MIE237 Term Test 2

2016-03-22

Examination Type B; Calculator Type 2 Permitted

50 Minutes; 40 Marks Available

Family Name:_____

Given Name:_____

Student Number:_____

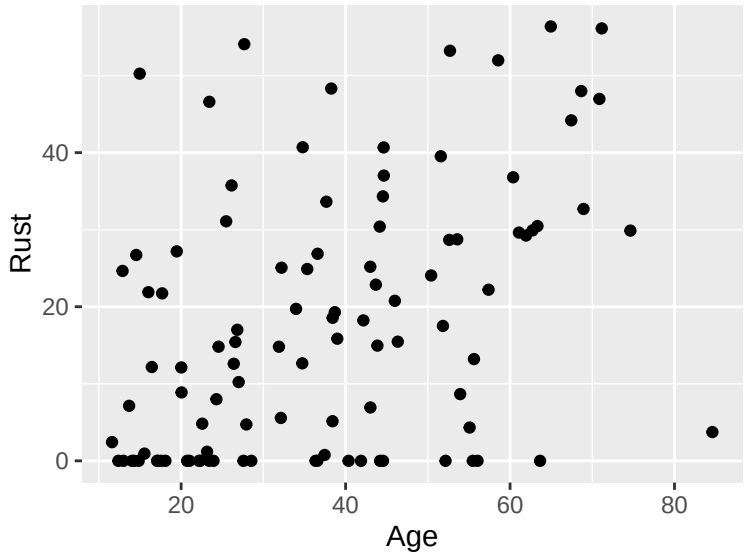
This test contains 9 pages. Pages 6–8 are tables. Page 9 is a formula sheet. You can detach the formula sheet if you like, but please don't detach the tables. (Detaching too many pages causes the test to fall apart.) You may use the backs of pages for rough work.

An electricity distribution company (a company that delivers electricity to homes and businesses) has accumulated a dataset related to 102 failed small transformers and wants to analyse some aspects of the data.

The dataset is similar in structure to the one from the first test, with one new variable added, similar to the first assignment. But the numbers are all different now.

The dataset has 5 variables: `ID, Manufacturer, Size, Age, Rust`. The variable `ID` contains the serial number of the transformer. The variable `Manufacturer` contains the manufacturer name, one of: `Camtran, Nema, Moloney`. The variable `Size` contains a description of the transformer's power rating. The variable `Age` contains the age in years of the transformer at the time of its failure. The variable `Rust` contains a number from 0 to 100 indicating the amount of corrosion that the transformer has.

1. **(25 marks total)** Here is a plot of the data along with the `R` output for the linear regression model fit with `Rust` as the output variable and `Age` as the input variable. Some of the entries have been removed and replaced with `XXX` (at least as many `X` as required).




```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  2.80224    3.58909   0.781     0.437
Age          XXXXXXX    0.08673   4.617 XXXXXXXXX
---

Residual standard error: 15.35 on XXXX degrees of freedom
Multiple R-squared:  XXXXXX,    Adjusted R-squared:  XXXXXX
F-statistic: XXXXX on XX and XXXX DF,  p-value: XXXXXXXX
```

a) **(5 marks)** Perform the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ **using an F distribution and estimating the p-value using the 0.05 and 0.01 probability F tables provided**.

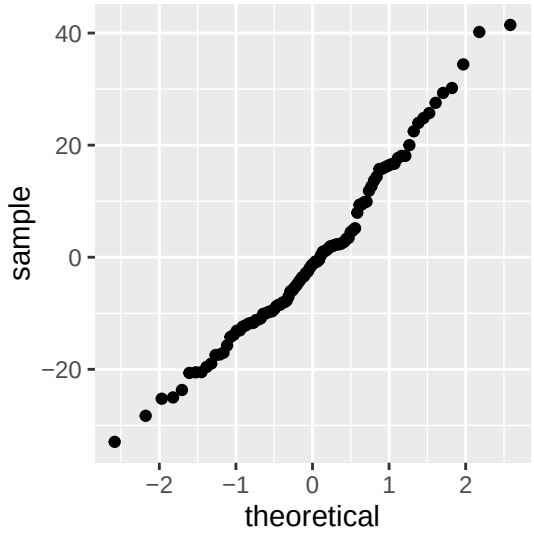
b) **(3 marks)** Now, give the *best possible* estimate for the p-value in a) using all available information in this midterm package (tables and formula sheet).

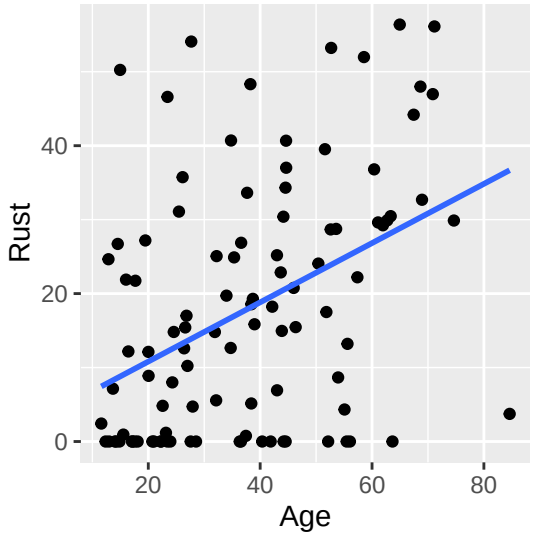
c) **(3 marks)** Produce a 95% confidence interval for β_1.

d) **(4 marks)** Denote the observed values of the 'Age' variable by $\{x_1, \ldots, x_n\}$. It turns out that $\overline{x} = 37.5$ and $S_{xx} = 31304.1$. Produce a 95% confidence interval for the mean value of 'Rust' at an 'Age' of 50.

e) **(3 marks)** Produce a 95% prediction interval for the value of 'Rust' for a transformer with an 'Age' of 15.

f) **(4 marks)** Here is a normal quantile plot of the residuals, along with a plot of the raw data with the fitted regression line included.





In light of these plots, which of the calculations done in a) to e) above do you think is probably the *least* accurate, and why?

g) **(3 marks)** Compute the sample correlation coefficient between the ‘Age‘ and ‘Rust‘ variables.

2. **(10 marks total)** Let's call the `Rust` variable y and the `Age` variable x_1. Two new variables x_2 and x_3 are added to the data. If the transformer was manufactured by Camtran, both x_2 and x_3 are set to 0. If the transformer was manufactured by Nema, x_2 is set to 1 and x_3 is set to 0. If the transformer was manufactured by Nema, x_2 is set to 0 and x_3 is set to 1.

The multiple regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$ is considered.

a) **(7 marks)**When the model is fit to the data, the value of MSE turns out to be 152.73. Perform the hypothesis test that answers the question "is there any linear relationship between the output variable and the input variables?". (You will need information from 1.a). If you are completely stuck you just can go ahead and use a total sum of squares of 30000 for a maximum of 4 out of 7 marks.)

b) **(3 marks)**The p-value for the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is 0.000000148. Give a brief practical conclusion you can make from this hypothesis test.

Upper tail probabilities for t_ν distributions $P(t_\nu \geq t)$													
df	0.3	0.2	0.15	0.1	0.05	0.025	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
11	0.540	0.876	1.088	1.363	1.796	2.201	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	0.538	0.870	1.079	1.350	1.771	2.160	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	0.536	0.866	1.074	1.341	1.753	2.131	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	0.534	0.863	1.069	1.333	1.740	2.110	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	0.533	0.861	1.066	1.328	1.729	2.093	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	0.532	0.859	1.063	1.323	1.721	2.080	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	0.532	0.858	1.061	1.321	1.717	2.074	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	0.532	0.858	1.060	1.319	1.714	2.069	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	0.531	0.857	1.059	1.318	1.711	2.064	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	0.531	0.856	1.058	1.315	1.706	2.056	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	0.531	0.855	1.057	1.314	1.703	2.052	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	0.530	0.855	1.056	1.313	1.701	2.048	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	0.530	0.854	1.055	1.311	1.699	2.045	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	0.530	0.854	1.055	1.310	1.697	2.042	2.147	2.278	2.457	2.581	2.750	3.030	3.646
31	0.530	0.853	1.054	1.309	1.696	2.040	2.144	2.275	2.453	2.576	2.744	3.022	3.633
32	0.530	0.853	1.054	1.309	1.694	2.037	2.141	2.271	2.449	2.571	2.738	3.015	3.622
33	0.530	0.853	1.053	1.308	1.692	2.035	2.138	2.268	2.445	2.566	2.733	3.008	3.611
34	0.529	0.852	1.052	1.307	1.691	2.032	2.136	2.265	2.441	2.562	2.728	3.002	3.601
35	0.529	0.852	1.052	1.306	1.690	2.030	2.133	2.262	2.438	2.558	2.724	2.996	3.591
36	0.529	0.852	1.052	1.306	1.688	2.028	2.131	2.260	2.434	2.555	2.719	2.990	3.582
37	0.529	0.851	1.051	1.305	1.687	2.026	2.129	2.257	2.431	2.551	2.715	2.985	3.574
38	0.529	0.851	1.051	1.304	1.686	2.024	2.127	2.255	2.429	2.548	2.712	2.980	3.566
39	0.529	0.851	1.050	1.304	1.685	2.023	2.125	2.252	2.426	2.545	2.708	2.976	3.558
40	0.529	0.851	1.050	1.303	1.684	2.021	2.123	2.250	2.423	2.542	2.704	2.971	3.551
41	0.529	0.850	1.050	1.303	1.683	2.020	2.121	2.248	2.421	2.539	2.701	2.967	3.544
42	0.528	0.850	1.049	1.302	1.682	2.018	2.120	2.246	2.418	2.537	2.698	2.963	3.538
43	0.528	0.850	1.049	1.302	1.681	2.017	2.118	2.244	2.416	2.534	2.695	2.959	3.532
44	0.528	0.850	1.049	1.301	1.680	2.015	2.116	2.243	2.414	2.532	2.692	2.956	3.526
45	0.528	0.850	1.049	1.301	1.679	2.014	2.115	2.241	2.412	2.529	2.690	2.952	3.520
46	0.528	0.850	1.048	1.300	1.679	2.013	2.114	2.239	2.410	2.527	2.687	2.949	3.515
47	0.528	0.849	1.048	1.300	1.678	2.012	2.112	2.238	2.408	2.525	2.685	2.946	3.510
48	0.528	0.849	1.048	1.299	1.677	2.011	2.111	2.237	2.407	2.523	2.682	2.943	3.505
49	0.528	0.849	1.048	1.299	1.677	2.010	2.110	2.235	2.405	2.521	2.680	2.940	3.500
50	0.528	0.849	1.047	1.299	1.676	2.009	2.109	2.234	2.403	2.519	2.678	2.937	3.496
60	0.527	0.848	1.045	1.296	1.671	2.000	2.099	2.223	2.390	2.504	2.660	2.915	3.460
80	0.526	0.846	1.043	1.292	1.664	1.990	2.088	2.209	2.374	2.486	2.639	2.887	3.416
100	0.526	0.845	1.042	1.290	1.660	1.984	2.081	2.201	2.364	2.475	2.626	2.871	3.390
120	0.526	0.845	1.041	1.289	1.658	1.980	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.170	2.326	2.432	2.576	2.807	3.291

	0.05 critical values for $F_{df1,df2}$ distributions. Columns: $df1$ Rows: $df2$								
	1	2	3	4	5	6	7	8	9
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124
41	4.079	3.226	2.833	2.600	2.443	2.330	2.243	2.174	2.118
42	4.073	3.220	2.827	2.594	2.438	2.324	2.237	2.168	2.112
43	4.067	3.214	2.822	2.589	2.432	2.318	2.232	2.163	2.106
44	4.062	3.209	2.816	2.584	2.427	2.313	2.226	2.157	2.101
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096
46	4.052	3.200	2.807	2.574	2.417	2.304	2.216	2.147	2.091
47	4.047	3.195	2.802	2.570	2.413	2.299	2.212	2.143	2.086
48	4.043	3.191	2.798	2.565	2.409	2.295	2.207	2.138	2.082
49	4.038	3.187	2.794	2.561	2.404	2.290	2.203	2.134	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880

	0.01 critical values for $F_{df1,df2}$ distributions. Columns: $df1$ Rows: $df2$								
	1	2	3	4	5	6	7	8	9
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067
31	7.530	5.362	4.484	3.993	3.675	3.449	3.281	3.149	3.043
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021
33	7.471	5.312	4.437	3.948	3.630	3.406	3.238	3.106	3.000
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981
35	7.419	5.268	4.396	3.908	3.592	3.368	3.200	3.069	2.963
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946
37	7.373	5.229	4.360	3.873	3.558	3.334	3.167	3.036	2.930
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915
39	7.333	5.194	4.327	3.843	3.528	3.305	3.137	3.006	2.901
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888
41	7.296	5.163	4.299	3.815	3.501	3.278	3.111	2.980	2.875
42	7.280	5.149	4.285	3.802	3.488	3.266	3.099	2.968	2.863
43	7.264	5.136	4.273	3.790	3.476	3.254	3.087	2.957	2.851
44	7.248	5.123	4.261	3.778	3.465	3.243	3.076	2.946	2.840
45	7.234	5.110	4.249	3.767	3.454	3.232	3.066	2.935	2.830
46	7.220	5.099	4.238	3.757	3.444	3.222	3.056	2.925	2.820
47	7.207	5.087	4.228	3.747	3.434	3.213	3.046	2.916	2.811
48	7.194	5.077	4.218	3.737	3.425	3.204	3.037	2.907	2.802
49	7.182	5.066	4.208	3.728	3.416	3.195	3.028	2.898	2.793
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718
80	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407

Simple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$

Analysis: $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

$\hat{\beta}_1 = S_{xy}/S_{xx}$

$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$

$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}$

Fitted value at x_i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

SS decomposition details:

$\sum_{i=1}^{n}(y_i - \overline{y})^2$	=	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	+	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
SST	=	SSR	+	SSE
$n-1$ d.f.	=	1 d.f.	+	$n-2$ d.f.

Test statistic for β_1:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

The denominator is called the "standard error" of $\hat{\beta}_1$

$(1-\alpha)\cdot 100\%$ C.I. for β_1 is:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{\frac{MSE}{S_{xx}}}$$

Alternate approach for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ uses (again... $T^2 = F$):

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F_{1,n-2}$$

$(1-\alpha)\cdot 100\%$ C.I. for mean response at x_0:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

$(1-\alpha)\cdot 100\%$ P.I. for new response at x_0:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Sample Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

In simple regression the square of r is equal to R^2

Multiple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$. Equivalently: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

No restriction on the x_i. Could be a function of one or more other input variables (esp. power or interaction of two terms)

SS decomposition is the same as in the simple case, except now the d.f. for SST, SSR, and SSE are $n-1$, k, and $n-(k+1)$ respectively.

Inference for individual β_i based on:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{MSE}\sqrt{c_{ii}}} \sim t_{n-(k+1)}$$

where c_{ii} is the $(i+1)^{st}$ diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$

Overall F test is based on

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F_{k,n-(k+1)}$$

No $T^2 = F$ relationship in multiple regression case.