Statistical Regression

Wednesday 24 April 2024 11:51 AM

→ Regression is a means of exploring the variation in some quantity

EXPLAINED UNEXPLAINED

→ E.g. ICE CREAM SALES – Temperature, Rainfall, school holidays, ?

Regression will be able to quantify how much can be explained & how much is unexplained

→ Why machine learning can be treated as a statistical inference problem?

e.g. lets try to explain the variation in ice cream sales using only one variable (daily temperature)

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i} \longrightarrow \text{Population regression equation}$$

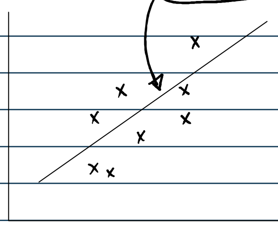
→ Regression aims to:

① Estimating the β's

② Quantifying the errors

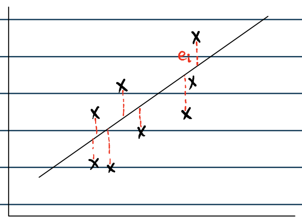
Question: Do we have all the data there is about daily temperatures and ice cream sales? NO, we will mostly be working with sample data, so we can never really know the true values of β_0 & β_1, we can only estimate them based on sample data.

→ Sample regression line is given by: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1$ (this describes the line itself)



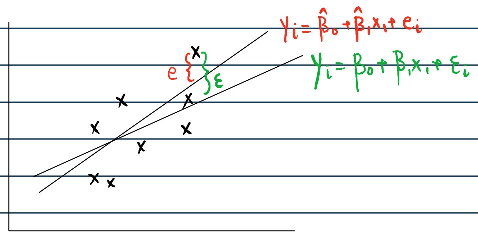
→ Each sample will yield a different sample regression line.

→ Value of y for each observation is given by: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + e_i$



→ $\underline{y_i = \overbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}^{f'(x)} + e_i}$ IS AN ESTIMATE OF $\underline{y_i = \overbrace{\beta_0 + \beta_1 x_i}^{f(x)} + \varepsilon_i}$

↳ Estimated relationship ↳ True relationship



$y = f'(x) + e_i$ (estimated relationship)

reducible – irreducible

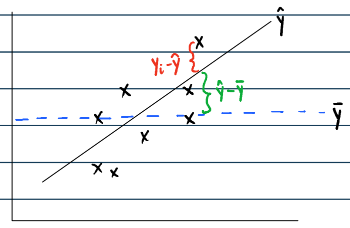
$f(x) - f'(x)$

→ Total variance (SST) $\sum(y_i - \bar{y})^2$

Sum of Squares due to Regression (SSR) (Explained) $\sum(\hat{y}_i - \bar{y})^2$

Sum of squares due to Error (SSE) (still unexplained) $\sum(y_i - \hat{y})^2$

→ Regression begins to look a lot like ANOVA i.e. the total sum of squares is partitioned b/w SSE & SSR.



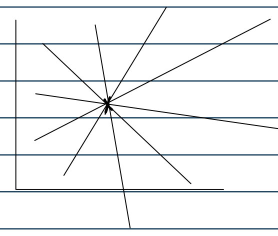
$$\boxed{SST = SSR + SSE}$$

→ coefficient of determination $(R^2) = \frac{SSR}{SST}$ (proportion of the variation in dependent variable explained by the independent variable)

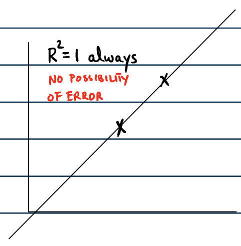
$(-\infty, 1]$

→ How many observations required to perform regression

one?



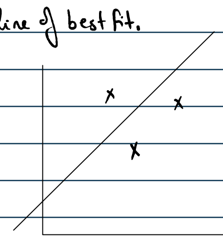
two?



→ With $R^2=1$ always, the strength of the relationship b/w x & y can't be assessed. we are not looking at the big picture.

→ Remember equation of regression, $y = \beta_0 + \beta_1 x_1 + \epsilon$.
Regression requires a possibility of error.

→ With 3 observations error is introduced, NOW we can find the line of best fit.



$n=3$

$df = 1 \; (n-(k+1))$

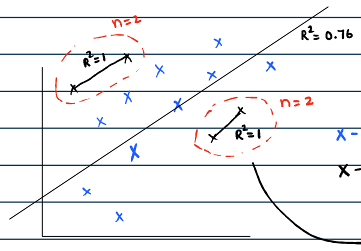
$(k+1)$ represents the number of parameters being estimated (including the intercept)

k → # of independent variables

n → # of observations

degree of freedom for the residuals

$$df\_total = df\_model + df\_residual$$
$$n-1 = k + n-(k+1)$$



x - The big picture (can be used to make predictions in the future) ✓

x - narrow vision (memorized data instead of generalized relationship) ✗

→ No point in making predictions about data that can't move/vary.

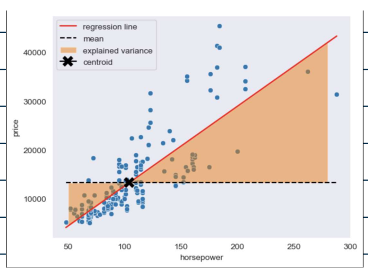
→ By adding more independent variables, degrees of freedom are reduced. Possibility of error in the model is reduced, i.e. R^2 continues to increase as we add more variables, we a fooling ourselves that the model is better, when in reality it's not. This is resolved by using the Adjusted R^2 which takes into account the reduced degrees of freedom.

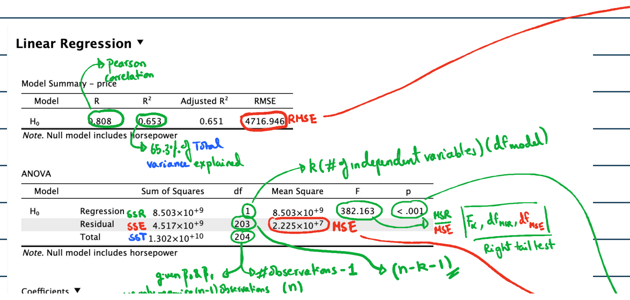
$$\boxed{Adj\, R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]}$$

→ If we are able to explain a lot of variation in the dependent variable when the data is free then we have a good model. More degrees of freedom (larger sample size) gives us a greater statistical power.

Regression Analysis

→ e.g. Car Price VS horsepower





Linear Regression

Model Summary - price

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.808	0.653	0.651	4716.946

Note. Null model includes horsepower

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	$8.503\times10^{+9}$	1	$8.503\times10^{+9}$	382.163	< .001
	Residual	$4.517\times10^{+9}$	203	$2.225\times10^{+7}$		
	Total	$1.302\times10^{+10}$	204			

Note. Null model includes horsepower

Coefficients

The standard error (RMSE) is the standard deviation of the error term, ϵ.

It is the average distance an observation falls from the regression line in units of the dependent variable.

We can think of s as a measure of how well the regression model makes prediction.

$$\boxed{RMSE = s = \sqrt{MSE}}$$

One of the assumptions for regression states that for a given x the error terms are normally distributed with $\mu = 0$ & $\sigma = RMSE$. Can also be used for outlier detection

H_0	Regression	SSR	$8.503\times10^{+9}$	1	$8.503\times10^{+9}$	382.163	< .001
	Residual	SSE	$4.517\times10^{+9}$	203	$2.225\times10^{+7}$ MSE		
	Total	SST	$1.302\times10^{+10}$	204			

Note. Null model includes horsepower

$\frac{MSR}{MSE}$ $\boxed{F_\alpha, df_{MSR}, df_{MSE}}$ Right tail test

given β_0 & β_1 we only require $(n-1)$ observations

observations − 1 (n)

$(n-k-1)$

Coefficients ▼

Model		Coef	Standard Error	[redacted]	t	p	95% CI Lower	95% CI Upper
H_0	(Intercept)	−3721.761	929.849		−4.003	< .001	−5555.163	−1888.360
	horsepower	163.263	8.351		19.549	< .001	146.796	179.730

$t = \frac{\text{coef}}{\text{std error}}$

$\boxed{t_{\alpha/2}, df_{MSE} \quad t_{1-\alpha/2}, df_{MSE}}$

$\boxed{H_0: \beta_k = 0 \quad H_a: \beta_k \neq 0}$

two tail test

$\boxed{\frac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}}}$

$\beta_k \pm t_{\alpha/2}\,\text{std error}$

model makes prediction.

$\boxed{RMSE = S = \sqrt{MSE}}$

Does a statistically significant linear relationship exist b/w the independent & dependent variables? Is the overall F-test or t-test (in simple linear regression these are actually the same thing) significant. $\boxed{F = t^2}$ only for SLR.

$\boxed{\begin{array}{l} H_0: \beta_1, \beta_2, \beta_3 \ldots \beta_k = 0 \text{ Expect Intercept} \\ H_a: \text{At least one regression coefficient is not equal to } 0 \end{array}}$

MSE (s^2) is an estimate of σ^2 the variance of the error, ε. In other words, how spread out the data points are from the regression line.

$\boxed{s^2 = MSE = \frac{SSE}{df_{SSE}}}$

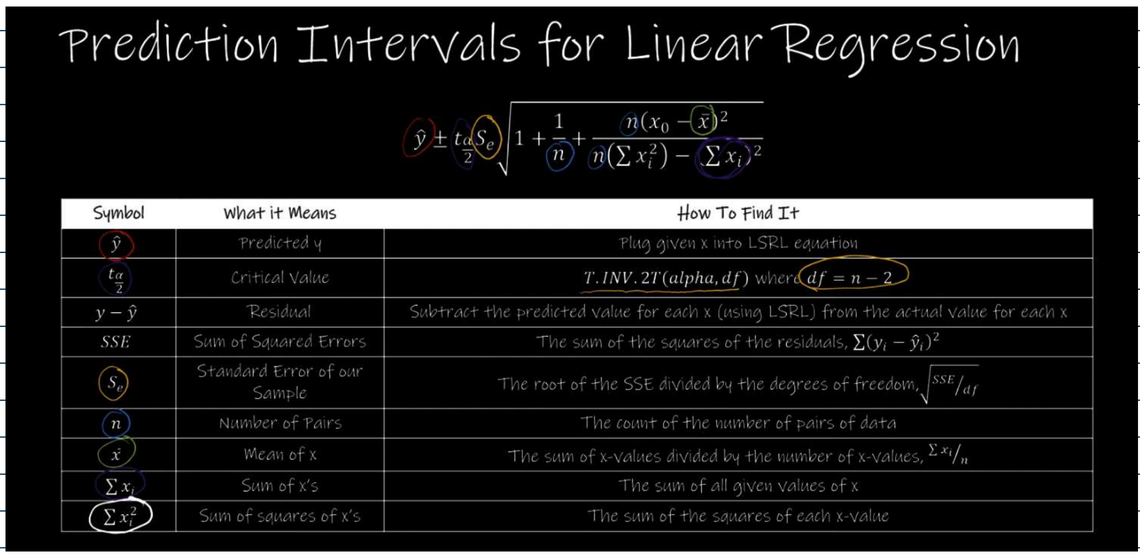
→ Would a regression analysis offer anything more than the $\bar{y}$ model? Using this nonregression model ($\bar{y}$) as a worst case, we can analyse the regression line to determine whether it adds a more significant amount of predictability of y than the $\bar{y}$ model.

→ If the slope is not different from 0, the regression line is doing nothing more than the $\bar{y}$ line in predicting y.

→ Remember that for each sample we will obtain a different slope value.
The Question: if all the pairs of data points for the population were available, would the slope of the regression line be different from 0?

→ For simple linear regression $\boxed{F = t^2}$

Note: We can only rely on the numbers above IF CERTAIN ASSUMPTIONS HOLD



Prediction Intervals for Linear Regression

$$\hat{y} \pm t_{\frac{\alpha}{2}} S_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x_i^2) - (\sum x_i)^2}}$$

Symbol	What it Means	How To Find It
$\hat{y}$	Predicted y	Plug given x into LSRL equation
$t_{\frac{\alpha}{2}}$	Critical Value	$T.INV.2T(alpha, df)$ where $df = n - 2$
$y - \hat{y}$	Residual	Subtract the predicted value for each x (using LSRL) from the actual value for each x
SSE	Sum of Squared Errors	The sum of the squares of the residuals, $\sum(y_i - \hat{y}_i)^2$
S_e	Standard Error of our Sample	The root of the SSE divided by the degrees of freedom, $\sqrt{SSE/_{df}}$
n	Number of Pairs	The count of the number of pairs of data
$\bar{x}$	Mean of x	The sum of x-values divided by the number of x-values, $\sum x_i/_n$
$\sum x_i$	Sum of x's	The sum of all given values of x
$\sum x_i^2$	Sum of squares of x's	The sum of the squares of each x-value