Predictive Modeling: Regression Models

Arun Kumar

Department of Mathematics, IIT Ropar

August 6, 2025

Correlation

Definition (Correlation)

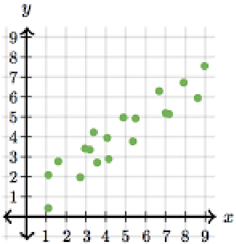
Correlation is a statistical measure that expresses the extent to which two variables are linearly related. Alternatively, correlation is a measure of linear association of two variables. Consider the data set of paired values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The correlation coefficient is given by

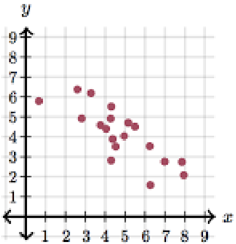
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

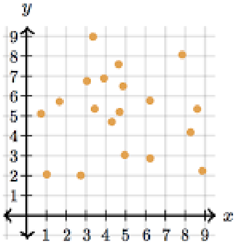
where

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- When $r > 0$, it is said that the sample data pairs are positively correlated.
- When $r < 0$, we say that they are negatively correlated.
- The sample correlation coefficient $r \in [-1, +1]$.
- The sample correlation coefficient r will equal $+1$ if, for some constant a, $y_i = a + bx_i, i = 1, \ldots, n$ where b is a positive constant.
- The sample correlation coefficient r will equal -1 if, for some constant a, $y_i = a + bx_i, i = 1, \ldots, n$ where b is a negative constant.


y
9
8
7
6
5
4
3
2
1
x
1 2 3 4 5 6 7 8 9


y
9
8
7
6
5
4
3
2
1
x
1 2 3 4 5 6 7 8 9


y
9
8
7
6
5
4
3
2
1
x
1 2 3 4 5 6 7 8 9

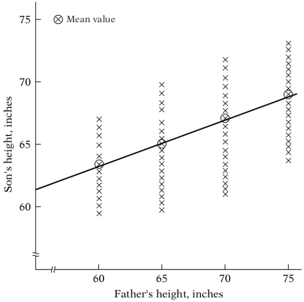
Simple Linear Regression

Simple linear regression is a statistical technique that allows us to study relationships between two continuous (quantitative) variables:

- the one variable is denoted by x, is called as independent, input, feature, explanatory or predictor variable.
- the another variable denoted by y, is called as dependent, output, response or predicted variable.

History of Regression

- The term regression was introduced by Francis Galton.
- Galton found that the average height of Children born of parents of a given height tended to “regress" towards the average height in the population as a whole.
- Alternatively, the height of the children of unusually tall or unusually short parents tends to move towards the average height of the population. In the words of Galton, this was “regression to mediocrity".



The Modern Interpretation of Regression

To predict the value of one dependent variable on one or several other independent variables.

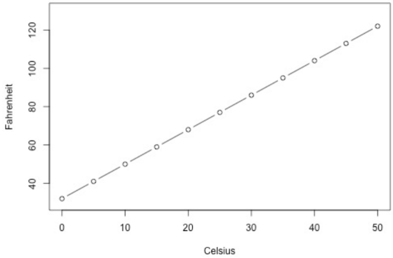
Example

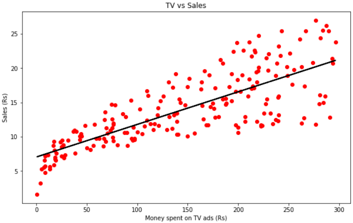
- Does expenditure depends on income
- Does crop yield depends on fertilizers, rain fall, sunshine, soil type etc
- Does technology equity price depends on some technological innovation

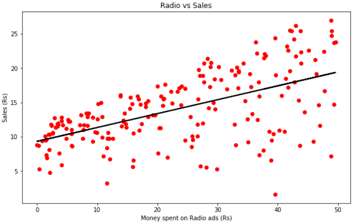
Statistical vs Deterministic Relationships

- In regression, we are concerned with the statistical not deterministic, dependence among variables.
- The dependence of crop yield on temperature, rainfall, sunshine and fertilizer is statistical in nature, since the feature variables, although certainly important, will not be able to predict the crop yield exactly because of error in measuring and host of other factors that affect the yield but are difficult to incorporate in the model.
- In deterministic phenomena, we study for example Newton's law, Ohm's law and Boyle's gas law etc.

Statistical vs deterministic relationships contd...







- Consider a set of n data points $\{(x_i, y_i)\}, 1 \leq i \leq n$.
- Suppose it is known that these points lie on a function of the form $y = f(x; m)$ where $f(\cdot)$ represents a function family and m represents a set of parameters.
- For example: $f(x)$ is a quadratic function of x, i.e. of the form $y = f(x) = ax^2 + bx + c$. In this case, $m = (a, b, c)$.

- In each case, we assume the function family form. But we do not know the function parameters, and would like to estimate these from $\{(x_i, y_i)\}, i = 1, 2, \cdots, n$.
- This is the problem of fitting a function to a set of points.

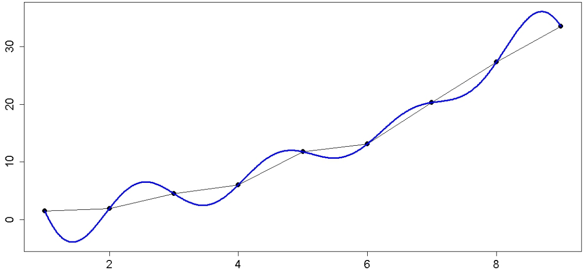


Figure: Piecewise Linear and Polynomial nterpolation

- In parametric regression, we want to find the parameters set m such that $f(x_i; m) \approx y_i,\ i = 1, 2, \cdots, n.$

- In interpolation, we want to fit some function such that $f(x_i; m) = y_i,\ i = 1, 2, \cdots, n.$

A model is said to be linear when it is linear in parameters. In such case $\frac{\partial y}{\partial \beta_j}$ should not depend on β_j's.

- $Y = \beta_0 + \beta_1 x$ is a linear model.
- $Y = \beta_0 x^{\beta_1}$ is a non-linear model.
- $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ can be transformed to linear model and hence is linear model.
- $Y = \beta_0 + \beta_1 x^{\beta_2}$ is non-linear model.

We consider the relationship

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where Y is the response variable and x is the predictor and ϵ is the error term. We can also write

$$\mathbb{E}(Y|x) = \beta_0 + \beta_1 x, \text{ since } \mathbb{E}(\epsilon) = 0.$$

Interpretation of β_0 and β_1

- For $x = 0$, we have $\mathbb{E}(Y|x = 0) = \beta_0$. Thus β_0 is the average value of Y when input variable $x = 0$.
- We have $\mathbb{E}(Y|x = 1) = \beta_0 + \beta_1$ and $\mathbb{E}(Y|x = 2) = \beta_0 + 2\beta_1$ which implies $\mathbb{E}(Y|x = 2) - \mathbb{E}(Y|x = 1) = \beta_1$.
- Thus β_1 is the change in the average value of Y per unit change in x.

Estimating β_0 and β_1

Suppose, we have n data points in the sample.

- y_i denotes the observed response for data point i
- x_i denotes the predictor value for data point i
- $\hat{y}_i$ is the corresponding predicted response.
- The difference $y_i - \hat{y}_i$ is called the prediction error.

We choose β_0 and β_1 such that the prediction error is minimum. That is, we need to find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Taking the derivative with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, setting to 0, and solving for $\hat{\beta}_0$ and $\hat{\beta}_1$ gives

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The resulted regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is often called as "least squares regression line“.

Significance and goodness of Fit

- The RSS is defined by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- The MSE is given by

$$MSE = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- R-squared is a goodness-of-fit measure for linear regression models.
- This statistic indicates the percentage of the variance in the dependent variable that the independent variable explains.
- Also, R-squared is defined by

$$R^2 = 1 - \frac{RSS}{TSS},$$

where RSS = residuals sum of squared and TSS = total sum of squared.

- Moreover, $R^2 = \rho^2$.

- We need t-value to be “large".
- We need p-value to be small preferably below 0.05.
- The F statistics should be “large".
- The significance F should be close to 0.
- R-squared close to 1 is better.

Linear Regression Diagnostics

1. The linear regression is sensitive to outliers and hence we assume that there are no outliers in the data.
2. The relationship between the independent and dependent variables is approximately linear.
3. The error term has 0 mean.
4. The error term has constant variance (also called homoscedasticity)
5. The errors are normally distributed.

Box and Whisker Plot for Outliers Detection

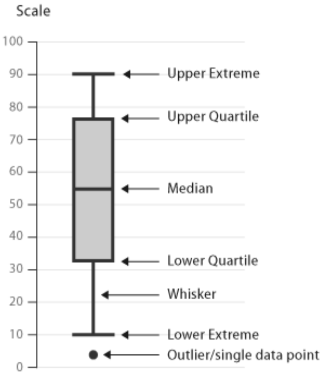


Figure: Box and Whisker Plot

Remark

The lower and upper extremes are the minimum and maximum values respectively, in the data set excluding outliers.

Outliers

The term “outlier"’ is not well defined. The definition may varies depending on the situation. In box plot (or box and whisker plot), the outliers are defined as any points which fall outside the interval $(Q_1 - 1.5 \times \mathrm{IQR}, Q_3 + 1.5 \times \mathrm{IQR})$, where IQR is inter quartile range.

Example

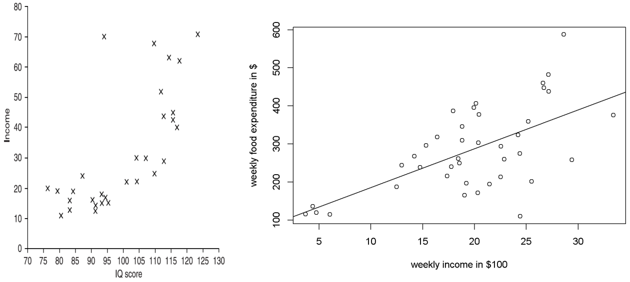
Find outlier if any in the data: 5, 2, 6, 3, 37, 4, 7, 4, 1, 8, 0.

Solution:

- The data in ascending order is: 0, 1, 2, 3, 4, 4, 5, 6, 7, 8, 37.
- $Q_1 = 2$ and $Q_3 = 7$ and IQR = $Q_3 - Q_1 = 5$.
- The outliers are the points beyond $(Q_1 - 1.5 \times \mathrm{IQR}, Q_3 + 1.5 \times \mathrm{IQR}) = (2 - 7.5, 7 + 7.5) = (-5.5, 14.5)$.
- Thus 37 is an outlier as per our definition.

Paired data

Sometimes a data set consists of pairs of values that have some relationship to each other. We often describe the *j*th pair by $(x_j, y_j), j = 1, \ldots, n$.



Testing Normality of Errors

The QQ plot is used to see how well a particular sample follows a particular theoretical distribution. The *q*-quantiles are values that partition a finite set of values into *q* subsets of (nearly) equal sizes. Essentially in Q-Q plot, the quantiles of data set and the theoretical distribution are plotted on *x* and *y* axis.

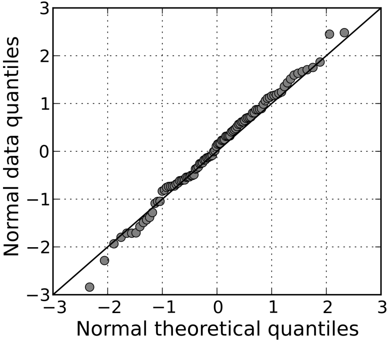


Figure: Q-Q Plot

Test of Normality

We can use the following tests before going for further analysis to check if our data is normal or not. Normality of data is the underlying assumption in several tests and models.

- Kolmogorov-Smirnov Test
- Anderson-Darling Test
- Jarque-Bera Test
- Shapiro-Wilk Test

```
from scipy import stats
from scipy.stats import anderson
result = anderson(data1.pH.values)
print('Statistic: %.3f' % result.statistic)
result.critical_values
```

```
Statistic: 11.972
```

```
: array([0.576, 0.655, 0.786, 0.917, 1.091])
```

Figure: Normality Test on Wine pH Data

Remark

The critical values are at the following significance levels 15%, 10%, 5%, 2.5%, 1%*.*

The test statistic value is large and hence we reject the null hypothesis that the wine pH data is normal at 1% level of significance.

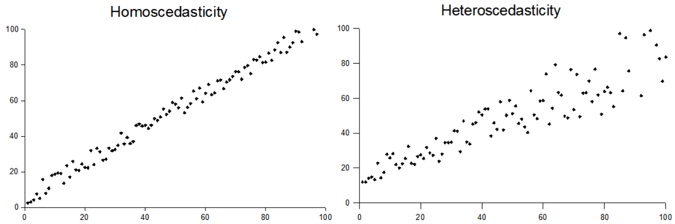


Figure: Homoscedasticity vs Heteroscedasticity

Remark

The presence of heteroscedasticity invalidates statistical tests of significance that assume that the modelling errors all have the same variance and hence it is a major concern in regression analysis.

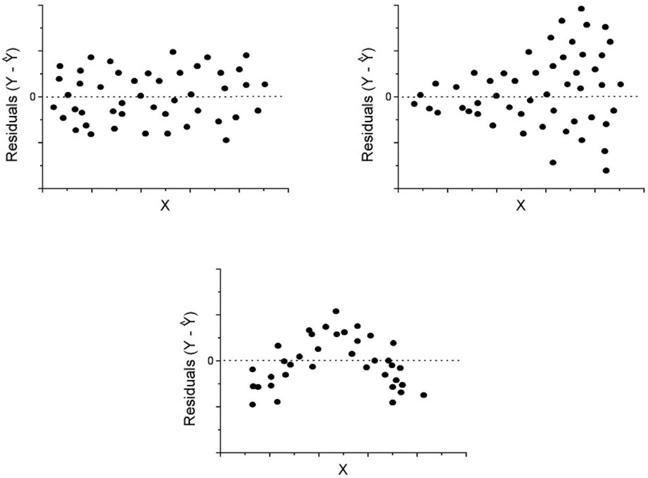
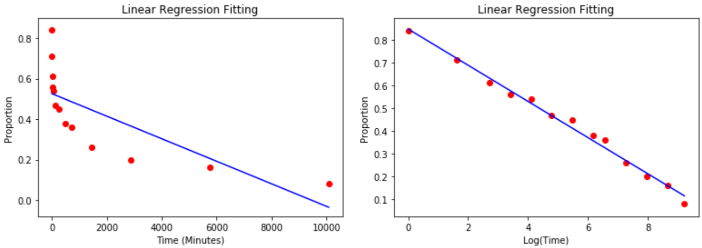


Figure: Residual plots: Top left) a perfect residual plot. Top right) represents heteroscedasticity of the data. Bottom) represents non-linear trend in the data

Transformations

- Transforming independent or dependent variables removes a number of model problems.
- Data transformation is a “trial and error" approach.
- For simple linear regression, one can easily see if we required transformation by looking at the scatter plot of x and y.
- However in MLR model, one can’t visualize it in a single plot and hence residual plots are used to check the appropriateness of the model.
- Data analysis is often called an **artful science!**.

The log transformation is important and helps in many cases. Transforming the predictor is appropriate when **non-linearity is the only problem**.



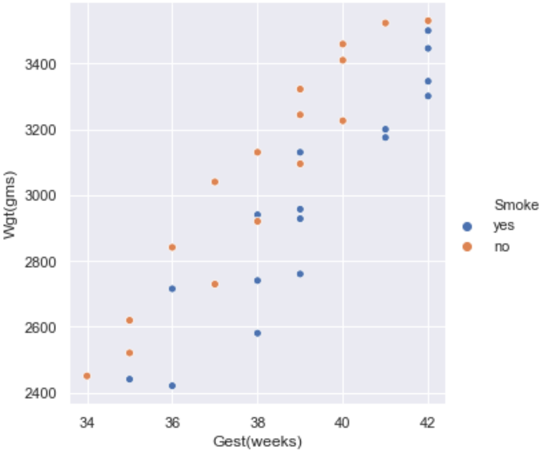
Remark

Transforming the response variable values should be considered **when non-normality or unequal variances or both** *are the problems with the model. The* **Box-Cox transformations** *which are a family of power transformations on response variable Y can help in achieving the goal of normal error term.*

Consider the data based on gestation period, birth weight and smoking habit.

	Wgt(gms)	Gest(weeks)	Smoke
0	2940	38	yes
1	3130	38	no
2	2420	36	yes
3	2450	34	no
4	2760	39	yes
5	2440	35	yes
6	3226	40	no
7	3301	42	yes
8	2729	37	no
9	3410	40	no
10	2715	36	yes
11	3095	39	no
12	3130	39	yes
13	3244	39	no
14	2520	35	no
15	2928	39	yes
16	3523	41	no
17	3446	42	yes
18	2920	38	no
19	2957	39	yes
20	3530	42	no
21	2580	38	yes
22	3040	37	no
23	3500	42	yes
24	3200	41	yes
25	3322	39	no
26	3459	40	no
27	3346	42	yes
28	2619	35	no
29	3175	41	yes
30	2740	38	yes
31	2841	36	no

The scatter plot based on gestation period, birth weight and smoking habit is



Multiple Linear Regression

The multiple linear regression for p predictors is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon.$$

Suppose, we have n data points such that

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n.
\end{aligned}$$

The MLR in matrix form can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

In general, we can write

$$Y = X\beta + \epsilon.$$

Matrix Differentiation

Let $y = f(x)$ where y is $m \times 1$ and x is $n \times 1$ vectors. Then, we denote

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

Proposition

Let $y = Ax$, where y is $m \times 1$, x is $n \times 1$, A is $m \times n$ and A does not depend on x, then

$$\frac{\partial y}{\partial x} = A.$$

Proof.

The ith element of y can be written as

$$y_i = \sum_{k=1}^{n} a_{ik} x_k.$$

Thus $\frac{\partial y_i}{\partial x_j} = a_{ij}$. Therefore $\frac{\partial y}{\partial x} = A$. □

Proposition

Let the scalar M be defined as

$$M = b^T Ac,$$

where b is $m \times 1$*, c is* $n \times 1$*, A is* $m \times n$ *and A does not depend on b and c, then*

$$\frac{\partial M}{\partial c} = b^T A \ \ and \ \ \frac{\partial M}{\partial b} = c^T A^T.$$

Proof.

Let $w^T = b^T A$, which implies $M = w^T c$ and hence $\frac{\partial M}{\partial c} = w^T = b^T A$. Further, M is a scalar and hence, we can write

$$M = M^T = c^T A^T b.$$

Thus $\frac{\partial M}{\partial b} = c^T A^T$. □

Proposition

Let the scalar M be defined as

$$M = b^T A b,$$

where b is $n \times 1$, A is $n \times n$ and A does not depend on b, then

$$\frac{\partial M}{\partial b} = b^T (A + A^T).$$

Proof.

We have

$$M = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} b_i b_j.$$

Differentiation with respect to b_k, leads to

$$\frac{\partial M}{\partial b_k} = \sum_{i=1}^{n} a_{ik} b_i + \sum_{j=1}^{n} a_{kj} b_j, \;\; k = 1, 2, \cdots, n.$$

Thus $\frac{\partial M}{\partial b} = b^T A^T + b^T A = b^T (A^T + A)$. $\square$

Remark

If A is symmetric then $\frac{\partial M}{\partial b} = 2b^T A$.

We want to minimize the sum of squared errors

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - Y^T X\beta + \beta^T X^T X\beta - \beta^T X^T Y \\ &= Y^T Y + \beta^T X^T X\beta - 2Y^T X\beta. \end{aligned}$$

Result related to derivative

If $M(b) = b^T A b$, where b is a $m \times 1$ vector and A is any $m \times m$ symmetric matrix, then

$$\frac{\partial M(b)}{\partial b} = 2b^T A.$$

We have

$$\frac{\partial S(\beta)}{\partial \beta} = 2\beta^T X^T X - 2Y^T X.$$

Further,

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X^T X \text{ (non-negative definite).}$$

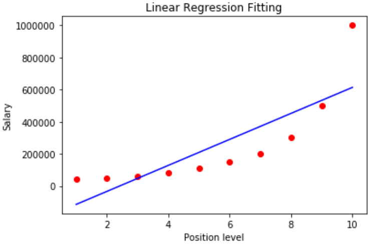
Putting $\frac{\partial S(\beta)}{\partial \beta} = 0$, we have

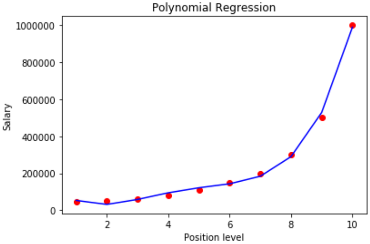
$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

provided $X^T X$ is invertible. $\hat{\beta}$ is called the ordinary least squares (OLS) estimator of β.

Polynomial Regression

Consider the following scatter plot of position level vs salary (US \$)





In this case linear regression is not an appropriate model. In this scenario, we can consider the polynomial regression such that

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon,$$

or higher degree polynomial.

We can estimate the coefficient in polynomial regression, using OLS technique discussed for MLR model and putting

- $x = x_1$
- $x^2 = x_2$
- $x^3 = x_3$
- $x^4 = x_4$.

Weighted Least Square

- The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

where the random errors are iid $N(0, \sigma^2)$.

- What happens if the ϵ_i's are independent but with unequal variance i.e. $\epsilon_i \sim N(0, \sigma_i^2)$? Also called heteroscedasticity.

- The ordinary least squares (OLS) estimates for β_j are unbiased, but no longer have the minimum variance.

- The weighted Least Squares (WLS) fixes the problem of heteroscedasticity

- For the model

$$y_i = \beta_0 + \beta_1 1x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_i^2)$?.

- The weighted Least Squares (WLS) is finding the estimates of β_j such that

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2}{\sigma_i^2}.$$

- In WLS, we focus more on minimizing errors of observation with smaller variances and focus less on minimizing errors of observations with larger variances.

- In general WLS is finding the estimates of β_j such that

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} w_i (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2,$$

where the weights $w_1, w_2, \cdots, w_n$ are known and $w_i > 0$ for all i.

The WLS estimates of β_j which minimize

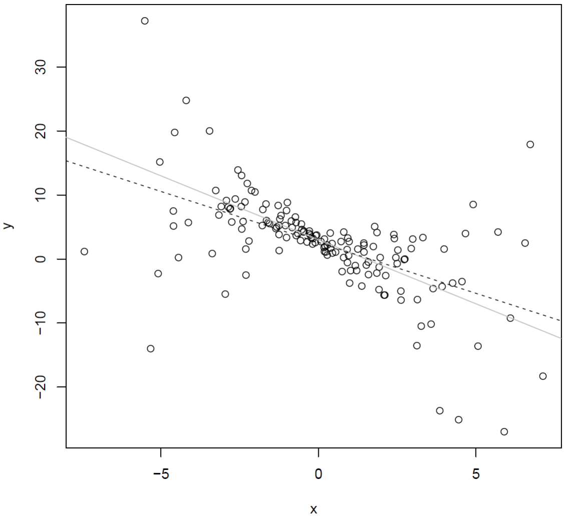
$$L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} w_i (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2,$$

is

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W Y,$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, W = \begin{pmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_n \end{pmatrix}$$


30
20
10
0
−10
−20
y
−5
0
5
x

THANK YOU!