Regression Modelling

Maxwell Munford
PhD Biomechanics, Mechanical Engineering

Objectives

- Correlation
- Linear Regression
- Multi-Variable Regression
- Interpretation and Application
- NOT focussing on theory

What Is Regression?

… and why do we use it?

4

What Is Regression?

What Is Regression?

- A statistical analysis that attempts to predict the effects of one or more variables on another variable
- Correlation is a mutual relationship between 2 or more variables

The Process

- Scatter plots
- Measure the degree of linearity between two variables
- Quantify this relationship

What do we assume?

- Assume a causal relationship
- Equation of line of best fit
- Test the significance
- Analyse residuals
- Predication

Analyse model

8

Why Do We Start With Scatter Plots?

x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y
55.3846	97.1795	32.33111	61.4111	53.36657	90.20803	50.48151	93.2227	38.33776	92.47272	58.21361	91.88189	57.61323	83.90517	51.14792	90.86741	55.99303	79.27726	51.20389	83.33978	47.6952	95.24119	52.87202	97.34322	65.81554	95.58837
51.5385	96.0256	53.42146	26.18688	52.80198	90.08806	50.28241	97.60998	35.75187	94.11677	58.19605	92.21499	51.27439	82.81798	50.51713	89.10239	50.03225	79.01307	58.97447	85.49982	44.60998	93.07584	59.01414	93.57487	65.67227	91.9334
46.1538	94.4872	63.9202	30.83219	47.05413	90.45894	50.1867	99.69468	32.76722	88.51829	58.71823	90.31053	50.75391	76.75413	50.20748	85.46005	51.28846	82.43594	51.87207	85.82974	43.85638	94.08587	56.37511	96.30515	39.00272	92.26184
42.8205	91.4103	70.28951	82.53365	42.44843	89.5077	50.32691	90.02205	33.72961	88.62227	57.27837	89.90761	37.02118	81.95447	50.06948	83.05767	51.17054	79.16529	48.17993	85.04512	41.57893	90.30357	37.8392	94.35944	37.7953	93.53240
40.7692	88.3333	34.11883	45.73455	42.70404	90.44263	50.45621	89.98741	37.23825	83.72493	58.08202	92.00815	42.88176	80.18477	50.56285	82.93782	44.37791	78.16463	41.6832	84.01794	49.17742	96.61053	39.88537	90.63466	35.5139	89.59915
38.7179	84.8718	67.67072	37.11095	37.37894	90.14414	50.46485	82.08923	36.0272	82.04078	57.48945	88.08529	37.15571	84.95411	50.28853	82.97525	45.01027	77.88086	37.89042	82.56749	42.65225	90.56064	44.07741	84.12585	39.21945	83.54348
35.641	79.8718	53.25913	97.47577	32.5307	70.1574	30.5016	82.30713	39.23928	79.26372	28.08874	63.51079	38.73181	83.78936	25.58348	82.91489	48.55982	78.78837	39.54897	80.8126	30.12332	81.1443	31.49703	67.91256	31.5882	77.02805
33.0769	77.5641	63.51498	25.10008	33.36526	70.45934	30.48957	77.72988	39.78452	82.26607	28.08547	63.5902	31.00053	82.57948	25.48358	82.82908	42.14227	76.88063	39.64957	82.66453	31.91552	79.79041	25.9526	73.03318	32.65778	80.11384
28.9744	74.4872	67.98054	80.95717	32.6561	70.05107	30.45025	79.83489	35.16603	84.15649	28.08727	63.12328	25.9855	74.34939	25.44353	82.8742	41.02697	76.40959	34.7506	80.01109	27.88772	75.45558	36.43472	62.29405	31.6416	77.61083
26.1538	71.4103	67.37247	29.7204	22.95509	70.4269	30.51839	68.25258	40.62212	78.5421	27.57803	62.82104	23.71457	75.09805	25.56511	82.9241	34.57531	72.72484	27.56084	72.84783	27.54823	75.16476	24.99748	75.44156	30.98714	63.94403
23.0769	66.4103	15.56075	80.06564	27.14984	70.21081	30.51166	66.09372	39.18191	79.81904	27.77992	63.51815	23.07732	71.72054	25.29884	82.82118	31.1686	69.24542	24.63554	71.61071	20.20978	67.51664	32.24628	70.10595	27.43963	65.74440
22.3077	61.7949	71.79077	71.06547	26.16691	70.50032	30.49808	62.38699	42.43089	75.13363	28.589	63.02408	24.93914	72.58029	27.55148	51.48739	32.64425	70.73256	20.95946	66.0415	24.44345	54.47366	18.10947	53.70811	30.57837	65.52195
22.3077	57.1795	70.24255	24.1096	26.33146	70.49456	30.62023	60.86008	43.08126	75.3467	28.73914	62.72086	17.8935	71.10969	27.53047	51.41485	26.66665	62.98097	20.68915	62.72131	24.63083	53.39574	21.7961	49.49953	28.8094	59.621
23.3333	52.9487	64.93744	81.5542	22.00371	50.49944	30.49219	55.87892	44.36072	70.47421	27.0246	62.90186	23.67308	64.87084	27.09557	52.0768	26.45926	62.63467	19.2882	62.06306	21.03797	49.8293	24.15049	47.18853	27.93952	50.98506
25.8974	51.0256	62.21352	21.47584	23.28475	50.47435	30.47463	52.32216	44.32478	71.0459	28.80134	63.38904	23.74907	63.55717	27.43925	51.71208	25.53801	60.63024	20.0245	61.34262	27.10624	54.71955	28.6032	44.22372	33.84995	46.28447
29.4872	51.0256	67.2694	18.70897	27.63404	50.50367	30.501	48.0697	47.55216	66.4604	27.18646	63.55873	32.21518	65.60096	27.87826	50.70891	26.32071	62.39628	35.46952	43.11588	28.08962	60.51221	20.932	51.64624	32.62846	54.06373
32.8205	51.0256	40.5702	79.37296	32.03711	50.50007	30.49442	45.77348	48.66492	64.961	29.28515	63.38361	29.43684	67.55453	27.33887	51.65305	26.16339	62.02681	36.89433	47.70656	31.13573	61.43191	35.43553	34.96104	32.43865	49.90303
35.3846	51.4103	74.74118	21.10164	29.33651	50.46189	30.49954	49.3821	46.28063	68.02842	39.40295	51.15086	40.0501	71.88903	27.6784	51.81195	25.21379	60.2734	39.05555	55.54697	37.70084	68.41841	44.37738	57.68474	35.83829	53.35149
40.2564	51.4103	71.76832	20.01106	40.81695	50.48776	50.14439	47.01428	50.56324	62.62383	28.81133	61.35785	37.5748	69.88874	52.63566	51.41855	26.85953	63.14183	46.95708	65.24041	43.39269	63.45759	46.31369	55.68559	37.14204	49.74199
44.1026	52.9487	76.16692	75.93617	42.73908	50.48316	49.9129	55.58784	52.63096	59.96394	34.30396	56.54213	49.31047	67.16876	52.02521	52.12301	31.76067	69.79766	37.31045	45.25875	51.2624	49.8293	46.38607	51.00216	34.64455	54.44413
46.6667	54.1026	65.62366	15.5828	51.86789	50.49327	50.33743	57.95531	54.63621	57.49384	29.60276	60.15735	47.34802	72.59161	52.88116	50.62155	39.79329	75.80796	40.00967	60.98658	43.28671	43.05686	42.49544	56.90368	35.6754	49.62564
50	55.2564	50.85063	13.9876	43.371	50.50084	50.41898	60.52491	51.34603	61.60055	49.11616	63.66	50.15741	71.16173	52.95261	50.07474	45.61744	78.11997	48.01439	65.71282	55.90592	54.25267	44.30052	58.13296	64.99459	49.46217
53.0769	55.641	33.02407	24.46783	53.0428	50.4875	50.20526	57.92422	57.13486	53.87797	39.61755	62.92519	58.21001	68.04454	52.52055	51.50244	53.77146	79.30293	53.70378	66.38948	50.19896	52.22503	46.98828	55.42064	66.12915	51.64373
56.6667	56.0256	39.70633	84.27529	56.61891	50.50337	50.43532	56.43932	54.17124	58.05981	43.23308	63.16522	56.60891	70.40895	52.34282	51.86195	56.45143	79.24711	63.0675	64.035	55.76155	54.03052	52.42152	47.86701	65.79065	55.94537
59.2308	57.9487	45.59648	9.763349	70.56829	50.49624	69.50579	59.28492	50.94238	62.09739	64.89275	65.81418	52.83382	72.55189	51.92759	52.25779	66.09372	77.01782	62.04803	63.84586	58.32314	60.03106	64.091	68.65099	67.50178	60.85022
61.2821	62.1795	42.96805	17.94546	67.87089	70.06803	69.53407	64.89298	66.51292	59.07128	62.49015	74.58429	61.81859	76.11282	52.71377	51.19794	56.92584	79.20632	59.83997	64.47676	70.17146	70.28222	63.01687	68.22526	67.62951	61.42913
61.5385	66.4103	52.49441	16.05111	58.82155	70.45578	69.51011	61.32962	64.39791	56.17488	68.98808	63.23215	61.72763	77.06924	50.4438	82.94183	58.98825	78.91726	55.16094	65.23731	60.60026	59.75331	67.30258	64.41346	63.91006	62.97232
61.7949	69.1026	46.08228	23.11046	60.18693	70.48355	69.52632	56.99528	68.68659	62.23734	62.10562	75.99087	62.19225	76.15682	50.2167	83.75234	57.90227	79.08087	61.27975	65.7664	63.76026	61.18564	67.22596	62.47641	66.54936	67.72559
57.4359	55.2564	74.24771	20.31419	59.43552	50.42063	49.94724	48.22958	65.02779	57.09626	32.46185	62.8819	58.11877	73.11896	52.18418	51.97525	64.81287	77.48694	60.83492	64.60377	54.3895	53.8119	50.30999	47.26788	64.05154	48.82009
54.8718	49.8718	64.56826	83.63963	49.58584	50.43498	50.0631	58.08504	53.3045	40.88384	41.3272	49.07025	58.27041	35.5969	52.7921	51.0734	60.34979	78.72945	61.52059	64.79186	55.2413	52.59357	45.41264	53.8708	68.19329	46.54354
52.5641	46.0256	74.02169	76.12827	41.77974	50.47064	50.4899	39.20541	52.94518	40.38463	44.00715	46.44967	59.76653	32.62125	52.58972	51.33809	48.89056	16.80654	36.91654	41.09525	49.07013	48.89694	54.02061	47.2699	65.27839	44.82035
48.2051	38.3333	62.39118	5.623071	46.12769	30.44941	49.68886	42.01851	45.22574	30.55523	44.07406	34.5532	52.53553	27.17412	52.02885	52.17684	47.54951	17.0714	38.5022	41.56716	46.01812	46.25772	49.84202	50.57655	34.99766	28.63299
49.4872	42.1795	74.18904	68.13358	38.03378	50.4909	50.33348	43.58867	50.8433	37.74463	44.00132	33.90421	50.89252	29.14528	52.72925	51.20177	30.88943	26.32904	48.66437	30.68067	44.75959	41.97461	50.7133	48.77229	33.92377	53.55869
51.0256	44.1026	28.23678	56.1396	44.54684	50.41591	50.44675	57.23059	49.65929	36.27344	45.0063	38.29902	49.96755	28.4602	52.88431	50.44144	28.97284	29.05047	50.28525	30.33792	51.60062	50.03247	55.05455	45.98627	66.00445	52.68463
45.3846	36.4103	75.77194	69.82923	46.23812	30.49561	50.42091	43.43796	47.14279	32.02506	44.44384	36.01908	51.34788	27.36426	52.5093	51.41621	35.16136	22.78685	42.27633	34.52764	36.46588	35.52419	37.5437	32.88402	37.02326	32.11848
42.8205	32.5641	75.85523	62.51704	49.73975	30.47272	50.37405	32.91824	41.27006	21.72182	42.17871	26.49212	37.55512	24.53766	50.86268	17.14563	45.87037	17.48939	54.03178	29.67235	33.3948	35.0168	39.7959	28.44322	34.69827	26.47535
38.7179	31.4103	65.97086	72.74486	39.11025	30.4787	30.47683	39.42461	47.13013	31.3099	44.04457	35.66224	39.19959	23.6805	50.89149	17.14132	32.83142	24.79754	37.32936	39.60204	37.58007	38.83157	28.52339	39.79667	34.86344	30.93921
35.1282	30.2564	21.77804	6.616625	39.14284	30.79286	30.50347	38.34856	40.91597	23.54159	41.64045	27.0931	36.38272	25.23801	25.85513	17.08191	39.48557	20.03305	41.38952	37.29606	36.61972	36.81141	31.39965	31.26526	31.09007	32.38375
32.5641	32.1795	67.7598	72.4212	34.62547	30.49974	30.50483	28.77149	38.66016	21.68955	41.93833	24.99152	41.10069	26.55261	26.02564	16.92501	33.44618	24.19064	40.07467	34.62369	36.16347	33.81931	39.24157	33.36142	34.92131	35.41844
30	36.7949	78.6172	52.57526	34.75421	30.49587	30.46222	32.75047	42.61255	24.19505	44.03915	33.55639	28.21781	28.28029	27.89317	50.66196	21.86358	46.05259	35.34968	47.14107	33.16776	31.24774	27.71406	42.25005	30.34336	32.52456
33.5897	41.4103	68.50771	15.45692	36.7537	50.5293	30.4913	43.56137	41.20904	24.12871	39.20672	51.53372	33.79707	28.69775	27.63997	51.3991	25.16587	35.76661	34.7637	47.6248	42.64498	40.15432	32.9607	35.82311	35.48333	44.49163
36.6667	45.641	74.88502	25.41661	37.01935	50.49267	30.49965	44.99512	50.64922	37.48379	28.70445	61.77753	43.91939	27.79784	27.89266	50.79528	22.95948	51.28771	37.02663	44.46229	47.22503	44.57565	35.40474	31.93916	33.23491	38.50248
38.2051	49.1026	66.4549	19.83663	41.41712	50.46977	30.48599	47.34795	52.1319	39.3646	31.70866	58.83775	44.4164	28.94769	52.79773	50.68604	25.58606	60.84658	36.45557	40.79184	48.39236	43.57313	51.64223	50.83834	35.43994	47.79752
29.7436	36.0256	77.3178	48.39835	22.81711	30.43604	30.50208	36.44121	39.79727	16.08355	42.81171	30.02045	28.0376	26.80382	27.58064	51.52476	21.31431	33.29717	35.53766	48.72939	27.24519	29.81951	34.20762	33.59381	33.87351	37.05603
29.7436	32.1795	58.91246	75.66776	34.8007	30.49793	30.47806	28.02058	43.54988	24.30721	43.30061	31.52643	31.68958	26.6664	26.4914	17.40539	33.0373	24.56446	20.40895	32.20303	30.40431	29.07954	36.52349	35.67105	28.56414	34.62644
30	29.1026	57.61745	8.194801	27.89189	30.48497	30.50116	38.31414	36.71466	17.38181	40.39863	16.34701	27.27961	19.25728	25.98532	17.20372	36.46889	21.87611	23.49571	25.32247	30.0808	27.58803	28.4192	21.31571	31.31955	21.66393
32.0513	26.7949	76.08823	59.67993	32.87064	30.50234	30.50477	26.41663	37.22461	17.98006	40.43569	20.23267	31.22188	26.44086	26.20142	17.09382	26.51355	32.91875	29.55754	21.36478	33.98063	35.03306	26.1605	23.04223	30.79617	23.51903
35.8974	25.2564	57.46605	1.504418	37.79405	30.49264	30.50912	22.82555	40.51245	22.96221	40.93655	16.913	33.54581	20.08071	25.85757	17.11384	36.02709	22.19544	33.00823	15.98507	34.25061	34.44159	40.43533	27.64007	34.23336	22.49591
41.0256	25.2564	79.42838	45.21079	35.66732	30.52204	30.51163	16.69401	40.22206	20.61462	39.66157	15.60936	39.73633	23.49481	50.70468	17.02374	42.68104	18.54254	53.9804	29.35099	51.14118	20.61832	34.54496	27.36224	36.63079	31.47683
44.1026	25.641	76.35652	10.41824	48.4423	30.50172	50.4483	21.64026	44.12496	23.36625	40.89926	20.79853	42.84184	20.3339	50.81198	17.11493	45.04108	17.73452	52.23431	29.71167	49.89959	19.56408	40.12373	30.07316	34.54862	26.11091
47.1795	28.718	64.40508	78.58418	58.66937	30.50684	50.17318	29.17786	42.34626	24.55833	41.96862	26.49707	47.09373	26.99347	50.56485	17.07778	47.35242	17.13475	59.50308	30.66967	55.58988	26.29925	42.18481	28.57367	37.51964	29.26763
49.4872	31.4103	40.63504	73.39475	47.91696	30.48673	49.92274	38.44047	44.18908	28.77263	40.38341	21.39123	51.45366	26.02294	50.9393	16.98102	53.84619	16.38732	41.16378	34.31576	55.48206	22.82747	40.00266	27.48751	62.4257	34.04423
51.5385	34.8718	43.94986	75.95872	51.89417	30.49476	50.3113	26.86049	45.81425	31.05698	56.53813	32.44425	49.36771	22.00558	50.45885	17.03858	55.13964	16.35325	48.99304	32.03036	63.57294	32.93862	57.12523	42.73902	64.32038	31.78563
53.5897	37.5641	30.99622	71.6944	51.74909	30.47299	50.39941	38.96384	46.64312	31.40722	52.97069	29.0402	48.25348	29.27086	52.90136	50.69056	52.18812	16.41822	59.26928	29.6407	59.02508	28.32209	60.37574	41.35646	65.45553	42.27357
55.1282	40.641	68.23077	80.8725	58.24655	50.44287	50.33736	40.91339	52.18989	39.3678	54.62095	30.34452	61.00296	30.01531	52.68495	51.29447	58.67266	16.66337	45.46918	33.83119	62.42378	33.4866	55.1059	46.14905	64.32563	41.61598
56.6667	42.1795	72.04639	12.91801	57.3107	50.48251	50.20232	47.21295	51.10032	38.0957	65.09904	27.24156	60.80656	27.86831	52.50009	51.59436	82.14431	33.19522	62.69127	30.29037	47.84991	48.55662	61.73727	38.58484	65.26434	35.71754
59.2308	44.4872	46.59277	84.97238	67.96805	50.45216	50.27493	49.6235	65.40083	42.61403	63.05599	29.7091	59.06743	26.78828	51.83564	52.33577	75.68657	24.74151	73.42867	48.57786	52.26856	52.80524	55.74124	46.65358	66.3887	38.84474
62.3077	46.0256	49.25722	81.8814	65.94932	50.49424	69.49438	52.61241	64.85204	43.13373	70.96014	41.2595	61.41463	31.03938	76.99541	52.04553	85.34496	52.88642	70.84643	52.28225	68.85102	59.57343	58.29641	41.5832	68.19142	45.16949
64.8718	46.7949	42.78176	12.99119	55.74709	50.47897	69.51934	47.38453	63.40381	44.84749	69.89582	43.45376	68.93437	31.61989	77.3106	51.74673	83.84465	58.61606	71.53902	45.52181	65.6764	35.10932	57.04255	42.41876	68.54066	48.43853
67.9487	47.9487	65.47595	14.27459	68.40301	50.49863	69.52918	57.84127	65.37155	42.69597	70.59589	41.96474	63.06524	34.23575	77.92604	50.31866	85.66476	45.54275	67.62087	38.00656	77.74689	47.37182	76.77561	55.2546	67.66143	51.06858
70.5128	53.718	71.96508	17.71024	76.69647	50.49383	69.49576	58.78507	61.90804	53.19035	69.64702	44.04445	72.27252	67.64444	77.25439	51.46182	77.931	68.69006	72.47095	51.12213	80.35297	47.8003	71.88708	59.26699	68.67552	50.46523
71.5385	60.641	32.14646	43.48171	67.21484	70.26003	69.19906	60.41896	71.82665	65.68623	77.39298	63.37146	75.35988	75.42161	76.24316	52.12369	73.03183	73.12057	64.81224	62.81092	60.33759	63.03496	73.78837	58.62081	69.08171	55.47409
71.5385	64.4872	31.8385	71.81217	72.30201	70.22044	69.55109	58.4011	72.65462	66.87828	64.40079	67.44872	72.83363	72.11469	77.08448	51.96714	70.1028	75.68321	60.85368	65.49915	61.65872	64.10002	75.13968	54.24455	69.04872	66.36603
69.4872	69.4872	31.00526	40.6825	66.76326	70.47386	69.50668	68.08447	76.88858	72.16547	63.86896	70.21374	70.81041	79.35046	75.22805	82.98566	57.14469	79.1941	67.7895	61.3637	71.54504	72.65772	64.59632	65.95251	67.30822	71.59034
46.9231	79.8718	80.47089	49.50215	51.6803	89.99703	50.44412	75.0133	43.74462	71.26796	56.59442	86.92701	45.80193	81.65827	50.65835	83.11448	46.96246	78.66656	41.60956	83.84869	49.04096	77.36258	46.90954	85.60226	38.7149	76.39698
48.2051	84.1026	71.96417	41.87428	53.09656	89.54024	50.46723	90.28146	39.05425	83.11902	56.53134	87.49981	45.83502	82.74259	50.20337	82.98266	50.56493	79.0773	53.00303	84.6748	51.88689	80.11361	46.9051	81.56564	35.54809	81.30498
50	85.2564	78.07942	93.13332	42.24229	89.70858	50.21088	80.94177	39.93167	79.93919	59.65216	87.80946	40.07881	80.24719	50.92955	82.84604	53.71968	79.30694	54.71417	84.04313	57.96898	85.35605	49.80699	82.08503	36.71067	86.07497
53.0769	85.2564	41.6776	30.20126	56.23388	89.63531	50.20541	90.64054	39.76283	76.40597	56.63651	85.6375	51.98292	82.20438	50.17867	83.18462	54.20023	79.325	44.29167	82.90944	54.85566	83.7834	48.86324	83.15836	64.91266	87.64588
55.3846	86.0256	65.9536	31.14741	54.68099	89.9646	50.43829	92.35476	41.32338	80.21187	58.67229	90.07716	57.33791	84.99952	50.4227	82.9099	47.8752	78.67585	49.19172	85.87623	54.1332	83.794	66.74819	84.00483	66.06414	89.20877
56.6667	86.0256	62.93446	31.91639	57.90269	89.6184	50.37352	82.21921	41.78732	83.98808	58.22161	90.41102	62.33623	83.24657	50.46422	82.93533	51.47346	79.15987	53.10138	84.54766	51.30883	80.13639	62.35977	90.24366	64.93593	87.07214
56.1538	82.9487	64.3738	28.86258	51.82026	89.83413	50.25368	81.74383	39.75565	79.6571	57.91466	89.9538	58.50591	81.74559	50.44927	83.69992	52.75122	79.29366	51.59985	84.81982	61.12039	88.70405	47.3749	83.47156	67.02655	76.22252
53.8462	80.641	72.50933	39.54013	56.8284	89.58366	50.25047	71.95139	40.74988	77.20466	55.31551	80.25186	56.24794	83.51395	49.92838	82.99367	48.56222	78.78238	54.37972	84.24036	55.56381	85.81038	47.44647	82.62212	65.31403	77.93874
51.2821	78.718	30.05229	96.61754	58.27228	69.90027	50.50706	73.95187	43.8569	77.60984	54.57573	77.53629	55.40652	80.26667	50.48801	83.09952	53.08124	79.29725	46.48077	83.51821	55.04412	83.54141	53.38528	78.87114	64.29522	84.12584
50	78.718	28.00332	46.67219	51.82813	70.35265	50.23107	80.64739	42.36782	73.63827	54.41305	78.22909	53.28397	84.2636	49.96491	83.70832	56.47538	79.16784	53.17466	84.26057	50.44415	78.19589	49.45271	79.9306	38.79801	74.61061
47.9487	77.5641	75.40123	88.63908	44.84998	70.3952	50.20774	73.1145	39.46296	78.45997	55.07451	79.81755	44.67676	80.23829	50.75211	82.90195	57.05091	79.03812	45.272	85.23708	43.82771	71.04705	54.80268	75.36398	34.67576	80.63761
29.7436	59.8718	38.98002	87.32216	24.27038	50.48557	30.49951	53.61253	44.51733	70.36935	29.43296	60.80178	30.01126	68.14676	27.42243	51.43888	37.6277	74.5575	36.0334	53.37169	37.79143	66.85464	37.08825	60.85723	30.71926	58.81320
29.7436	62.1795	65.21991	84.68295	34.74247	70.32993	30.52036	65.10041	41.63609	76.34922	29.42965	63.06946	29.74521	74.40453	27.67409	51.30411	32.2627	70.21664	28.27115	52.84023	35.29271	59.782	32.37596	68.51081	34.10515	62.1358

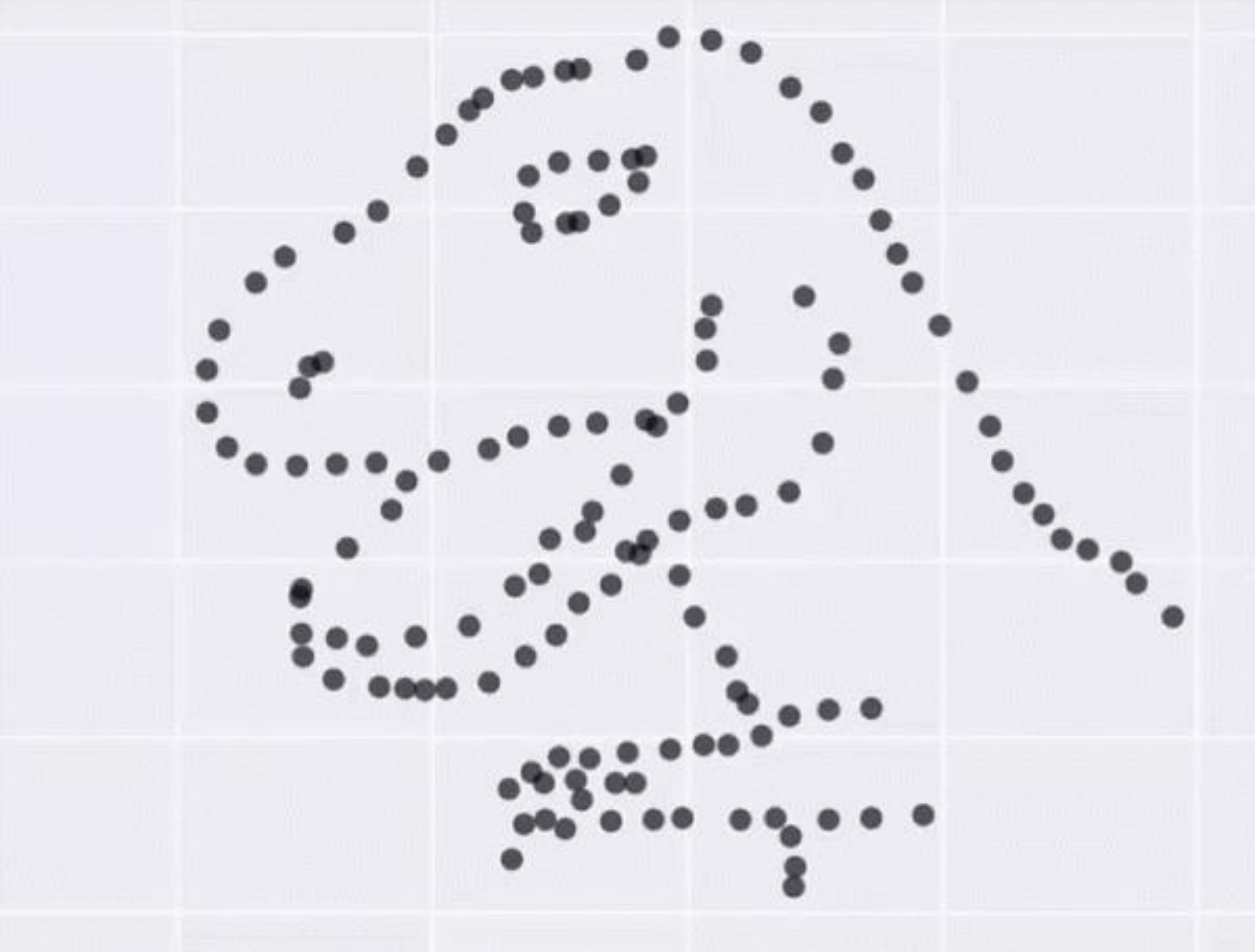
54.26
X Mean

47.83
Y Mean

16.76
X Standard Deviation

26.93
Y Standard Deviation

-0.06
Correlation Coefficient



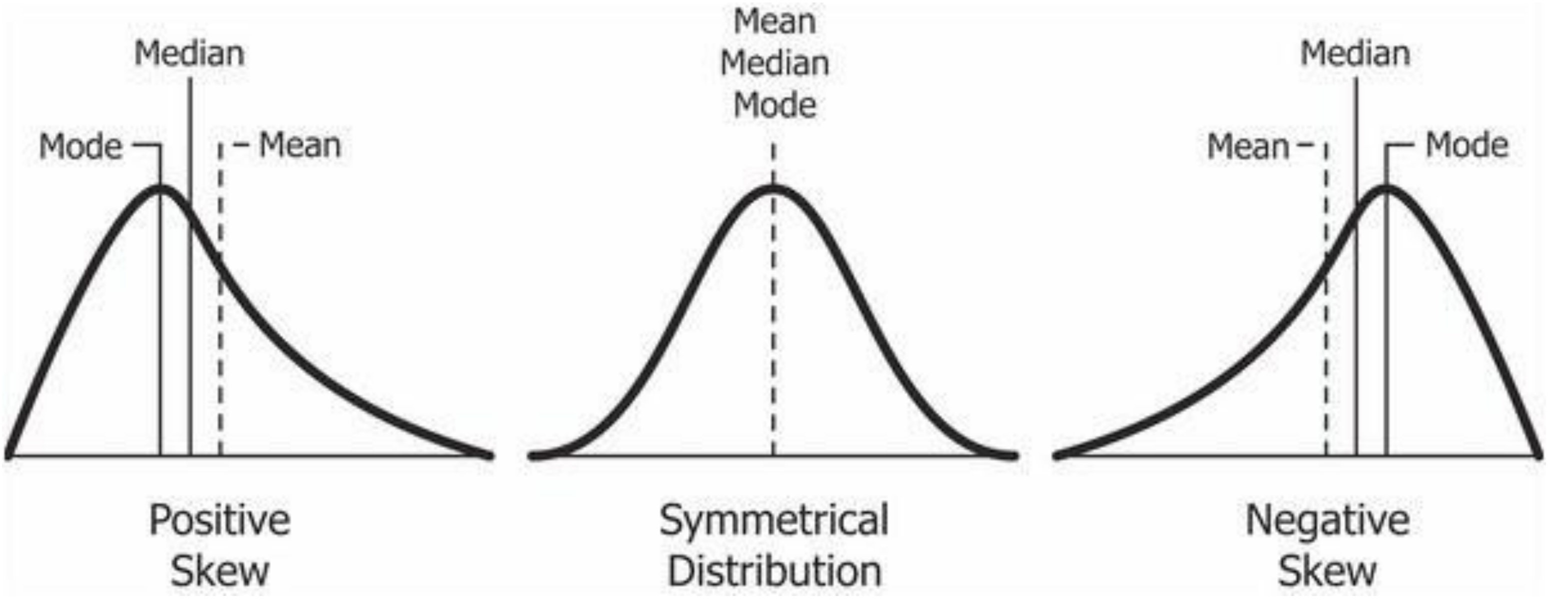
Scatter Plots

- Plot a scatter diagram and look for evidence of linear trend

Skew

- Skewness, in statistics, is the degree of distortion from the symmetrical bell curve in a probability distribution.
- Can be positive, negative or zero, to a varying degree.
- Helps to consider the extremes of data, not just the averages, by standardising data.

Skew


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

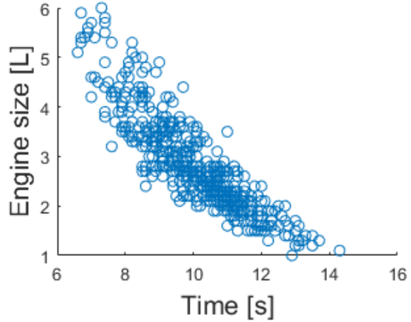
Car Emissions Data Task

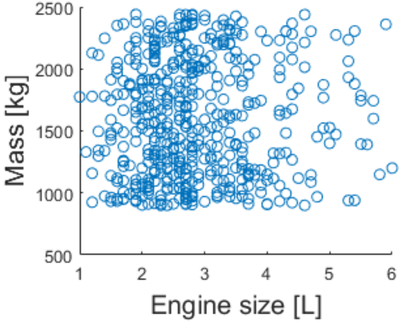
Regression Analysis

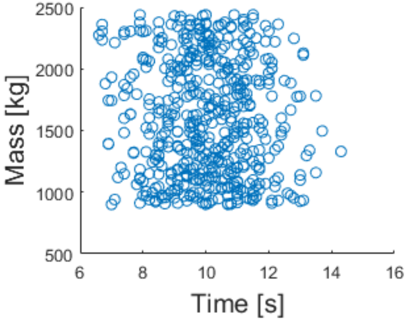
Generate and evaluate a model for fuel efficiency in terms of time, vehicle mass, engine size, fuel type and colour

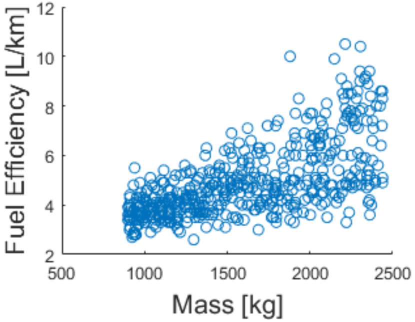
Scatter Plots

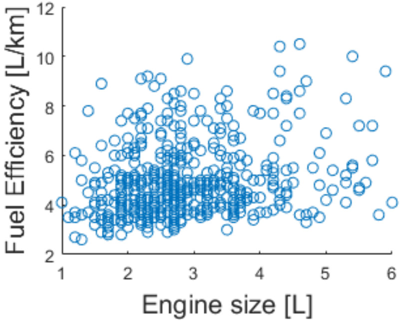
- Looking for possible correlation

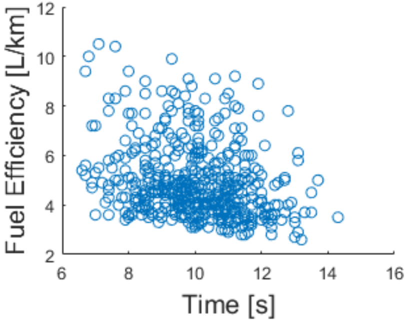






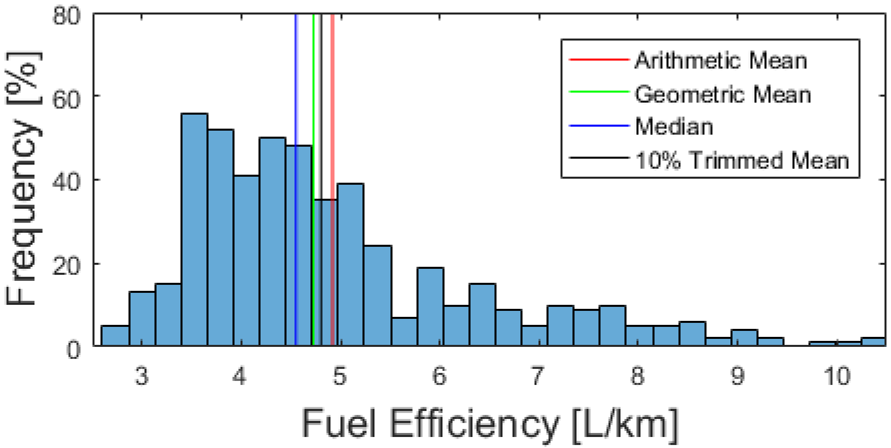






Histogram

- Looking for skew in the response variable
- Positive skew => mode < median < mean



Initial Thoughts...

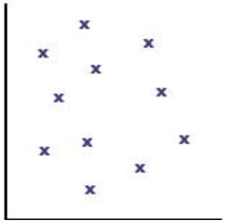
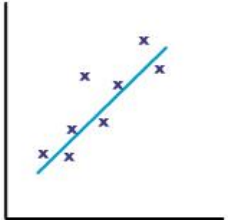
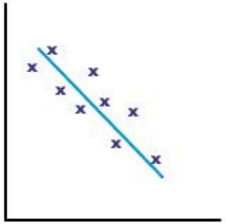
- Correlation exists between fuel efficiency, vehicle mass, acceleration time and engine size and engine size and acceleration time.
- Weaker correlation between fuel efficiency and engine size
- Positive skew, so we should standardise by the geometric mean, $x' = \frac{x - \bar{x}_{Geometric}}{\sigma_{Geometric}}$

Correlation

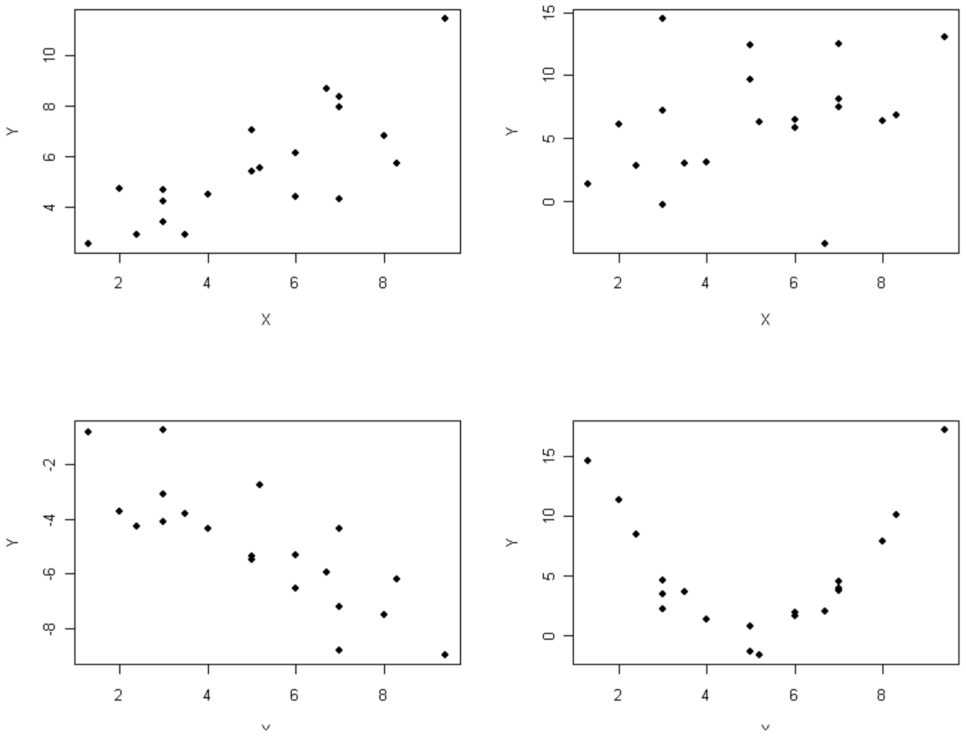
- Having decided that x_i and y_i are paired
- We want to study the relationship between them

- The Correlation coefficient measures the degree of linear relation between x and y

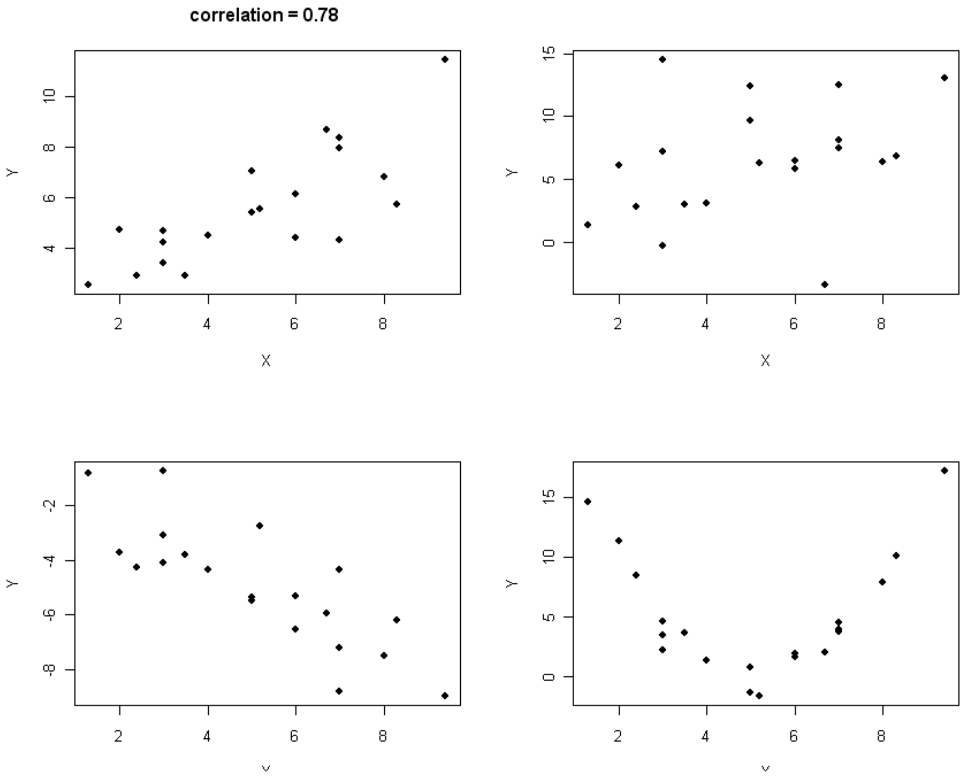
Correlation

R	Results	Correlation	Example
0	Uncorrelated	No correlation	
>0	Positively correlated	$y = \beta x + \alpha$	
<0	Negatively corerlated	$y = -\beta x + \alpha$	

Correlation

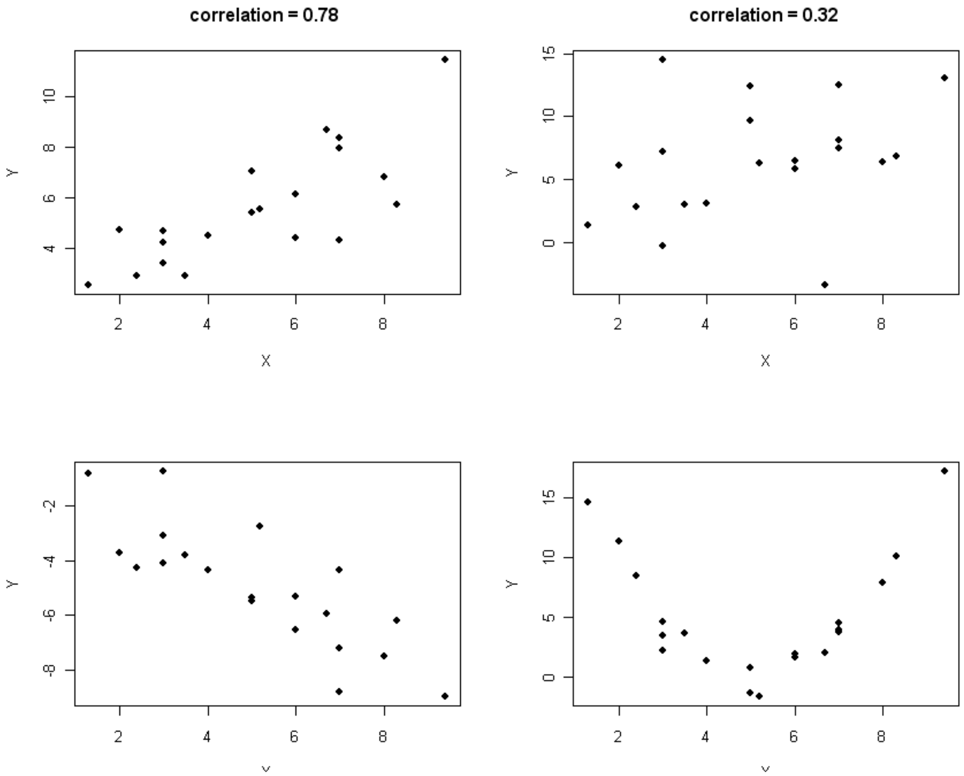


Correlation

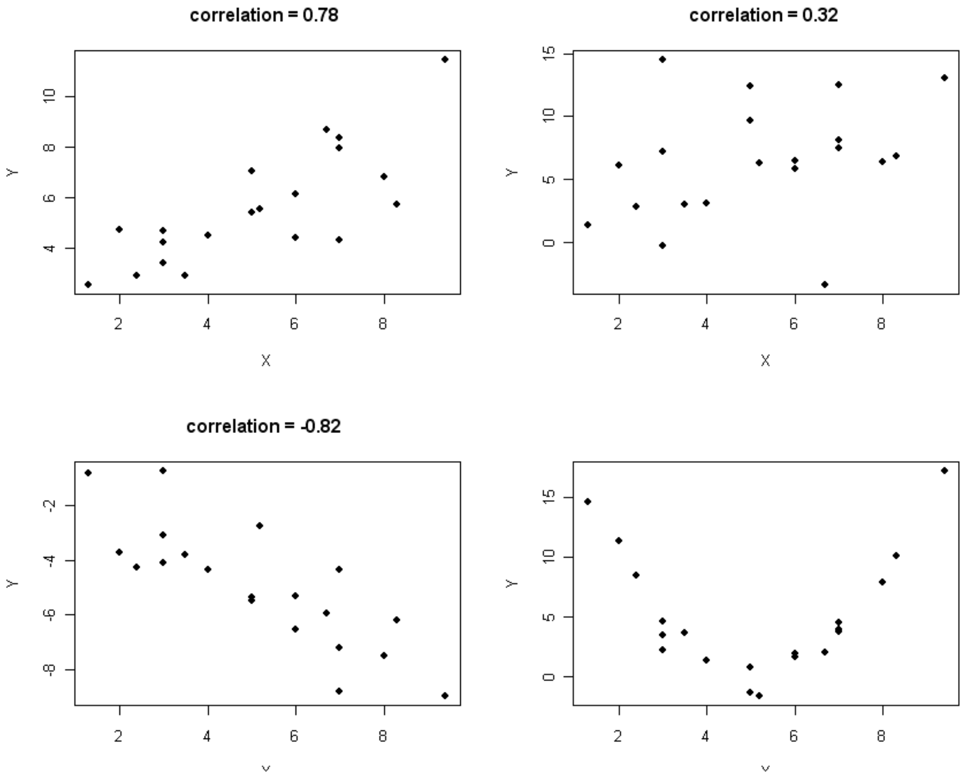


22

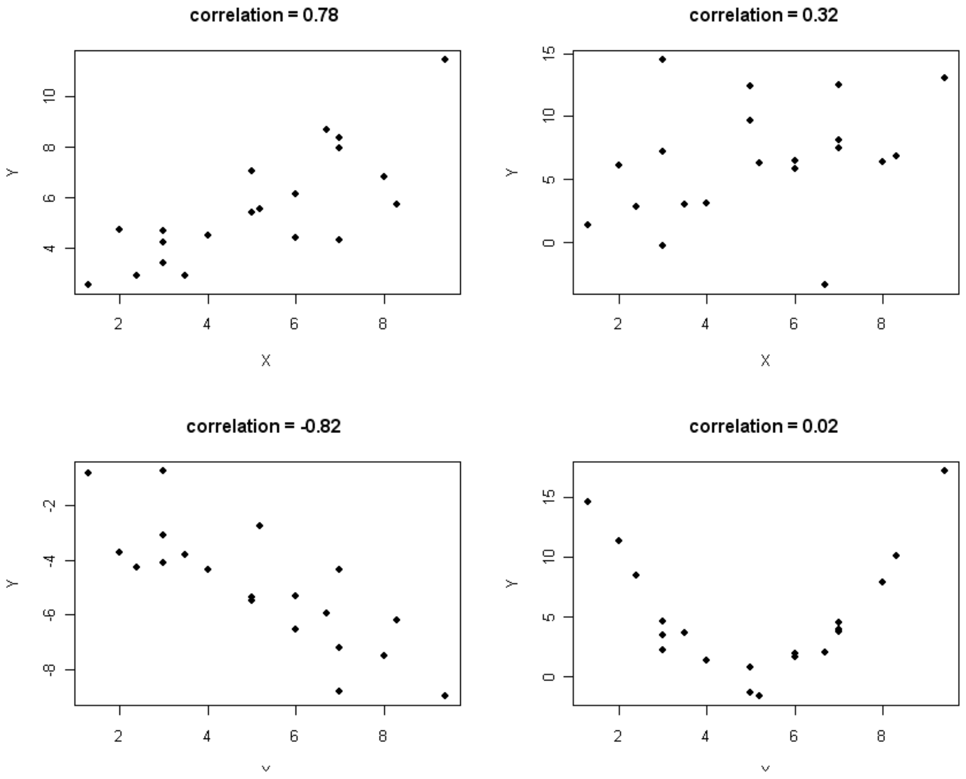
Correlation



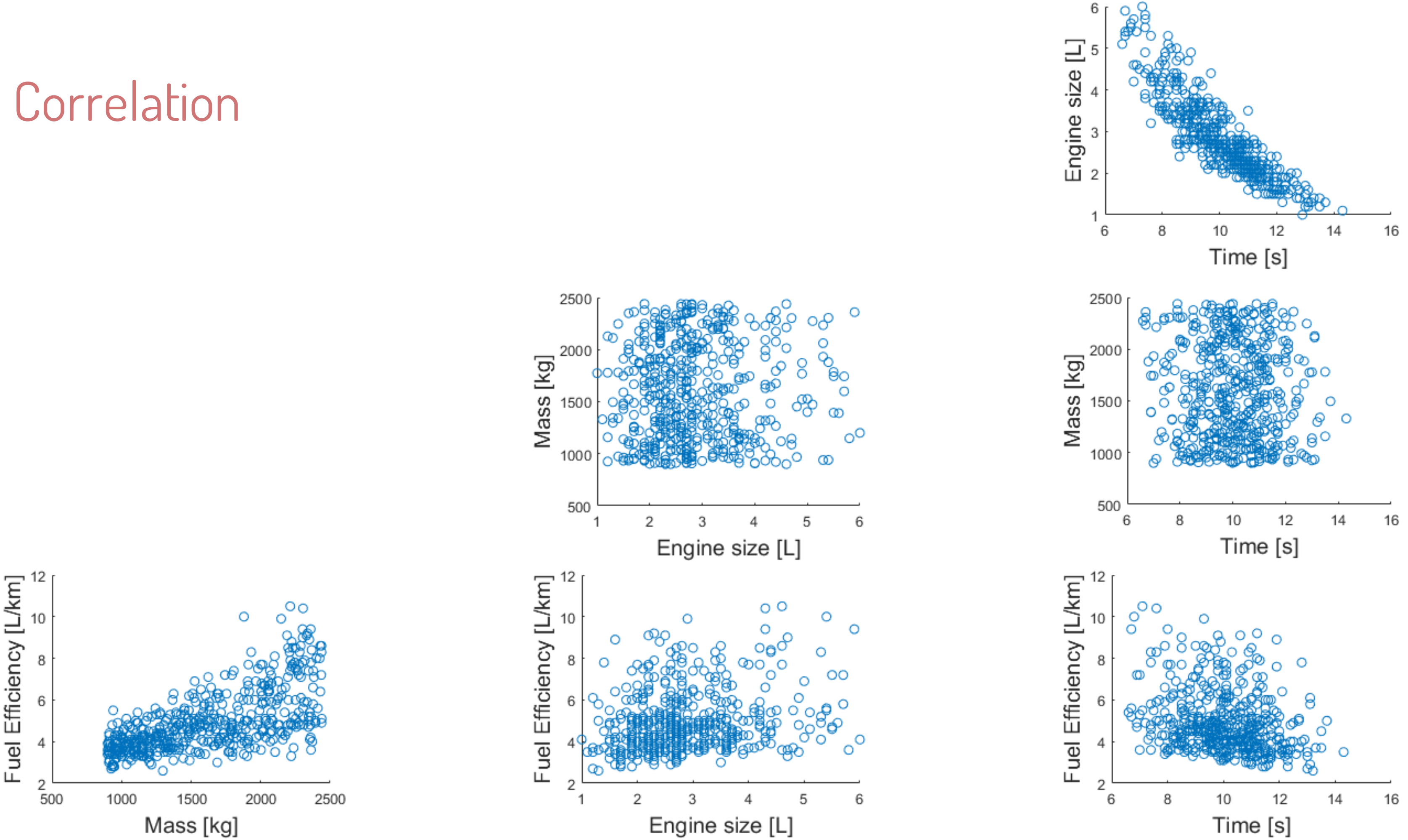
Correlation



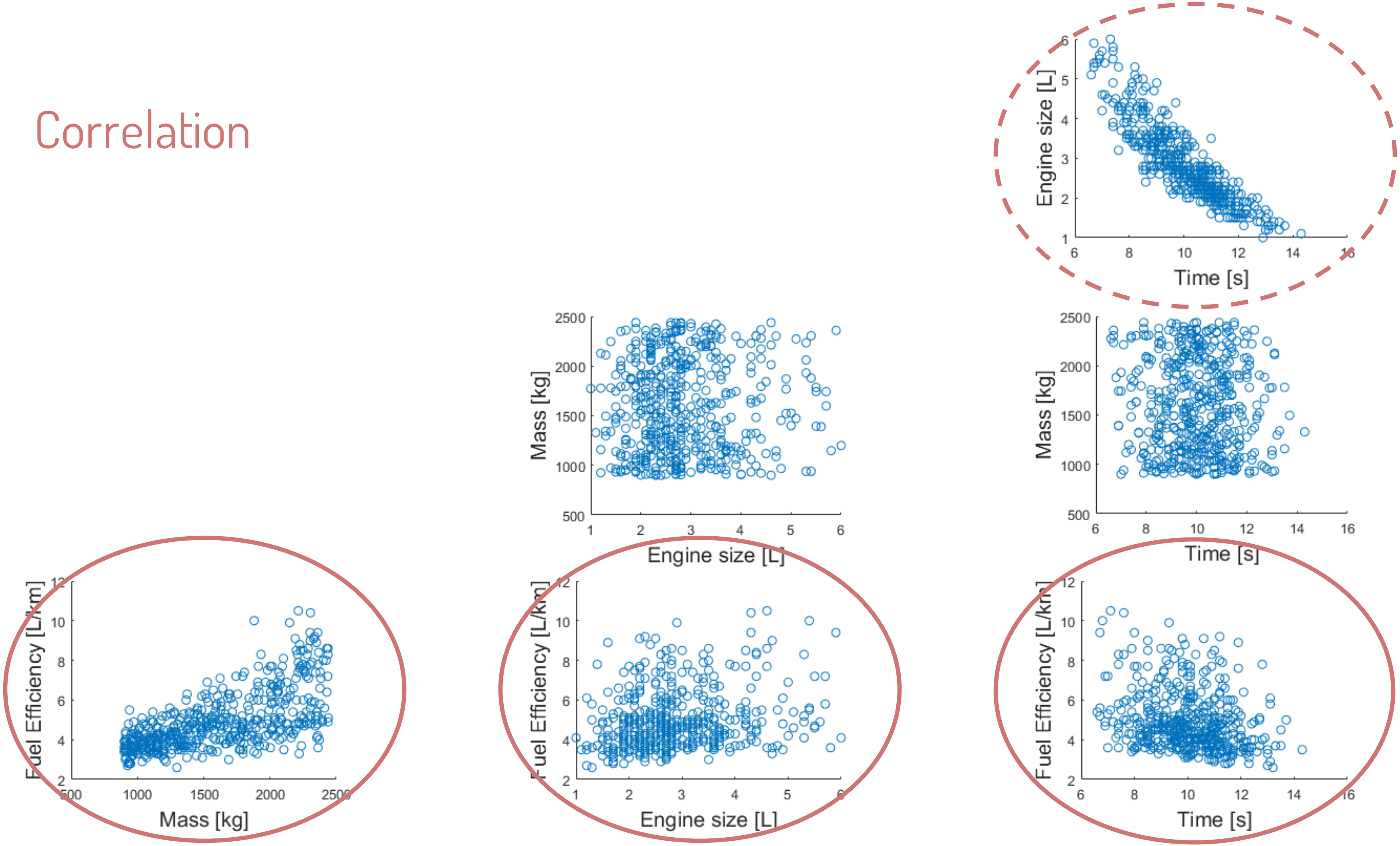
Correlation



Correlation



Correlation

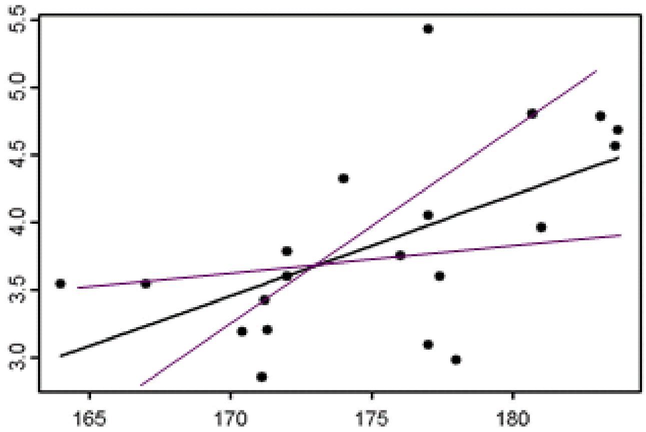


Correlation

- Not necessarily interested in the correlation coefficient
- Where does weaker correlation exist
- Interaction term may be needed between engine size and time

Regression

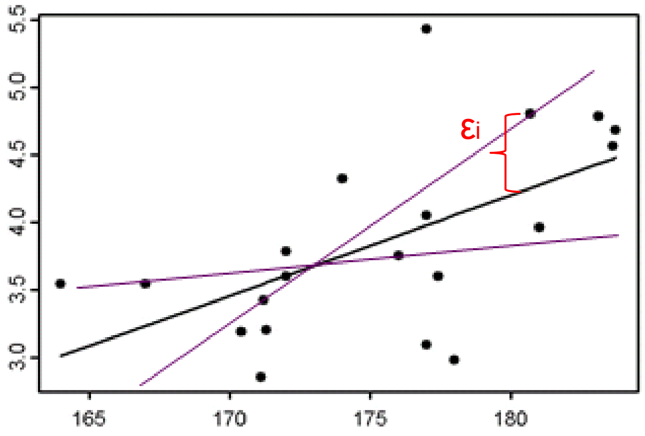
- $y_i = \alpha + \beta x_i$
- Let a = estimated α
  b = estimated β



- a = average fuel consumption at height = 0
- b = increase in fuel consumption for 1cm increase in height

Regression: Residuals

- Our model is not perfect
- $y_i = \alpha + \beta x_i + \epsilon_i$



- Residuals are the error between our model predictions and the actual data
- We assume these are normally distributed

Regression: Least Squares

- Find the line of best fit (estimate α and β) to minimise this error
- $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$
- MATLAB, Python and excel all have libraries to solve this

Regression: Least Squares

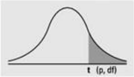
- $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

- This is an optimisation in terms of α and β
- The resulting model predicts y values given x

- $y = a + bx$

Regression: Confidence Interval

- How accurate are our estimated values of a and b?
- Confidence intervals based on the standard error
- 95% CI in b calculated from t scores
- $(b \; - \; t\,SE(b)\;,\; b + \; t\,SE(b) \;)$



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	—	—	80%	90%	95%	98%	99%	99.9%

Warning

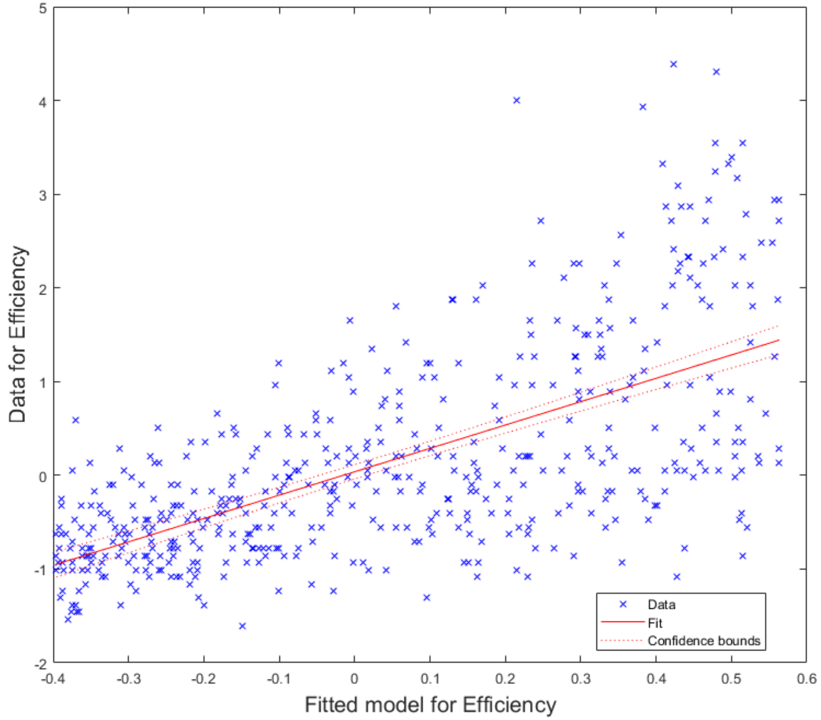
- Do not use your model to predict data outside the range of values in the domain of x
- Be cautious of overfitting

Regression with a Single Predictor Variable

- Model with one predictor:
- $Efficiency = \alpha + \beta Mass$

Regression with a Single Predictor Variable

- Model with one predictor:
- $Efficiency = \alpha + \beta Mass$
- Rsquared = 0.4157
- MSE = 0.7402
- AIC = 12705



What do we think of this model?

- Rsquared = 0.4157
- MSE = 0.7402
- AIC = 1271

Multiple Variable Linear Regression?

Multiple Explanatory Variables

- Multiple variables which simultaneously affect output variable
- Interpretation can become increasingly difficult with more variables
- Prevent cofounding and reduce residual variation

Multiple Explanatory Variables

- $y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_2 x_{2,i} + \cdots + \epsilon_i$

- i = number of observations
- $x_{1,i}$ = i^{th} observation of the 1st variable
- $x_{2,i}$ = i^{th} observation of the 2nd variable
- β_1 = the increase in y for a unit increase in x_1

Categorical Variables

- Takes one of a limited number of values
- Binary variables take values 0 or 1

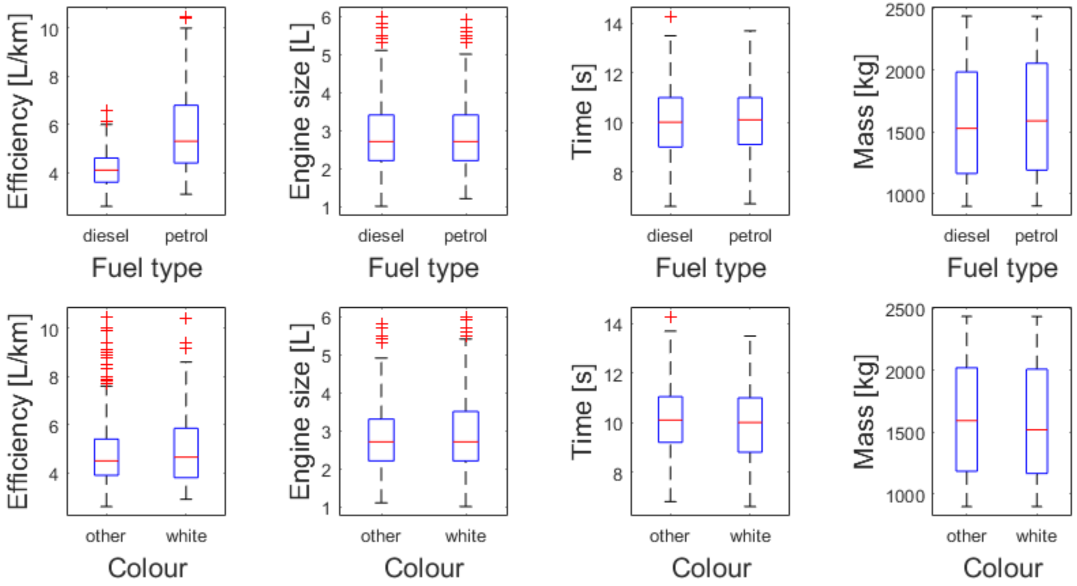
Binary Variables

- $y_i = \alpha + \beta_1 x_i + \epsilon_i$

- This model fits the mean of y for each category of x
- α = mean y_i in 1st group
- $\alpha + \beta$ = mean y_i in 2nd group
- β = difference in y_i between groups

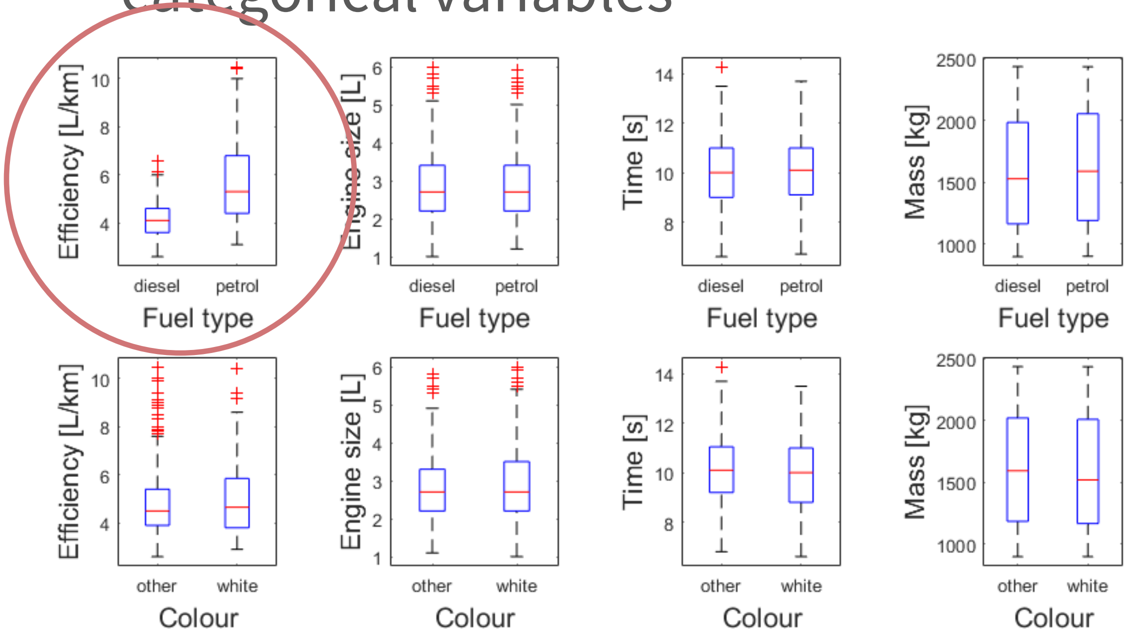
Categorical Variables

- Box plots of our data against known categorical variables



Categorical Variables

- Box plots of our data against known categorical variables



Categorical Variables

- Fuel efficiency looks likely to depend on Fuel type to some degree
- Fuel type could be a suitable predictor variable in the model

Regression: Least Squares with Multiple Variables

- $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1,i} - \beta_2 x_{2,i} - \beta_2 x_{2,i})^2$

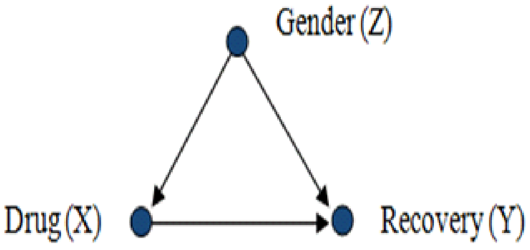
- This is an optimisation in terms of α and $\beta_{1,2,3}$ ...
- The resulting model predicts y values given x

Regression with Multiple Variables

- Regression coefficients (*b*) report the effect of each variable while holding all others at their average values

Confounding

- Consider a researcher attempting to assess the effectiveness of drug X, from population data in which drug usage was a patient's choice.
- Data show that gender differences influence a patient's choice of drug as well as their chances of recovery (Y).
- In this scenario, gender z confounds the relationship between X and Y since Z is a cause of both X and Y.



Model Interactions

- $y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 (x_{2,i}\, x_{3,i}) + \cdots + \epsilon_i$

- $x_{3,i}$ is a categorical variable
- β_2 = increase in y_i for unit increase in x_2
- $\beta_3 + \beta_4$ = increase in y_i for unit increase in x_2 for an observation in the categorical variable x_3

Model Interactions

- $$ArmLength_i = \alpha + \beta_1 Age_i + \beta_2 Height_i + \beta_3 Gender_i + \beta_4 (Height_i Gender_i) + \epsilon_i$$

- Gender is the categorical variable
- β_2 = increase in $ArmLength$ for unit increase in $Height$
- $\beta_2 + \beta_3$ = increase in $ArmLength$ for unit increase in $Height$ for a $Female$

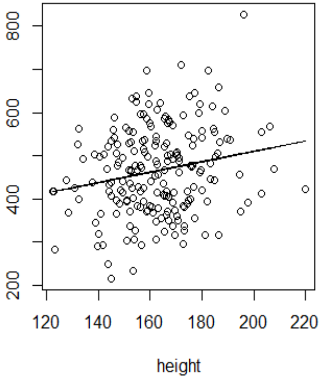
Model Interactions

- If the interaction term is significant then we should include both the interaction and individual terms

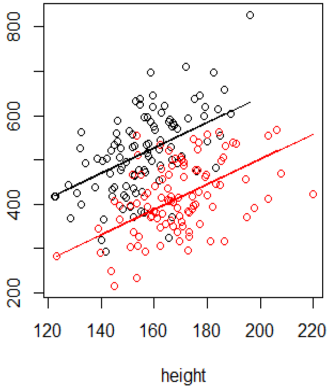
- The effect of gender is the different between males and females in this model on the value of height
- This must be found considering the interaction term

Model Interactions

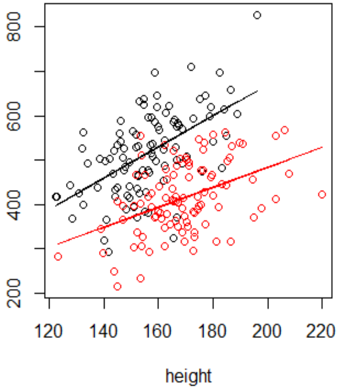
- Assuming the effect of height is significant we can have three models depending on whether gender and the interaction are significant.



$$ArmLength_i = \alpha + \beta_1 Age_i + \beta_2 Height_i$$



$$ArmLength_i = \alpha + \beta_1 Age_i + \beta_2 Height_i + \beta_3 Gender_i$$



$$ArmLength_i = \alpha + \beta_1 Age_i + \beta_2 Height_i + \beta_3 Gender_i + \beta_4 (Height_i Gender_i)$$

Improve on the Previous Model by Including more Terms

- Model with one predictor:
- $Efficiency = \alpha + \beta_1 Mass + \beta_2 FuelType$

- Rsquared = 0.6540
- MSE = 0.4383
- AIC = 1010

Model Checking

Model checking should be performed to avoid erroneous extrapolation of data trends

Residuals

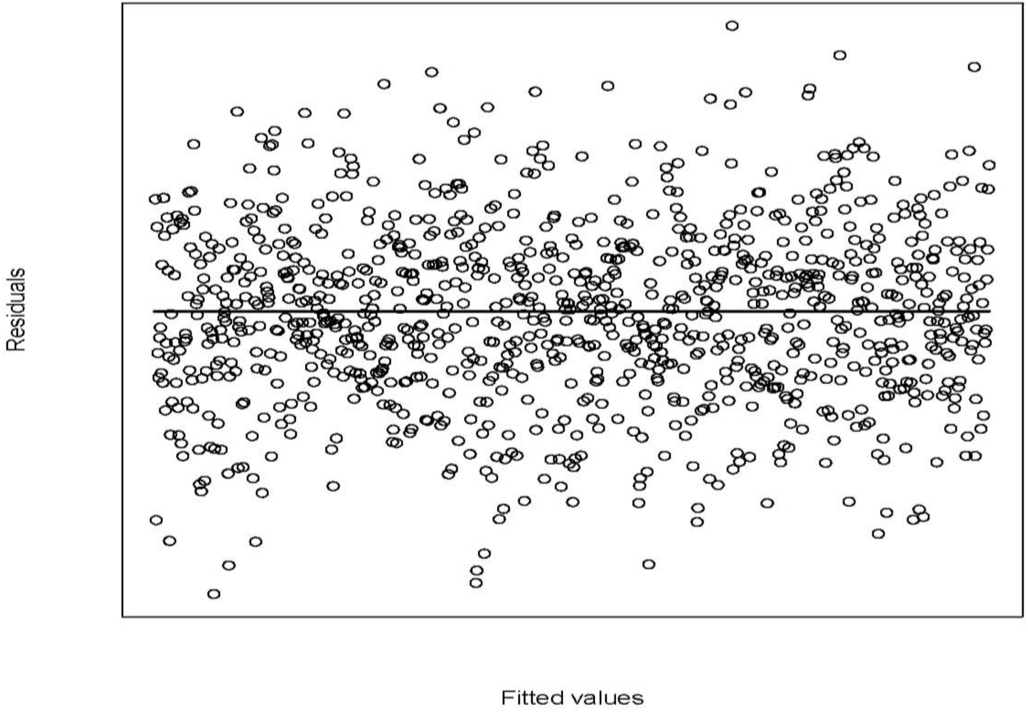
- Residuals are the error between the observed and predicted data
- Look for trends or patterns in the residuals which indicate an assumption is not valid

Residual Analysis

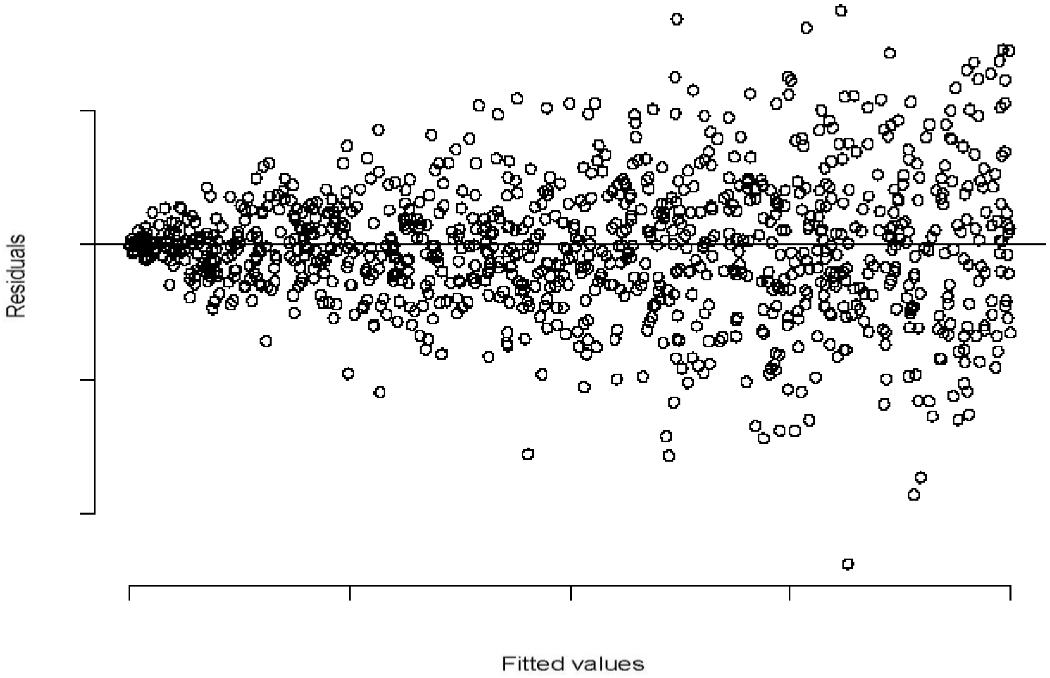
- Scatter plots of residuals against fitted values help identify:
  - Non-constant variance
  - Violation of the linearity assumption
  - Potential outliers

- If these assumptions are valid then you will see no trend

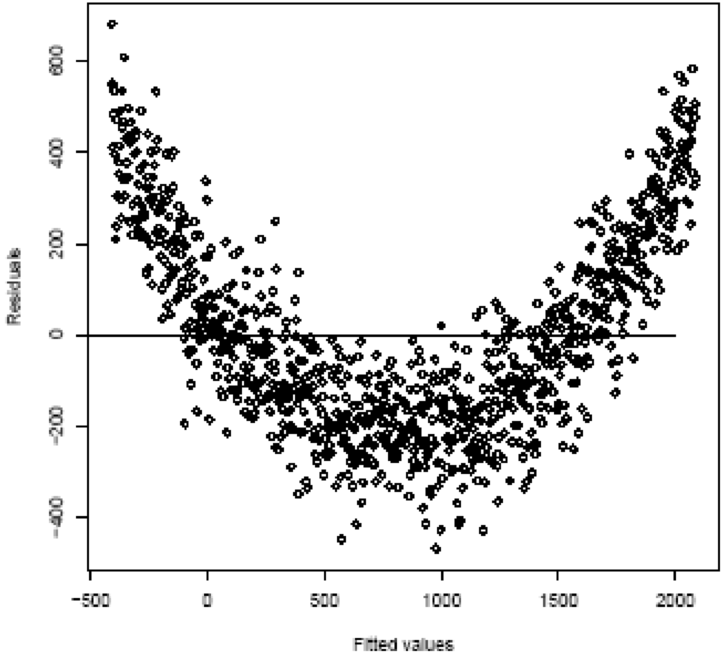
Residual Scatter Plot: Satisfactory



Residual Scatter Plot: Non-constant Variance



Residual Scatter Plot: Non-Linear Relationship



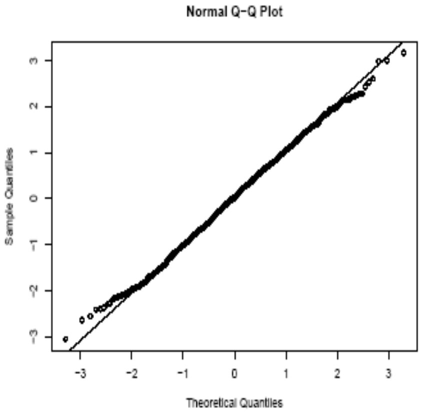
Residual Analysis

- Standardise residuals by dividing by their standard deviation
- They should now be Normal with mean = 0 and variance = 1

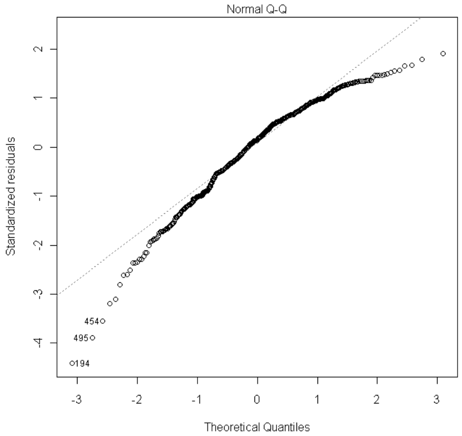
- Box plots – symmetric?
- Proportion of standardised residuals inside percentiles
- Q-Q plot – x = y graph should form

Q-Q Plots and Skew

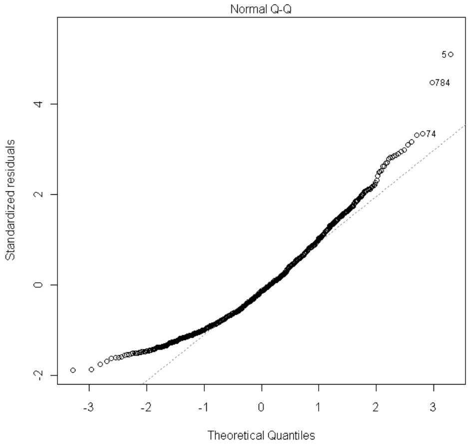
- Satisfactory



- Negative Skew



- Positive Skew



Skew

- Median < Geometric Mean < Arithmetic Mean
- Positive skew => Standardise by the geometric mean

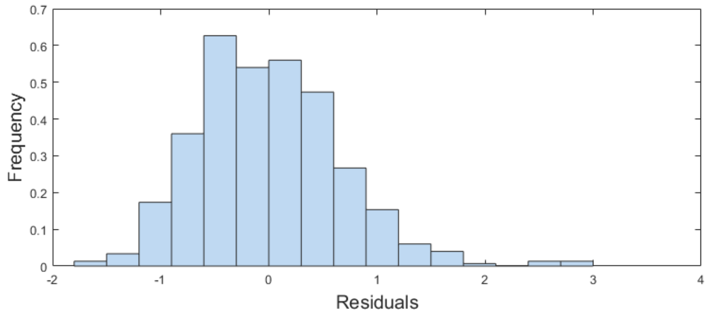
- Negative skew=> Standardise

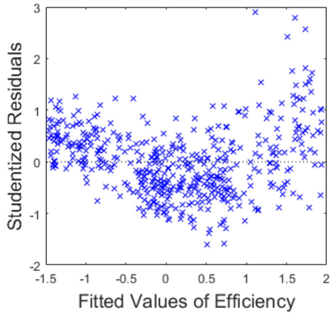
Identifying Outliers

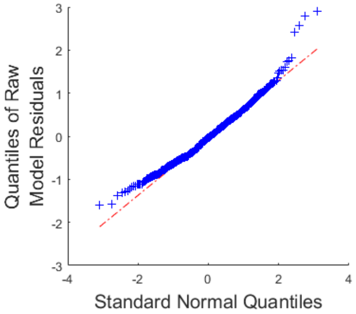
- Any points which stand out as having larger residuals than other values should be checked
- Cook's Distance is given by most software and measures the influence of each individual point on the model

Residual Analysis

- $Efficiency = \alpha + \beta_1 Mass + \beta_2 FuelType$

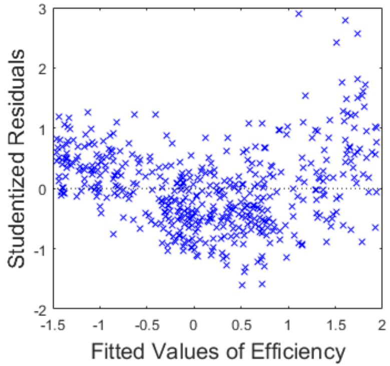




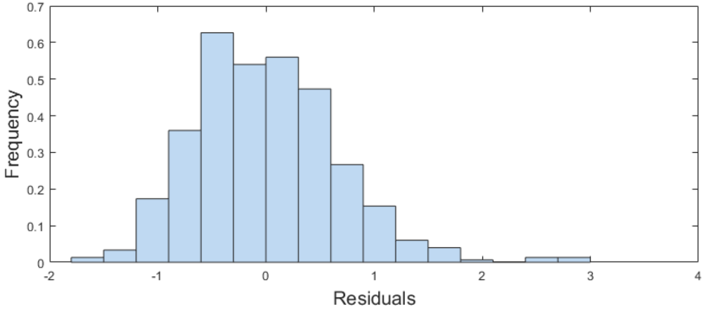


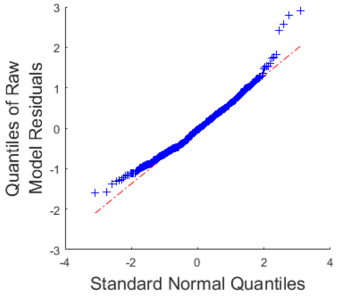
Residual Analysis

- Possibly a non-linear relationship?



- Maybe very slightly positively skewed?





Model Fitting

Model checking should be performed to avoid erroneous extrapolation of data trends

Method

Scatter Plots, Box Plots & Histograms
Look for linearly related variables, skewness

Fit Model
Based on which variables you suspect are linearly related from scatter diagrams.

Interaction Terms
These may improve the model of your data

Residual Analysis
Look for non-constant variance, non-linearity and potential outliers

Adjust Model
You may transform variables before fitting a model or include additional terms

MATLAB Script

Model Fitting

- Given data on car emissions in terms of vehicle mass, acceleration time, engine size, fuel type and colour.
- Fit models seeking to optimise for:
  - Error
  - AIC – relative loss of information

Search for Best Models

- Open file: Regression_Analysis_Car_Emissions
- Use the scatter and box plots to pick a starting model

How to Use the Script

- Standardise by the arithmetic or geometric mean

```
94
95 %%%%%%%%%%%%%%%%%%%%%
96 % 'G' for geometric standardisation or 'A' for arithmetic standardisation
97 standard = Standardise(cont,FieldNames,ArMean,ArStd,GeoMean,GeoStd,'A');
98 %%%%%%%%%%%%%%%%%%%%%
99
```

- Edit line 115 with your trial model

```
113 %%%%%%%%%%%%%%%%%%%%%%%%%
114 % Insert Model Here
115 model = ['Efficiency ~ Mass + FuelType'];
116
117 % Other Examples for models
118 % model = ['Efficiency ~ Mass + EngineSize^2'];
119 % AccelTime
120 % EngineSize
121 % Mass
122 % FuelType
123 % Mass:FuelType % Example of an interaction term
124 %%%%%%%%%%%%%%%%%%%%%%%%%
```

Search for Best Models

- Keep note of the MSE, Rsquared and AIC
- Seek to increase Rsquared
- Seek to reduce MSE and AIC

Things to Try

- Increasing the number of terms
- Including categorical variables
- Using interaction terms (FuelType:AccelerationTime)
- Standardising Method
- Raising terms to a power

How did you do?

Prizes for the ‘best’ model

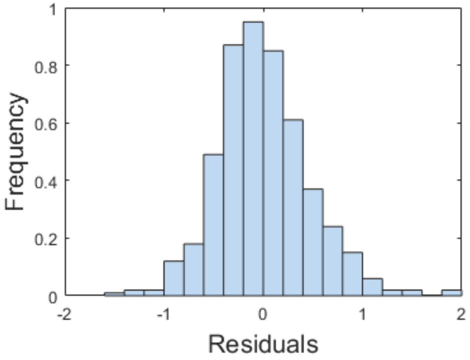
An Optimum Model?

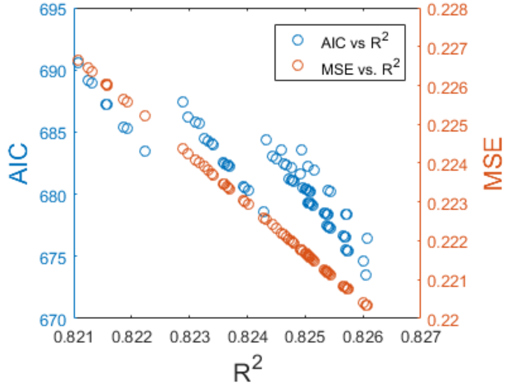
Optimum Model

- Running the script:

```
% Make Model:
% [Rmax,mdl,Criteria] = MakeModel(tbl,model);
[CountOpt,Rmax,mdl,Criteria] = MakeOptimumModel(tbl,3)
```

- Normally distributed residuals
- AIC and MSE clearly decrease with Rsquared





Model Simplicity

- Model 1
  - Rsquared = 0.8261
  - MSE = 0.220
  - AIC = 673

$$Efficiency = \alpha + \beta_1 Mass + \beta_2 EngineSize + \beta_3 AccelerationTime + \beta_4 MassFuelType^2$$

- Model 2
  - Rsquared = 0.8261
  - MSE = 0.220
  - AIC = 676

$$Efficiency = \alpha + \beta_1 Mass^3 + \beta_2 EngineSiz + \beta_3 AccelerationTime^2 + \beta_4 MassFuelType^2 + \beta_4 EngineSize\,FuelType$$

Further Reading

What are next steps

Next Steps

- K-means cluster classification
- Bootstrapping (Confidence Intervals)
- Introductory courses to Machine Learning (Stanford, coursera)

79

How did we do?

- Please let us know what you thought of this course and how we can improve it
- https://forms.gle/HAdcNGkFK5fkCimk6