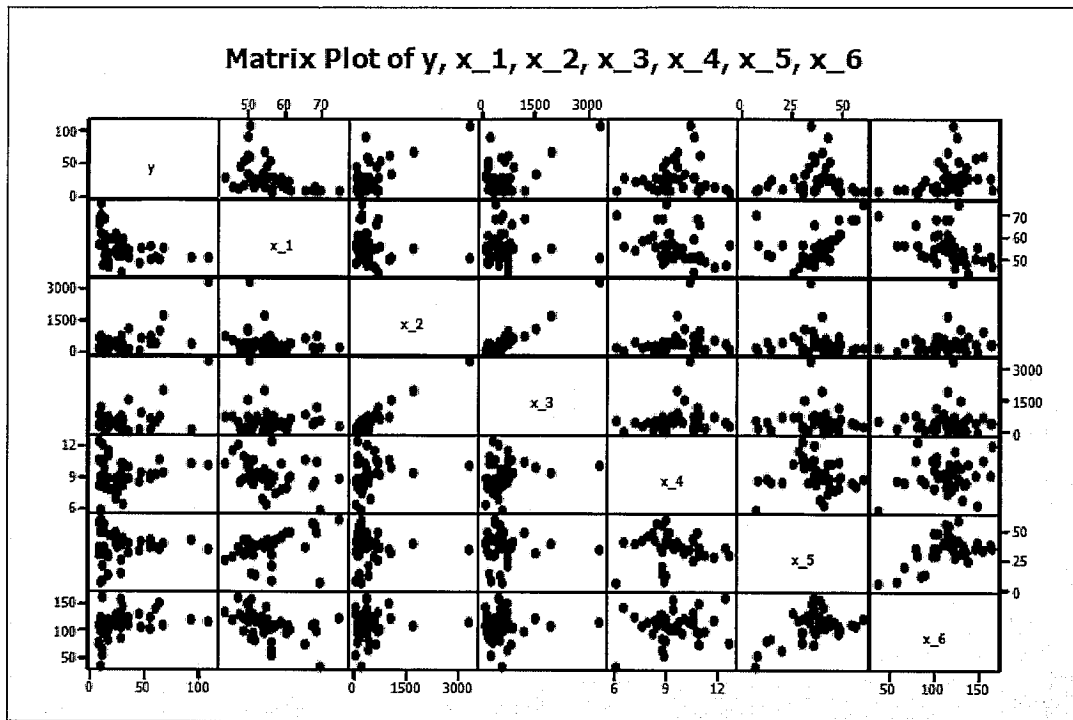


13.76

a.



The plots of y versus the explanatory variables don't indicate that any higher order terms are needed.

b.

Correlations: $x_1, x_2, x_3, x_4, x_5, x_6$

	x_1	x_2	x_3	x_4	x_5
x_2	-0.190				
x_3	-0.063	0.955			
x_4	-0.350	0.238	0.213		
x_5	0.386	-0.032	-0.026	-0.013	
x_6	-0.430	0.132	0.042	0.164	0.496

There is a serious collinearity issue with variables x_2 and x_3 (correlation = 0.955).

c.

Regression Analysis: y versus $x_1, x_2, x_3, x_4, x_5, x_6$

The regression equation is

$$y = 112 - 1.27 x_1 + 0.0649 x_2 - 0.0393 x_3 - 3.18 x_4 + 0.512 x_5 - 0.052 x_6$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	111.73	47.32	2.36	0.024	
x_1	-1.2679	0.6212	-2.04	0.049	3.764
x_2	0.06492	0.01575	4.12	0.000	14.704
x_3	-0.03928	0.01513	-2.60	0.014	14.341
x_4	-3.181	1.815	-1.75	0.089	1.256
x_5	0.5124	0.3628	1.41	0.167	3.405
x_6	-0.0521	0.1620	-0.32	0.750	3.444

S = 14.6360 R-Sq = 67.0% R-Sq(adj) = 61.1%

The VIF's are very large for x_2 and x_3 (larger than 10) which indicates collinearity problems. This is confirmed by the correlation of 0.955 between x_2 and x_3 .

13.77

a. Best Subsets Regression: y versus x₁, x₂, x₃, x₄, x₅, x₆
 Response is y

Vars	R-Sq	R-Sq(adj)	Mallows		x x x x x x						
			Cp	S	1	2	3	4	5	6	
1	41.6	40.1	23.1	18.170		X					
1	24.4	22.4	40.8	20.671			X				
1	18.8	16.7	46.5	21.420	X						
1	13.7	11.4	51.8	22.088						X	
2	58.6	56.5	7.6	15.489		X	X				
2	51.6	49.1	14.8	16.752	X	X					
2	49.8	47.2	16.6	17.060		X				X	
2	42.1	39.1	24.5	18.318		X			X		
3	61.7	58.6	6.4	15.096		X	X			X	
3	61.3	58.1	6.9	15.191	X	X	X				
3	59.3	56.0	8.9	15.569		X	X		X		
3	59.3	56.0	8.9	15.570		X	X	X			
4	64.0	60.0	6.1	14.853	X	X	X		X		
4	63.3	59.2	6.8	14.991	X	X	X	X			
4	62.9	58.8	7.2	15.068	X	X	X			X	
4	62.8	58.7	7.2	15.081		X	X	X		X	
5	66.9	62.1	5.1	14.447	X	X	X	X	X		
5	65.0	60.0	7.0	14.843	X	X	X	X	X		
5	64.0	58.8	8.1	15.063	X	X	X		X	X	
5	62.9	57.6	9.2	15.284		X	X	X	X	X	
6	67.0	61.1	7.0	14.636	X	X	X	X	X	X	

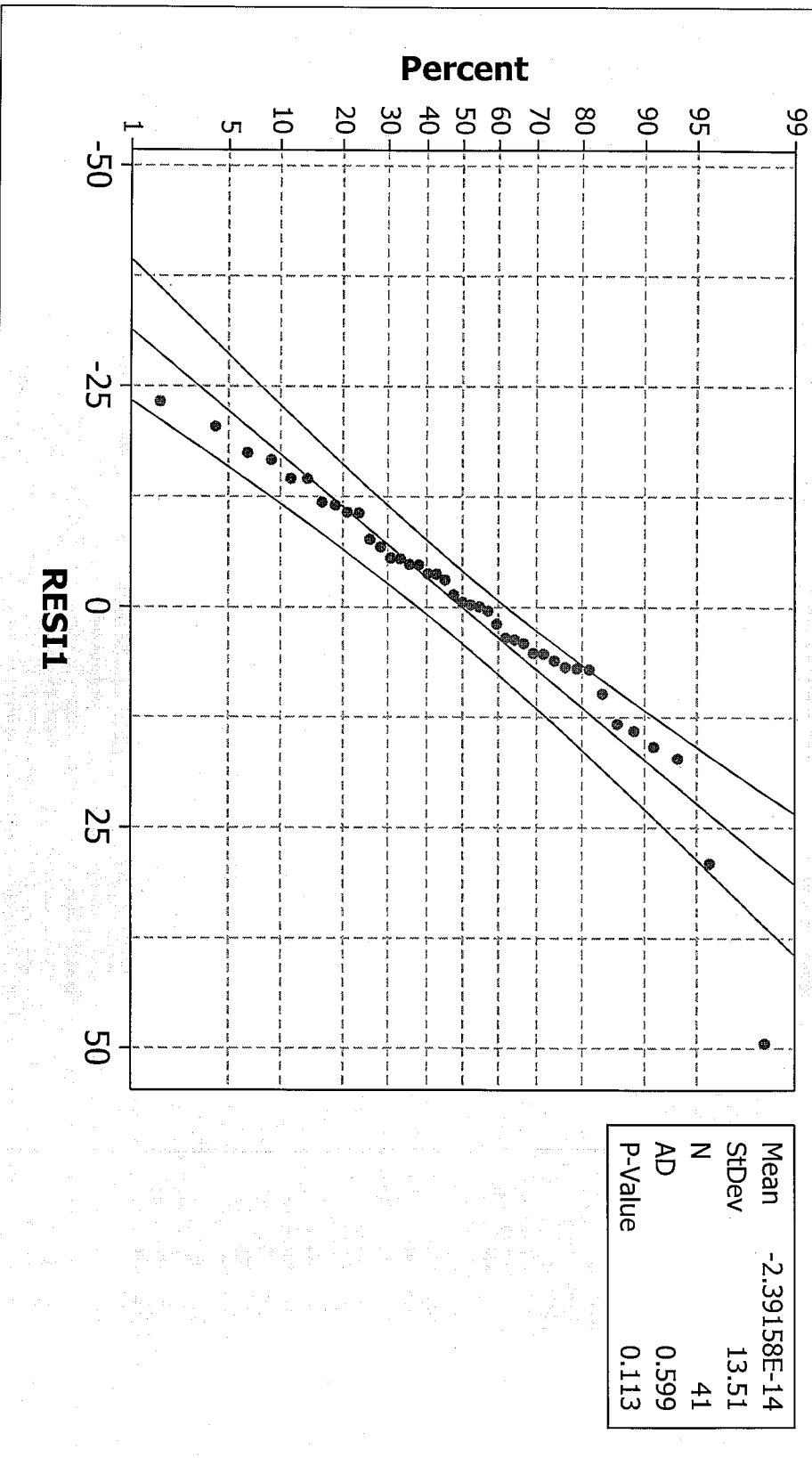
MSE = s²

- b. The model with 5 parameters (all but x₆) with $R^2_{adj} = 62.1\%$ gives the highest R^2_{adj} and a C_p closest to the number of parameters.
- c. The variables chosen in the five variable model are number of manufacturing enterprises employing 20 or more workers, population size, average annual temperature, average annual wind speed, and average annual precipitation.

13.78(a)

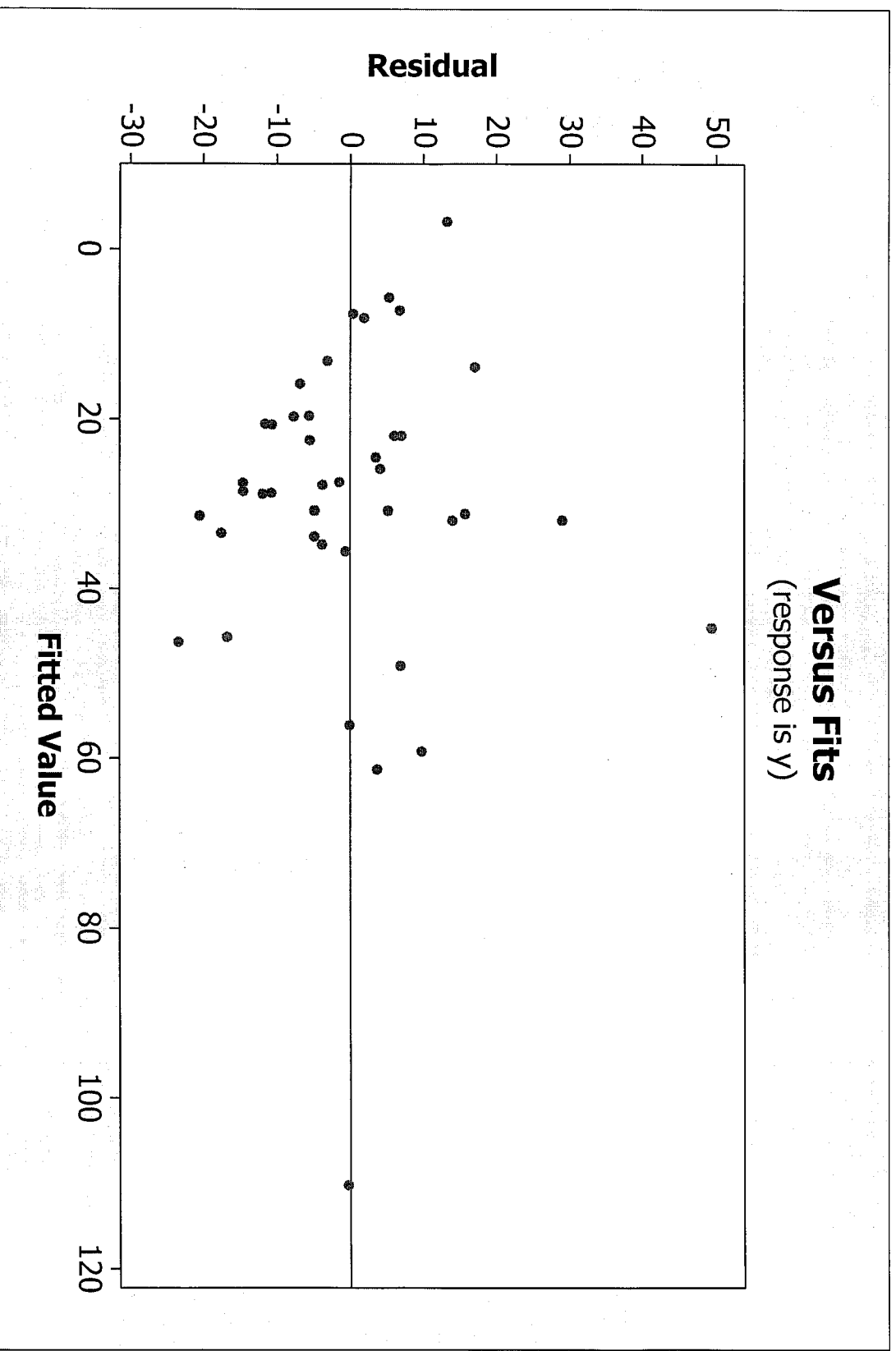
Probability Plot of RES11

Normal - 95% CI



Normality assumption is OK except for one point.

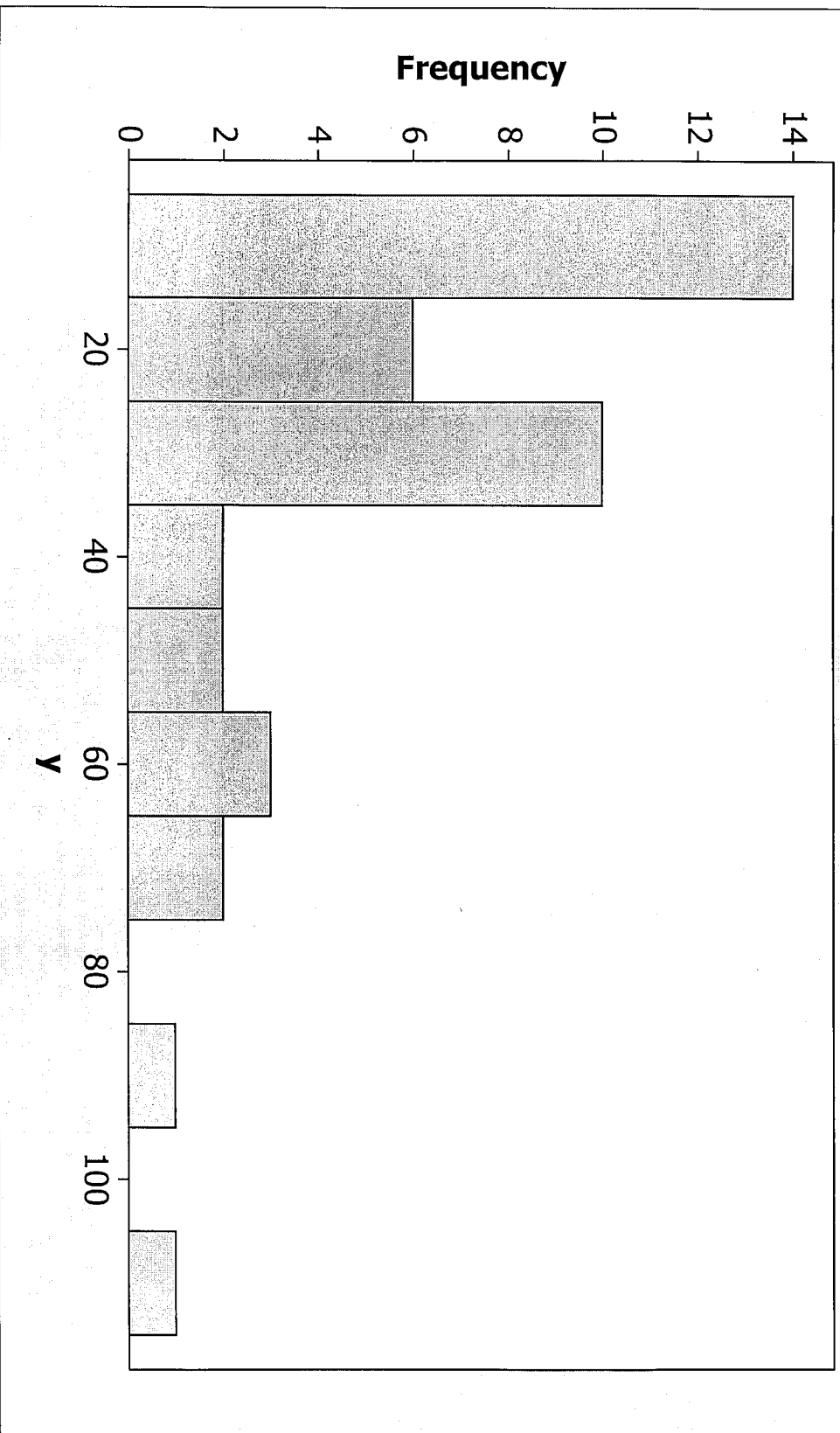
13.78 (b)



points are not randomly distributed around 0, suggesting that the constant variance assumption may be violated.

13.78 (c)

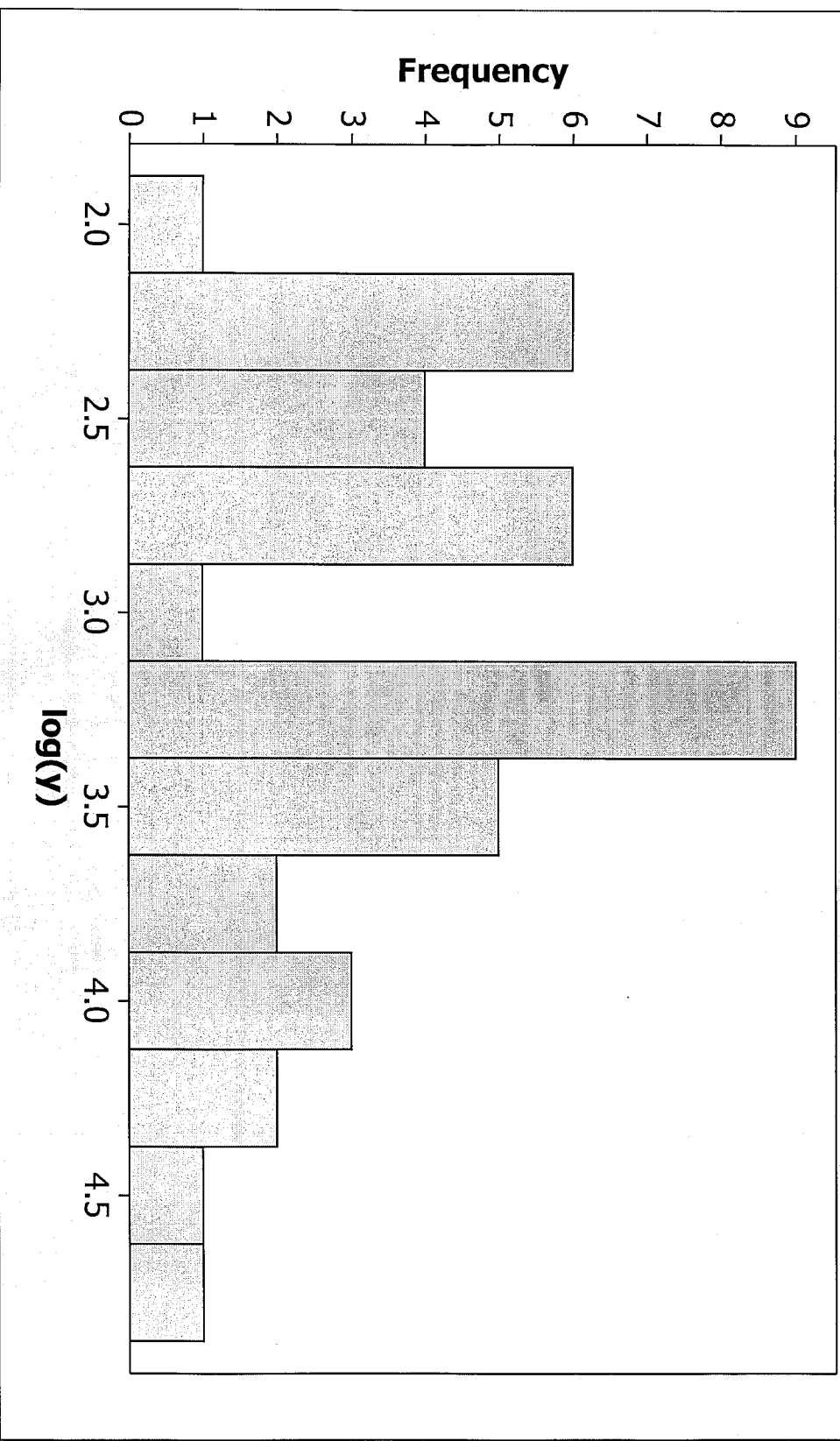
Histogram of y



The dist'n of y is skewed to the right, suggest if a $\log(y)$ (natural logarithm) transformation.

13.78 (c) (optional)

Histogram of $\log(Y)$



Histogram of $\log(Y)$, showing that the skewness is corrected when a \log transformation is used.

13.78 (c)

Stepwise Regression: log(y) versus x_1, x_2, x_3, x_4, x_5, x_6

Backward elimination. Alpha-to-Remove: 0.05

Response is log(y) on 6 predictors, with N = 41

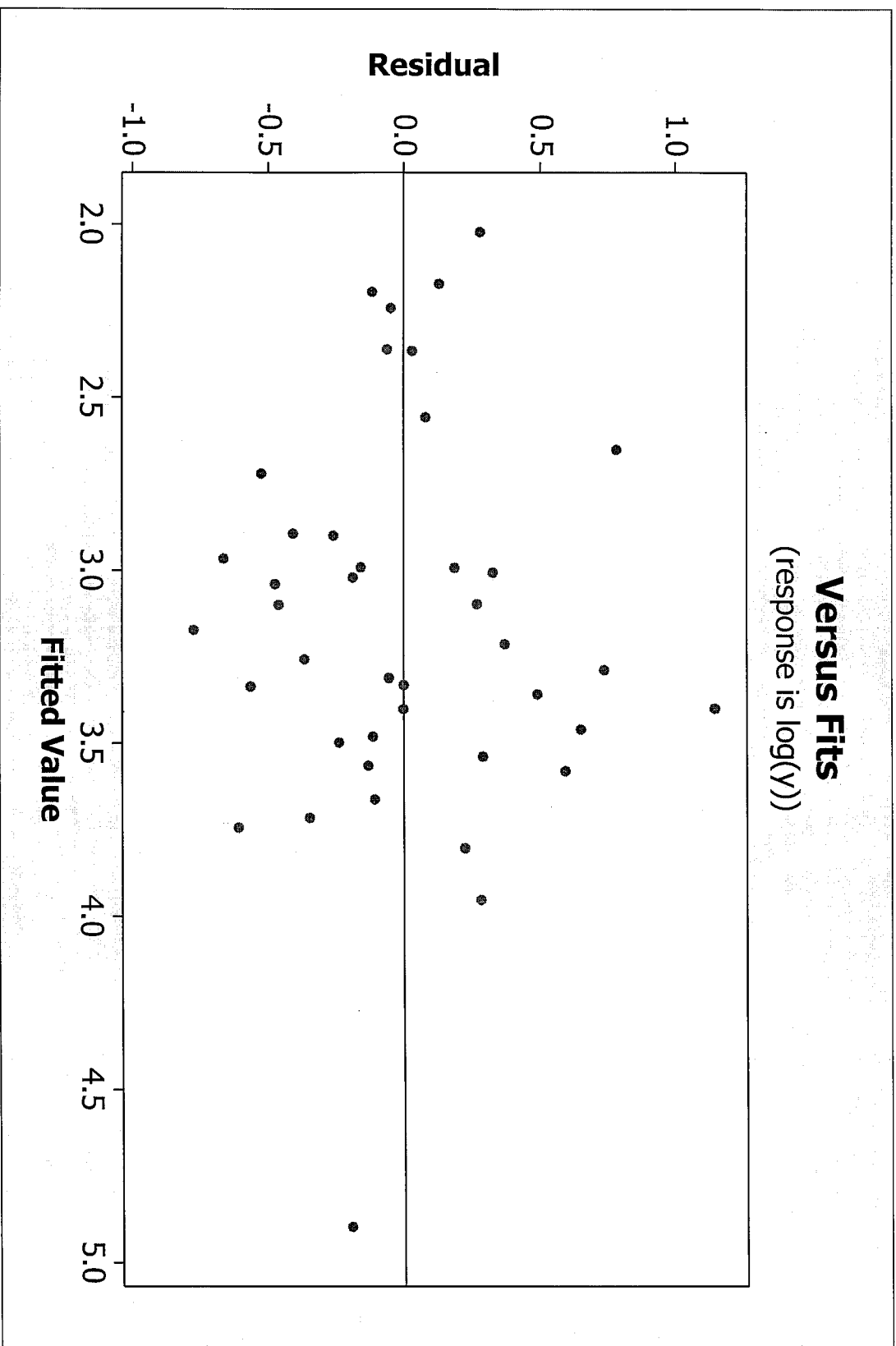
Step	1	2	3
Constant	7.253	7.350	7.763
x_1	-0.060	-0.061	-0.070
T-Value	-3.15	-4.81	-6.04
P-Value	0.003	0.000	0.000
x_2	0.00126	0.00126	0.00056
T-Value	2.62	2.66	4.24
P-Value	0.013	0.012	0.000
x_3	-0.00071	-0.00071	
T-Value	-1.53	-1.55	
P-Value	0.136	0.130	
x_4	-0.170	-0.171	-0.180
T-Value	-3.05	-3.16	-3.30
P-Value	0.004	0.003	0.002
x_5	0.0174	0.0181	0.0200
T-Value	1.56	2.75	3.03
P-Value	0.127	0.009	0.005
x_6	0.0004		
T-Value	0.09		
P-Value	0.931		
S	0.448	0.442	0.450
R-Sq	65.41	65.40	63.03
R-Sq(adj)	59.31	60.46	58.92
Mallows Cp	7.0	5.0	5.3

model selected:

$$\log y = 7.763 - 0.070 x_1 + 0.00056 x_2 - 0.180 x_4 + 0.02 x_5$$

When y is properly transformed, only one of the variables x₂ and x₃ is selected since they are highly correlated (which means these two variables contain largely the same information).

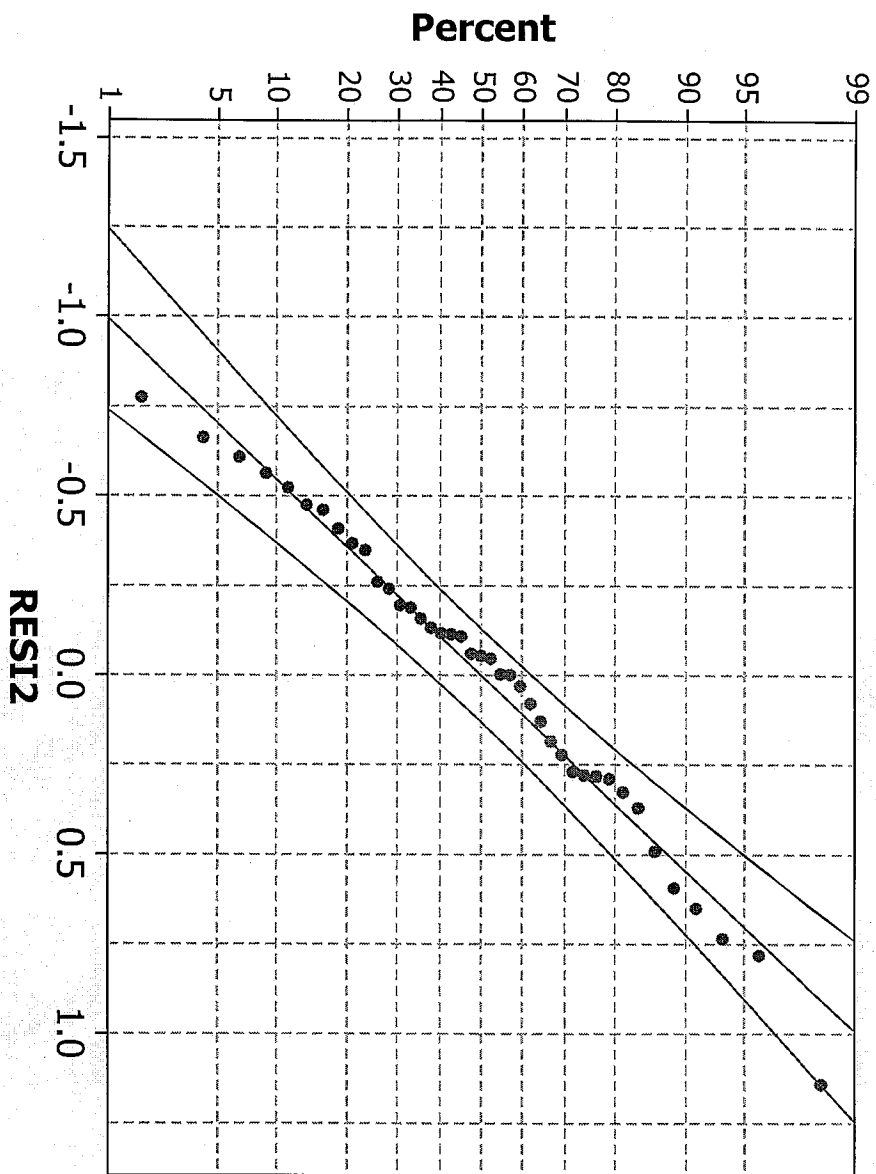
13.78 (e) (optional)
MS of $\log(y)$, the constant variance assumption is satisfied.



13.78 (c) (optional)

Probability Plot of RESI2

Normal - 95% CI

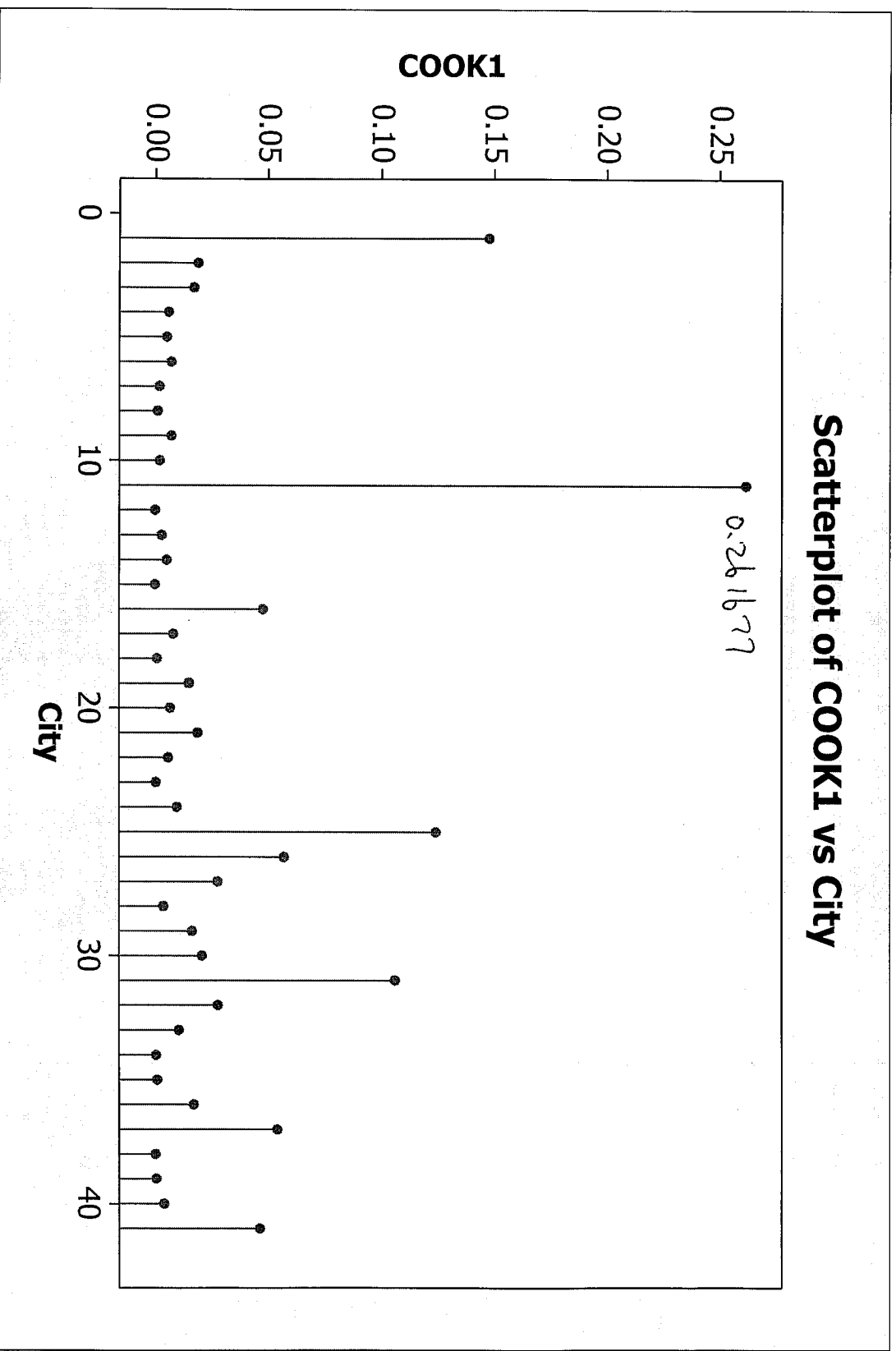


Mean	-2.42624E-15
StdDev	0.42270
N	41
AD	0.249
P-Value	0.732

Using $\log(y)$, the residuals are more normally distributed than before.

13.79 (a)

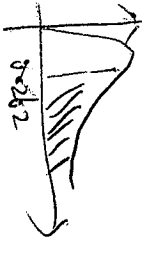
Look for influential observations (high influence on β estimates)



Compare 0.262 to $F(p, n-p)$

or $F(5, 36)$.

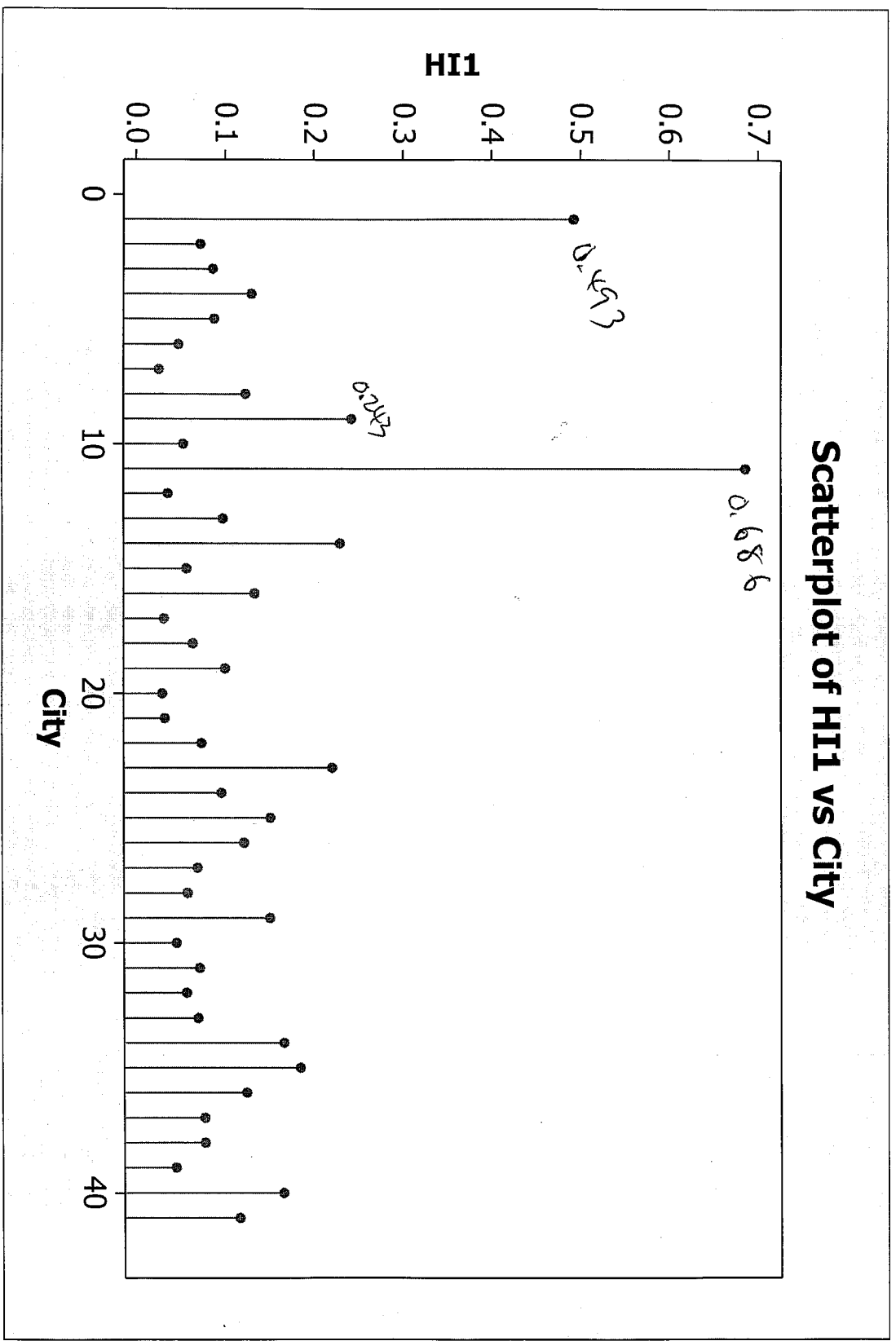
the tail area is $1 - 0.97 = 0.03$



$k = 4$, $p = 5$. $n = 41$.

\Rightarrow not influential

13.79 (a) look for observations w/ high leverage (outlying w.r.t. x values)



Scatterplot of HI1 vs City

> $\frac{2(k+1)}{n} = \frac{2(4+1)}{41} = 0.244$ these two observations (#1 & #11) have high leverage.

Or: > 0.5 high leverage
 < 0.5 moderate leverage
 Observations #1 & #11 have high or moderately high leverage.

13.79

(a) To determine whether observation #11 (with cook's distance of .262) is influential:

Cumulative Distribution Function

F distribution with 5 DF in numerator and 36 DF in denominator

x	P(X <= x)
0.262	0.0691688

Tail probability is $1-0.07=0.93$, which is greater than 0.50. So not influential.

Inverse Cumulative Distribution Function

F distribution with 5 DF in numerator and 36 DF in denominator

P(X <= x)	x
0.5	0.886833

Since 0.262 is less than 0.887, hence the tail area is larger than 0.50. Not influential.

(b)

Model with all data:

Regression Analysis: log(y) versus x_1, x_2, x_4, x_5

The regression equation is

$$\log(y) = 7.76 - 0.0699 x_1 + 0.000556 x_2 - 0.180 x_4 + 0.0200 x_5$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	7.7628	0.9032	8.59	0.000	
x_1	-0.06994	0.01159	-6.04	0.000	1.385
x_2	0.0005556	0.0001310	4.24	0.000	1.075
x_4	-0.18039	0.05462	-3.30	0.002	1.202
x_5	0.020032	0.006622	3.03	0.005	1.200

S = 0.450103 R-Sq = 63.0% R-Sq(adj) = 58.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	12.4356	3.1089	15.35	0.000
Residual Error	36	7.2933	0.2026		
Total	40	19.7289			

Model after observations #1 and #11 are deleted (n=41-2=39):

Regression Analysis: log(y) versus x_1, x_2, x_4, x_5

The regression equation is

$$\log(y) = 7.77 - 0.0752 x_1 + 0.000688 x_2 - 0.172 x_4 + 0.0241 x_5$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	7.7662	0.9123	8.51	0.000	
x_1	-0.07521	0.01343	-5.60	0.000	1.615
x_2	0.0006880	0.0002316	2.97	0.005	1.080
x_4	-0.17192	0.05823	-2.95	0.006	1.127
x_5	0.024135	0.008350	2.89	0.007	1.562

S = 0.454557 R-Sq = 57.7% R-Sq(adj) = 52.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	9.5734	2.3934	11.58	0.000
Residual Error	34	7.0252	0.2066		
Total	38	16.5986			

The beta values do not change much when the two observations with high leverage are deleted. The biggest changes are in beta_2 from 0.00056 to 0.00069 (23% change) and in beta_4 from 0.020 to 0.024 (20% change). Note that the two models would give the same conclusions with respect to the effect of x's on y.

13.80

Regression Analysis: log(y) versus x_1, x_2, x_4, x_5

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	2.6472	0.1021	(2.4400, 2.8543)	(1.7111, 3.5832)

- (a) Estimate of the average level of sulfur dioxide is 2.65.
- (b) The 95% C.I. is (2.44, 2.85).
- (c) The value of x_2=150 falls outside the range of the data used in the model. This is an extrapolation and should be interpreted with caution.