Regression Evaluation Measures

Machine Learning

Dr. Adnan Abid

Courtesy Super Data Science

Regressions

Simple Linear Regression

$$y = b_0 + b_1{}^*x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + \ldots + b_n{}^*x_n$$

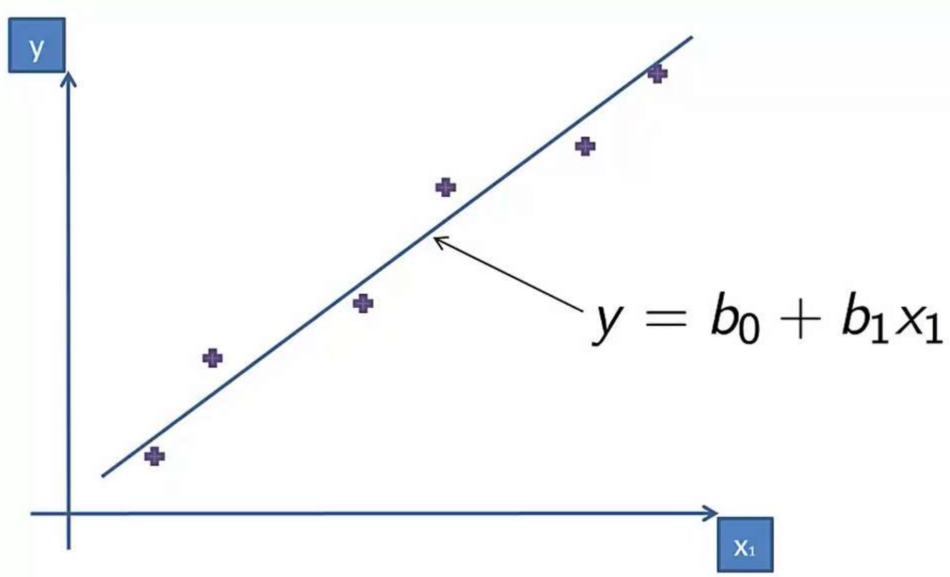
Regressions

Simple Linear Regression: $y = b_0 + b_1 x_1$

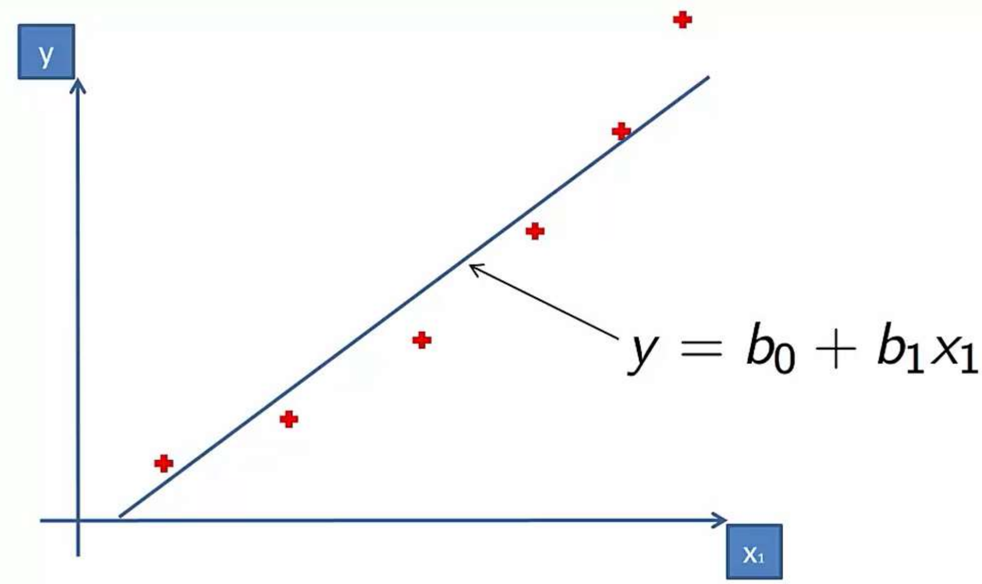
Multiple Linear Regression: $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

Polynomial Linear Regression: $y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

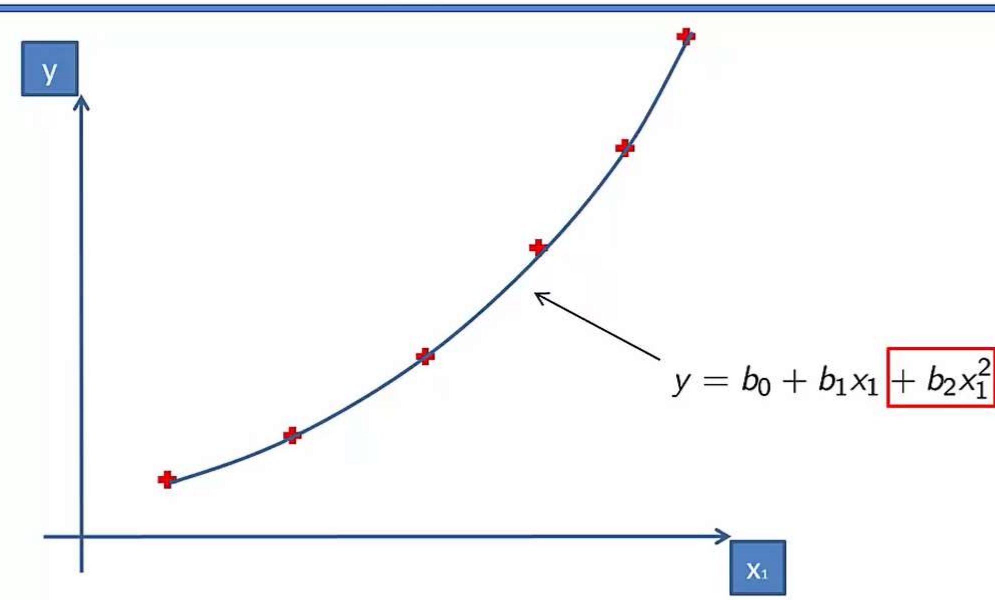
Simple Linear Regression



Simple Linear Regression



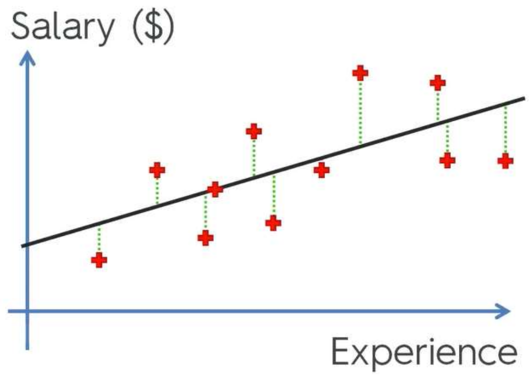
Polinomial Regression



Ordinary Least Squared

R Squared

Simple Linear Regression:

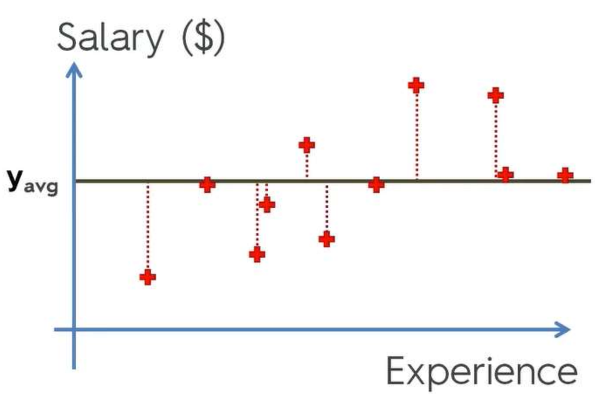


$$SS_{res} = SUM\ (y_i - \hat{y_i})^2$$

R Squared

R Squared

Simple Linear Regression:

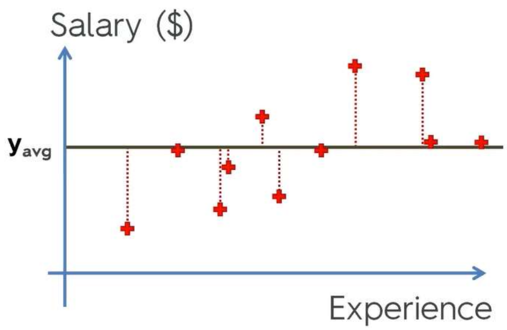


$$SS_{res} = SUM\ (y_i - \hat{y_i})^2$$

$$SS_{tot} = SUM\ (y_i - y_{avg})^2$$

R Squared

Simple Linear Regression:



$SS_{res} = SUM\ (y_i - \hat{y}_i)^2$

$SS_{tot} = SUM\ (y_i - y_{avg})^2$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- Intuition is that we compare the SS_{res} with SS_{tot} which is the difference from the average line (which anyone can draw)
- IDEALLY R^2 will be 1 when SS_{res} is 0, i.e. all the predicted points lie exactly on the test data. However, generally it is never the case.
- Also R^2 can be negative when SS_{tot} is less than SS_{res}, which will be a seldom case, when the predictor performs even worse than the average line
- Normally, it is between 0 and 1, where closer to 1 is better.

Adjusted R Squared

Adjusted R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$

SS_{res} -> Min

- R^2 will never decrease
- Particularly, in case of multiple regression, when we add another variable to our model, it somehow effects the model, and tried to adjust the SS_{res}.
- In fact, it will help finding a coefficient for the new variable so that it helps minimizing the SS_{res} (Otherwise, it may assign 0 to the coefficient, though this is not the case, as there exists a random correlation between new independent variable and dependent variable)

Adjusted R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$ ← $\boxed{+ b_3 * x_3}$

SS_{res} -> Min → R^2 will never decrease

Problem:

- Since, R^2 will never decrease
- In fact, it will increase by adding new variables.
- This is a biased behavior, so we need to improve it.
- So we need Adjusted R^2

OLS Regression Results

Dep. Variable:	y	R-squared:	0.947
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	849.8
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	3.50e-32
Time:	23:58:33	Log-Likelihood:	-527.44
No. Observations:	50	AIC:	1059.
Df Residuals:	48	BIC:	1063.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.903e+04	2537.897	19.320	0.000	4.39e+04	5.41e+04
x1	0.8543	0.029	29.151	0.000	0.795	0.913

OLS Regression Results

Dep. Variable:	y	R-squared:	0.950
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	450.8
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	2.16e-31
Time:	23:57:42	Log-Likelihood:	-525.54
No. Observations:	50	AIC:	1057.
Df Residuals:	47	BIC:	1063.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.698e+04	2689.933	17.464	0.000	4.16e+04	5.24e+04
x1	0.7966	0.041	19.266	0.000	0.713	0.880
x2	0.0299	0.016	1.927	0.060	-0.001	0.061

OLS Regression Results

Dep. Variable:	y	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	296.0
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	4.53e-30
Time:	23:56:50	Log-Likelihood:	-525.39
No. Observations:	50	AIC:	1059.
Df Residuals:	46	BIC:	1066.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.012e+04	6572.353	7.626	0.000	3.69e+04	6.34e+04
x1	0.8057	0.045	17.846	0.000	0.715	0.897
x2	-0.0268	0.051	-0.526	0.602	-0.130	0.076
x3	0.0272	0.016	1.655	0.105	-0.006	0.060

OLS Regression Results

Dep. Variable:	y	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.946
Method:	Least Squares	F-statistic:	217.2
Date:			
Time:			
No. O			
Df Re			
Df Mo			
Covar			

const
x1
x2
x3
x4

OLS Regression Results

Dep. Variable:	y	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	169.9
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	1.34e-27
Time:	23:54:59	Log-Likelihood:	-525.38
No. Observations:	50	AIC:	1063.
Df Residuals:	44	BIC:	1074.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.013e+04	6884.820	7.281	0.000	3.62e+04	6.4e+04
x1	198.7888	3371.007	0.059	0.953	-6595.030	6992.607
x2	-41.8870	3256.039	-0.013	0.990	-6604.003	6520.229
x3	0.8060	0.046	17.369	0.000	0.712	0.900
x4	-0.0270	0.052	-0.517	0.608	-0.132	0.078
x5	0.0270	0.017	1.574	0.123	-0.008	0.062

Adjusted R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - number of regressors
n – sample size

- In Adjusted R^2 calculation we create a competition between the magnitude of improvement brought by the new variable in $1 – (1-R^2)$ and the denominator n-p-1, which penalizes the result with an addition of a new variable.
- If addition of new independent variable improves R^2 value, then $1-R^2$ will decrease, which will help improving adjusted R^2.
- While, by adding new variable (n-1)/(n-p-1) will increase, thereby increasing $(1-R^2)[(n-1)/(n-p-1)]$, and hence decreasing adjusted R^2.
- Here p is the number of independent variables
- THEREFORE, **ADJUSTED R^2** IS THE PREFERRED MEASURE FOR EVALUATION OF REGRESSION MODELS