




Inspire…Educate…Transform.

Supervised models

Model Building using Linear Regression

Dr. Anand Jayaraman

Feb 26, 2017

Thanks to Dr.Sridhar Pappu for the material

Output Analysis - Recap

What is the total variation and its explainable and unexplainable components?

$$SST = SSR + SSE$$

$$SST = \sum (y_i - \bar{y})^2 \qquad SSR = \sum (\hat{y}_i - \bar{y})^2 \qquad SSE = \sum (y_i - \hat{y}_i)^2$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

How much of total variation can be explained by variation in independent variables?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$1 - \frac{SSE}{SST} = 1 - \frac{93.08}{474.93}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

What is the correlation between actual and expected values?

$\sqrt{R^2}$: Correlation between y and $\hat{y}$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

How much of total variation can be explained by variation in independent variables (IVs) that *actually*

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$R^2 - (1 - R^2)\frac{k}{n-k-1}$$

$$1 - \frac{MSE}{MST}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333	33.923809521		

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

What is the “average” deviation of the actual values from the expected values?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$\sqrt{MSE}$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

What is the average of the squared errors?

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

$$MSE = \frac{SSE}{df_{error}}$$

Output Analysis - Recap

Is the model significant?

SUMMARY OUTPUT	

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$F = \frac{MSR}{MSE}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

F Table for α = 0.05

/	df_1=1	2	3	4	5	6	7	8	9	10	12
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876

Output Analysis – Recap

What do regression coefficients

A coefficient is the slope of the linear relationship between the dependent variable (DV) and the **independent contribution** of the independent variable (IV), i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

How much will the variation be between the estimated coefficient and the corresponding true population

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$SE_{b_1} = \frac{SE}{\sqrt{\sum\left(x_{1_i} - \bar{x}_1\right)^2}} \sqrt{1 - R^2{}_{(x_1, x_2 x_3)}}$$

R^2 with x_1 as dependent and other Xs as independent

ANOVA

	df	SS	MS	F	Significance F
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	b_0	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	b_1	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	b_2	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	b_3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

Are the coefficients significant?

SUMMARY OUTPUT	

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$t = \frac{b_i - \beta_{i_{null}}}{SE_{b_i}} \quad \beta_{i_{null}} = 0$$

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	381.8467141	127.282238	15.0408794
Residual	11	93.08661926	8.462419933	
Total	14	474.9333333		

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

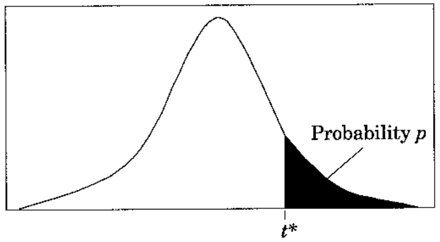


Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Output Analysis - Recap

What are the confidence intervals for the coefficients?

$$b_i - t_{\left(\frac{\alpha}{2},\nu\right)} * SE_{b_i} \le \beta_i \le b_i + t_{\left(\frac{\alpha}{2},\nu\right)} * SE_{b_i}$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	381.8467141	127.282238	15.0408794
Residual	11	93.08661926	8.462419933	
Total	14	474.9333333		

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

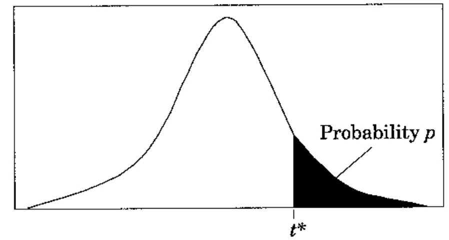


Table B t distribution critical values

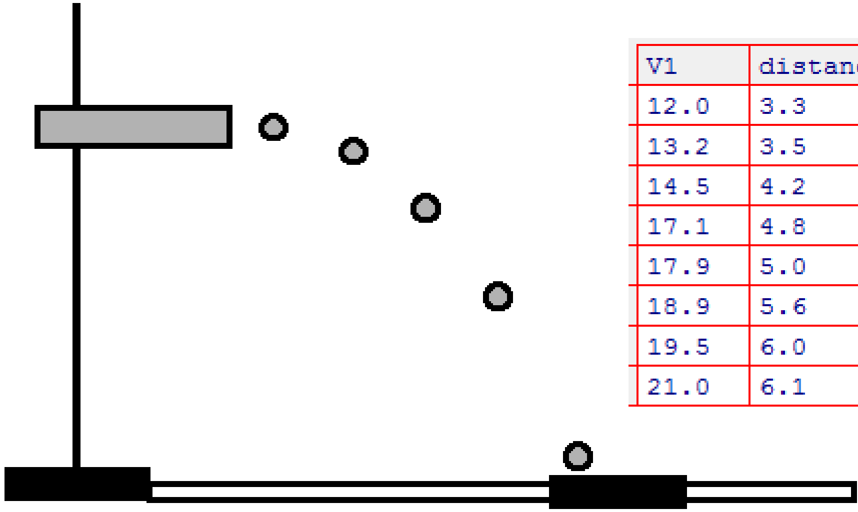
	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	[illegible]	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	[illegible]	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	[illegible]	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Linear Regression through Origin

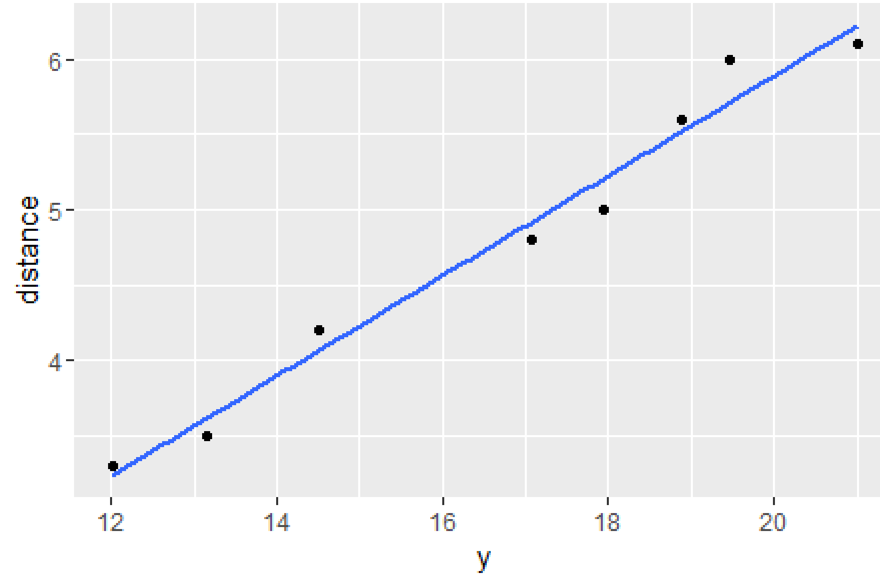
In some physical examples, we might know based on physical intuition that when x=0, y should also be 0.

In those instances it might make sense to try force the regression line to go through the origin.

Linear Regression through Origin



V1	distance
12.0	3.3
13.2	3.5
14.5	4.2
17.1	4.8
17.9	5.0
18.9	5.6
19.5	6.0
21.0	6.1



- Best fit line
- Distance = 0.33 V – 0.74
- However, we know when V=0, Distance =0
- We can force the intercept to be zero by

```
> lmout <-lm(distance~speed + 0)
```

Caution: Regression through Origin

With Intercept

```
> summary(lm(dist~speed,data=cars))

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Through origin

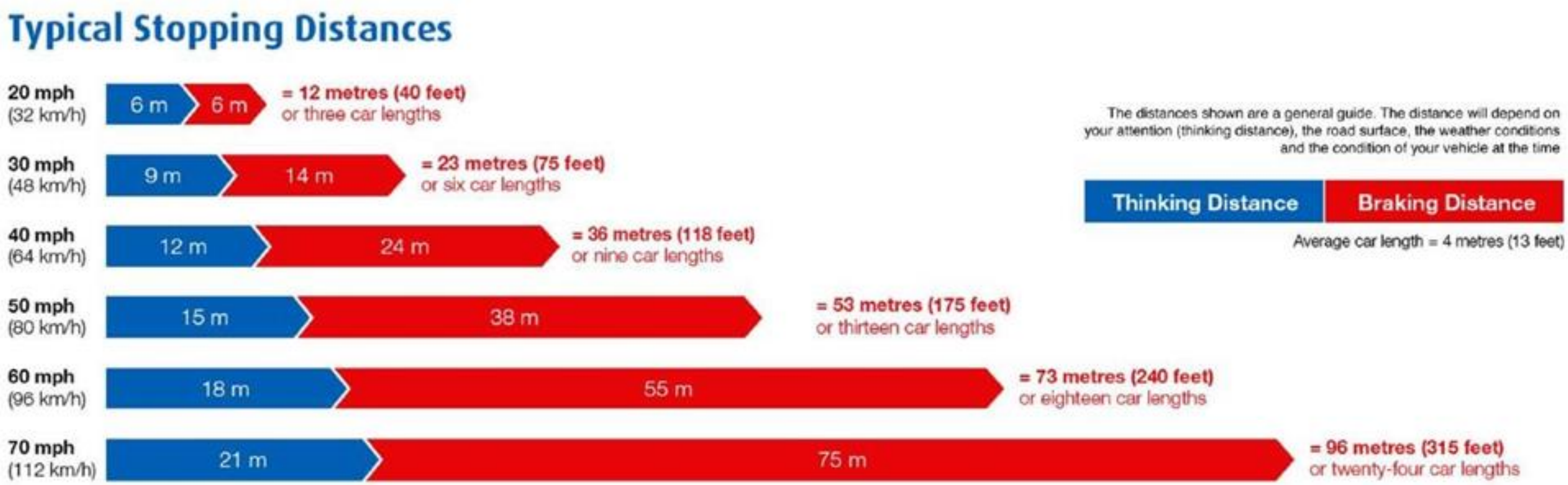
```
> summary(lm(dist~speed+0,data=cars))

Call:
lm(formula = dist ~ speed + 0, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-26.183 -12.637  -5.455   4.590  50.181

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
speed   2.9091     0.1414   20.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.26 on 49 degrees of freedom
Multiple R-squared:  0.8963,	Adjusted R-squared:  0.8942
F-statistic: 423.5 on 1 and 49 DF,  p-value: < 2.2e-16
```

Do not believe the R^2 when the fit is forced through the origin.
R^2 over-states the quality of fit when you force the intercept to be zero! Pay attention instead to the Residual Standard Error. Note that here, the Residual Standard Error is higher after forcing the fit to go through origin.

Handling Simple Non-linearity

Non-linear transformation of data - Example

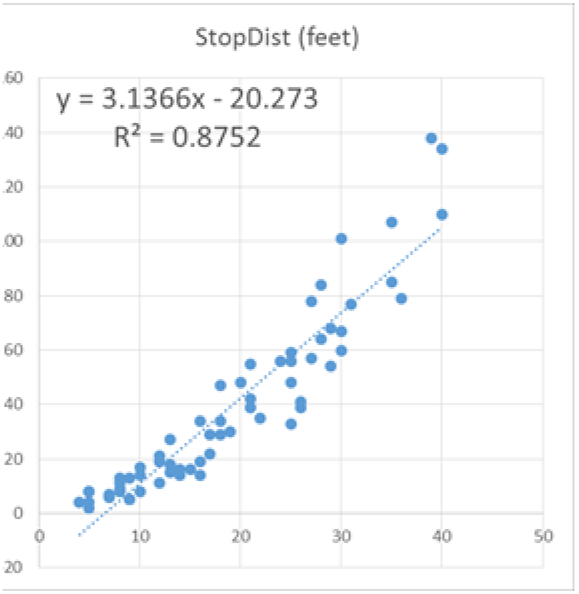
American Automobile Association (AAA) publishes data that looks at the relationship between average stopping distance and the speed of car.

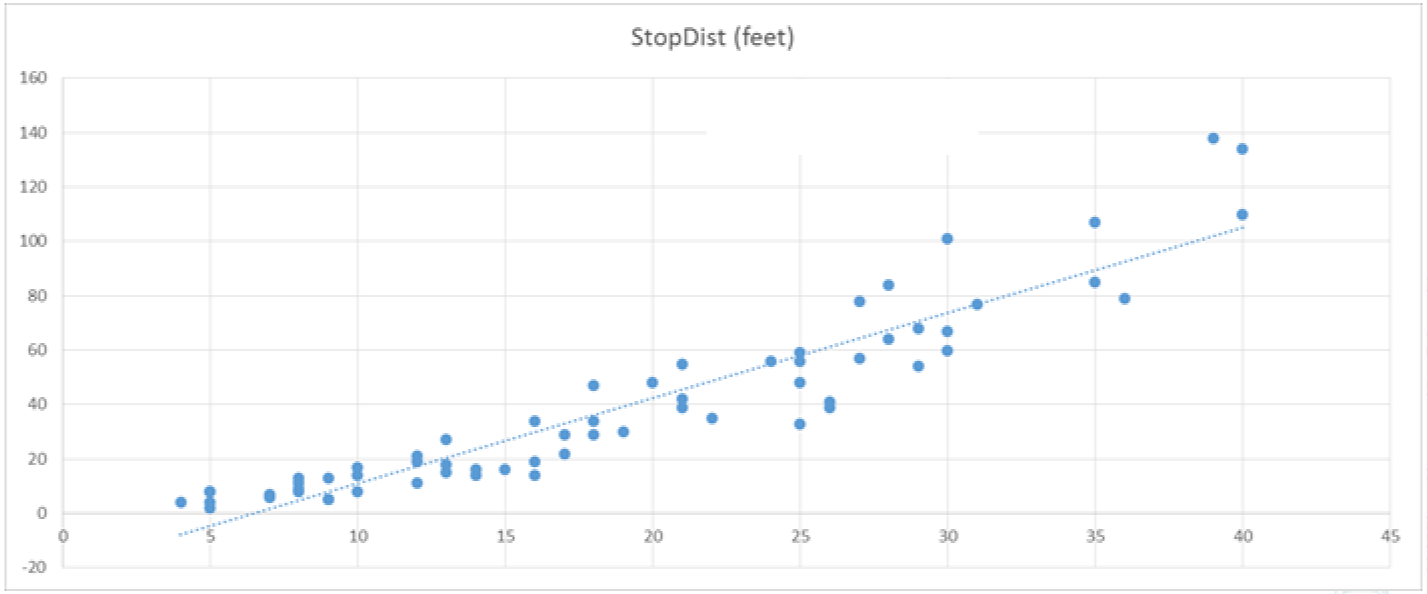


Non-linear transformation of data

American Automobile Association (AAA) publishes data that looks at the relationship between average stopping distance and the speed of car.

Does the estimated regression line fit the data well?





Using Domain knowledge

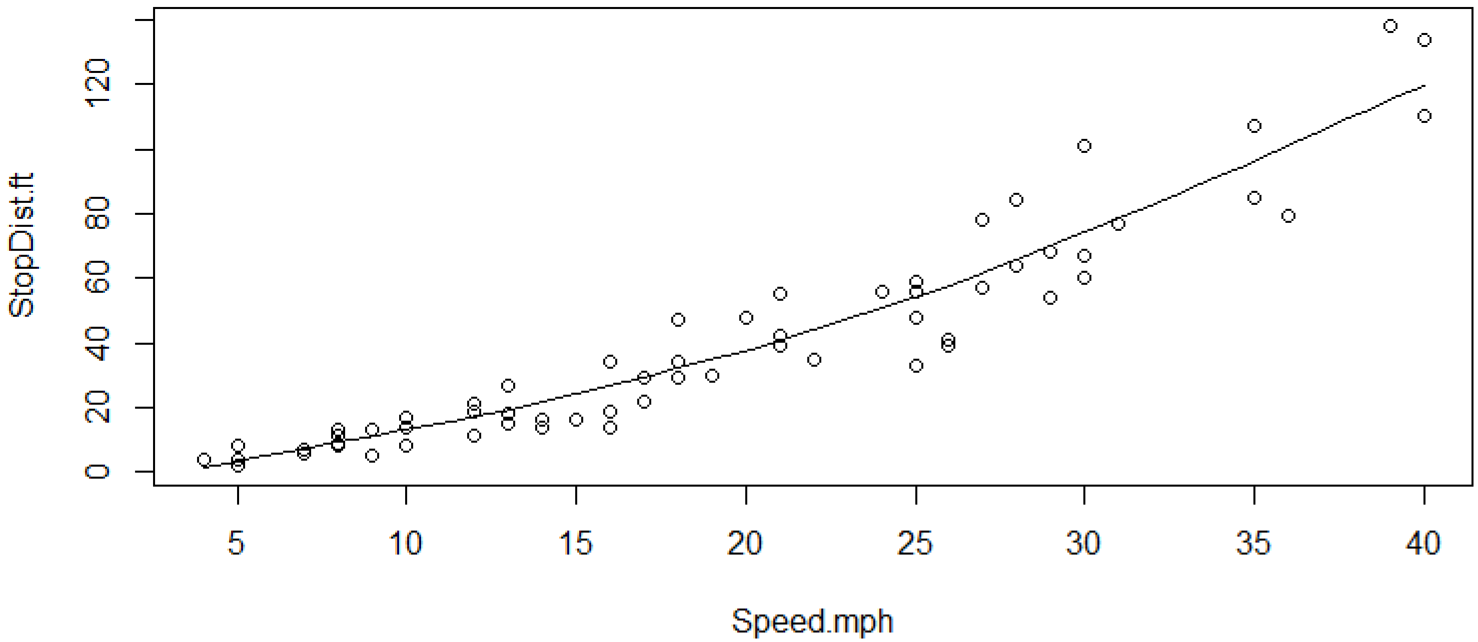
- Basic physics equations, show that stopping distance D and Speed V is related as

$$D \propto V^2$$

Or

$$\sqrt{D} \propto V$$

The plot of the data with a smoothing line shows the non-linear structu

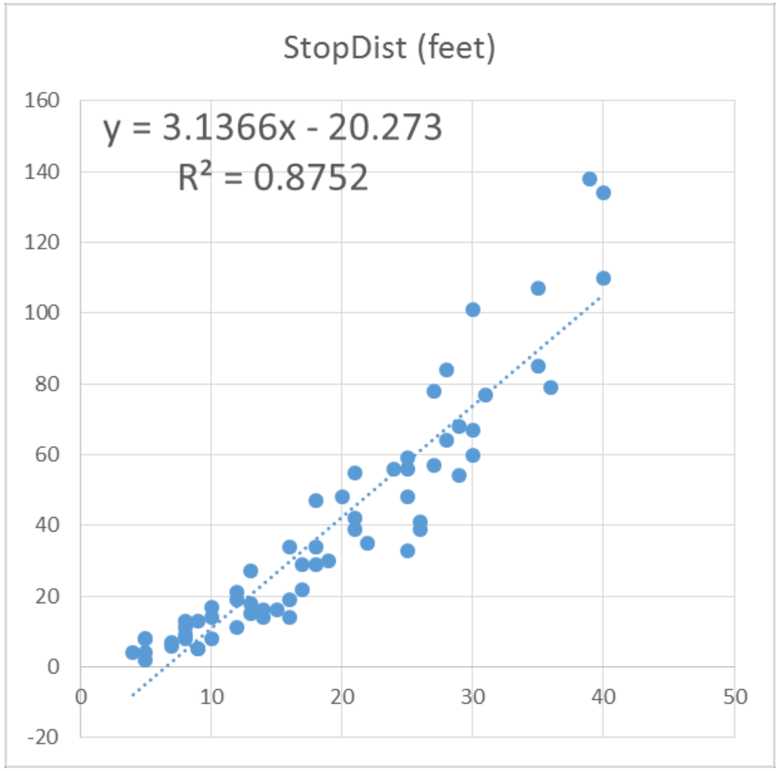


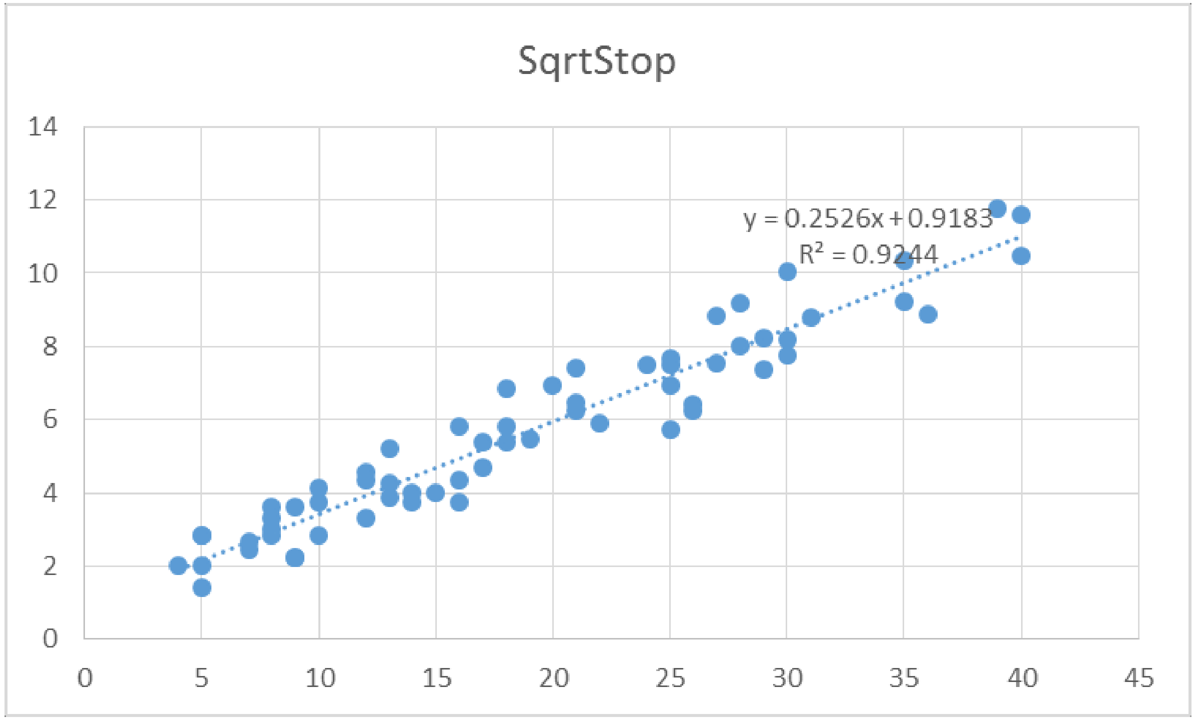

```
> scatter.smooth(Speed.mph,StopDist.ft,family="gaussian")
```

The smoothing line is created by Local Linear Regression (or “loess”) method.

Transformed data fits better

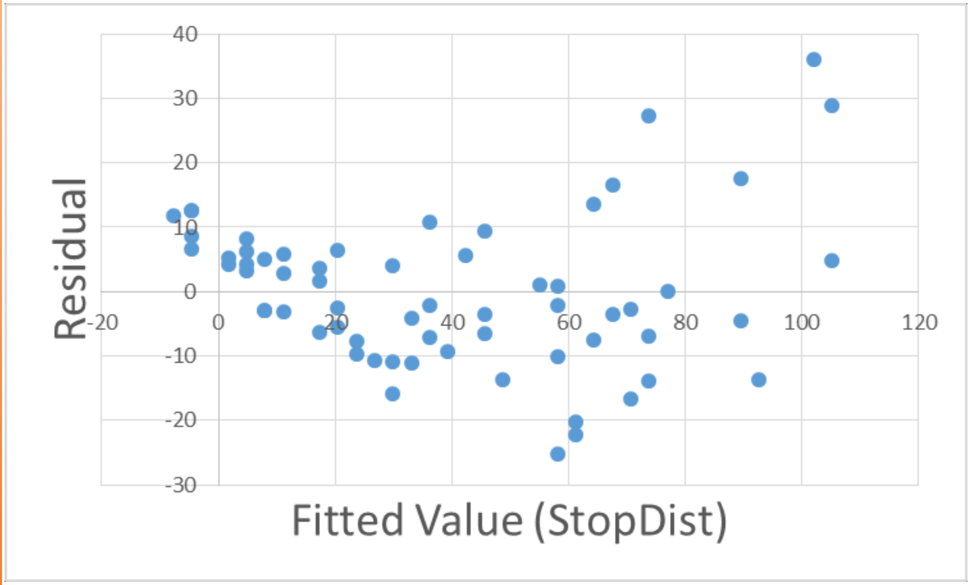
A large R^2 by itself doesn't imply that the linear model is correct.

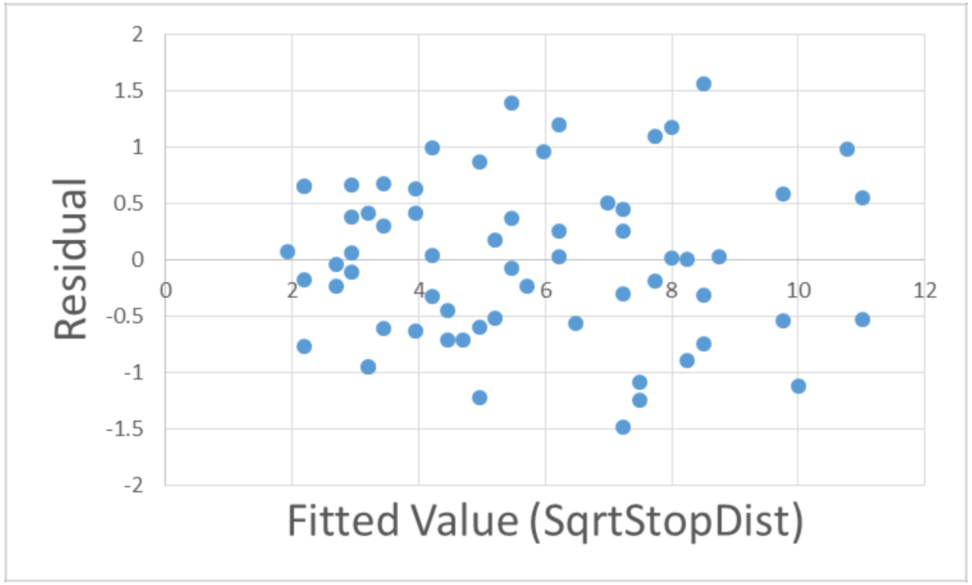




Transformed data fits better

Residuals show better homoscedacity for the transformed data

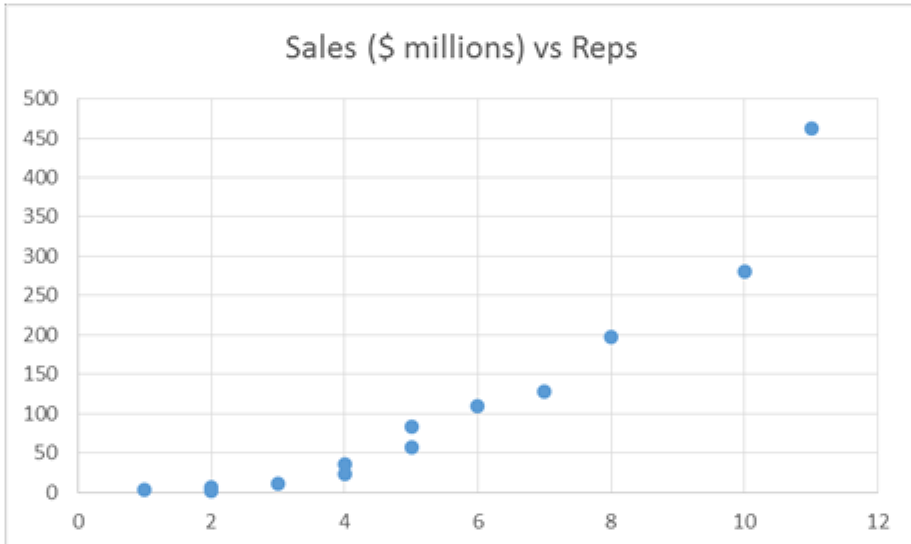


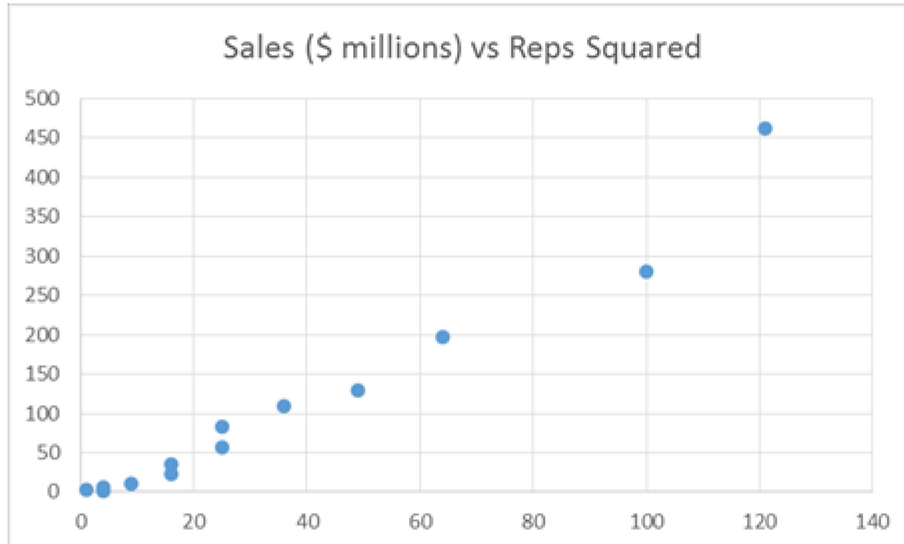


Moral: Linearity might not be applicable always. Use domain knowledge when available.

Nonlinear Models – Polynomial Regression - Excel

Sales volume versus # of sales reps and # of sales reps squared

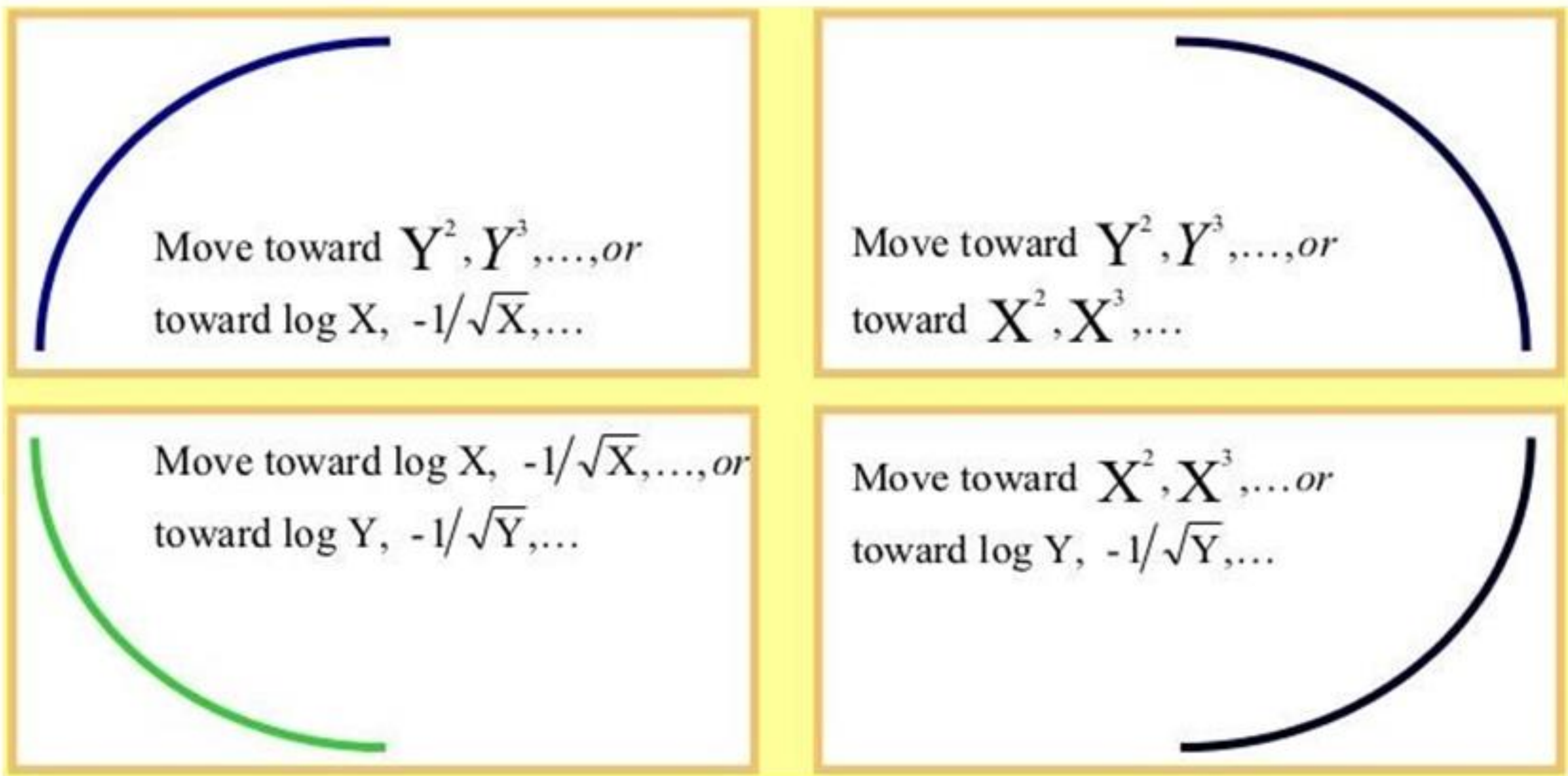




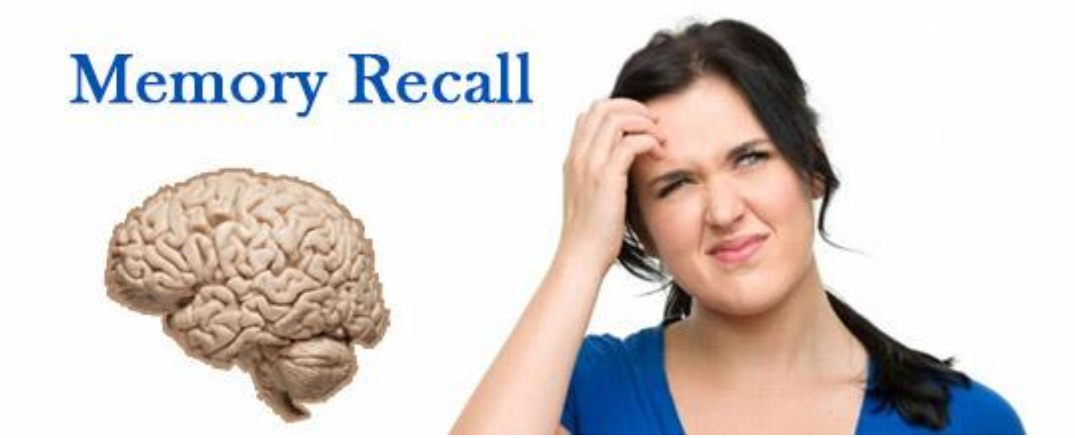
Tukey's Ladder of Transformations

Ladder for x		
Up ladder	Neutral	Down ladder
$\ldots, x^4, x^3, x^2, x$	$\sqrt{x}, x, logx$	$-\frac{1}{\sqrt{x}}, -\frac{1}{x}, -\frac{1}{x^2}, -\frac{1}{x^3}, \ldots$
Ladder for y		
Up ladder	Neutral	Down ladder
$\ldots, y^4, y^3, y^2, y$	$\sqrt{y}, y, logy$	$-\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}, \ldots$

Tukey's Four-Quadrant Approach



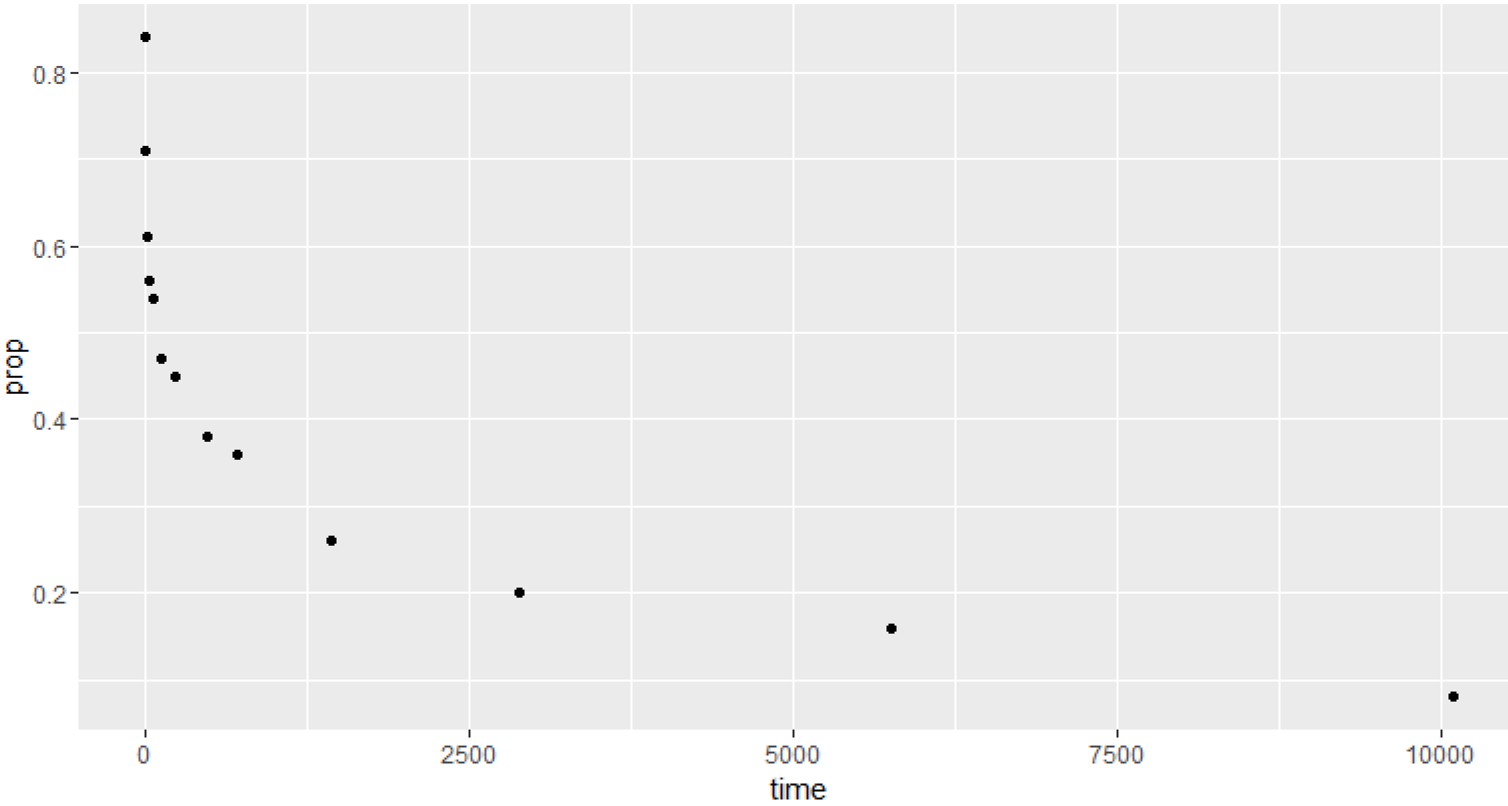
Nonlinear Transformation Example



For a study on memory retention, 13 volunteers were asked to memorize a list of disconnected items. The subjects were asked to recall the items at various times up to a week later.

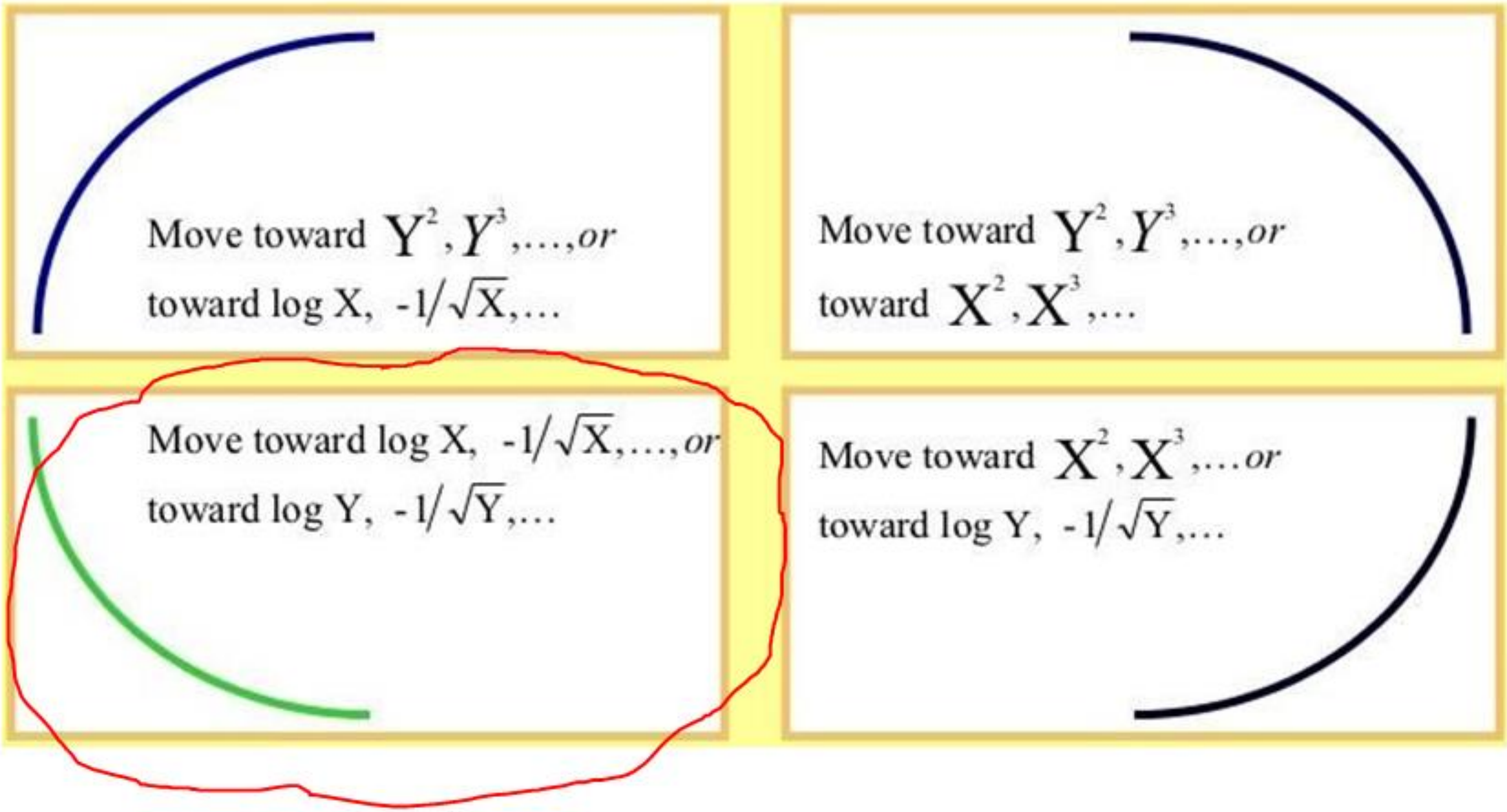
The proportion of items (y = *prop*) correctly recalled at various times (x = *time*, in minutes) since the list was memorized were recorded.

Nonlinear Transformation Example

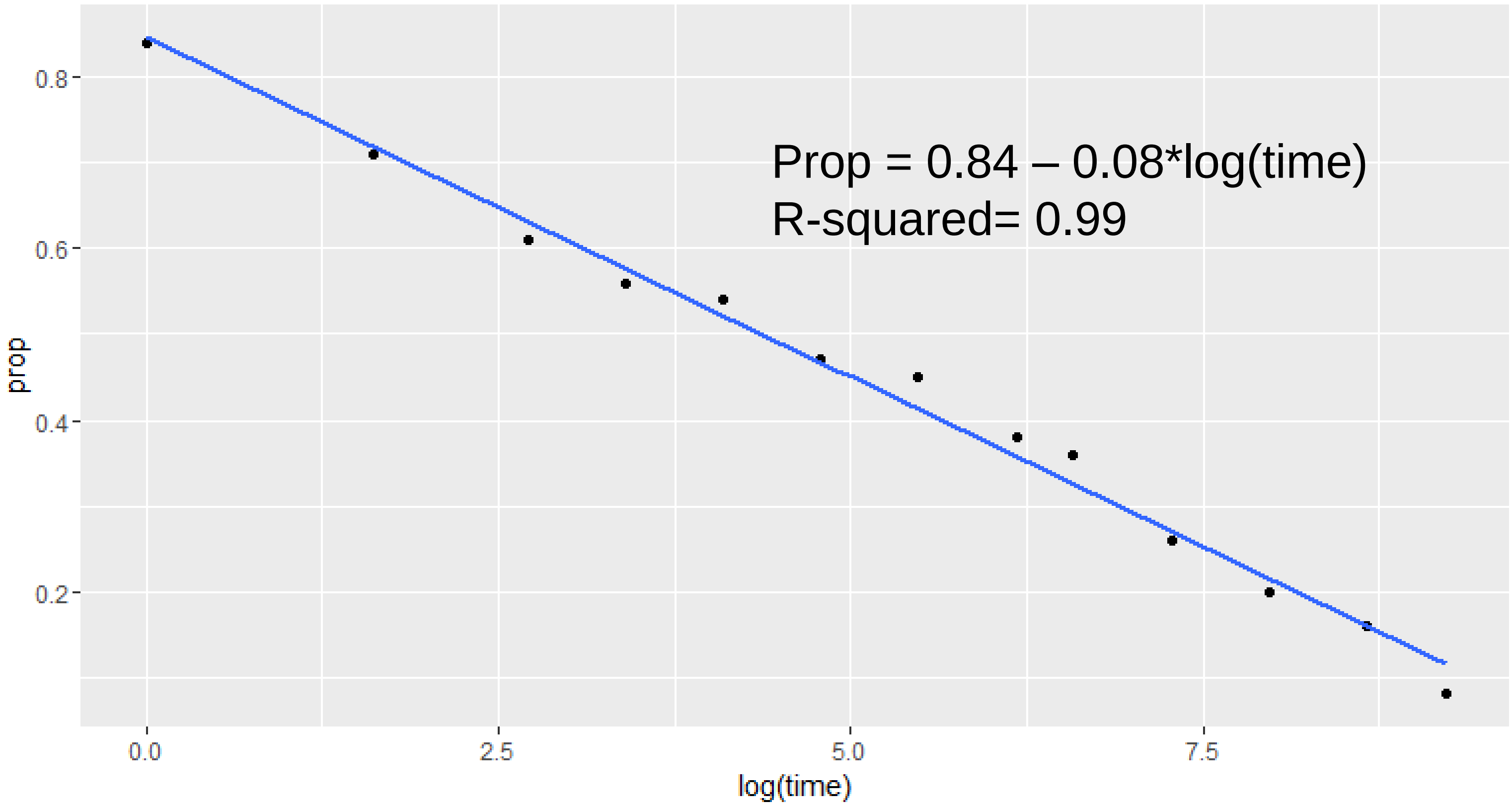


time	prop
1	0.84
5	0.71
15	0.61
30	0.56
60	0.54
120	0.47
240	0.45
480	0.38
720	0.36
1440	0.26
2880	0.20
5760	0.16
10080	0.08

Tukey's Four-Quadrant Approach



Nonlinear Transformation Example



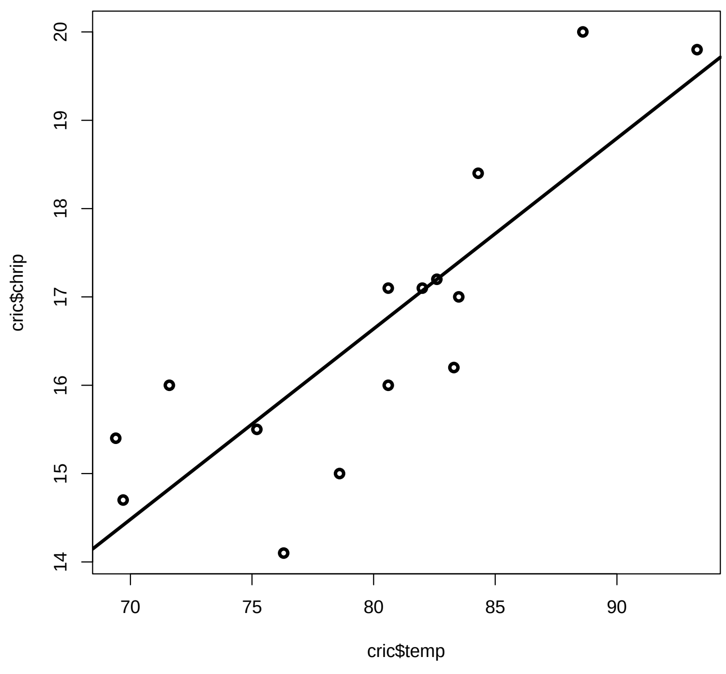
Cricket Chirps

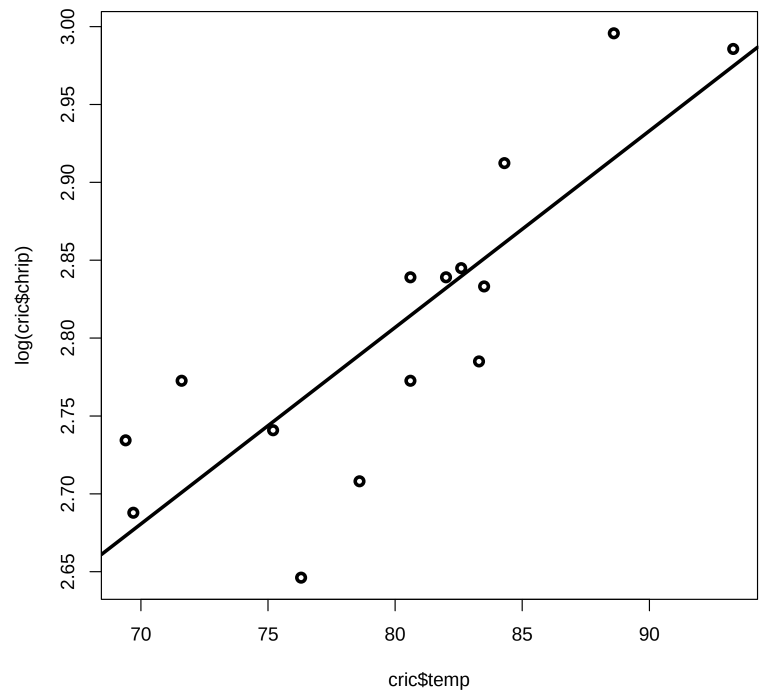
- http://www.almanac.com/content/cricket-chirps-natures-thermometer

Simple Model vs Complex : Cricket chirp vs temperature

Crickets are cold-blooded and their metabolism is influenced by the surrounding temperature. So external temperature has an important effect on their behavior, such as chirp frequency.

Consider two plots: chirps vs temperature (left) and log(chirps) vs temperature (right). Both they show more or less linear behaviour. In these cases the simplest of the models (linear on temperature) that fits should be preferred.





Nonlinear Models – Polynomial Regression

For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon$

How is this a special case of the general linear model?

Replace ${x_1}^2$ with x_2, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Multiple linear regression assumes a linear fit of the regression coefficients and regression constant, but not necessarily a linear relationship of the independent variable values.

Nonlinear Models – With Interaction

Interaction can be examined as a separate independent variable in regression.

For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

Nonlinear Models – Without Interaction - Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.687213365
R Square	0.47226221
Adjusted R Square	0.384305911
Standard Error	4.570195728
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	2	224.2930654	112.1465327	5.369282452	0.021602756
Residual	12	250.6402679	20.88668899		
Total	14	474.9333333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	50.85548009	3.790993168	13.41481713	1.38402E-08	42.59561554	59.11534464
Stock 2 ($)	-0.118999968	0.19308237	-0.616317112	0.54919854	-0.539690313	0.301690376
Stock 3 ($)	-0.07076195	0.198984841	-0.35561478	0.728301903	-0.504312675	0.362788775

Nonlinear Models – With Interaction - Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

CATEGORICAL PREDICTORS

Categorical Variables

Categorical variables such as gender, geographic region, occupation, marital status, level of education, economic class, religion, buying/renting a home, etc. can also be used in multiple regression analysis.

If there are n categories, n-1 dummy variables need to be inserted into the regression analysis.

Indicator (Dummy) Variables

If a survey question asks about the region of country your office is located in, with North, South, East and West as the options, the **recoding** can be done as follows:

Region
North
East
North
South
West
West
East

North	West	South
1	0	0
0	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	0

Indicator (Dummy) Variables - Excel

Consider the issue of sex discrimination in the salary earnings of workers in some industries. If there is discrimination, how much is one gender earning more than the other?

Indicator (Dummy) Variables - Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.943391358
R Square	0.889987254
Adjusted R Square	0.871651797
Standard Error	0.096791578
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	2	0.909488418	0.454744209	48.5391351	1.77279E-06
Residual	12	0.112423316	0.00936861		
Total	14	1.021911733			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.732060612	0.235584356	7.35218859	8.82767E-06	1.218766395	2.245354829
Age (10 years)	0.111220164	0.072083424	1.542936758	0.148795574	-0.045836124	0.268276453
Sex (1=Male, 0=Female)	0.458684065	0.053458498	8.58018991	1.82311E-06	0.342208003	0.575160126

Separate equation for each gender

Feature Selection and Model Building

Feature Selection

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 =
gear	Number of forward gears
carb	Number of carburetors

Mtcars database.

Feature Selection

Does Adding more explanatory variables result in a better fit?

Mpg =f(wt,hp)

```
> summary(lm(mpg~wt+hp,data=mtcars))

Call:
lm(formula = mpg ~ wt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-3.941 -1.600 -0.182  1.050  5.854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148 
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Mpg=g(wt,hp,qsec)

```
> summary(lm(mpg~wt+hp+qsec,data=mtcars))

Call:
lm(formula = mpg ~ wt + hp + qsec, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8591 -1.6418 -0.4636  1.1940  5.6092 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 27.61053    8.41993   3.279  0.00278 ** 
wt          -4.35880    0.75270  -5.791 3.22e-06 ***
hp          -0.01782    0.01498  -1.190  0.24418    
qsec         0.51083    0.43922   1.163  0.25463    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,    Adjusted R-squared:  0.8171 
F-statistic: 47.15 on 3 and 28 DF,  p-value: 4.506e-11
```

Adding an extra variable *qsec*, impacts the significance level of slope coefficient for *hp*

Feature Selection

```
> summary(lm(mpg~.,data=mtcars))

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

If we use all the available variables, none of them show up as being significant!

- How do we decide which variables are the best ones to fit the data?

Model Building: Search Procedures

Suppose a model to predict the world crude oil production (barrels per day) is to be developed and the predictors used are:

- US energy consumption (BTUs)
- Gross US nuclear electricity generation (kWh)
- US coal production (short-tons)
- Total US dry gas (natural gas) production (cubic feet)
- Fuel rate of US-owned automobiles (miles per gallon)

What does your intuition say about how each of these variables would affect the oil production?

Model Building: Search Procedures

Two considerations in model building:

- Explaining most variation in dependent variable
- Keeping the model simple AND economical

Quite often, the above two considerations are in conflict of each other.

If 3 variables can explain the variation nearly as well as 5 variables, the simpler model is better. Search procedures help choose the more attractive model.

Search Procedures: All Possible Regressions

All variables used in all combinations. For a dataset containing *k* independent variables, 2^k-1 models are examined. In the example of the oil production, 31 models are examined.

Tedious, Time-Consuming, Inefficient, Overwhelming.

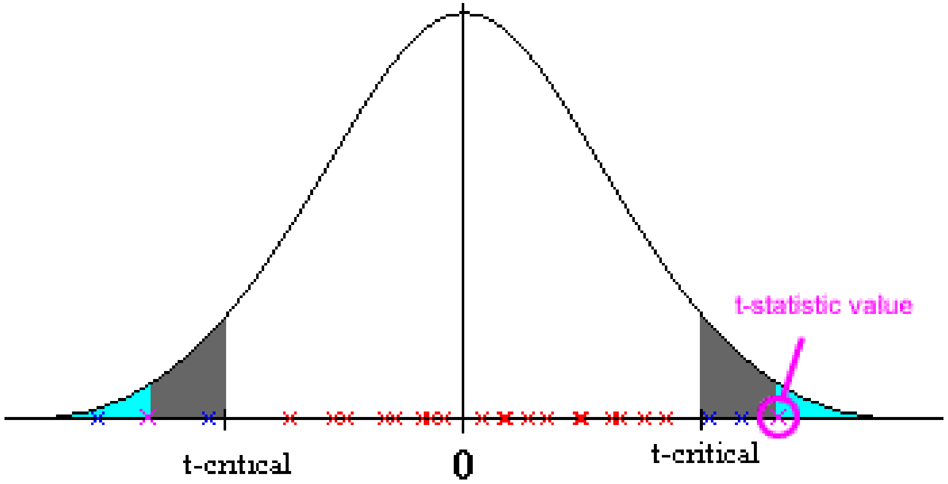
Search Procedures: Stepwise Regression

Starts a model with a single predictor and then adds or deletes predictors one step at a time.

- Step 1
  - Simple regression model for each of the independent variables one at a time.
  - Model with largest absolute value of *t* selected and the corresponding independent variable considered the best single predictor, denoted x_1.
  - If no variable produces a significant *t*, the search stops with no model.

Why LARGEST absolute *t* value and not the SMALLEST?

*Visualize the normal (or **t**) distribution, recall hypothesis testing, think of what the null hypothesis is and then understand what the largest and smallest absolute **t** values mean in terms of the distance from the null value.*



Search Procedures: Stepwise Regression

- Step 2
  - All possible two-predictor regression models with x_1 as one variable.
  - Model with largest absolute t value in conjunction with x_1 and one of the other k-1 variables denoted x_2.
  - Occasionally, if x_1 becomes insignificant, it is dropped and search continued with x_2.
  - If no other variables are significant, procedure stops.
- The above process continues with the 3rd variable added to the above 2 selected and so on.

Search Procedures: Stepwise Regression - Excel

Step 1

Dependent Variable	Independent Variable	*t* Score	*p*-value	R^2
Oil production	Energy consumption	11.77	1.86e-11	85.2%
Oil production	Nuclear	4.43	0.000176	45.0
Oil production	Coal	3.91	0.000662	38.9
Oil production	Dry gas	1.08	0.292870	4.6
Oil production	Fuel rate	3.54	0.00169	34.2

$$y = 13.075 + 0.580x_1$$

Search Procedures: Stepwise Regression - Excel

Step 2

Dependent Variable, y	Independent Variable, x_1	Independent Variable, x_2	t Score of x_2	p-value	R^2
Oil production	Energy consumption	Nuclear	-3.60	0.00152	90.6%
Oil production	Energy consumption	Coal	-2.44	0.0227	88.3
Oil production	Energy consumption	Dry gas	2.23	0.0357	87.9
Oil production	Energy consumption	Fuel rate	-3.75	0.00106	90.8

$$y = 7.14 + 0.772x_1 - 0.517x_2$$

t value for Energy Consumption is now at 11.91 and still significant (2.55e-11).

Search Procedures: Stepwise Regression - Excel

Step 3

Dependent Variable, *y*	**Independent Variable, x_1**	**Independent Variable, x_2**	**Independent Variable, x_3**	***t* Score of x_3**	***p*-value**
Oil production	Energy consumption	Fuel rate	Nuclear	-0.43	0.67210
Oil production	Energy consumption	Fuel rate	Coal	1.71	0.10225
Oil production	Energy consumption	Fuel rate	Dry gas	-0.46	0.65038

No *t* ratio is significant at $\alpha = 0.05$. No new variables are added to the model.

Search Procedures: Forward Selection

Same as stepwise, but once a variable is entered into the model, it is not re-examined in further steps.

When independent variables are correlated in forward selection, their overlapping information can limit the potential predictability of two or more variables in combination.

Search Procedures: Backward Elimination

Starts with a full model including all predictors and removes the **non-significant predictor** with the lowest absolute *t* value (highest *p* value).

Builds a new model with previously selected significant predictors and follows the same process.

Search Procedures: Backward Elimination

Step 1: Full Model

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.8357	4.64	0.000
Nuclear	-0.00654	-0.66	0.514
Coal	0.00983	1.35	0.193
Dry gas	-0.1432	-0.32	0.753
Fuel rate	-0.7341	-1.34	0.196

Search Procedures: Backward Elimination

Step 2: Four Predictors

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.7853	9.85	0.000
Nuclear	-0.004261	-0.64	0.528
Coal	0.010933	1.74	0.096
Fuel rate	-0.8253	-1.80	0.086

Search Procedures: Backward Elimination

Step 3: Three Predictors

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.75394	11.94	0.000
Coal	0.010479	1.71	0.102
Fuel rate	-1.0283	-3.14	0.005

Search Procedures: Backward Elimination

Step 4: Two Predictors

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.77201	11.91	0.000
Fuel rate	-0.5173	-3.75	0.001

All variables are significant. Process stops.

Feature Selection

- The same search process can be done with R^2 instead of t-values. That could lead potentially to a different set of variables.

- In R, a commonly used search method is *stepAIC* which tries to minimize AIC (Akaike Information Criteria)

Multicollinearity - Excel

Two or more independent variables are highly correlated.

	Energy consumption	Nuclear	Coal	Dry gas	Fuel rate
Energy consumption	1				
Nuclear	0.856	1			
Coal	0.791	0.952	1		
Dry gas	0.057	-0.404	-0.448	1	
Fuel rate	0.791	0.972	0.968	-0.423	1

Multicollinearity

Sign of estimated regression coefficient when interacting may be opposite of the signs when used as individual predictors.

For example, fuel rate and coal production are highly correlated (0.968).

$$\hat{y} = 44.869 + 0.7838(fuel\ rate)$$

$$\hat{y} = 45.072 + 0.0157(coal)$$

$$\hat{y} = 45.806 + 0.0277(coal) - 0.3934(fuel\ rate)$$

Multicollinearity

Multicollinearity can lead to a model where the model (*F* value) is significant but all individual predictors (*t* values) are insignificant.

```
> summary(lm(mpg~.,data=mtcars))

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Multicollinearity

- Stepwise regression prevents this problem to a great extent.

- Variance Inflation Factor (VIF): A regression analysis is conducted to predict an independent variable by the other independent variables. The independent variable being predicted becomes the dependent variable in this analysis.

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

VIF > 10 or ${R_i}^2$>0.90 for the largest VIFs indicates a severe multicollinearity.

MTcars example

```
> mtcarsLmOut <- lm(mpg~. , data=mtcars)
> vif(mtcarsLmOut)
      cyl      disp        hp      drat        wt      qsec        vs        am      gear      carb 
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487  5.357452  7.908747 
> 
```

StepAIC results in a truncated model

```
> mtcarsStepOut <- stepAIC(mtcarsLmOut)

> mtcarsStepOut

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Coefficients:
(Intercept)           wt         qsec           am  
      9.618       -3.917        1.226        2.936  

> vif(mtcarsStepOut)
      wt     qsec       am 
2.482952 1.364339 2.541437 
```

PUTTING IT ALL TOGETHER

Bike Sharing Program Data

We are provided hourly rental data spanning two years. You must predict the total count of bikes rented during each hour, using only information available prior to the rental period.

Bike Sharing Data

datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday - whether the day is considered a holiday
workingday - whether the day is neither a weekend nor holiday
weather - 1: Clear, Few clouds, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated
count - number of total rentals

Bike Sharing Data

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
01-01-2011 00:00	1	0	0	1	9.84	14.395	81	0	3	13	16
01-01-2011 01:00	1	0	0	1	9.02	13.635	80	0	8	32	40
01-01-2011 02:00	1	0	0	1	9.02	13.635	80	0	5	27	32
01-01-2011 03:00	1	0	0	1	9.84	14.395	75	0	3	10	13
01-01-2011 04:00	1	0	0	1	9.84	14.395	75	0	0	1	1
01-01-2011 05:00	1	0	0	2	9.84	12.88	75	6.0032	0	1	1
01-01-2011 06:00	1	0	0	1	9.02	13.635	80	0	2	0	2
01-01-2011 07:00	1	0	0	1	8.2	12.88	86	0	1	2	3
01-01-2011 08:00	1	0	0	1	9.84	14.395	75	0	1	7	8
01-01-2011 09:00	1	0	0	1	13.12	17.425	76	0	8	6	14
01-01-2011 10:00	1	0	0	1	15.58	19.695	76	16.9979	12	24	36
01-01-2011 11:00	1	0	0	1	14.76	16.665	81	19.0012	26	30	56
01-01-2011 12:00	1	0	0	1	17.22	21.21	77	19.0012	29	55	84
01-01-2011 13:00	1	0	0	2	18.86	22.725	72	19.9995	47	47	94
01-01-2011 14:00	1	0	0	2	18.86	22.725	72	19.0012	35	71	106
01-01-2011 15:00	1	0	0	2	18.04	21.97	77	19.9995	40	70	110
01-01-2011 16:00	1	0	0	2	17.22	21.21	82	19.9995	41	52	93
01-01-2011 17:00	1	0	0	2	18.04	21.97	82	19.0012	15	52	67
01-01-2011 18:00	1	0	0	3	17.22	21.21	88	16.9979	9	26	35
01-01-2011 19:00	1	0	0	3	17.22	21.21	88	16.9979	6	31	37
01-01-2011 20:00	1	0	0	2	16.4	20.455	87	16.9979	11	25	36
01-01-2011 21:00	1	0	0	2	16.4	20.455	87	12.998	3	31	34
01-01-2011 22:00	1	0	0	2	16.4	20.455	94	15.0013	11	17	28
01-01-2011 23:00	1	0	0	2	18.86	22.725	88	19.9995	15	24	39

Which variables are useful for prediction?
Identify the nature of each variable (categorical/numerical).

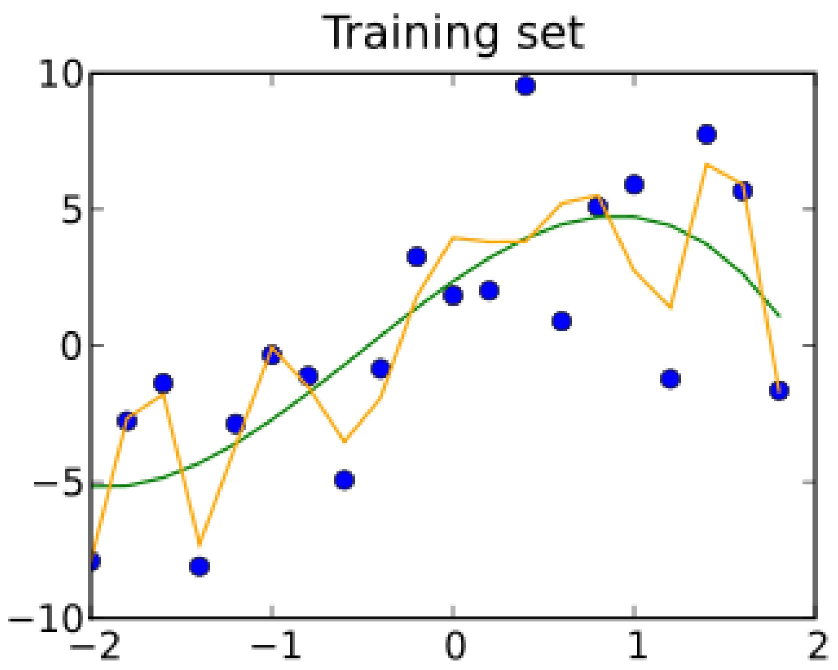
First Attempt

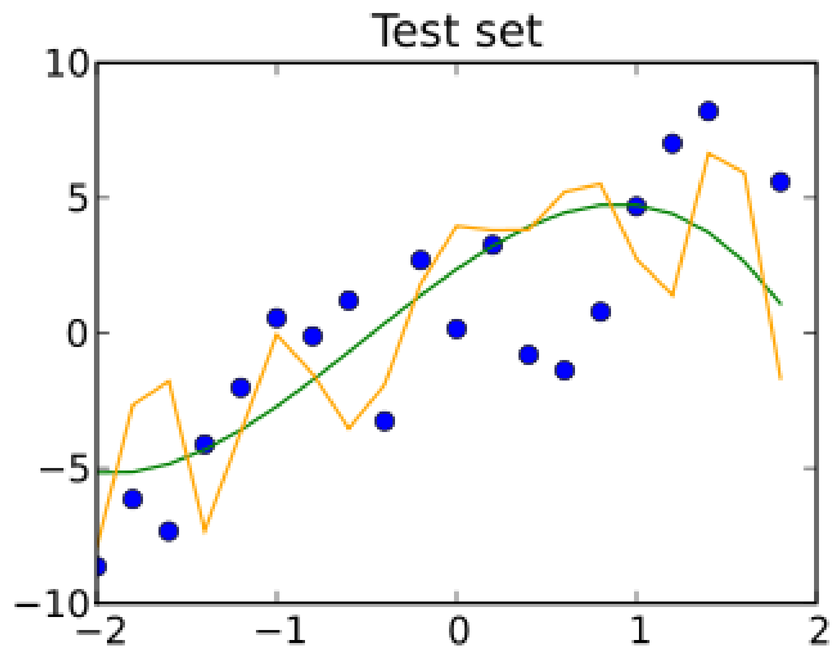
```
> lmbike0 <- lm(count~ season+holiday+workingday+weather+temp+atemp+humidity+windspeed,data=bike)
> summary(lmbike0)

Call:
lm(formula = count ~ season + holiday + workingday + weather +
    temp + atemp + humidity + windspeed, data = bike)

Residuals:
    Min      1Q  Median      3Q     Max
-335.81 -102.67  -31.95   66.44  677.02

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 135.79052    8.71016  15.590  < 2e-16 ***
season       22.75882    1.42662  15.953  < 2e-16 ***
holiday      -9.15872    9.27009  -0.988 0.323181
workingday   -1.14953    3.31527  -0.347 0.728795
weather       5.93872    2.61924   2.267 0.023389 *
temp          1.84737    1.14210   1.618 0.105796
atemp         5.63120    1.05057   5.360 8.49e-08 ***
humidity     -3.05684    0.09262 -33.003  < 2e-16 ***
windspeed     0.77762    0.19999   3.888 0.000102 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 155.8 on 10877 degrees of freedom
Multiple R-squared:  0.2609,    Adjusted R-squared:  0.2604
F-statistic:   480 on 8 and 10877 DF,  p-value: < 2.2e-16
```

Diagnostic Hints

- Coefficients that tend to infinity (or *NA*) could be a sign that an input is perfectly correlated with a subset of your responses. Or put another way, it could be a sign that this input is only really useful on a subset of your data, so perhaps it is time to segment the data.

Need for Segmenting the Data

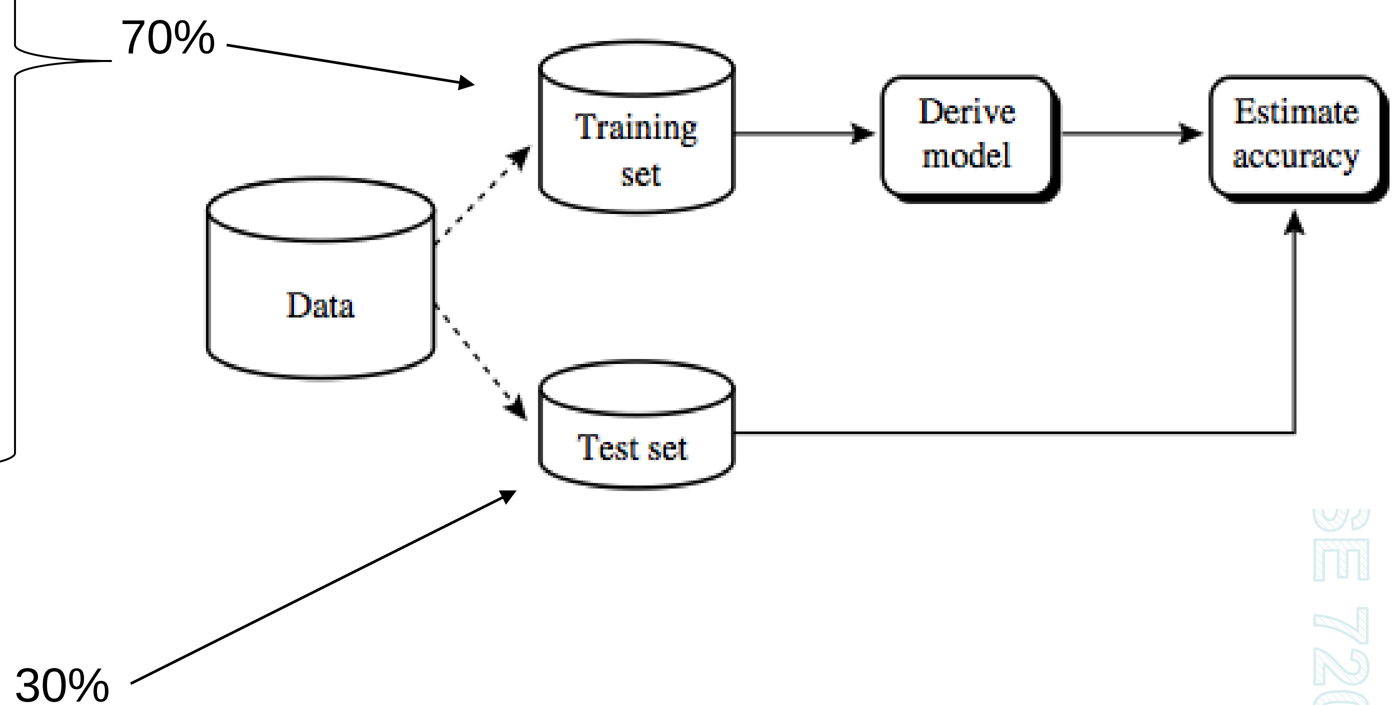




MSE1 = 4
MSE2 = 9

MSE1 = 15
MSE2 = 13

mpg	cyl	disp	hp	drat
21	6	160	110	3.9
21	6	160	110	3.9
22.8	4	108	93	3.85
21.4	6	258	110	3.08
18.7	8	360	175	3.15
18.1	6	225	105	2.76
14.3	8	360	245	3.21
24.4	4	146.7	62	3.69
22.8	4	140.8	95	3.92
19.2	6	167.6	123	3.92
17.8	6	167.6	123	3.92
16.4	8	275.8	180	3.07
17.3	8	275.8	180	3.07
15.2	8	275.8	180	3.07
10.4	8	472	205	2.93
10.4	8	460	215	3
14.7	8	440	230	3.23
32.4	4	78.7	66	4.08
30.4	4	75.7	52	4.93
33.9	4	71.1	65	4.22
21.5	4	120.1	97	3.7
15.5	8	318	150	2.76
15.2	8	304	150	3.15
13.3	8	350	245	3.73
19.2	8	400	175	3.08
27.3	4	79	66	4.08
26	4	120.3	91	4.43
30.4	4	95.1	113	3.77
15.8	8	351	264	4.22
19.7	6	145	175	3.62
15	8	301	335	3.54
21.4	4	121	109	4.11



Evaluating the Accuracy of Forecast

- Root mean-square error is a commonly used metric

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

- The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient.
- One can compare the RMSE to observed variation in measurements of a typical point.
- Other metrics such as Root mean-square log-error are also used, depending on the situation

RMSE for our First attempt fit

```
> #Lets extract the predictions of the model for the TestData
> OutputForTest0 <- predict(lmbike0,newdata=TestData)
>
> #Lets compute the root-mean-square error between actual and predicted
>  Error0<-rmse(TestData$count,OutputForTest0)
> Error0
[1] 155.5974
```

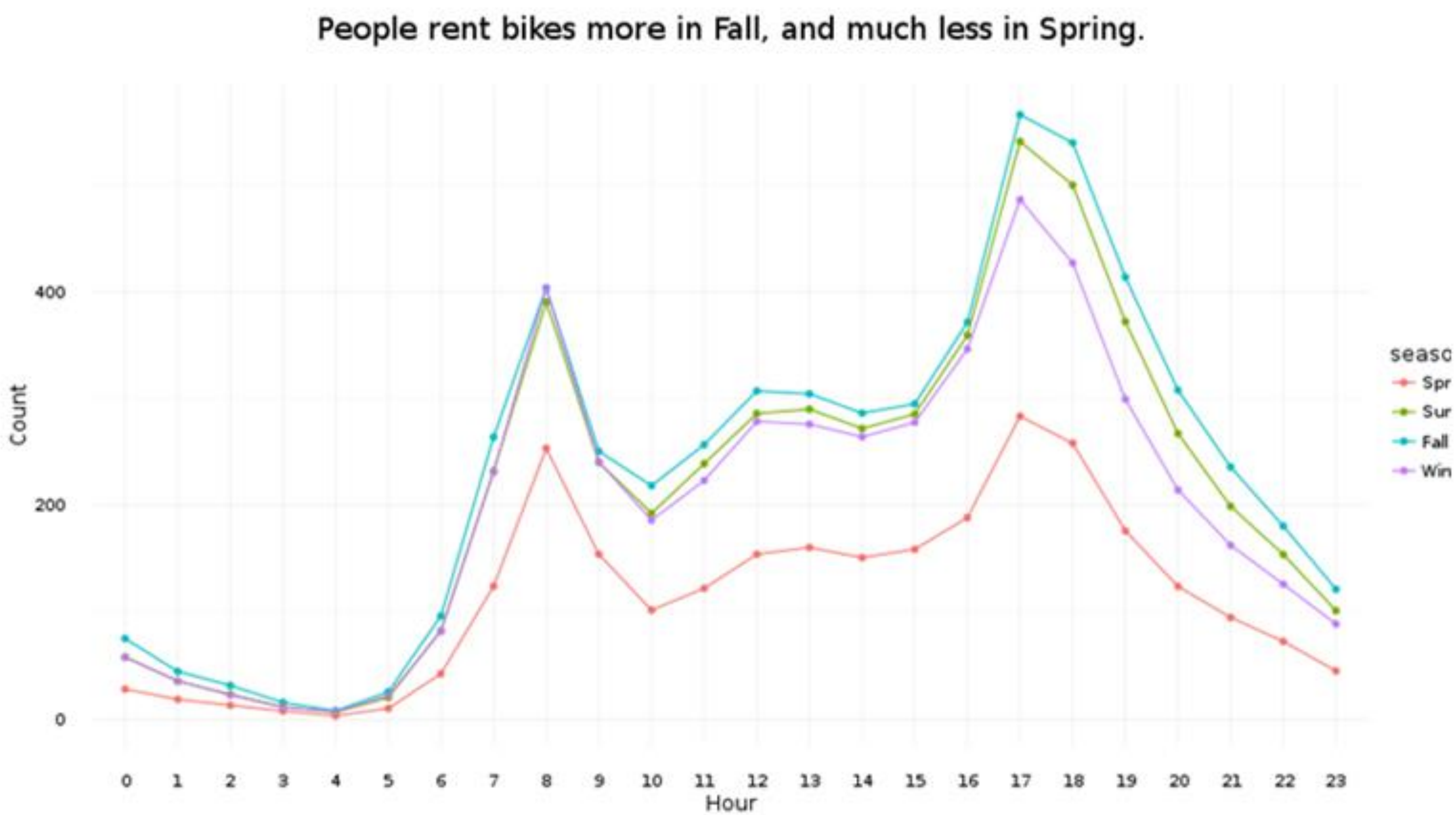
Caution: You might get slightly different numbers on your attempt, depending on the exact split of Training vs Testing data.

Bike Share Data

Lets extract other useful information

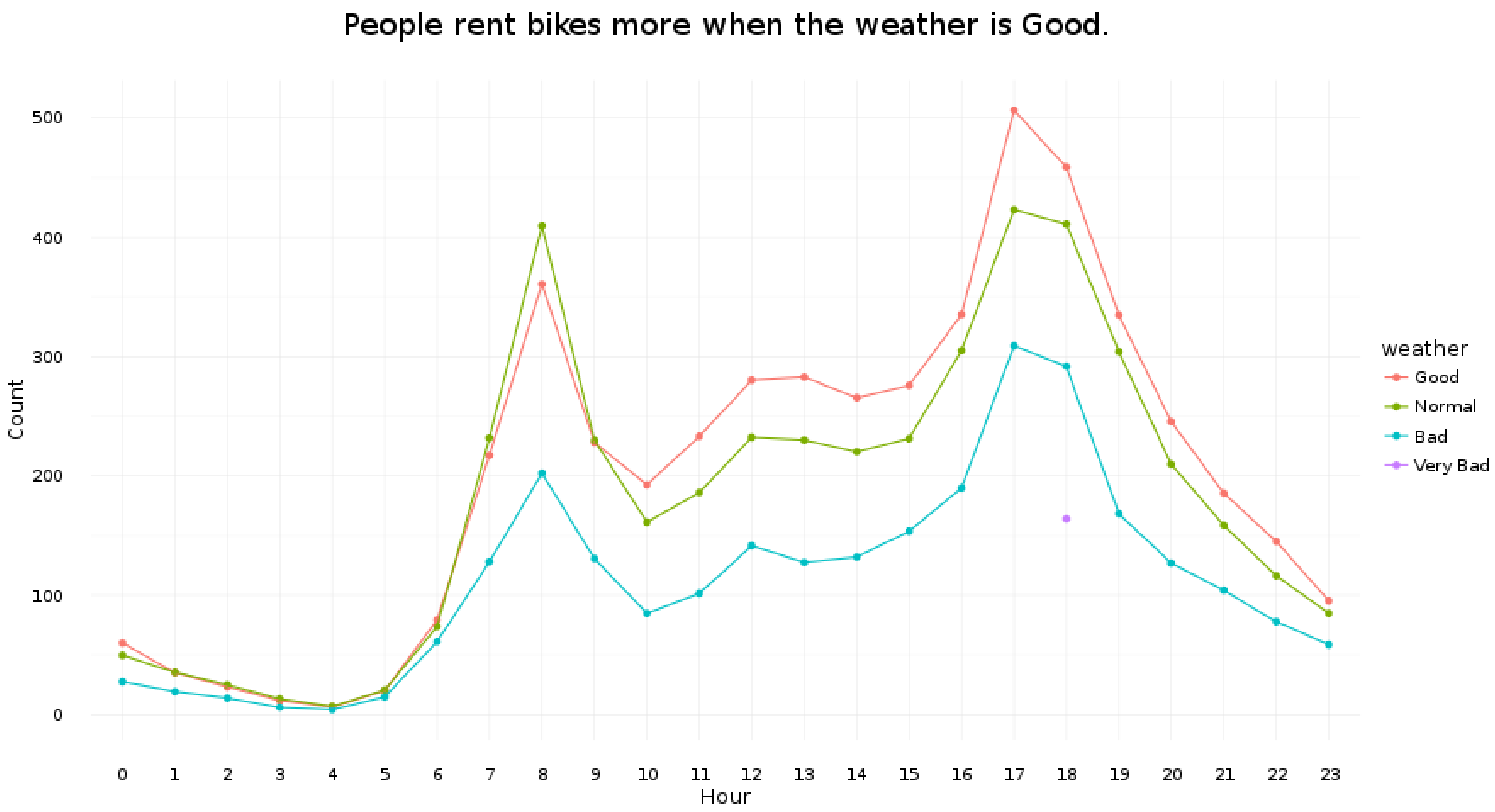
```
> bike <- read.csv("BikeShare.csv")
> str(bike)
'data.frame':   10886 obs. of  12 variables:
 $ datetime  : Factor w/ 10886 levels "2011-01-01 00:00:00",..: 1 2 3
 $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
 $ weather   : int  1 1 1 1 1 2 1 1 1 1 ...
 $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
 $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
 $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
 $ windspeed : num  0 0 0 0 0 ...
 $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
 $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
 $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
> #create day of week column
> bike$day <- weekdays(as.Date(bike$datetime))
> bike$day <- factor(bike$day)
>
> #Now lets extract date and time from the datetime stamp
> bike$time <- substring(bike$datetime,12,20)
>
```

Understanding the data



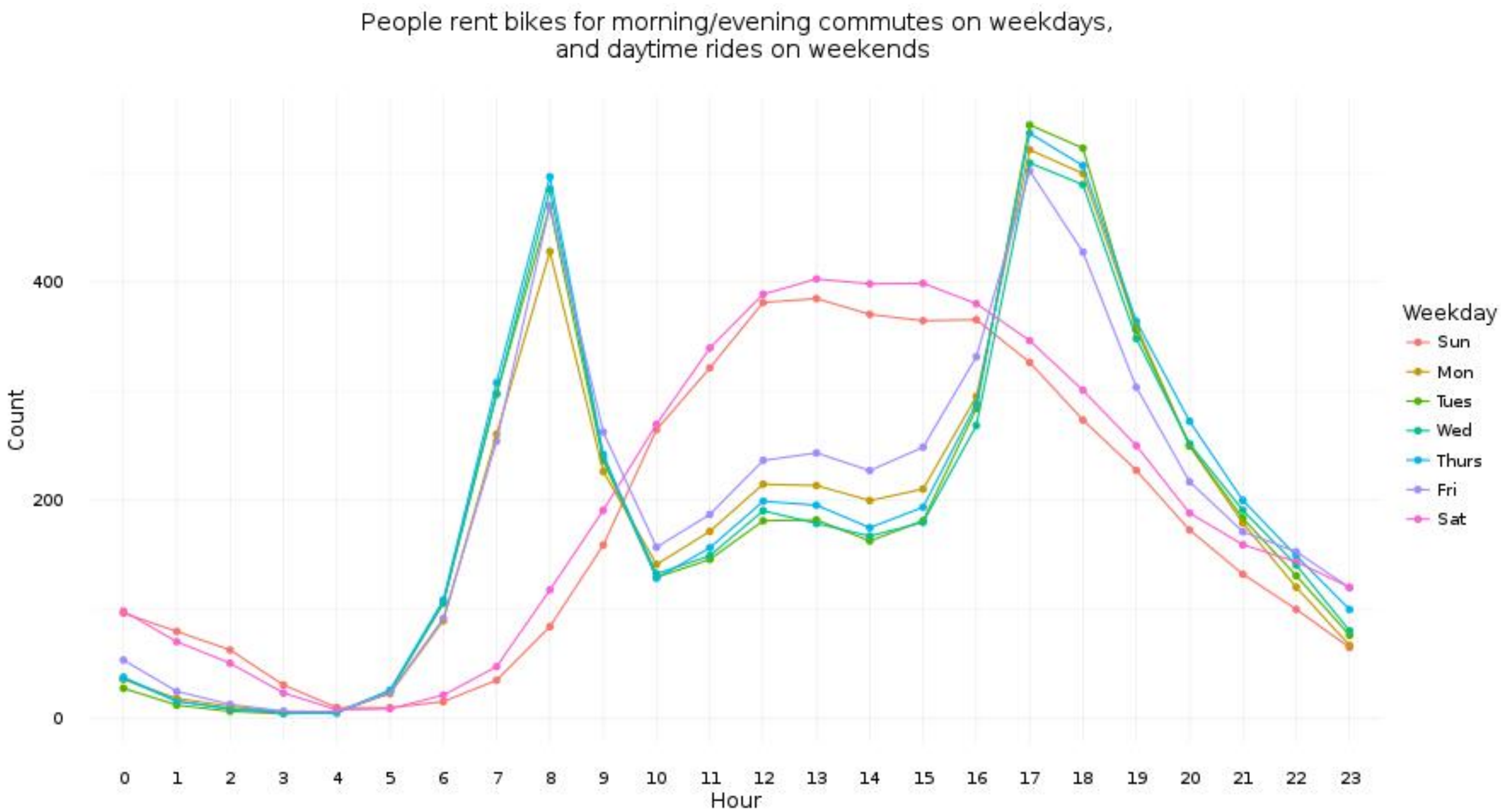

```
Warning message:
In cbind(hournum = c(1L, 2L, 3L, 4L, 5L, 6L, 7L, 8L, 9L, 10L, 11L,  :
  number of rows of result is not a multiple of vector length (arg 1)
> ggplot(bike,aes(x=hournum,y=count,colour=season)) +geom_point(data=season_summary,aes(group=season))
> ggplot(bike,aes(x=hournum,y=count,colour=season)) +geom_point(data=season_summary,aes(group=season)
+ ) +geom_line(data=season_summary,aes(group=season))
> 
```

Understanding the data



See: https://www.kaggle.com/h19881812/bike-sharing-demand/data-vizualization/code
for details on creating the plots

Understanding the data



See: https://www.kaggle.com/h19881812/bike-sharing-demand/data-vizualization/code for details on creating the plots

Add more Features and try again

Clearly hour of the day matters. So does the day of the week. Lets add these predictors (features) and redo the regression.

```
#second attempt
lmbike1 <- lm(count~ season+holiday+workingday+weather+temp+atemp+humidity+windspeed+ day +
time,data=bike[inTrain,])
```

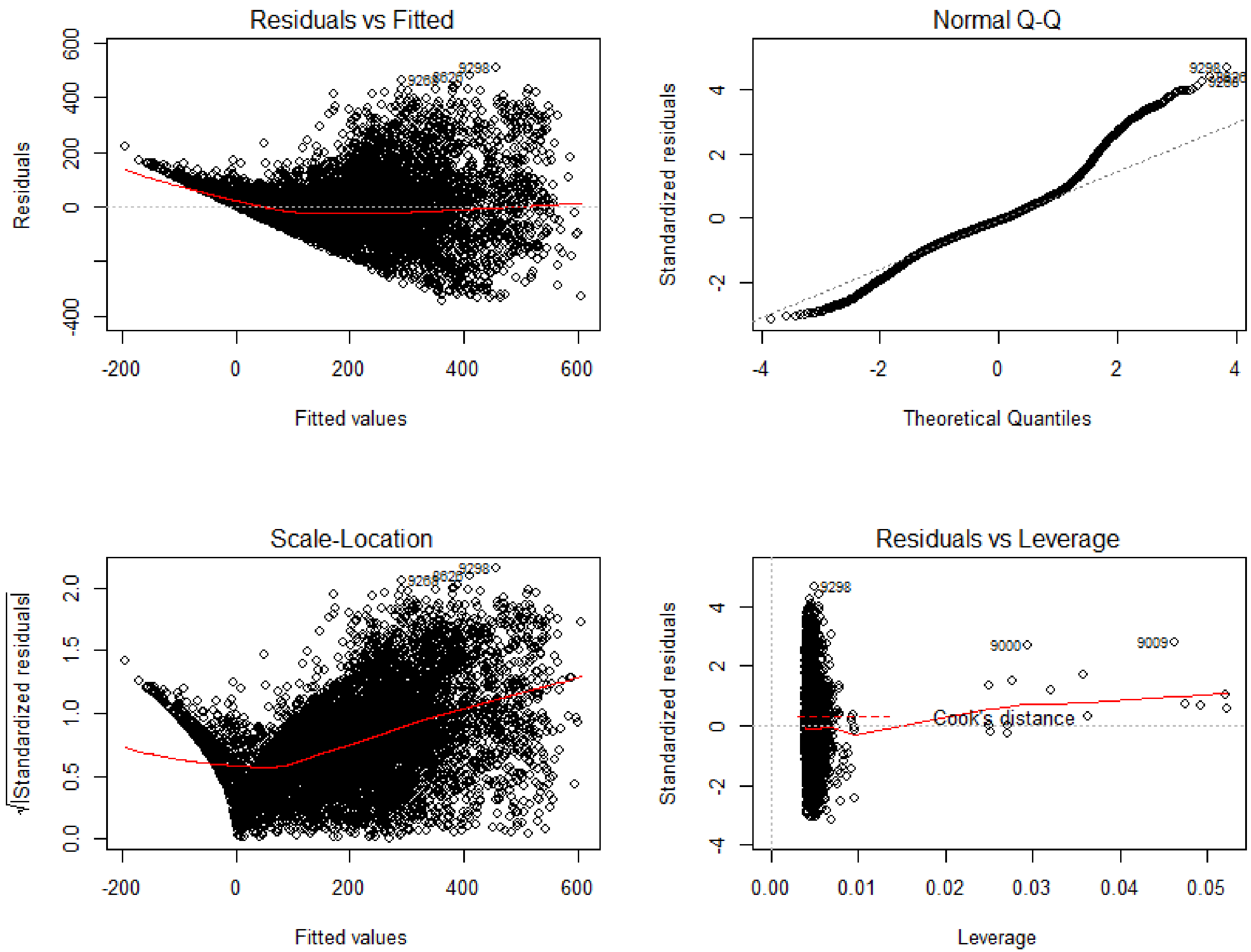
This gives us a much higher R-squared number

```
Residual standard error: 109.7 on 8125 degrees of freedom
Multiple R-squared:  0.6312,    Adjusted R-squared:  0.6293
F-statistic: 347.6 on 40 and 8125 DF,  p-value: < 2.2e-16
```

```
> #Lets compute the root-mean-square error between actual and predicted
>  Error1<-rmse(TestData$count,OutputForTest1)
>
> Error1
[1] 110.5511
>
```

We also get a smaller RMSE, indicating a better forecast in the test set

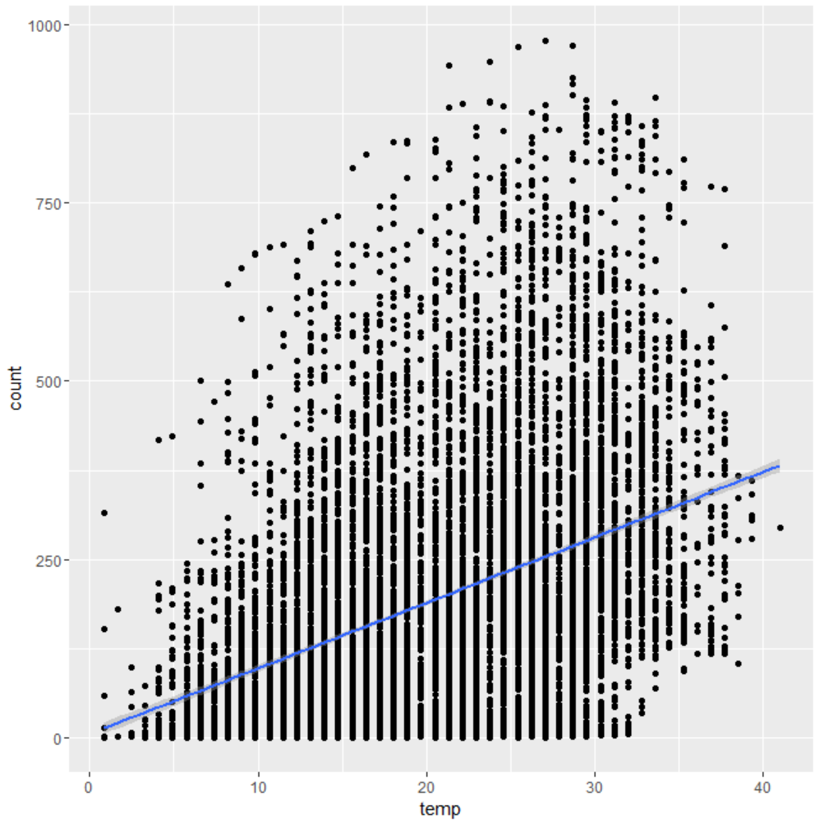
Analyze the Residuals



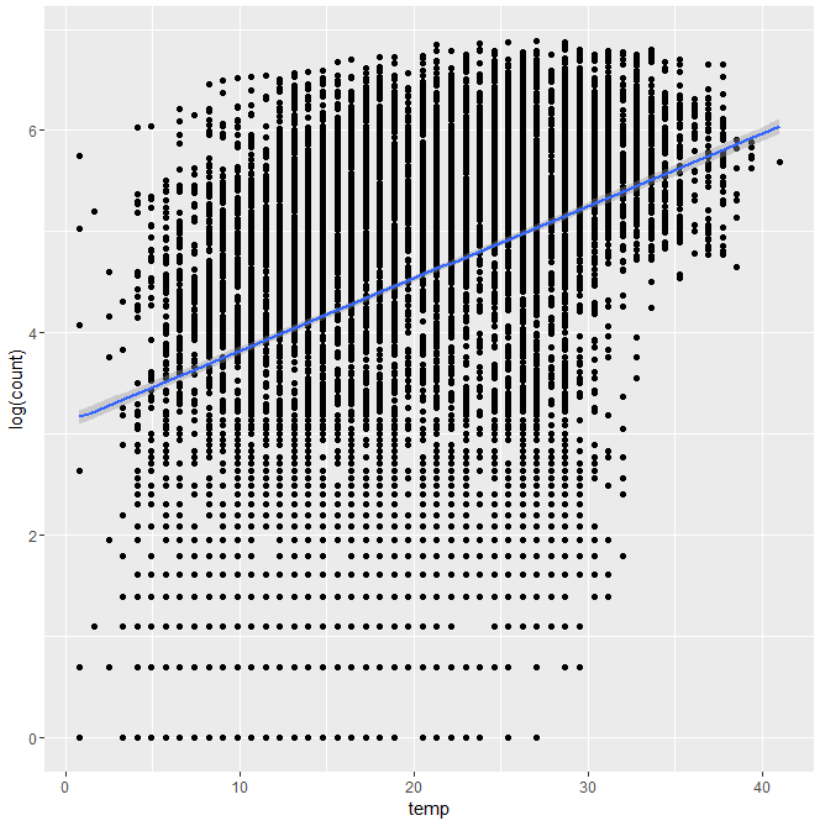
Non-linearity?

Rental Count vs Temperature

Count vs Temp



Log(Count) vs Temp



```
Call:
lm(formula = log(count) ~ season + weather + temp + atemp + humidity +
    windspeed + day + time, data = bike[inTrain, ])

Coefficients:
 (Intercept)       season2       season3       season4      weather2      weather3      weather4
    3.049276      0.344755      0.258877      0.552742     -0.022959     -0.549171      0.148066
        temp         atemp      humidity     windspeed     dayMonday   daySaturday     daySunday
    0.022266      0.017387     -0.003721     -0.004090     -0.183813      0.012058     -0.129645
 dayThursday    dayTuesday  dayWednesday time01:00:00  time02:00:00  time03:00:00  time04:00:00
   -0.107724     -0.206598     -0.184436     -0.634864     -1.198660     -1.706250     -2.009790
time05:00:00  time06:00:00  time07:00:00  time08:00:00  time09:00:00  time10:00:00  time11:00:00
   -0.947903      0.296741      1.279704      1.889773      1.593652      1.230197      1.320859
time12:00:00  time13:00:00  time14:00:00  time15:00:00  time16:00:00  time17:00:00  time18:00:00
    1.507606      1.492103      1.409758      1.427881      1.708407      2.113903      2.050704
time19:00:00  time20:00:00  time21:00:00  time22:00:00  time23:00:00
    1.781563      1.479375      1.230289      0.971963      0.590783
```

```
Residual standard error: 0.6687 on 8126 degrees of freedom
Multiple R-squared:  0.7991,    Adjusted R-squared:  0.7981
F-statistic: 828.6 on 39 and 8126 DF,  p-value: < 2.2e-16
```

```
> summary(TestData$count)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    42.0   145.0   192.1   283.2   977.0
> summary(OutputForTest3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.306  42.960 140.500 171.300 263.800 814.800
>
```

CSE 7202c

International School of Engineering

Plot 63/A, Floors 1&2, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering