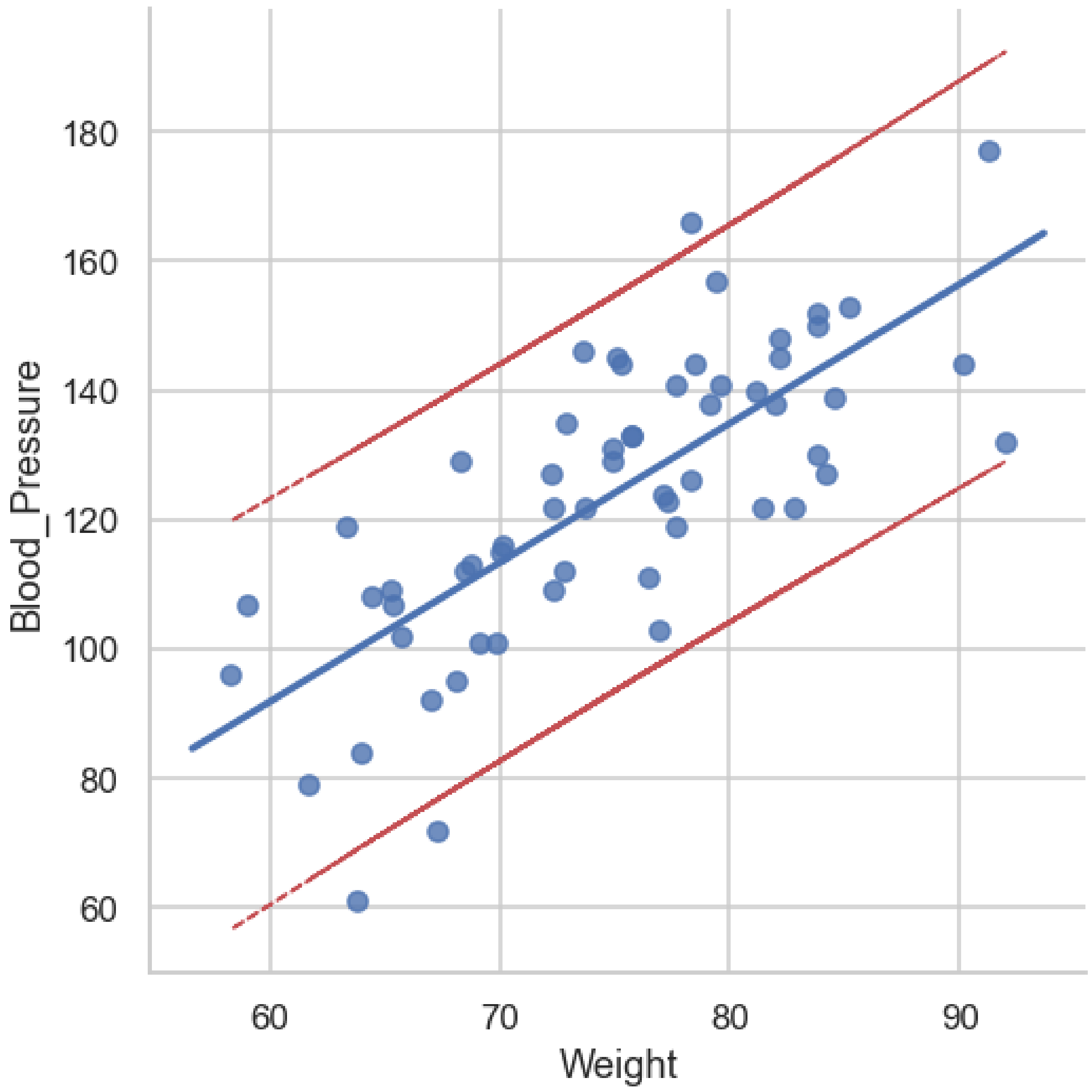
Prediction Interval for Future Values

- Can predict the range of possible values of y for a new independent value of x not used in the regression model
- The prediction interval describes the spread of the observations around the mean value: $\hat{y}_i$
- The prediction interval is wider than the confidence interval
- The interval widens with distance from the mean value of x, but is not so obvious to see

```
In [810]: predictions = model.get_prediction()
     ...: pred_df=predictions.summary_frame(alpha=0.05)
     ...: pred_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 60 entries, 0 to 59
Data columns (total 6 columns):
# Column Non-Null Count Dtype
--- ------ -------------- -----
0 mean 60 non-null float64
1 mean_se 60 non-null float64
2 mean_ci_lower 60 non-null float64
3 mean_ci_upper 60 non-null float64
4 obs_ci_lower 60 non-null float64
5 obs_ci_upper 60 non-null float64
dtypes: float64(6)
memory usage: 2.9 KB

#add prediction interval lines to the plot instead confidence interval lines
In [811]: sns.lmplot(x='Weight', y='Blood_Pressure',data=df,ci=0)
     ...: plt.plot(df['Weight'], pred_df['obs_ci_lower'], 'r--', lw=1)
     ...: plt.plot(df['Weight'], pred_df['obs_ci_upper'], 'r--', lw=1)
Out[811]: [<matplotlib.lines.Line2D at 0x140cecb20>]
```

95% Prediction Interval



Interpolation and Extrapolation

- *Interpolation*
  - Making a prediction for *Y* within the range of values of the predictor *X* in the sample used in the analysis
  - Generally this is fine

- *Extrapolation*
  - Making a prediction for *Y* outside the range of values of the predictor *X* in the sample used in the analysis
  - No way to check linearity outside the range of values sampled, not a good idea to predict outside this range

Correlation and Regression

- Correlation only indicates the strength of the relationship between two variables, it does not give a description of the relationship or allow for prediction

- The t-test of the null hypothesis that the correlation is zero is exactly equivalent to that for the hypothesis of zero slope in the regression analysis

```
corr= pearsonr(df["Blood_Pressure"], df["Weight"])
corr
Out[812]: (0.7480937239999523, 6.444357622401537e-12)
```

Correlation and Regression

- For correlation both variables must be random, for regression *X* does not have to be random
- Correlation is often over used
- One role for correlation is in generating hypotheses, remember correlation is based on one number, limit to what can be inferred with one number

Summary I

- Simple linear regression- describe and predict linear relationship

- Least squares regression

- Assumptions:
  - Linearity: A linear relationship between the dependent variable and the independent variables.
  - Normal Distribution: Residuals are normally distributed with mean zero.
  - Constant Variance: The variance of the residuals are similar across the values of the independent variables.
  - i.i.d: Residuals are independently and identically distributed –random scatter.

Summary II

- Need to be familiar with:
  - Regression coefficient (slope)
  - Intercept
  - Residuals – Normal$(0, \sigma^2)$
  - Fitted Value
  - R^2 (coefficient of determination)
  - Residual standard deviation

- Confidence and Prediction Intervals

- Interpolation and Extrapolation