DataSoc
THE DATA SOCIETY

INTRO TO MULTIVARIATE REGRESSION ANALYSIS

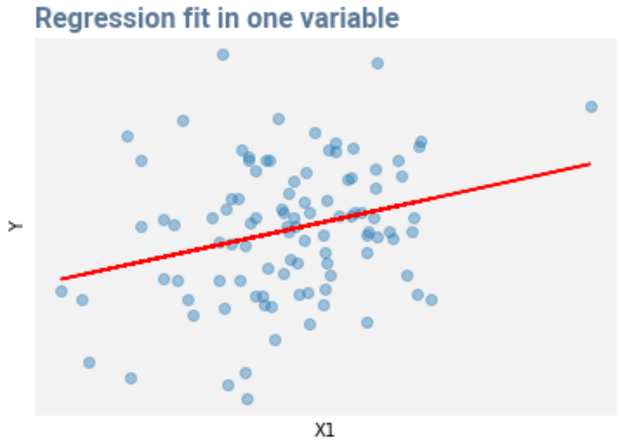
Applied Data Science Crash Course

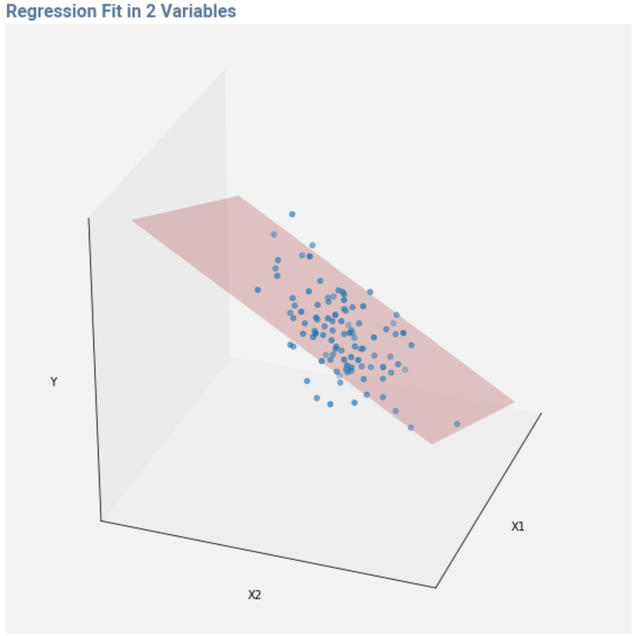
https://youtu.be/AOEKMYk9RiQ

INTRODUCTION TO LINEAR REGRESSION

2D Case

A simple analysis of the rate at which one variable increases with respect to another.





3D Case

Exploring how 2 variables contribute to the impact on a third variable, this forms a prediction surface rather than a line.

N Dimensional Case

Difficult to visualise, but a general case in which many variables of interest can contribute to the dependent variable prediction. This surface is known as a hyper-plane.

1 is for the **intercept** beta_0 in the model

X_i,j denotes the **i^th** observation and the **j^th feature**

$$\begin{bmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

Predicted values for each fo the **n** observations.

The (n x p+1) **design matrix** with n observations and p features for each observation

The **weights** on each of the **p features**, also known as the parameters of the model.

A linear regression is a formula relating a simple weighted combination of independent variables to predict the outcome

Shift / Intercept

Slope dY/dX1

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Model describes the slope of the regression in each dimension, the epsilon denotes that the model is not a perfect description, there is some noise in the data

$$\tilde{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Prediction based on inputs for each of the p-variables.

Assumption: A straight line is appropriate to describe the relationship between x and y

A LINEAR REGRESSION IS THE SOLUTION TO AN OVER-CONSTRAINED SYSTEM OF EQUATIONS - THUS NO PREDICTION IS PERFECT -> THERE IS A PREDICTION ERROR

System of Equations

A regression tries to explain the behaviour of the dependent variable through only one or two variables, this necessarily means that we can't perfectly satisfy the regression equation for each data point

This **system of equations** can only be **solved if n <= p.** If there are more equations than unknowns there is no solution.

$$Y_1 = \beta_0 + \beta_1 X_{1,1} + \cdots + \beta_p X_{1,p}$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_{n,1} + \cdots + \beta_p X_{n,p}$$

But we can try to find a solution which gets as close as possible up to some error.

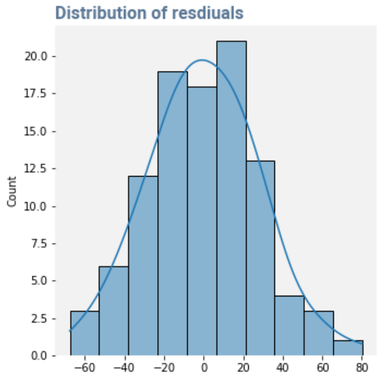
$$Y_1 = \beta_0 + \beta_1 X_{1,1} + \cdots + \beta_p X_{1,p} + \varepsilon_1$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_{n,1} + \cdots + \beta_p X_{n,p} + \varepsilon_n$$

The notion of error

The parameters we choose will try to minimise the overall **error** in the system, and this is equivalent to **minimising the vertical distance between the predicted and the true value.**

Independent Errors

In statistics we assume that the **errors are independent** of one another and that they are randomly **drawn from the same distribution**. This is a generic assumption which enables us to check whether we may have left out important information.



$$\varepsilon \sim N(0, \sigma^2)$$

We expect errors to come from a **normal** / bell curve **distribution,** which is common for a white noise, which contains no additional information.

Assumption: Independent Errors drawn from same distribution

The Method we employ to find the parameter values must be the method which minimises the total error in the predictions

The Sum of Square Errors **(SSE)**

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 = (Y - X\beta)^T (Y - X\beta)$$

Can be compactly written in terms of **matrix multiplication.**

$$SSE = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$$

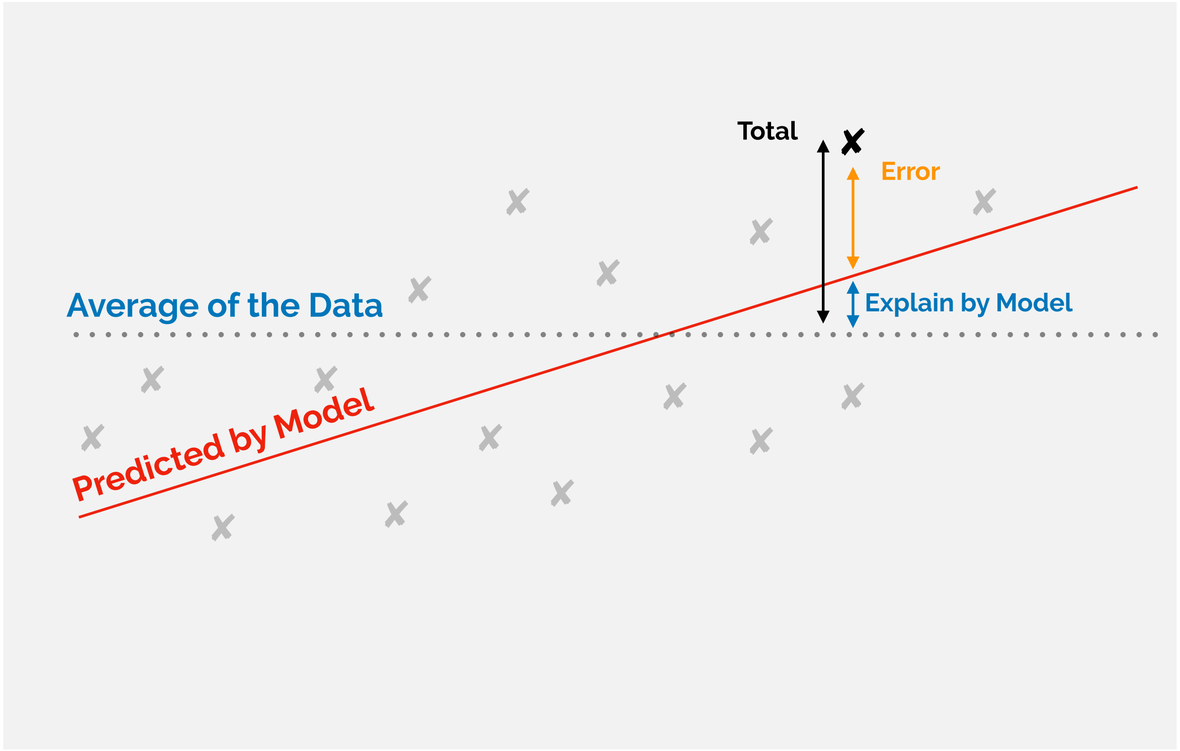
Taking derivatives wrt the vector of parameters we can find the values of the parameters which **minimise the SSE.**

$$\frac{\partial SSE}{\partial \beta} = -2X^T Y + 2X^T X\beta = 0$$

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

See that it is crucial that an XTX is invertible.

A NOTE ON EXPLAINED VARIATION OR 'R-SQUARED'



The R Squared - or the % of variance in the data explained by the model above a model which makes a prediction of the mean of observations.

The total variation in the data is the sum of squares of the *total arrow* in the diagram over all observations. By definition this is the (unscaled) variance of the data (no division by n-1).

The residual variation in the data is the sum of squared errors (SSE) / (n-1) as previous defined.

Hence **Total Variation = Explained Variance + Error Variance**

More commonly we simply work with unscaled sums of squares (ie. the variance multiplied by n-1.

Hence **SST = SSM + SSE** where SSM is Explained Variance * n-1 (Sum of Square Model)

Hence **R^2 = SSM/SST = 1 - SSE/SST** the latter is easiest to calculate directly from model output.

IN GENERAL WE WILL CONSIDER MORE THAN 2 VARIABLES AND MANY DATA POINTS WE USE MATRIX NOTATION AND CONDUCT THE MINIMISATION OF THE MATRIX EQUATION

Encoding Data in Matrices

Recalling our exploration of Numpy and Pandas we know how to store data in matrices and we know how to do matrix multiplication. Thus we can encode our regression equation into matrix form.

$$\begin{bmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\tilde{Y} = X\beta$$

$$Y = X\beta + \varepsilon$$

Recall error introduced as we can't get a perfect fit through all the points

Likelihood vs Method of Moments Parameter Estimation

Previously we obtained parameter estimates by minimising the **SSE.** This is called a **method of moments** and has no statistical basis. A more rigorous approach is to assume the errors are a white noise drawn from a normal distribution and minimise the **likelihood** of observing the errors. This will yield the same result here, however for more complex model derivations likelihood based methods yield better estimates.

Likelihood; The probability of observing the errors.

Independence Assumption — PDF of Normal Distribution

$$L(\beta) = P(\varepsilon_1,...,\varepsilon_n) = \prod_{i=1}^{n} P(\varepsilon = \varepsilon_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\varepsilon_i^2}{2\sigma^2}) = [\frac{1}{\sqrt{2\pi\sigma^2}}]^n \exp(-\sum_{i=1}^{n} \frac{\varepsilon_i^2}{2\sigma^2})$$

$$L(\beta) = [\frac{1}{\sqrt{2\pi\sigma^2}}]^n \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2$$

The Log-Likelihood.

$$\ell(\beta) = n\log[\frac{1}{\sqrt{2\pi\sigma^2}}] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 = n\log[\frac{1}{\sqrt{2\pi\sigma^2}}] - \frac{1}{2\sigma^2} SSE$$

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2} \frac{\partial SSE}{\partial \beta} = 0$$

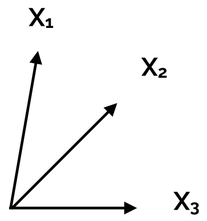
We see minimising log likelihood is same as minimising SSE!

Matrix Inversion

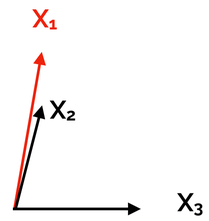
A practical note on this implementation is that matrix **XTX is only invertible** when the the columns / features are linearly independent. I.e there is no **Multicollinearity** in the feature space.

Assumption: No Multicollinearity

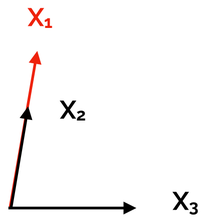
If X1, X2, X3 are any columns or rows of the design matrix (either observations or features) then if any Xi is a multiple of Xj then we have multicollinearity.



Independent Vectors, no issues.



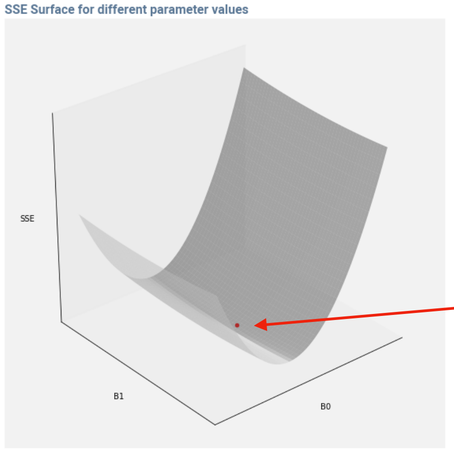
Some features are **very similar**. This will lead to **large errors** in parameter estimates.



Multicollinearity - Two or more features are scaled version of each other. **Parameters cannot be found.**

Visualising the 2D Minimisation Problem

The Cost Function for the 2d case is in fact a parabolic function, by setting the gradient to zero we find the point of zero slope in all directions which directly corresponds to the minimum of the bowl



Minimising values of the two parameters are at the bottom of the bowl (where gradient is zero)

ASSUMPTIONS WE HAVE MADE ARE CRITICAL TO GETTING REASONABLE ESTIMATES

Assumption: A straight line is appropriate to describe the relationship between x and y

STOP

Everything else is meaningless unless this is satisfied. either in the original variables or under some transformation.

Assumption: No Multicollinearity

STOP

If you get a large condition number > 50 or so, then start to worry, you may have features that say the same thing, these will lead to large numerical errors in parameter estimates and high variance as the matrix is almost not invertible.

Assumption: Independent Errors drawn from same distribution

Observations / Errors are independent → If they are **not independent you need to use a more complicated model** as there may be a hierarchy or time series behaviour in the data. Or aggregate data.

Errors have zero mean → If they do not, **you are missing a categorical variable** / intercept.

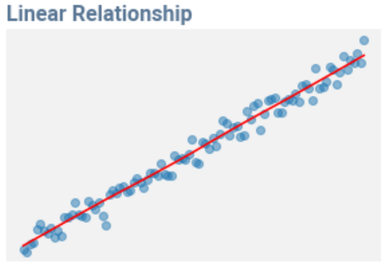
Errors have constant constant variance → If they do not there may be **other variables or a transformation of an existing variable required.**

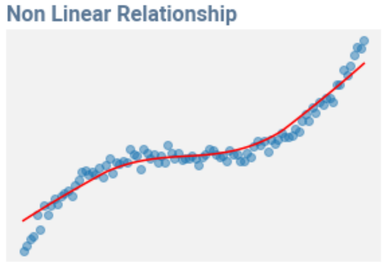
Errors are normally distributed → Not the end of the world, unless you have a small sample size < 30 observations. Though **this may be telling you you are missing features in your model.**

PLOTS TO CHECK THE ASSUMPTIONS

Linearity

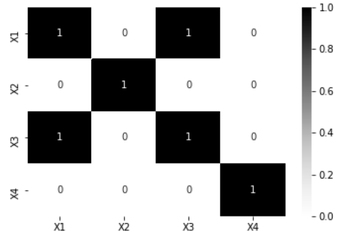
Plot each variables against Y, use a loess line to asses whether this is accurate.





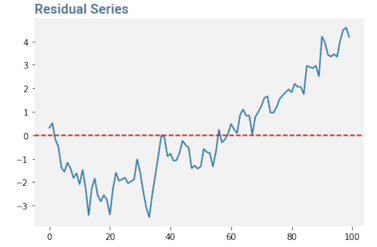
Multicollinearity

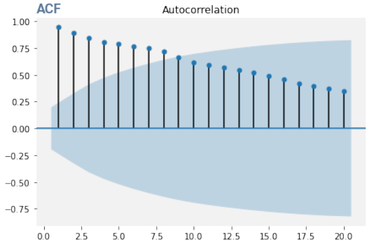
Correlation greater than 0.6 is probably something to watch out for. We can also check Variance Inflation Factor (In a few slides)



Independence

Plot the autocorrelation function and check that all lags < cut off.





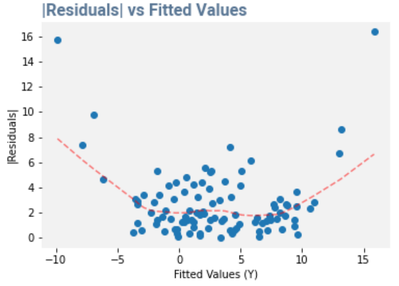
Zero Mean

Despite this being an assumption of the model, it is always true up to some numerical error, a quick proof here shows this to be true!

$$\tilde{\varepsilon} = Y - X\tilde{\beta}$$
$$\tilde{\beta} = (X^T X)^{-1} X^T Y$$
$$X^T X\tilde{\beta} = X^T Y$$
$$0 = X^T (Y - X\tilde{\beta})$$
$$0 = X^T\tilde{\varepsilon} = \begin{bmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{1,n} \\ \vdots & \ddots & \vdots \\ X_{p,1} & \cdots & X_{p,n} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \vdots \\ \tilde{\varepsilon}_n \end{bmatrix}$$
$$\frac{1}{n}\sum \tilde{\varepsilon}_i = 0$$

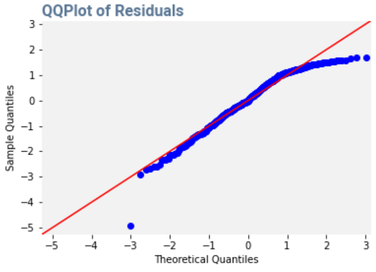
Constant Variance

Plot Abs(residuals) vs fitted values and check for widening / narrowing.



QQ - Plot

Plots the quantiles of normal vs observed distribution, should be a straight 45 degree line with some veer off in the tails.



TESTS TO CHECK THE ASSUMPTIONS

Some of the below are statistical tests, i.e there is some transformation of the observed residuals / data which should satisfy a known distribution under the modelling assumptions. If we find that the probability of observing the transformed qty is small then the hypothesis is likely false and the modelling assumptions are violated.

Linearity

We can't test for this directly but it will definitely show up in the residuals if its not true.

Independence

The Ljung Box Test can be applied to the residuals, if the test statistic does not come from the assumed distribution then autocorrelation is present.

Multicollinearity

We can use variance inflation factor coefficient. The procedure works by removing all variables VIF > 5 one by one and recomputing the model + VIF. Additionally a large condition number is usually present, indicating a small change in Y would lead to a large change in Beta and thus and unstable solution.

$$X_i = \beta_1 X_1 + \cdots + \beta_n X_n$$

$$VIF_i = \frac{1}{1 - R_i^2}$$

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

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

$$\kappa(X^{-T}(X^T X)) = \left\|(X^T X)^{-1} X^T\right\| \left\|X^{-T}(X^T X)\right\|$$

Zero Mean

As previously stated in simple regression we won't need to check for this at all. But a t-test could be used to confirm this.

Constant Variance

Usually we simply visually check for this.

Normality Check

Jacque Bera statistic can be used to test for normality

P-Values: These are outputted by all statistical tests, and are between 0 and 1 stating the probability that the null hypothesis is true (i.e that data is zero mean or that no autocorrelation is present.) A value smaller than 0.05 is usually cause to reject this hypothesis, i.e one of our assumptions is void.

Note: Usually these tests are stated with the null hypothesis are the assumption we would make in the absence of evidence to the contrary.

ADDITIONAL GOTCHA'S – INFLUENTIAL POINTS

Some of the below are statistical tests, i.e there is some transformation of the observed residuals / data which should satisfy a known distribution under the modelling assumptions. If we find that the probability of observing the transformed qty is small then the hypothesis is likely false and the modelling assumptions are violated.

Outlier

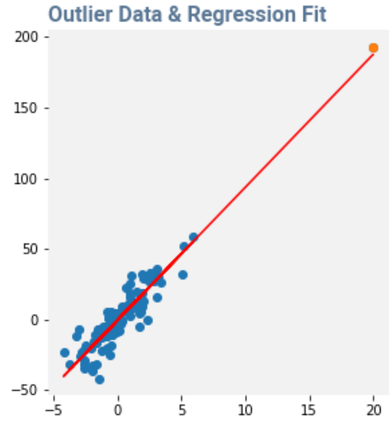
Any residual (and corresponding observation) which is more than 3 standard deviations from the mean. A more rigorous test con be performed by checking Bonferroni p=values.

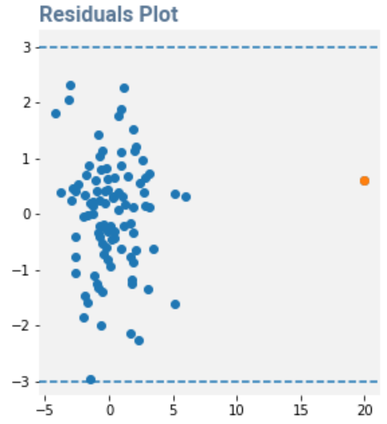
Influential Point

A point which is both an outlier and a leverage point will significantly skew the regression slope towards the outlier, this is undesirable as the outlier does not reflect the majority of the observations

Leverage Point

Any observation which when perturbed would have a substantial effect on the slope of the regression line.





THE HAT MATRIX

The hat matrix is the matrix that takes any point and projects it onto the fitted line as shown in the diagram, where the red line is where the points (X,Y) and projected onto (X,Y_hat)

Parameter Estimates

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

Fitted Values

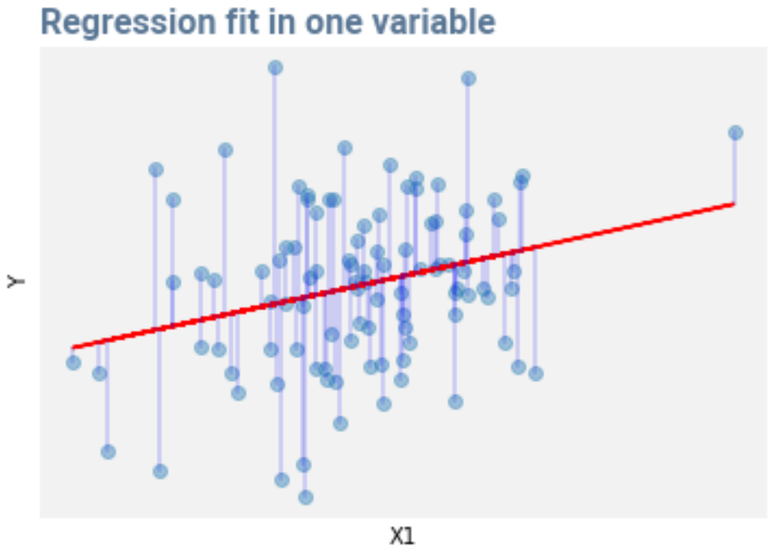
$$\tilde{Y} = X\beta = X(X^T X)^{-1} X^T Y$$

Relabel

$$\tilde{Y} = HY$$

Hat Matrix

$$H = X(X^T X)^{-1} X^T$$



Leverage

The diagonals of the hat matrix **H_ii** , are the leverage for observation i in the fit.

Why is the diagonal a leverage metric?

$$\tilde{y}_i = (HY)_i$$

$$\frac{\partial \tilde{y}_i}{\partial y_i} = h_{ii}$$

The rate of change of the fitted value with respect to the observed value is the diagonal of the hat matrix. Thus if the diagonal is large then a small change in the value of the observed value could massively change the fitted value.

Leverage is thus a measure of 'self sensitivity'.

RESIDUALS VS ERRORS

Errors

$$Y = X\beta + \varepsilon$$

Under Modelling Assumptions the errors are **independently identically distributed. with variance:**

$$Var(\varepsilon) = \sigma^2$$

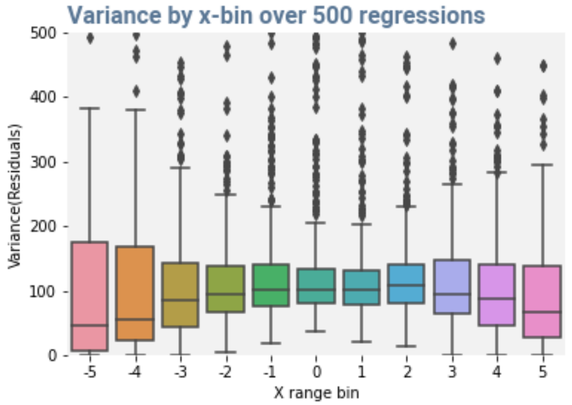
Residuals

$$\tilde{\varepsilon} = Y - \tilde{Y} \qquad \tilde{Y} = X\tilde{\beta} \qquad \tilde{\beta} \approx \beta$$

However the least squares fit to the data we recall guarantees that the mean of the residuals is zero, thus the **residuals are clearly dependent**

$$\tilde{\varepsilon} = Y - X\tilde{\beta}$$
$$\tilde{\beta} = (X^T X)^{-1} X^T Y$$
$$X^T X\tilde{\beta} = X^T Y$$
$$0 = X^T (Y - X\tilde{\beta})$$
$$0 = X^T\tilde{\varepsilon} = \begin{bmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{1,n} \\ \vdots & \ddots & \vdots \\ X_{p,1} & \cdots & X_{p,n} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \vdots \\ \tilde{\varepsilon}_n \end{bmatrix}$$
$$\frac{1}{n}\sum \tilde{\varepsilon}_i = 0$$

It also makes intuitive sense that the **variance of residuals must be larger for extreme values** as the central points won't change much irrespective of what slope we fit through the data. Below I ran 500 regressions with draws from the same data generating process and compare the variance of the residuals in slices along the x axis.



Conclusion: The residuals at extremities have higher variance and therefore would be more likely to be considered an outlier by the fact they are usually higher variance. Thus instead we must the variance by the leverage of the data point / studentize the residuals.

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

TESTING FOR OUTLIERS, INFLUENTIAL AND LEVERAGE POINTS

Outlier

Any residual (and corresponding observation) which is more than 3 standard deviations from the mean. A more rigorous test con be performed by checking if studentised residuals have Bonferroni p=values < 0.05/m where m is the number of observations.

Influential Point

A point which is both an outlier and a leverage point will significantly skew the regression slope towards the outlier, this is undesirable as the outlier does not reflect the majority of the observations

Leverage Point

Any observation which when perturbed would have a substantial effect on the slope of the regression line.

$$t_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$$

Studentised residuals are t-distributed with n-m-1 degrees of freedom where m is the number of parameters in the model

We can use **Cooks distance** to measure this, which calculates the change in the residual error in the model when a particular data point is removed. (The formula for this is not particularly insightful)

Leverage points aren't necessarily a problem unless they are also influential or outliers which both require investigation. If a point has high leverage you may find that you have insufficient data for this region of the feature space.

INTERPRETING MODEL FIT OUTPUTS

P-Values

Identify whether a coefficient / feature in the model is non zero - ie adds some additional information relevant to accurate prediction.

Parameter Values (Mean)

Parameter Values (StdError)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	5342.
Date:	Tue, 07 Sep 2021	Prob (F-statistic):	7.39e-111
Time:	10:53:13	Log-Likelihood:	-216.18
No. Observations:	100	AIC:	442.4
Df Residuals:	95	BIC:	455.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.7916	0.219	21.884	0.000	4.357	5.226
X1	4.1788	0.262	15.960	0.000	3.659	4.699
X2	-0.0949	0.107	-0.883	0.379	-0.308	0.118
X3	-0.0810	0.134	-0.604	0.547	-0.347	0.185
X4	-2.9793	0.033	-89.965	0.000	-3.045	-2.914

Omnibus:	0.854	Durbin-Watson:	2.200
Prob(Omnibus):	0.652	Jarque-Bera (JB):	0.873
Skew:	-0.048	Prob(JB):	0.646
Kurtosis:	2.552	Cond. No.	23.1

Explained Variance

Always use Adj. R-Squared as it penalises for extra parameters in the model

AIC

AIC is an information criterion - lower is better, usually used instead of R^2 when accuracy is desired over explainability.

Durbin Watson

Quick way to identify autocorrelation in residuals if the value is <1.5 or >2.5.

JB-Score

Also a p-value (< 0.05 suggest reside not normally distributed)

Condition Number

Anything lower than 50 is probably fine

FEATURE ENGINEERING

Smoothing

Taking a moving average or multiple to isolate trends and reversions

Volatility Scaling

Normalising a series by volatility in a rolling window to get stable feature

Bucketing

When data is extremely noisy or sparse grouping into categories may be only viable option.

Log / Exp

If data is particularly skewed towards a particular value the these will help fix issues in model residuals

Standardising

Centring and rescaling data to avoid numerical issues with numerical solvers

Implied Values

A known formula relating to the data may be used to back out other values which may be significant.

SURVEY OF OTHER RELATED METHODS

Auto-Regressions

Where dependence is on the history of the last p points in a time series.

Arch / Garch

Where we have a white noise process with volatility which changes over time. Many packages for AR-Garch type behaviour

Vector Autoregression

Where dependence is on the history of the last p points in a time series and that of several other time series

Exogenous Variables

ARX process uses a time series and fixed attribute variables in forecast

ARMA

Where dependence is on the history of the last p points in a time series and a the error in the last q forecasts. A cheap (non rigorous) trick to get back to an AR is by subtracting a moving average from the time series. Or use a large p.

Seasonality / Trend

These usually have to be removed from the process, however for more complex seasonal behaviour we can also express these as a high order AR / SAR model.

Bayesian Methods

All methods discussed can be expressed in a bayesian manner, where prior beliefs and uncertainty about model parameters can be incorporated prior to model fitting. Eg. a regression which takes an initial parameter mean value from a previous fit at a past time period.

GLMs

For recessions with binary outcomes (logistic regression) or rate target ('poisson regressions) or multiple outcomes. The regression framework can be extended, however a straight regression is preferable when we can avoid this type.

Cox Models

When a base rate of some event may be rescaled by some exogenous factors we can perform this type of regression.

Splines

When complex trends must be removed from observations before a useful model can be fitted we can use cubic splines

Penalised Methods

We can add penalisation terms to Loss Function to reduce the size of model weights and thus select from a large set of features.

PCA

We can reduce the number of variables in a problem by rotating the feature space to obtain a subset of uncorrelated factors.

Neural Nets

For problems with high signal to noise ratio (eg. images or speech) where complex signals can be extracted.