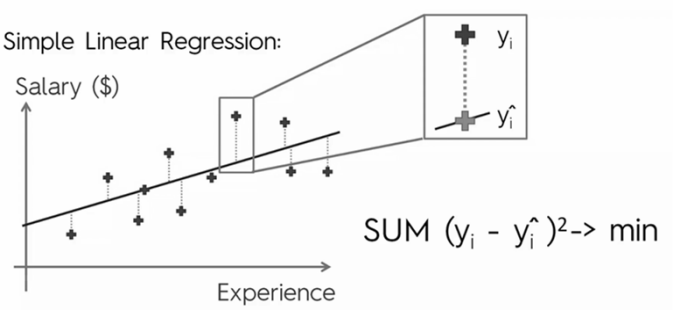
MODEL EVALUATION-REGRESSION

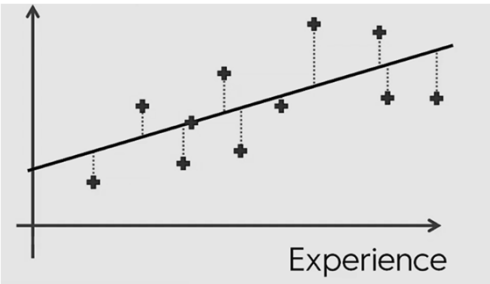
MOHAN M J

INTRODUCTION

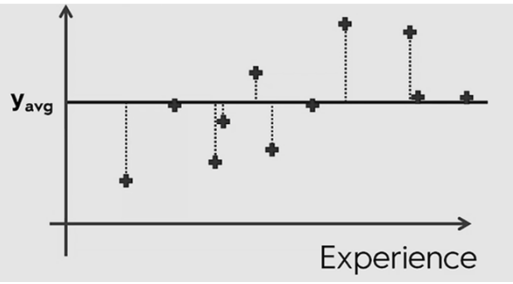


R SQUARED

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



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



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

ADJUSTED R SQUARED

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

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

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$ ← $+ b_3 * x_3$

Problem:

SS_{res}-> Min

ADJUSTED R SQUARED

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

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

$y = b_0 + b_1 * x_1$

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

Problem:

$+ b_3 * x_3$

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

ADJUSTED R SQUARED

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

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

p - number of regressors
n – sample size

INTUITION

- Higher the value of R squared better the model is
- R squared will give the goodness of fit
- R squared will never decrease when we add more features
- Maximum value is 1
- Adjusted R squared will give us the idea on whether the independent variable is helping the model or not
- If the Adj R square increases that means the variable is helping
- Useful in backward elimination method

BACKWARD ELIMINATION

- Step 1: Select significance level (p = 0.05)
- Step 2: Fit the model
- Step 3: Consider the predictor with highest p value. If p value >0.05 go to Step 4. Otherwise model is ready
- Step 4: Remove the predictor
- Step 5: Fit the model without the variable. And go to Step 3

```
# Building the optimal model using Backward Elimination
import statsmodels.formula.api as sm
X = np.append(arr = np.ones((50, 1)).astype(int), values = X, axis = 1)
X_opt = X[:, [0, 1, 2, 3, 4, 5]]
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()
```

Dep. Variable:	y	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	169.9
Date:	Sun, 08 Apr 2018	**Prob (F-statistic):**	1.34e-27
Time:	23:20:07	**Log-Likelihood:**	-755.64
No. Observations:	50	**AIC:**	1523.
Df Residuals:	44	**BIC:**	1535.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.013e+06	6.88e+05	7.281	0.000	3.62e+06	6.4e+06
x1	1.988e+04	3.37e+05	0.059	0.953	-6.6e+05	6.99e+05
x2	-4188.7019	3.26e+05	-0.013	0.990	-6.6e+05	6.52e+05
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

```
X_opt = X[:, [0, 1, 3, 4, 5]]
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()
```

Dep. Variable:	y	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.946
Method:	Least Squares	**F-statistic:**	217.2
Date:	Sun, 08 Apr 2018	**Prob (F-statistic):**	8.49e-29
Time:	23:20:07	**Log-Likelihood:**	-755.64
No. Observations:	50	**AIC:**	1521.
Df Residuals:	45	**BIC:**	1531.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.011e+06	6.65e+05	7.537	0.000	3.67e+06	6.35e+06
x1	2.202e+04	2.9e+05	0.076	0.940	-5.62e+05	6.06e+05
x2	0.8060	0.046	17.606	0.000	0.714	0.898
x3	-0.0270	0.052	-0.523	0.604	-0.131	0.077
x4	0.0270	0.017	1.592	0.118	-0.007	0.061

```
X_opt = X[:, [0, 3, 4, 5]]
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()
```

Dep. Variable:	y	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.948
Method:	Least Squares	**F-statistic:**	296.0
Date:	Sun, 08 Apr 2018	**Prob (F-statistic):**	4.53e-30
Time:	23:20:07	**Log-Likelihood:**	-755.64
No. Observations:	50	**AIC:**	1519.
Df Residuals:	46	**BIC:**	1527.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.012e+06	6.57e+05	7.626	0.000	3.69e+06	6.34e+06
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

```
X_opt = X[:, [0, 3, 5]]
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()
```

Dep. Variable:	y	**R-squared:**	0.950
Model:	OLS	**Adj. R-squared:**	0.948
Method:	Least Squares	**F-statistic:**	450.8
Date:	Sun, 08 Apr 2018	**Prob (F-statistic):**	2.16e-31
Time:	23:20:07	**Log-Likelihood:**	-755.79
No. Observations:	50	**AIC:**	1518.
Df Residuals:	47	**BIC:**	1523.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.698e+06	2.69e+05	17.464	0.000	4.16e+06	5.24e+06
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

```
X_opt = X[:, [0, 3]]
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()
```

Dep. Variable:	y	**R-squared:**	0.947
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	849.8
Date:	Sun, 08 Apr 2018	**Prob (F-statistic):**	3.50e-32
Time:	23:20:07	**Log-Likelihood:**	-757.70
No. Observations:	50	**AIC:**	1519.
Df Residuals:	48	**BIC:**	1523.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	4.903e+06	2.54e+05	19.320	0.000	4.39e+06	5.41e+06
x1	0.8543	0.029	29.151	0.000	0.795	0.913

Omnibus:	13.727	**Durbin-Watson:**	1.116
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	18.536
Skew:	-0.911	**Prob(JB):**	9.44e-05
Kurtosis:	5.361	**Cond. No.**	1.65e+07

THANK YOU