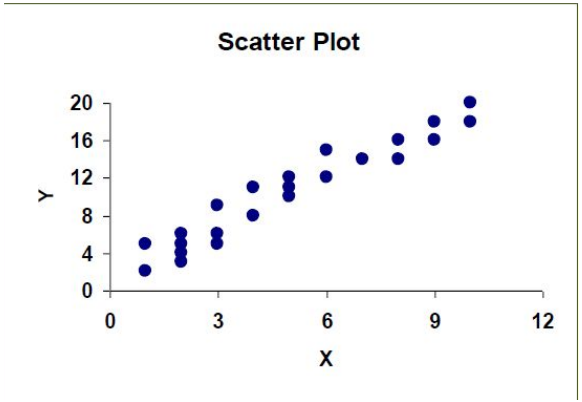
Regression

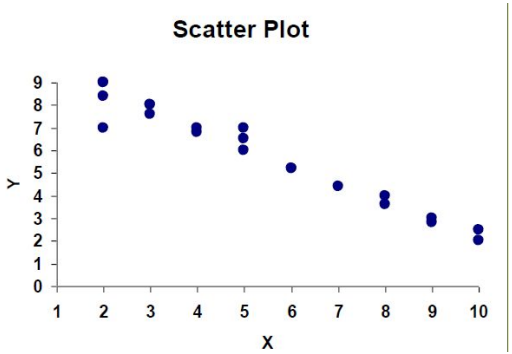
Prof. (Dr.) Honey Sharma

Correlation

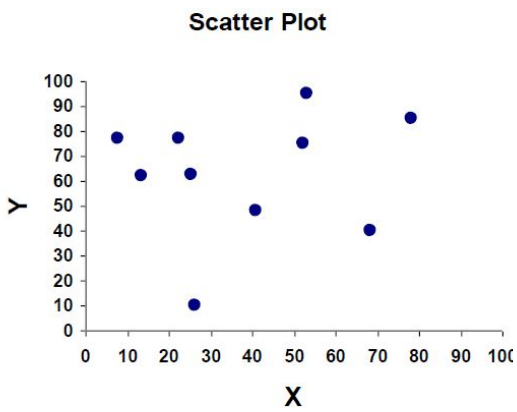
Correlation analysis is a technique to identify the relationship between two variables. Type and degree of relationship between two variables.



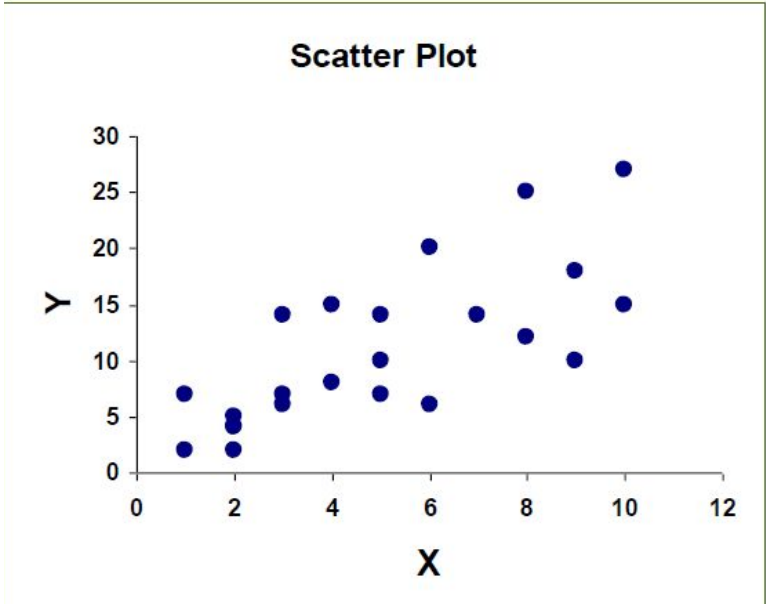
Positive Correlation: y increases as x increases & vice versa



Negative Correlation: y decreases as x increases & vice versa



No Correlation: Random Distribution of points



Is there any correlation ?

Measure of Correlation

Coefficient of Correlation (Symbol : r)

Range : -1 to 1
Sign : Type of correlation
Value : Degree of correlation

Examples:

r = 0.6, 60 % positive correlation
r = -0.82, 82% negative correlation
r = 0, No correlation

Correlation helps To check whether two variables are related If related Identify the type & degree of relationship

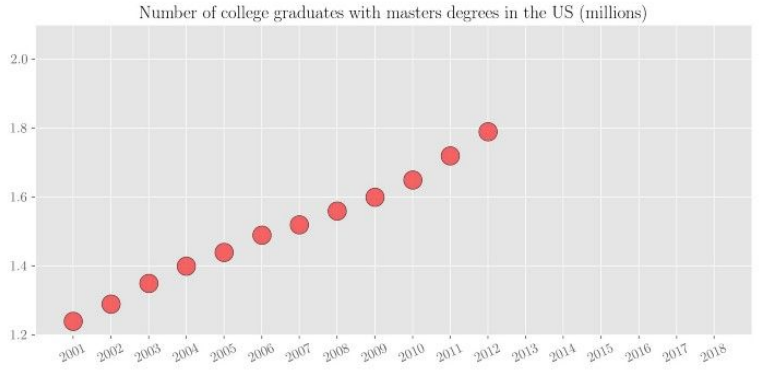
Regression

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Regression

The scatter plot below shows the number of college graduates in the US from the year 2001 to 2012.

Now based on the available data, what if someone asks you how many college graduates with master's degrees will there be in the year 2018? It can be seen that the number of college graduates with master's degrees increases almost linearly with the year. So by simple visual analysis, we can get a rough estimate of that number to be between 2.0 to 2.1 million. Let's look at the actual numbers.



Regression in machine learning consists of mathematical methods that allow data scientists to predict a continuous outcome (y) based on the value of one or more predictor variables (x).

In simpler words, it means fitting a function from a selected family of functions to the sampled data under some error function.

Mathematically Speaking

Estimate a function $f_\beta(.)$ (parameterized with β) given data points (x_i, y_i) $\forall i \in \{0, 1, ..., n-1\}$ under a loss function $\sum_i l(f(x_i), y_i)$

Evaluating a Regression Algorithm: VARIANCE

Variance is the amount by which the estimate of the target function changes if different training data were used. The target function f establishes the relation between the input and the output variables.

When a different dataset is used the target function needs to remain stable with little variance because, for any given type of data, the model should be generic.

To avoid false predictions, we need to make sure the variance is low. For that reason, the model should be generalized to accept unseen features.

Evaluating a Regression Algorithm: BIAS

Bias is the algorithm's tendency to consistently learn the wrong thing by not taking into account all the information in the data. For the model to be accurate, bias needs to be low. If there are inconsistencies in the dataset like missing values, less number of data tuples or errors in the input data, the bias will be high and the predicted value will be wrong.

Evaluating a Regression Algorithm: Accuracy and Error

Accuracy and error are the two other important metrics.

The error is the difference between the actual value and the predicted value estimated by the model.

Accuracy is the fraction of predictions our model got right.

For a model to be ideal, it's expected to have low variance, low bias and low error.

Bias and variance are always in a trade-off. When bias is high, the variance is low and when the variance is low, bias is high. The former case arises when the model is too simple with a fewer number of parameters and the latter when the model is complex with numerous parameters. We require both variance and bias to be as small as possible, and to get to that the trade-off needs to be dealt with carefully, then that would bubble up to the desired curve.

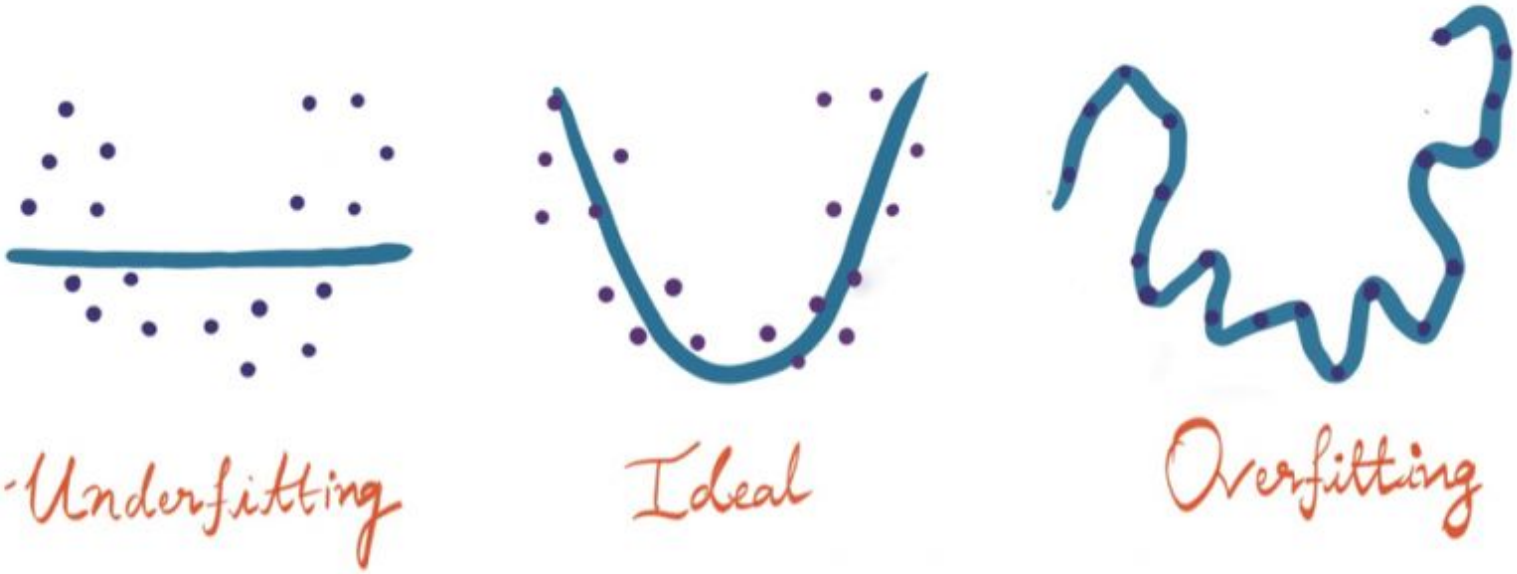
To achieve this, we need to partition the dataset into train and test datasets.

The model will then learn patterns from the training dataset and the performance will be evaluated on the test dataset. To reduce the error while the model is learning, there is an error function.

If the model memorizes/mimics the training data fed to it, rather than finding patterns, it will give false predictions on unseen data.

The curve derived from the trained model would then pass through all the data points and the accuracy on the test dataset is low. This is called overfitting and is caused by high variance.

On the flip side, if the model performs well on the test data but with low accuracy on the training data, then this leads to underfitting.


Underfitting
Ideal
Overfitting

Simple Linear Regression

It is a very straigh forward simple linear approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as $Y \approx \beta 0 + \beta 1X$.

You might read "≈" as "is approximately modeled as".

For example, X may represent TV advertising and Y may represent sales. Then we can regress sales onto TV by fitting the model

$$\text{sales} \approx \beta 0 + \beta 1 \times \text{TV}.$$

β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model.

Once we have used our coefficient parameter training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients.

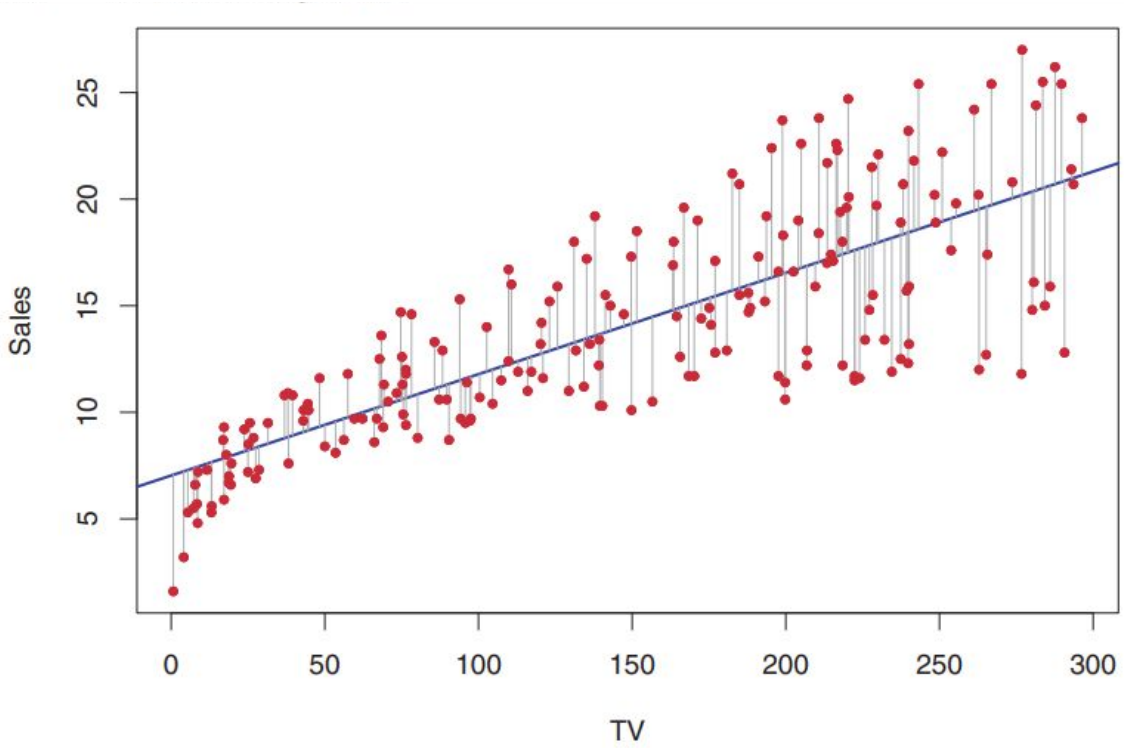
We want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to then data points. There are a number of ways of measuring closeness. However, by far the most common approach involves minimizing the least squares criterion

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*—this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad (3.3)$$


Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV

Out[27]:

OLS Regression Results

Dep. Variable:	mydata.Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	605.4
Date:	Thu, 03 Mar 2022	Prob (F-statistic):	8.13e-99
Time:	10:26:28	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	774.7
Df Residuals:	196	BIC:	787.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.6251	0.308	15.041	0.000	4.019	5.232
mydata.TV	0.0544	0.001	39.592	0.000	0.052	0.057
mydata.Radio	0.1070	0.008	12.604	0.000	0.090	0.124
mydata.Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012

Omnibus:	16.081	Durbin-Watson:	2.251
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.655
Skew:	-0.431	Prob(JB):	9.88e-07
Kurtosis:	4.605	Cond. No.	454.

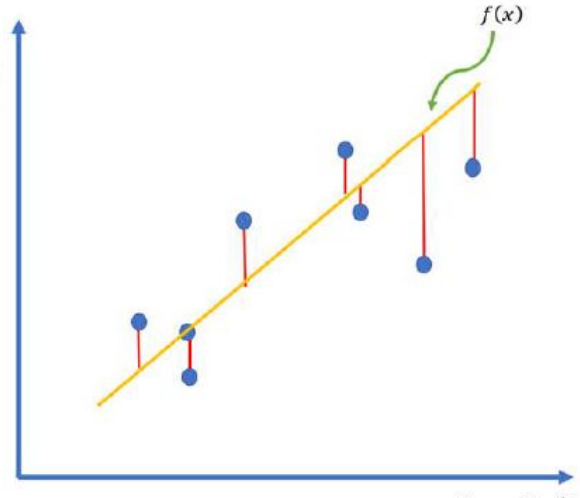
Cost Function

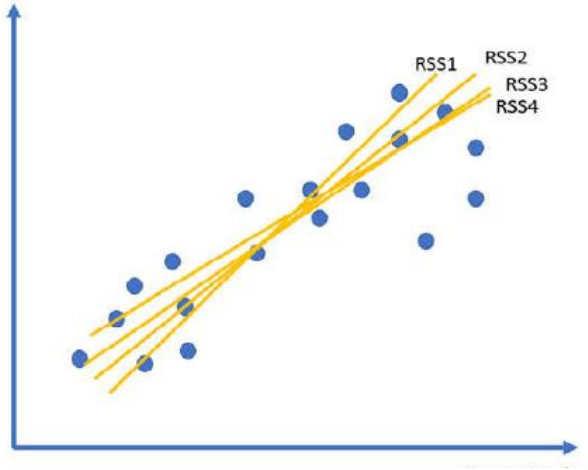
The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. **Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.**

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

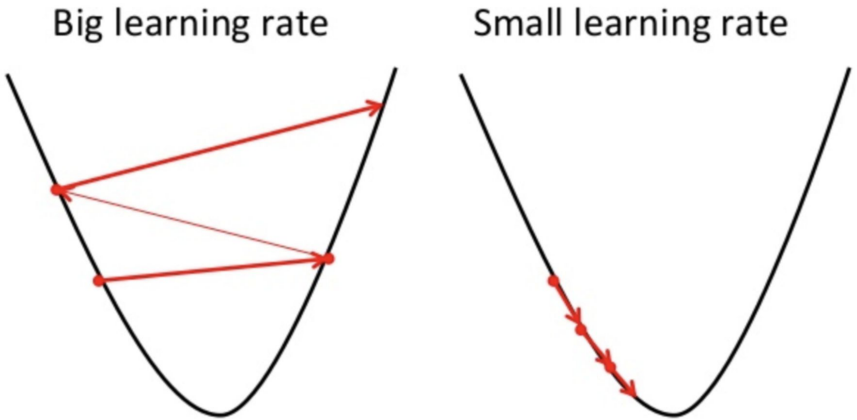
This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function. Now, **using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.**





Gradient Descent

Gradient descent is a method of updating a_0 and a_1 to reduce the cost function(MSE). The idea is that we start with some values for a_0 and a_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.



To update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to a_0 and a_1.

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)^2$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i) \implies \frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i) \cdot x_i \implies \frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

$$a_0 = a_0 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \cdot x_i$$

The partial derivatives are the gradients and they are used to update the values of a_0 and a_1. Alpha is the learning rate which is a **hyperparameter** that you must specify. **A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that you could overshoot the minima. One Solution for this is dynamic larning rate. Start with higher rate and slowly decrease it.**

DRAWING THE BEST-FIT LINE

Imagine, you're given a set of data and your goal is to draw the best-fit line which passes through the data. This is the step-by-step process you proceed with:

- Consider your linear equation to be **y = a_0 + a_1 * x**
- Adjust the line by varying the values of a_0 and a_1, i.e., the coefficient and the bias.
- Come up with some random values for the coefficient and bias initially and plot the line.
- Since the line won't fit well, change the values of a_0 and a_1. This can be done using the 'gradient descent algorithm'.

For any set of data, a_0 & a_1 can be calculated

Regression model $y = a_0 + a_1 * x$ can be build

But all the models may not be useful

Multiple Linear Regression

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. For example, in the Advertising data, we have examined the relationship between sales and TV advertising. How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

One option is to run three separate simple linear regressions, each of which uses a different predictor.

However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory.

First of all, it is unclear how to make a single prediction of sales given levels of the three advertising media budgets, since each of the budgets is associated with a separate regression equation.

Second, each of the three regression equations ignores the other two media in forming estimates for the regression coefficients.

We will see shortly that if the media budgets are correlated with each other in the 200 markets that constitute our data set, then this can lead to very misleading estimates of the individual media effects on sales.

Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors.

We can do this by giving each predictor a separate slope coefficient in a single model.In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + e$$

where Xj represents the jth predictor and βj quantifies the association between that variable and the response.

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in (3.19) are unknown, and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p. \tag{3.21}$$

The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression. We choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned} \tag{3.22}$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

We notice that the multiple regression coefficient estimates for TV and radio are pretty similar to the simple linear regression coefficient estimates. However, while the newspaper regression coefficient estimate was significantly non-zero, the coefficient estimate for newspaper in the multiple regression model is close to zero.

This illustrates that the simple and multiple regression coefficients can be quite different.

Some Important Question

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. Is at least one of the predictors X1, X2,...,Xp useful in predicting the response?
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Test for the significance of Regression Slope Coefficient

State the Hypotheses

If there is a significant linear relationship between the independent variable x and the dependent variable y, the slope will not equal zero.

Ho: A_1 = 0 (A_1 x intercept of population)

Ha: A_1 ≠ 0

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Significance level (\alpha)

Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

Find the test statistic and the corresponding p-value.
The test statistic is t = (a_1-A_1) / SE_a_1 with n-2 degrees of freedom.

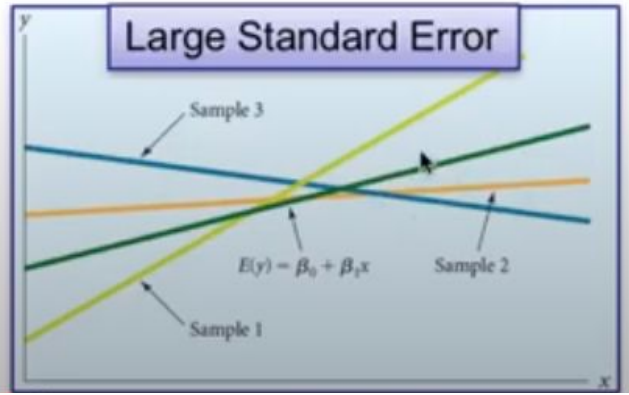
P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the t Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom.

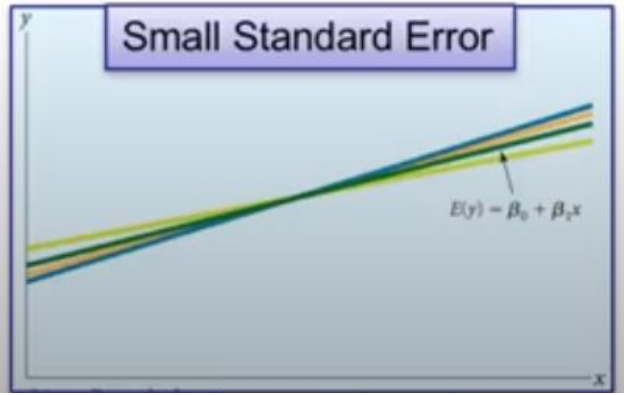
If p value < \alpha, then H0 is rejected & y can be modeled with x

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x_i - \bar{x})^2}}$$

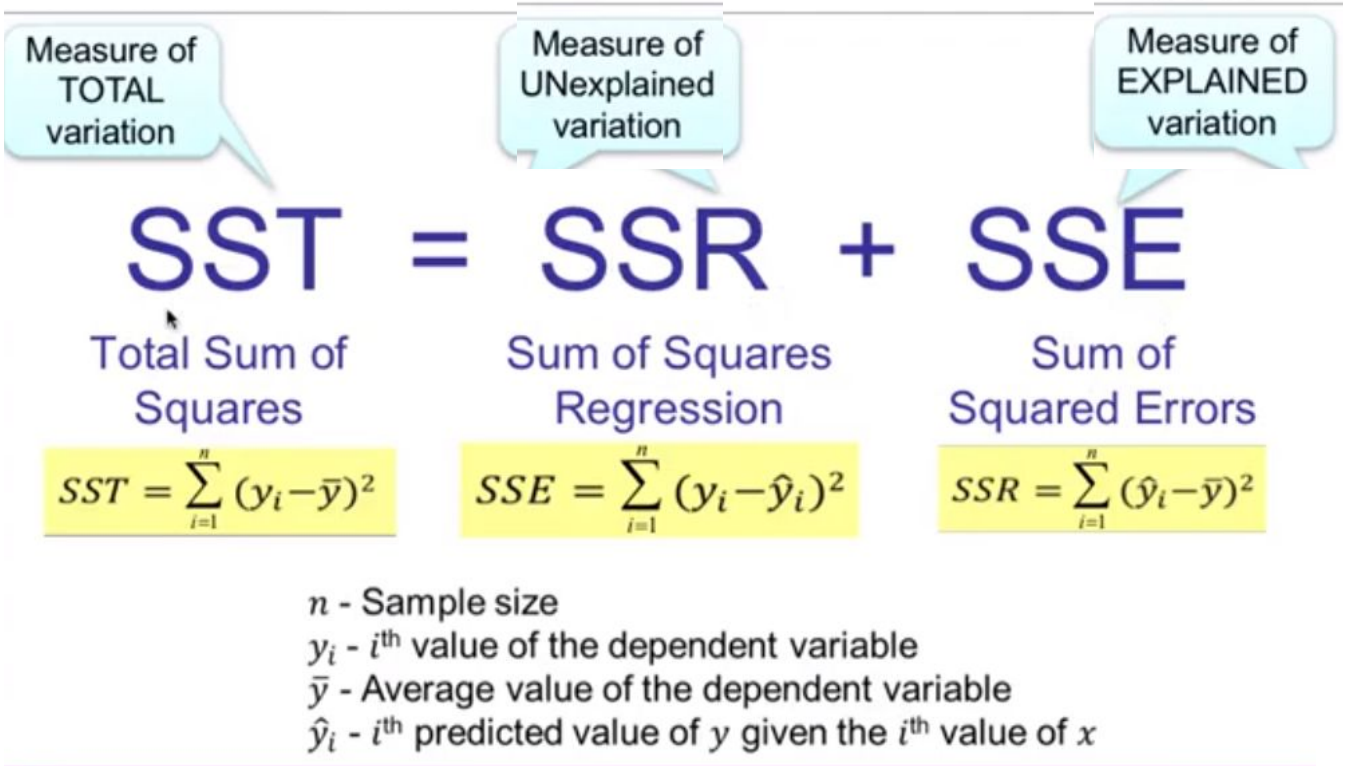
s_{b_1} - Standard deviation of the regression slope (*standard error of the slope*)

s_ε - Sample standard error of the estimate (the measure of deviation of the actual y-values around the regression line)





Sum of Squares Measure of variation ```



Coefficient of Regression R^2

It measures the proportion of the variation in your dependent variable explained by all of your independent variables in the model.

It assumes that every independent variable in the model helps to explain variation in the dependent variable. In reality, some independent variables (predictors) don't help to explain dependent (target) variable. In other words, some variables do not contribute in predicting target variable.

Mathematically, R-squared is calculated by dividing sum of squares of residuals (SSR) by total sum of squares (SST) and then subtract it from 1.

$$R^2 = 1 - (SSR/SST)$$

R-Squared is also called coefficient of determination. It lies between 0% and 100%. A r-squared value of 100% means the model explains all the variation of the target variable. And a value of 0% measures zero predictive power of the model. Higher R-squared value, better the model.

Adjusted R-Squared

It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

$$Adjusted - R^2 = 1 - \frac{SSR/df_e}{SST/df_t}$$

Here degree of freedom df_e=n(number of independent variables

df_t=n-p-1, p is number of predictors

Difference between R-square and Adjusted R-square

- Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable.
- Adjusted r-squared value always be less than or equal to r-squared value.

F-Statistics & p Value

The F value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. In other words, the model has no predictive capability. Basically, the f-test compares your model with zero predictor variables (the intercept only model), and decides whether your added coefficients improved the model or not.

H_0: Proposed Model= Model with only intercept value

For a significance level of 0.05:

- **If the p-value associated with the F-statistic is ≥ 0.05:**

  Then there is no relationship between ANY of the independent variables and dependent variables

- **If the p-value associated with the F-statistic < 0.05:**

  Then, AT LEAST 1 independent variable is related to dependent variable

t-Test

Whenever we perform linear regression, we want to know if there is a statistically significant relationship between the predictor variable and the response variable.

We test for significance by performing a t-test for the regression slope. We use the following null and alternative hypothesis for this t-test:

H0: $\beta 1 = 0$ (the slope is equal to zero)

HA: $\beta 1 \neq 0$ (the slope is not equal to zero)

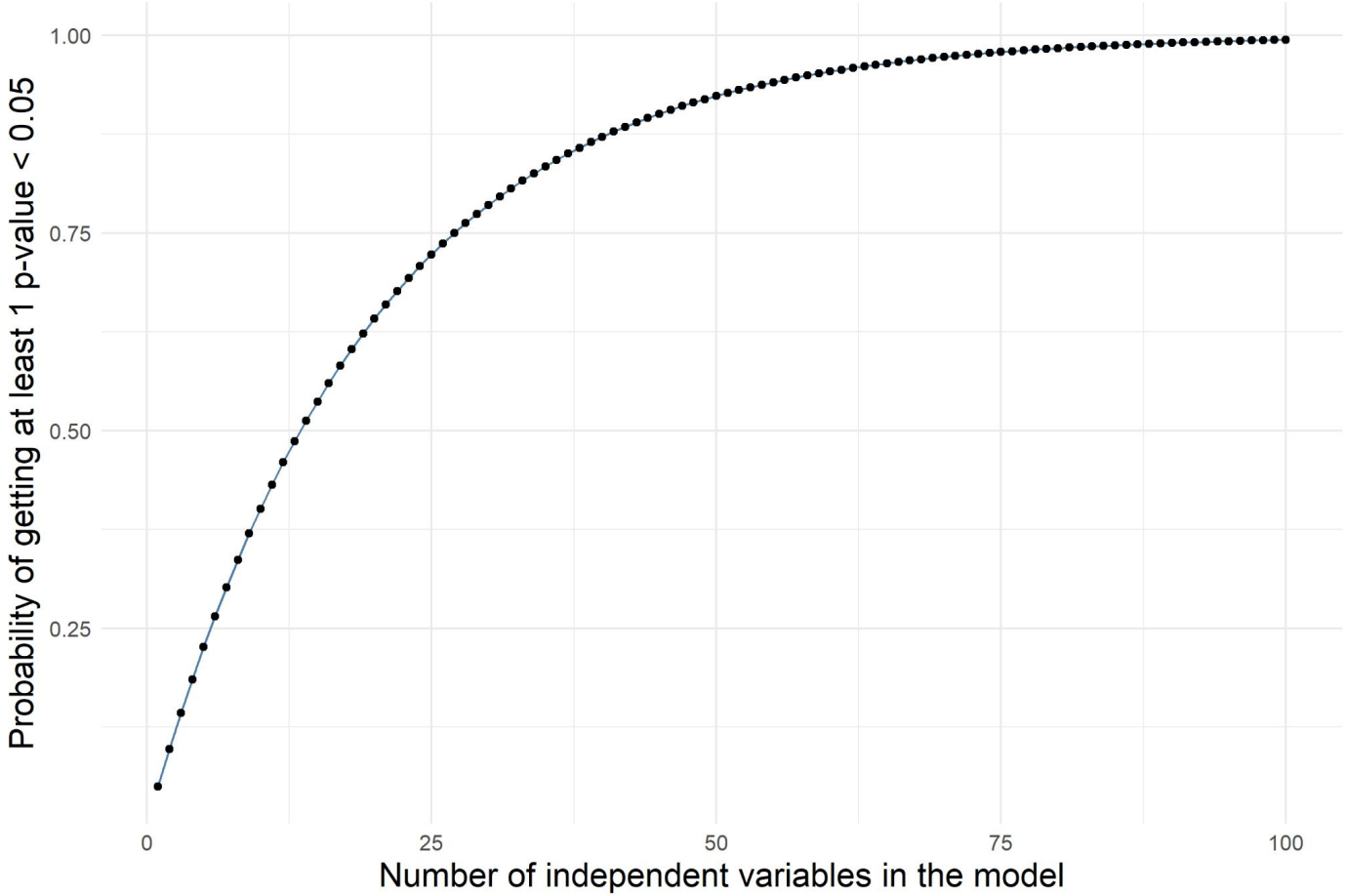
If the p-value that corresponds to t is less than some threshold (e.g. $\alpha = .05$) then we reject the null hypothesis and conclude that there is a statistically significant relationship between the predictor variable and the response variable.

Why do we need a global test? Why not look at the p-values associated with each coefficient?

This is because each coefficient's p-value comes from a separate statistical test that has a 5% chance of being a false positive result (assuming a significance level of 0.05).

For instance, if we take the example in which, we have 4 independent variables (X1 through X4) and each of them has a 5% risk of yielding a p-value < 0.05 just by chance (when in reality they're not related to Y).

The more variables we have in our model, the more likely it will be to have a p-value < 0.05 just by chance.


Probability of getting at least 1 p-value < 0.05
1.00
0.75
0.50
0.25
0
25
50
75
100
Number of independent variables in the model

What if the F-statistic has a statistically significant p-value but none of the coefficients does?

This may happen when all independent variables are correlated variables.

Because of correlation present in the variables, the effect of each of them was diluted and therefore their p-values were ≥ 0.05, when in reality they are related to the dependent variables.

CONCLUSION:

When it comes to the overall significance of the linear regression model, always trust the statistical significance of the F-statistic over that of each independent variable.

Log-Likelihood

The log-likelihood value of a regression model is a way to measure the goodness of fit for a model. The higher the value of the log-likelihood, the better a model fits a dataset.

The log-likelihood value for a given model can range from negative infinity to positive infinity.

The actual log-likelihood value for a given model is mostly meaningless, but it's useful for comparing two or more models.

In practice, we often fit several regression models to a dataset and choose the model with the highest log-likelihood value as the model that fits the data best.

When calculating log-likelihood values, it's important to note that adding more predictor variables to a model will almost always increase the log-likelihood value even if the additional predictor variables aren't statistically significant.

This means you should only compare the log-likelihood values between two regression models if each model has the same number of predictor variables.

The Akaike information criterion (AIC)

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from.

In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data.

AIC is calculated from:

- the number of independent variables used to build the model.
- the maximum likelihood estimate of the model (how well the model reproduces the data).

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.

$$AIC = 2K - 2ln(L)$$

K is the number of independent variables used and L is the log-likelihood estimate (a.k.a. the likelihood that the model could have produced your observed y-values). The default K is always 2, so if your model uses one independent variable your K will be 3, if it uses two independent variables your K will be 4, and so on.

To use AIC for model selection, we simply choose the model giving smallest AIC over the set of models considered. If a model is more than 2 AIC units lower than another, then it is considered significantly better than that model.

Bayesian Information Criterion (BIC)

BIC is a method for scoring and selecting a model.

Like AIC, it is appropriate for models fit under the maximum likelihood estimation framework. The BIC statistic is calculated as follows:

$$BIC = -2 * LL + \log(N) * k$$

Where log() has the base-e called the natural logarithm, LL is the log-likelihood of the model, N is the number of examples in the training dataset, and k is the number of parameters in the model.

The model with the lowest BIC is selected.

The quantity calculated is different from AIC, although can be shown to be proportional to the AIC. Unlike the AIC, the BIC penalizes the model more for its complexity, meaning that more complex models will have a worse (larger) score and will, in turn, be less likely to be selected.

Omnibus/Prob(Omnibus)

One of the assumptions of OLS is that the residuals are normally distributed. Omnibus test is performed in order to check this.

Here, the null hypothesis is that the residuals are normally distributed. Prob(Omnibus) is supposed to be close to the 1 in order for it to satisfy the OLS assumption.

In this case Prob(Omnibus) is 0.062, which implies that the OLS assumption is not satisfied. Due to this, the coefficients estimated out of it are not Best Linear Unbiased Estimators(BLUE).

Skew/ Kurtosis

Skew – a measure of data symmetry. We want to see something close to zero, indicating the residual distribution is normal. Note that this value also drives the Omnibus. This result has a small, and therefore good, skew.

Kurtosis – a measure of "peakiness", or curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.

Durbin-watson

Another assumption of OLS is of homoscedasticity. This implies that the variance of errors is consistent across the data. A value between 1 to 2 is preferred.

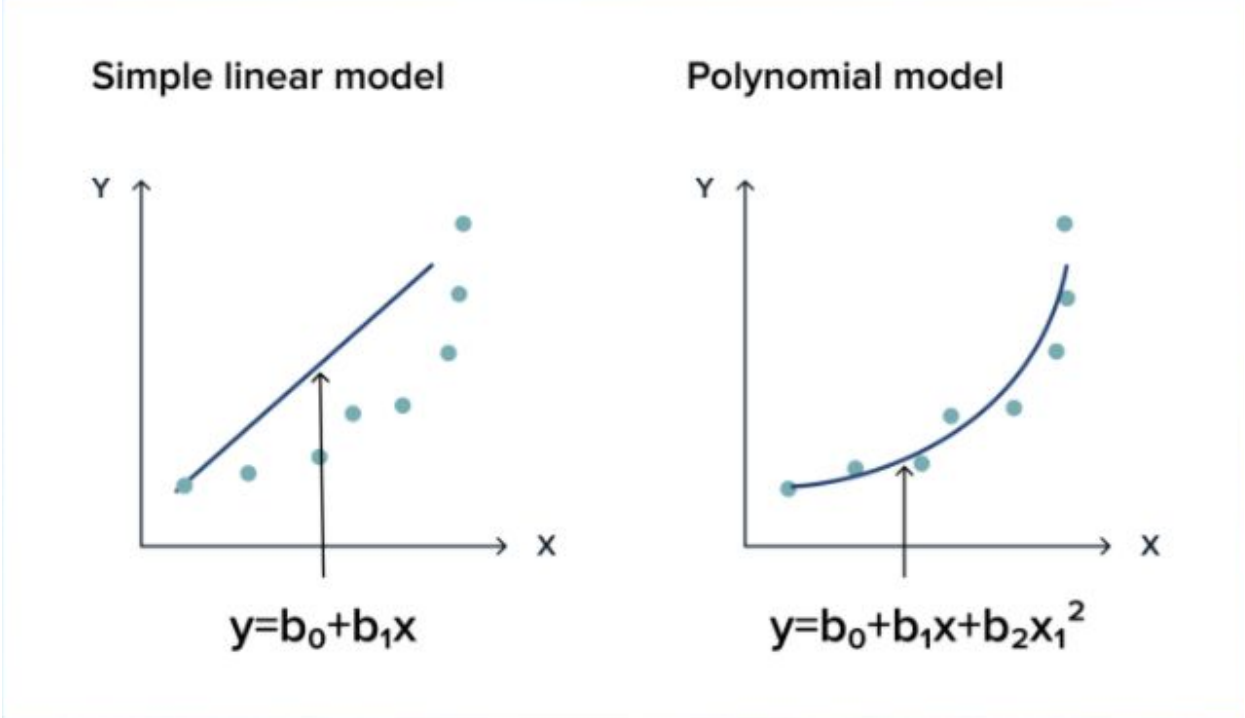
Prob(Jarque-Bera)

It is in line with the Omnibus test. It is also performed for the distribution analysis of the regression errors. It is supposed to agree with the results of Omnibus test. A large value of JB test indicates that the residuals are not normally distributed.

Condition Number

This test measures the sensitivity of a function's output as compared to its input (characteristic #4). When we have multicollinearity, we can expect much higher fluctuations to small changes in the data, hence, we hope to see a relatively small number, something below 30.

Polynomial Regression



Polynomial Regression

The standard way to extend linear regression to settings in which the relationship between the predictors and the response is non-linear has been to replace the standard linear model with a polynomial function

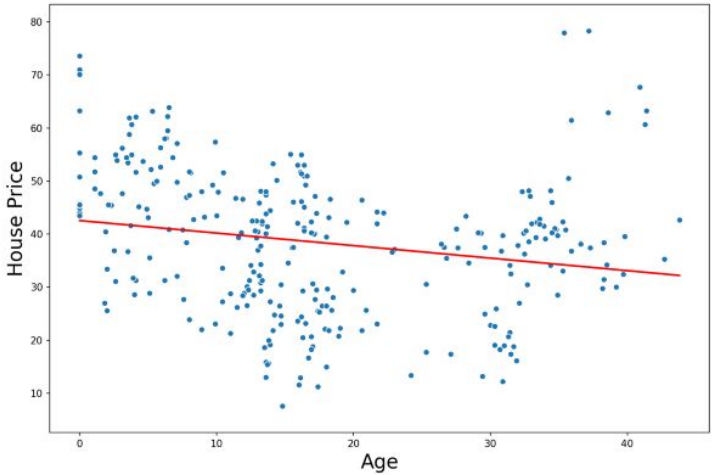
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,$$

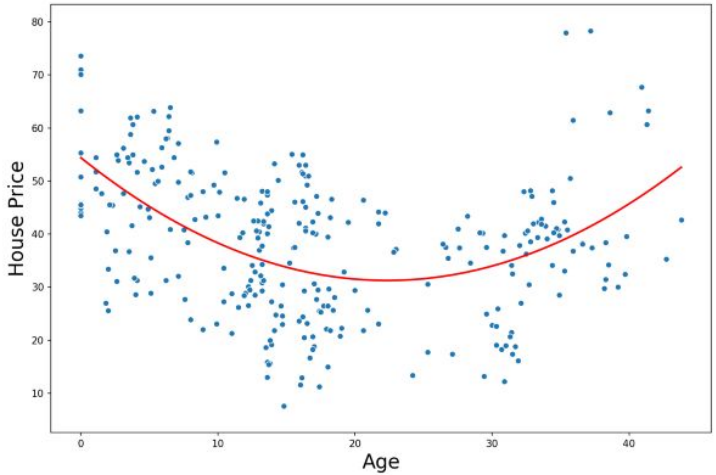
This approach is known as polynomial regression. For large regression enough degree d, a polynomial regression allows us to produce an extremely non-linear curve.

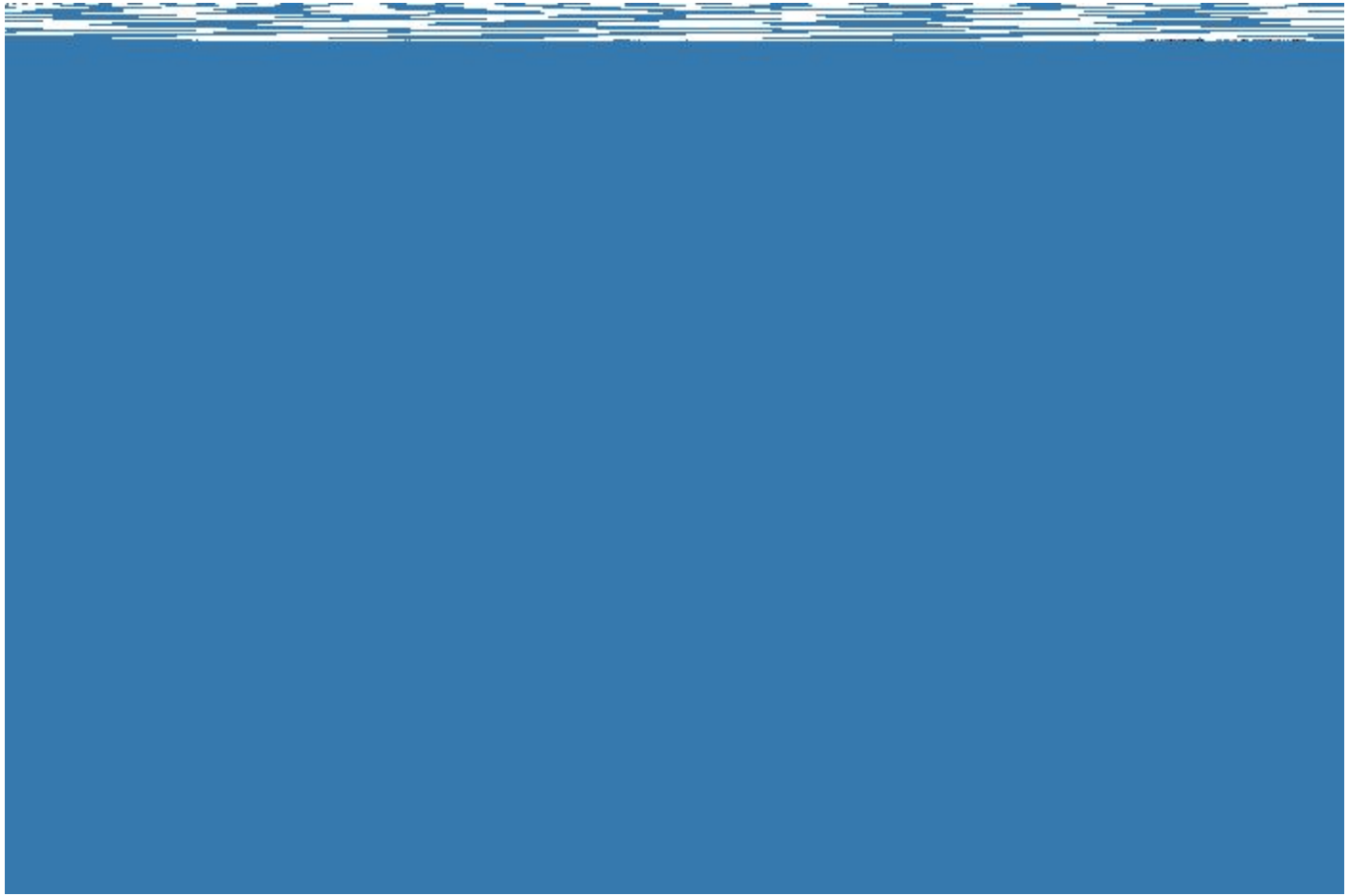
Notice that the coefficients can be easily estimated using least squares linear regression because this is just a standard linear model with predictors as power of x^i.

Generally speaking, it is unusual to use d greater than 3 or 4 because for large values of d, the polynomial curve can become overly flexible and can take on some very strange shape

Polynomial Regression can improve the accuracy of your models but, if used incorrectly, overfitting can occur. We want to avoid this as it would leave you with a model that does not perform well in the future.







Another way to visualise this is by looking at the MSE on the training and test set. the train MSE tends to decreases as you increase the degree. This means the model is becoming more and more accurate on the training set. The test MSE tells a different story. The test MSE is a minimum when n=2 and then tends to increase after that. This means the model is performing worse and worse on the test set. In other words, as we increase the degree the model is becoming more overfitted.