Name _____ Spring, 2011

Applied Statistics Comprehensive Examination
Regression Methods & Linear Models

Calculators are permitted on this part of the examination.

1. (10 points) Recall the model for multiple regression using matrix notation: $y = X\beta + \varepsilon$, where $\varepsilon \sim N(0,\sigma^2 I)$ and β is estimated by $\hat{\beta} = (X^T X)^{-1} X^T y$. Show that $\hat{\beta}$ is an unbiased estimate of β and find $Var(\hat{\beta})$.

2. (40 points) Data are gathered on 25 brands of cigarettes including weight (g), tar content (mg), nicotine content (mg), and Carbon Monoxide (CO) content (mg). The dependent variable is CO. A regression analysis of all three independent variables can be found on pages 3 - 7 of this exam.
   a. (5 points) Conduct the global hypothesis test for the model.
   b. (5 points) Which, if any, of the four individual model coefficients are statistically significant? Give evidence.
   c. (10 points) What assumptions are required for a regression analysis? Give evidence of whether they hold for these data.
   d. (5 points) Identify any points that are influential. Give evidence.

   A regression analysis of all one and two independent variable models can be found on pages 8 - 13 of this exam.
   e. (15 points) Considering all information presented to you in this problem, which model is appropriate in identifying variables that are associated with CO? Explain how you arrived at your decision.

3. (20 points) Consider a 3 by 3 factorial design with these theoretical means:

$\mu_{11} = 8$	$\mu_{12} = 0$	$\mu_{13} = 10$
$\mu_{21} = 6$	$\mu_{22} = ?$	$\mu_{23} = ?$
$\mu_{31} = 4$	$\mu_{32} = ?$	$\mu_{33} = ?$

   a. (5 points) Assuming that there is no interaction, find the missing theoretical means.
   b. (5 points) Specify any 2 degree of freedom contrast in rows assuming a cell means model. Use the following order for the parameter vector:
   $$\mu' = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{31}, \mu_{32}, \mu_{33}).$$
   c. (5 points) Specify the Population Marginal Means for columns assuming a fixed effects model with terms for rows, columns and interaction.
   d. (5 points) Suppose there are 3 observations per cell and a fixed effects model with terms for rows, columns and interaction is fit to the data. Determine how many degrees of freedom are associated with the Mean Squared Error and justify your answer.

4. (30 points) Consider the following incomplete block design with 6 treatments and 3 blocks: (A, B, D, E), (B, C, E, F), (A, C, D, F). Assume a mathematical model with fixed effects for treatments and blocks and no interaction between treatments and blocks.
    a. (10 points) Write the design matrix X, with no restrictions on the parameters.
    b. (10 points) Determine if the design is connected for treatments and justify your answer.
    c. (10 points) Determine if the treatment contrast $\tau_A - \tau_C$ is estimable and justify your answer.

Regression Analysis: CO versus Weight, Nicotine, Tar

The regression equation is

CO = 3.20 - 0.13 Weight - 2.63 Nicotine + 0.963 Tar

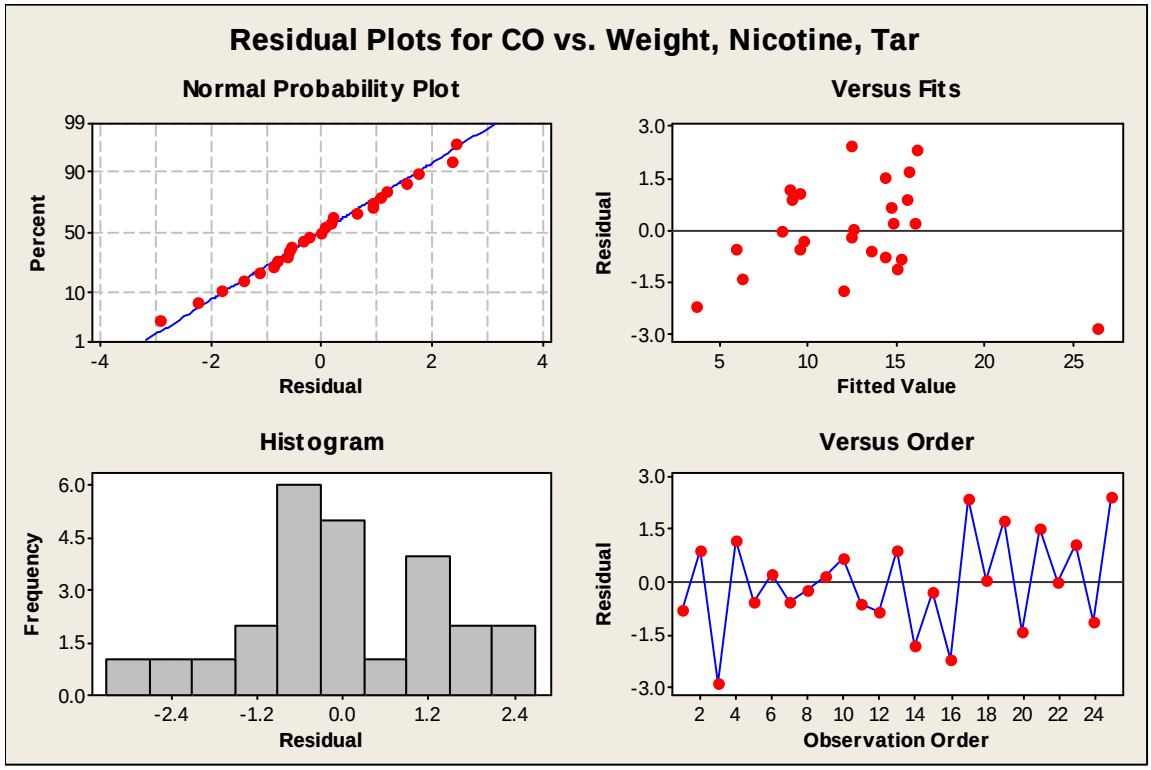
Predictor	Coef	SE Coef	T	P	VIF
Constant	3.202	3.462	0.93	0.365	
Weight	-0.130	3.885	-0.03	0.974	1.334
Nicotine	-2.632	3.901	-0.67	0.507	21.900
Tar	0.9626	0.2422	3.97	0.001	21.631

S = 1.44573 R-Sq = 91.9% R-Sq(adj) = 90.7%

Mallows Cp = 4.0

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	495.26	165.09	78.98	0.000
Residual Error	21	43.89	2.09		
Total	24	539.15			



Obs	Weight	CO	Fit	SE Fit	Residual	St Resid
1	0.99	13.600	14.383	0.597	-0.783	-0.59
2	1.09	16.600	15.671	0.554	0.929	0.70
3	1.17	23.500	26.393	1.030	-2.893	-2.85RX
4	0.93	10.200	9.018	0.418	1.182	0.85
5	0.95	5.400	5.973	0.535	-0.573	-0.43
6	0.89	15.000	14.788	0.513	0.212	0.16
7	1.03	9.000	9.539	0.567	-0.539	-0.41
8	0.92	12.300	12.518	0.436	-0.218	-0.16
9	0.94	16.300	16.111	0.453	0.189	0.14
10	0.89	15.400	14.745	0.518	0.655	0.49
11	0.96	13.000	13.606	0.353	-0.606	-0.43
12	0.93	14.400	15.247	0.696	-0.847	-0.67
13	0.97	10.000	9.084	0.423	0.916	0.66
14	1.12	10.200	11.976	0.728	-1.776	-1.42
15	0.85	9.500	9.807	0.565	-0.307	-0.23
16	0.79	1.500	3.720	0.783	-2.220	-1.83
17	0.92	18.500	16.130	0.646	2.370	1.83
18	1.04	12.600	12.545	0.735	0.055	0.04
19	0.96	17.500	15.760	0.634	1.740	1.34
20	0.91	4.900	6.310	0.498	-1.410	-1.04
21	1.01	15.900	14.370	0.323	1.530	1.09
22	0.98	8.500	8.496	0.442	0.004	0.00
23	0.97	10.600	9.538	0.380	1.062	0.76
24	0.95	13.900	15.025	0.391	-1.125	-0.81
25	1.12	14.900	12.449	0.687	2.451	1.93

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: CO

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	495.25781	165.08594	78.98	<.0001
Error	21	43.89259	2.09012		
Corrected Total	24	539.15040			

Root MSE	1.44573	R-Square	0.9186
Dependent Mean	12.52800	Adj R-Sq	0.9070
Coeff Var	11.53996		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.20219	3.46175	0.93	0.3655
Tar	1	0.96257	0.24224	3.97	0.0007
Nicotine	1	-2.63166	3.90056	-0.67	0.5072
Weight	1	-0.13048	3.88534	-0.03	0.9735

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: CO

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	-0.7827	-0.5850	0.1704	1.3690	-0.2651
2	0.9289	0.6869	0.1470	1.2981	0.2852
3	-2.8926	-3.5537	0.5075	0.3484	-3.6073
4	1.1815	0.8480	0.0835	1.1515	0.2560
5	-0.5726	-0.4179	0.1372	1.3606	-0.1666
6	0.2121	0.1532	0.1258	1.3840	0.0581
7	-0.5388	-0.3969	0.1536	1.3917	-0.1691
8	-0.2177	-0.1542	0.0908	1.3305	-0.0487
9	0.1888	0.1343	0.0984	1.3433	0.0444
10	0.6553	0.4765	0.1283	1.3329	0.1828
11	-0.6057	-0.4235	0.0597	1.2473	-0.1067
12	-0.8470	-0.6592	0.2315	1.4513	-0.3618
13	0.9164	0.6537	0.0855	1.2213	0.1999
14	-1.7762	-1.4597	0.2534	1.0860	-0.8504
15	-0.3068	-0.2253	0.1526	1.4200	-0.0956
16	-2.2202	-1.9440	0.2933	0.8608	-1.2525
17	2.3698	1.9511	0.1999	0.7567	0.9751
18	0.0547	0.0429	0.2585	1.6386	0.0253
19	1.7404	1.3671	0.1925	1.0529	0.6674
20	-1.4097	-1.0408	0.1188	1.1171	-0.3822
21	1.5299	1.0905	0.0499	1.0154	0.2499
22	0.004285	0.003038	0.0934	1.3407	0.0010
23	1.0620	0.7535	0.0691	1.1675	0.2053
24	-1.1251	-0.8015	0.0732	1.1558	-0.2253
25	2.4508	2.0727	0.2261	0.7212	1.1204

Output Statistics

	DFBETAS			
Obs	Intercept	Tar	Nicotine	Weight
1	-0.0101	-0.2300	0.2278	-0.0266
2	-0.1861	0.1115	-0.1161	0.2000
3	1.3517	0.3859	-1.0067	-0.7090
4	0.0586	-0.1465	0.1244	-0.0380
5	0.0132	0.0511	-0.0179	-0.0429
6	0.0437	0.0068	0.0023	-0.0450
7	0.0795	0.1164	-0.0965	-0.0835
8	-0.0222	0.0259	-0.0296	0.0244
9	0.0230	0.0140	-0.0063	-0.0242
10	0.1398	0.0393	-0.0118	-0.1407
11	-0.0207	0.0439	-0.0530	0.0259
12	-0.1284	-0.3179	0.2955	0.0880
13	-0.0261	0.0418	-0.0750	0.0696
14	0.6577	-0.1862	0.2934	-0.7497
15	-0.0678	0.0436	-0.0464	0.0663
16	-0.8265	-0.1810	0.3169	0.5782
17	0.4468	-0.4355	0.5969	-0.5762
18	-0.0086	-0.0217	0.0216	0.0062
19	0.1694	0.5817	-0.5363	-0.1026
20	-0.0766	0.1083	-0.0437	0.0080
21	-0.0539	0.0360	-0.0240	0.0602
22	-0.0002	-0.0004	0.0002	0.0004
23	-0.0248	-0.0926	0.0660	0.0475
24	-0.0947	-0.1043	0.0757	0.0874
25	-0.8652	0.2507	-0.3805	0.9789

Sum of Residuals	0
Sum of Squared Residuals	43.89259
Predicted Residual SS (PRESS)	89.47755

Regression Analysis: CO versus Weight, Nicotine

The regression equation is
CO = 1.61 + 0.06 Weight + 12.4 Nicotine

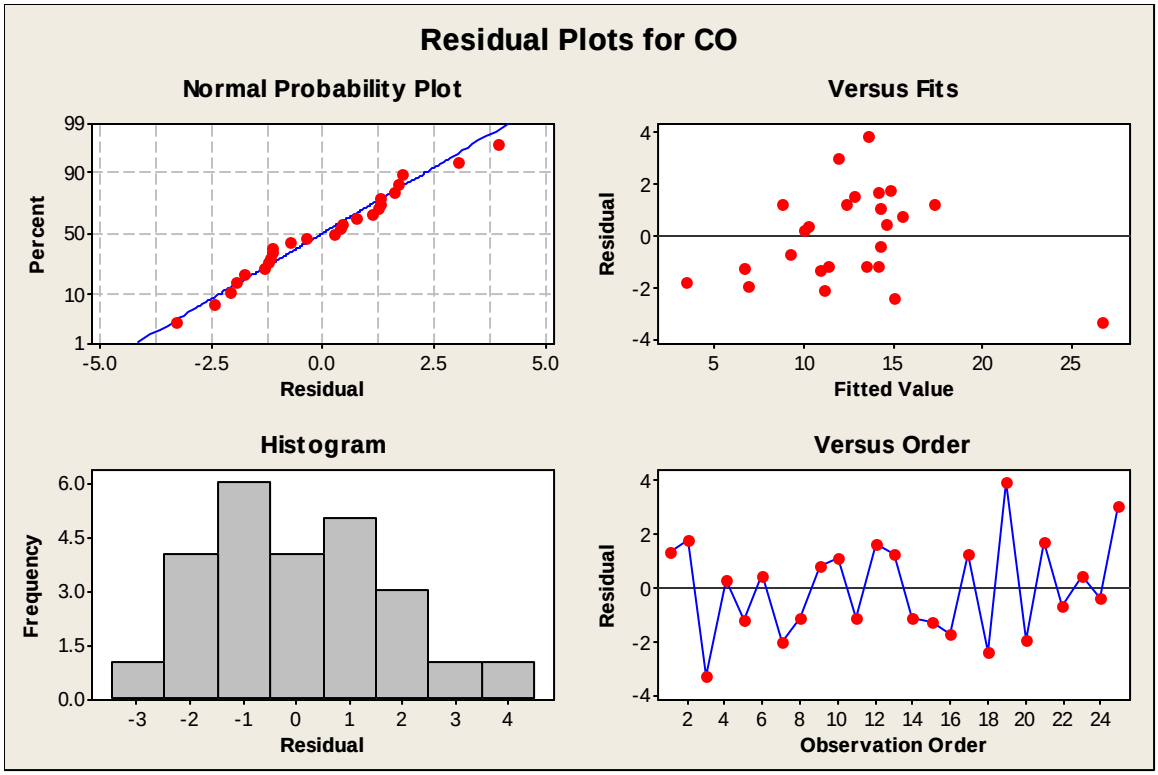
Predictor	Coef	SE Coef	T	P	VIF
Constant	1.614	4.447	0.36	0.720	
Weight	0.059	5.024	0.01	0.991	1.334
Nicotine	12.388	1.245	9.95	0.000	1.334

S = 1.86954 R-Sq = 85.7% R-Sq(adj) = 84.4%

Mallows Cp = 17.8

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	462.26	231.13	66.13	0.000
Residual Error	22	76.89	3.50		
Total	24	539.15			



Regression Analysis: CO versus Weight, Tar

The regression equation is
CO = 3.11 - 0.42 Weight + 0.804 Tar

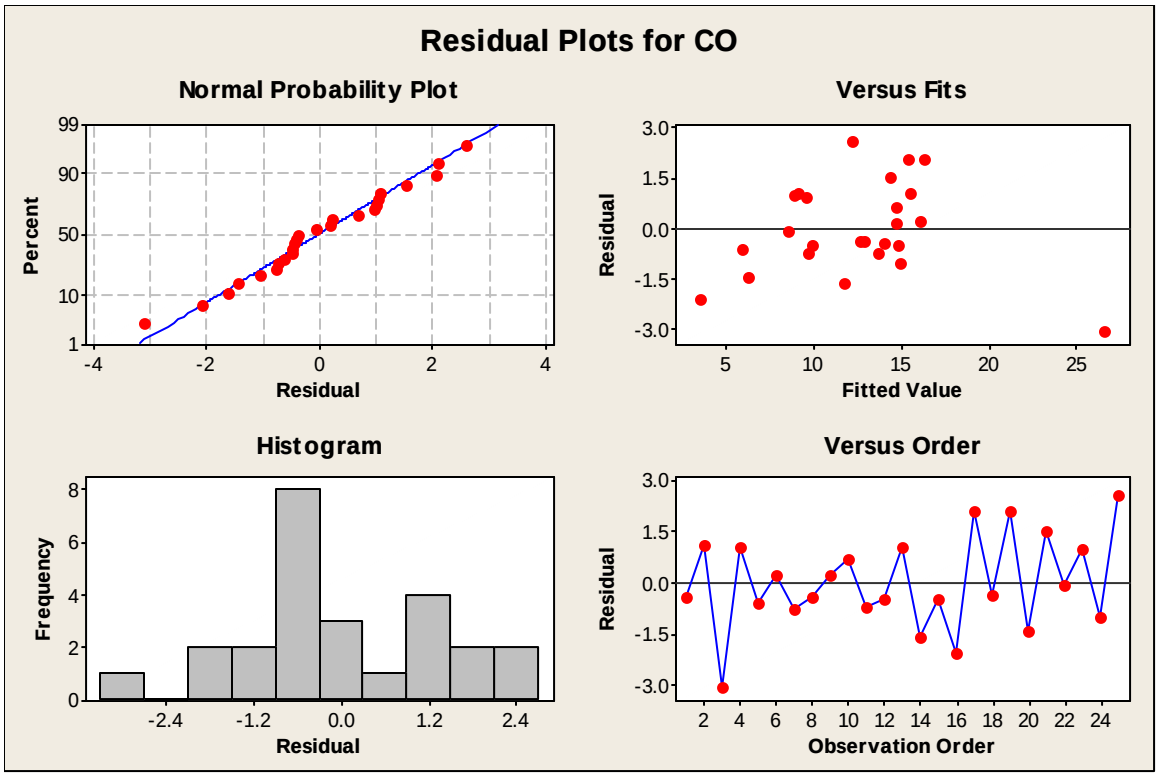
Predictor	Coef	SE Coef	T	P	VIF
Constant	3.114	3.416	0.91	0.372	
Weight	-0.423	3.813	-0.11	0.913	1.317
Tar	0.80419	0.05904	13.62	0.000	1.317

S = 1.42771 R-Sq = 91.7% R-Sq(adj) = 90.9%

Mallows Cp = 2.5

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	494.31	247.15	121.25	0.000
Residual Error	22	44.84	2.04		
Total	24	539.15			



Regression Analysis: CO versus Nicotine, Tar

The regression equation is
CO = 3.09 - 2.65 Nicotine + 0.962 Tar

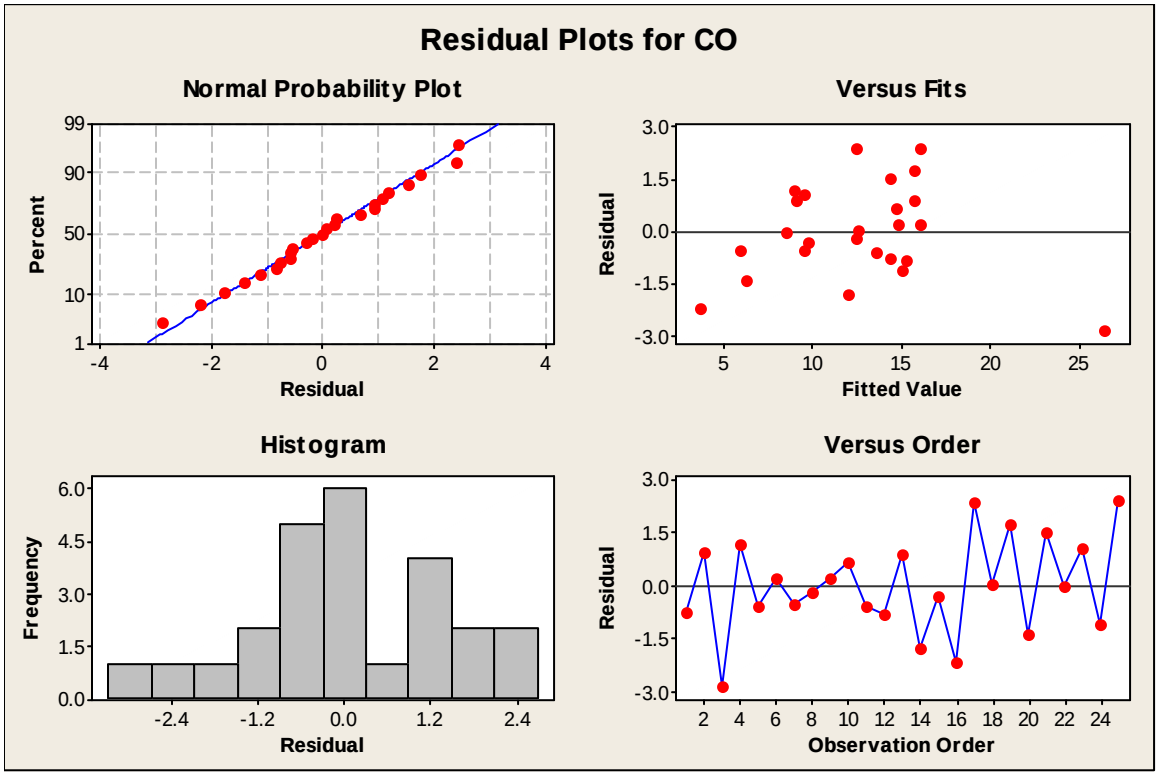
Predictor	Coef	SE Coef	T	P	VIF
Constant	3.0896	0.8438	3.66	0.001	
Nicotine	-2.646	3.787	-0.70	0.492	21.627
Tar	0.9625	0.2367	4.07	0.001	21.627

S = 1.41252 R-Sq = 91.9% R-Sq(adj) = 91.1%

Mallows Cp = 2.0

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	495.26	247.63	124.11	0.000
Residual Error	22	43.89	2.00		
Total	24	539.15			



Regression Analysis: CO versus Nicotine

The regression equation is
CO = 1.66 + 12.4 Nicotine

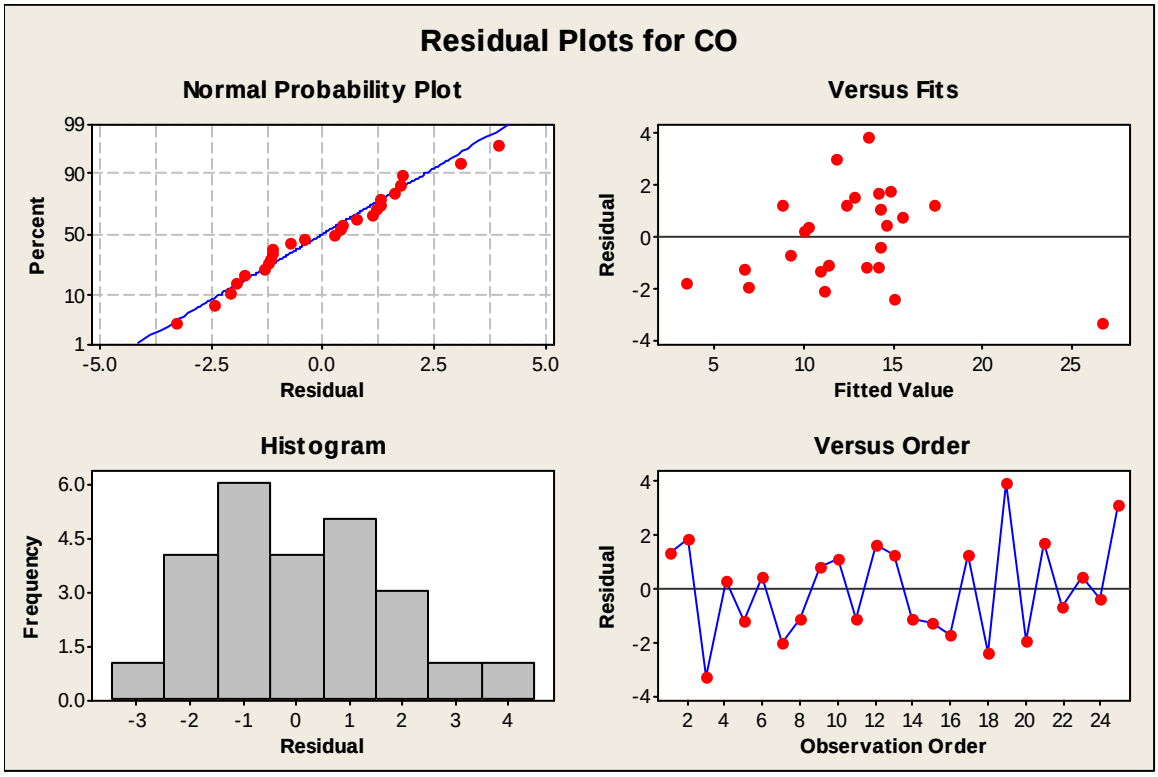
Predictor	Coef	SE Coef	T	P	VIF
Constant	1.6647	0.9936	1.68	0.107	
Nicotine	12.395	1.054	11.76	0.000	1.000

S = 1.82845 R-Sq = 85.7% R-Sq(adj) = 85.1%

Mallows Cp = 15.8

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	462.26	462.26	138.27	0.000
Residual Error	23	76.89	3.34		
Total	24	539.15			



Regression Analysis: CO versus Tar

The regression equation is
CO = 2.74 + 0.801 Tar

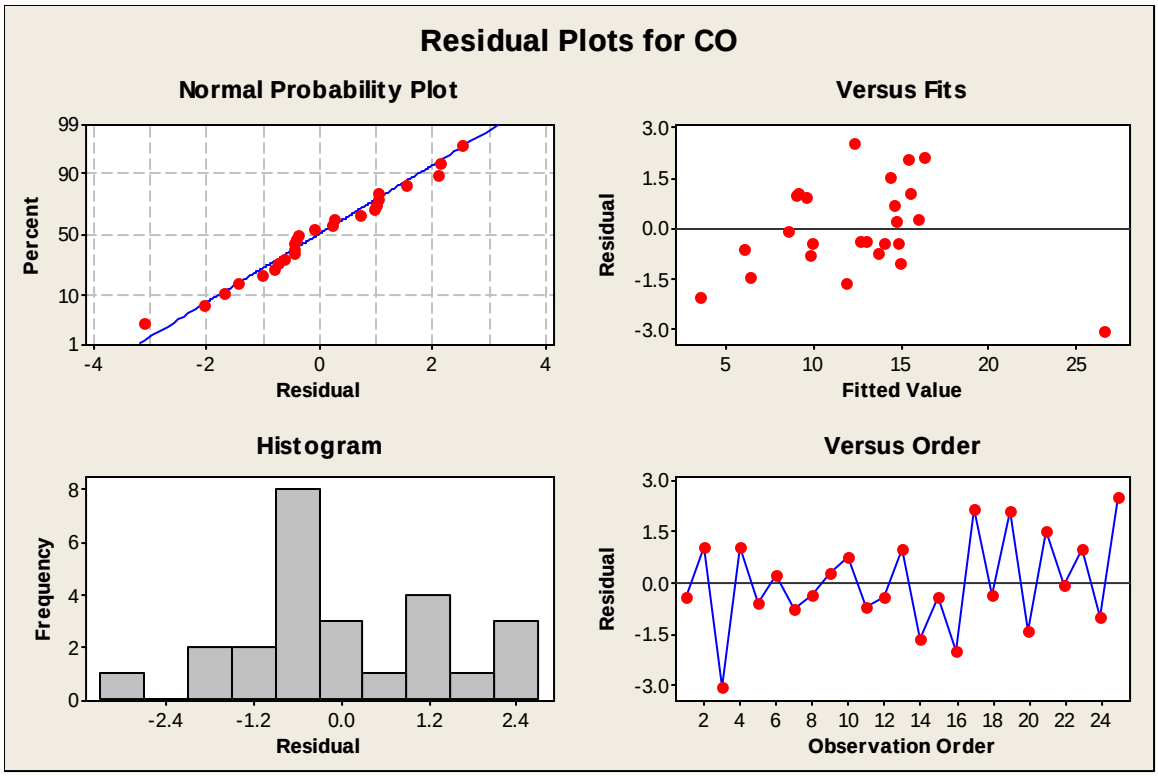
Predictor	Coef	SE Coef	T	P	VIF
Constant	2.7433	0.6752	4.06	0.000	
Tar	0.80098	0.05032	15.92	0.000	1.000

S = 1.39672 R-Sq = 91.7% R-Sq(adj) = 91.3%

Mallows Cp = 0.5

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	494.28	494.28	253.37	0.000
Residual Error	23	44.87	1.95		
Total	24	539.15			



Regression Analysis: CO versus Weight

The regression equation is
CO = - 11.8 + 25.1 Weight

Predictor	Coef	SE Coef	T	P	VIF
Constant	-11.795	9.722	-1.21	0.237	
Weight	25.068	9.980	2.51	0.019	1.000

S = 4.28898 R-Sq = 21.5% R-Sq(adj) = 18.1%

Mallows Cp = 181.4

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	116.06	116.06	6.31	0.019
Residual Error	23	423.09	18.40		
Total	24	539.15