Assignment for Applied Regression Analysis

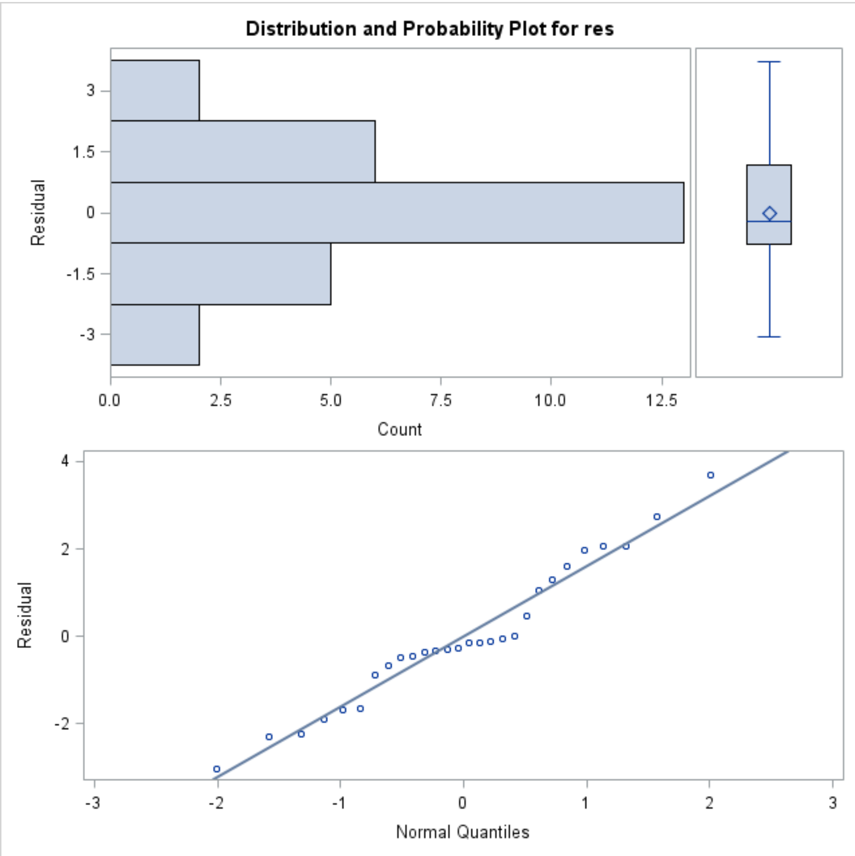
朱强强
17064001
Applied Statistics

4.2

Consider the multiple regression model fit to the National Football League team performance data in Problem 3.1.

a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

Solution



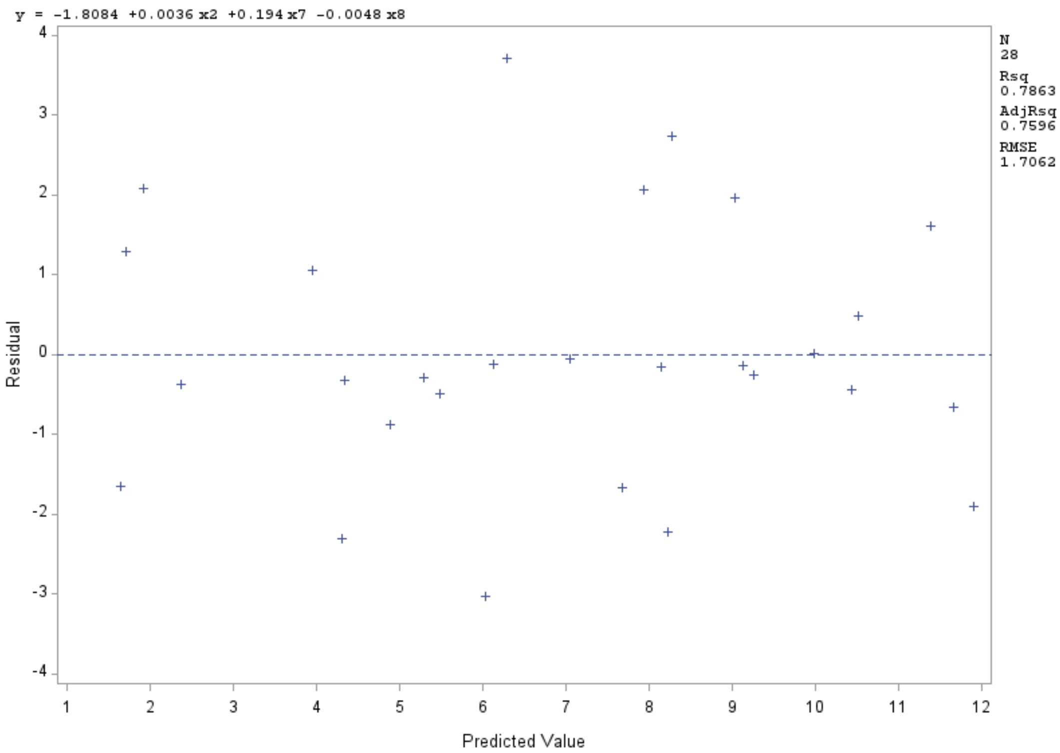
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.965083	Pr < W	0.4566
Kolmogorov-Smirnov	D	0.176242	Pr > D	0.0242
Cramer-von Mises	W-Sq	0.108293	Pr > W-Sq	0.0858
Anderson-Darling	A-Sq	0.515849	Pr > A-Sq	0.1828

From the table above, the result of Shapiro-Wilk test shows that the residuals conform to the normality assumption.

```
proc import datafile='E:\Applied Regression Analysis\SAS\data-table-B1.csv' out = file_data;
    getnames=yes;
run;
proc reg data=file_data;
    model y=x2 x7 x8/r;
    output out=residual_data residual=res student=stu_res press=press_res rstudent=rstu_res h=h;
run;
proc univariate data=residual_data normal plot;
    var res;
run;
```

b. Construct and interpret a plot of the residuals versus the predicted response.

Solution

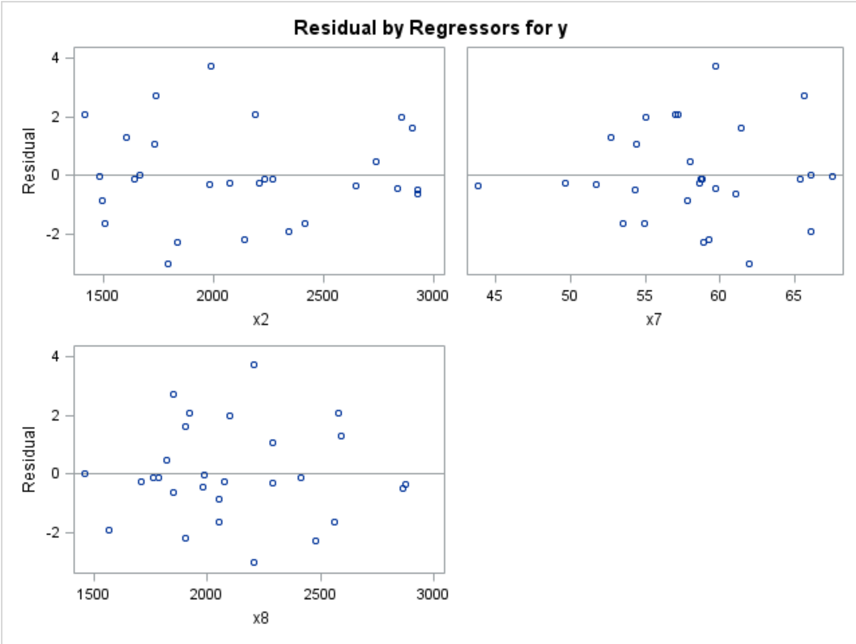


From the plot of the residuals above, we can clearly see that the residuals distribute equably and can be contained in a horizontal band, then there are no obvious model defects.

```
proc reg data=file_data;
    model y=x2 x7 x8/r;
    plot r.*p.;
run;
```

c. Construct plots of the residuals versus each of the regressor variables. Do these plots imply that the regressor is correctly specified?

Solution

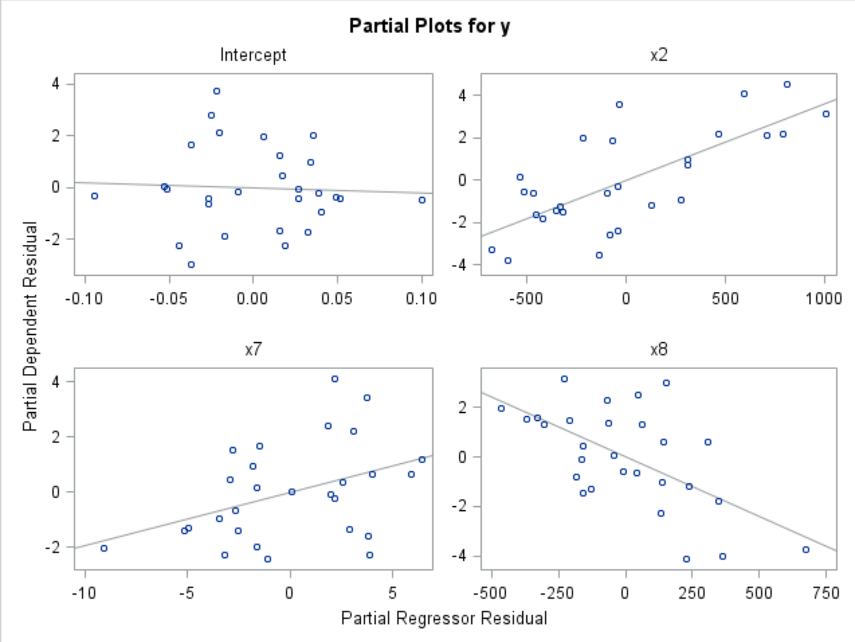


From the first plot and the third plot, we can see that residuals distribute equably and do not show any obvious trend, which indicates that the regressor x_2 and x_8 are correctly specified. However, the patterns in the second plot indicate that the variance of residuals is not constant, and the outward-opening funnel pattern in this figure implies that the variance of residuals increases as x_7 increases. Thus, the regressor x_7 is not correctly specified.

```
proc reg data=residual_data;
    model y=x2 x7 x8;
run;
```

d. Construct the partial regression plots for this model. Compare the plots with the plots of residuals versus regressors from part c above. Discuss the type of information provided by these plots.

Solution




```
proc reg data=residual_data;
    model y=x2 x7 x8/partial;
run;
```

e. Compute the studentized residuals and the R-student residuals for this model. What information is conveyed by these scaled residuals?

Solution

Obs	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	res	stu_res	h	press_res	rstu_res
1	10	2113	1985	38.9	64.7	4	868	59.7	2205	1917	3.70494	2.23185	0.05343	3.91407	2.45435
2	11	2003	2855	38.8	61.3	3	615	55	2096	1575	1.96135	1.22562	0.12033	2.22964	1.23922
3	11	2957	1737	40.1	60	14	914	65.6	1847	2175	2.72895	1.70263	0.11758	3.09259	1.77759
4	13	2285	2905	41.6	45.3	-4	957	61.4	1903	2476	1.61071	1.02977	0.15962	1.91665	1.03112
5	10	2971	1666	39.2	53.8	15	836	66.1	1457	1866	0.00939	0.00612	0.19222	0.01163	0.00600
6	11	2309	2927	39.7	74.1	8	786	61	1848	2339	-0.65572	-0.41888	0.15825	-0.77899	-0.41156
7	10	2528	2341	38.1	65.4	12	754	66.1	1564	2092	-1.90405	-1.20684	0.14497	-2.22689	-1.21899
8	11	2147	2737	37	78.3	-1	761	58	1821	1909	0.47978	0.29933	0.11753	0.54367	0.29357
9	4	1689	1414	42.1	47.6	-3	714	57	2577	2001	2.07450	1.33803	0.17432	2.51246	1.36163
10	2	2566	1838	42.3	54.2	-1	797	58.9	2476	2254	-2.30597	-1.44176	0.12130	-2.62429	-1.47681
11	7	2363	1480	37.3	48	19	984	67.5	1984	2217	-0.05515	-0.03647	0.21456	-0.07021	-0.03570
12	10	2109	2191	39.5	51.9	6	700	57.2	1917	1758	2.06178	1.25109	0.06712	2.21011	1.26675
13	9	2295	2229	37.4	53.6	-5	1037	58.8	1761	2032	-0.13650	-0.08385	0.08972	-0.14996	-0.08210

14	9	1932	2204	35.1	71.4	3	986	58.6	1709	2025	-0.25816	-0.16067	0.11315	-0.29110	-0.15737
15	6	2213	2140	38.8	58.3	6	819	59.2	1901	1686	-2.21969	-1.33537	0.05092	-2.33878	-1.35870
16	5	1722	1730	36.6	52.6	-19	791	54.4	2288	1835	1.05013	0.64499	0.08946	1.15331	0.63695
17	5	1498	2072	35.3	59.3	-5	776	49.6	2072	1914	-0.28955	-0.19694	0.25746	-0.38995	-0.19295
18	5	1873	2929	41.1	55.3	10	789	54.3	2861	2496	-0.48529	-0.36501	0.39284	-0.79927	-0.35832
19	6	2118	2268	38.2	69.6	6	582	58.7	2411	2670	-0.12736	-0.07900	0.10721	-0.14265	-0.07735
20	4	1775	1983	39.3	78.3	7	901	51.7	2289	2202	-0.33168	-0.20646	0.11353	-0.37416	-0.20230
21	3	1904	1792	39.7	38.1	-9	734	61.9	2203	1988	-3.03697	-1.86994	0.09396	-3.35192	-1.98052
22	3	1929	1606	39.7	68.8	-21	627	52.7	2592	2324	1.28993	0.81727	0.14431	1.50747	0.81144
23	4	2080	1492	35.5	68.8	-8	722	57.8	2053	2550	-0.88464	-0.55106	0.11476	-0.99932	-0.54290
24	10	2301	2835	35.3	74.1	2	683	59.7	1979	2110	-0.44172	-0.27654	0.12364	-0.50404	-0.27115
25	6	2040	2416	38.7	50	0	576	54.9	2048	2628	-1.67085	-1.01859	0.07573	-1.80775	-1.01942
26	8	2447	1638	39.9	57.1	-8	848	65.3	1786	1776	-0.15040	-0.09406	0.12174	-0.17124	-0.09209
27	2	1416	2649	37.4	56.3	-22	684	43.8	2876	2524	-0.36901	-0.26213	0.31928	-0.54209	-0.25698
28	0	1503	1503	39.3	47	-9	875	53.5	2560	2241	-1.64873	-1.04875	0.15105	-1.94209	-1.05103

They can identify the potential outliers.

```
proc print data=residual_data;
run;
```

4.20

Myers Montgomery and Anderson - Cook discuss an experiment to determine the influence of five factors:

x_1—acid bath temperature

x_2—cascade acid concentration

x_3—water temperature

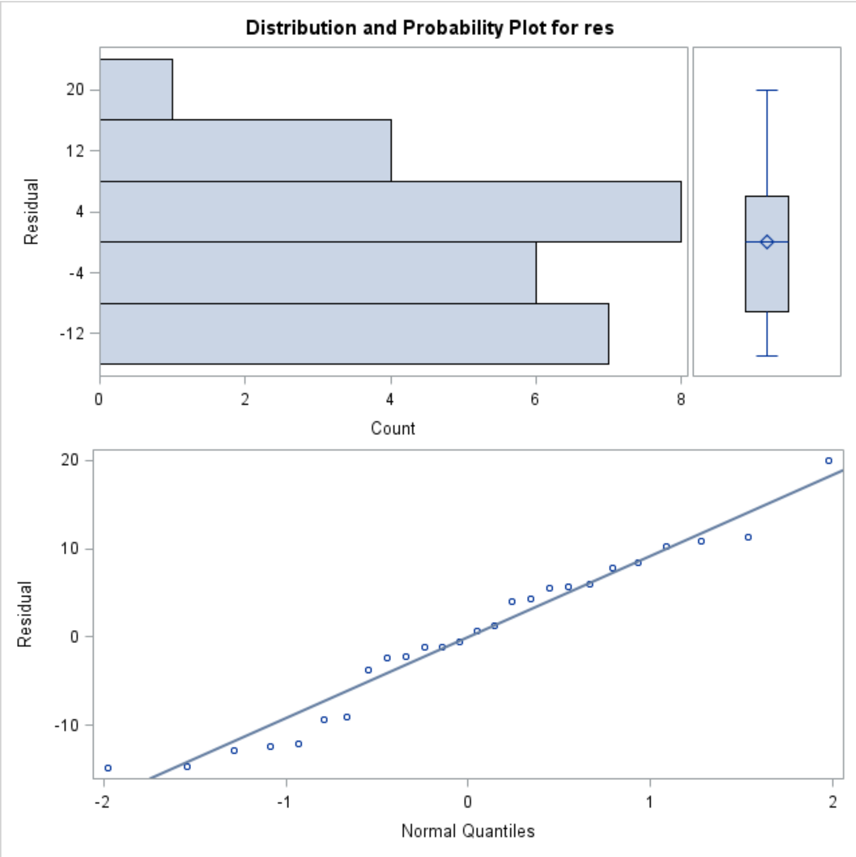
x_4—sulfide concentration

x_5—amount of chlorine bleach

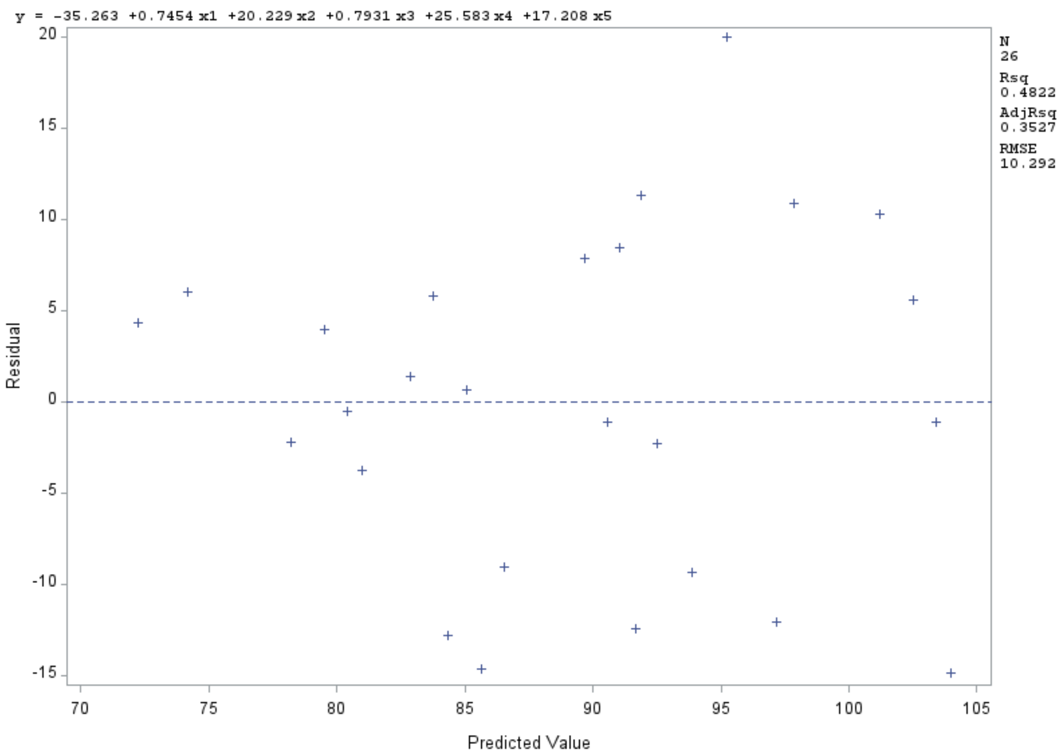
on an appropriate measure of the whiteness of rayon (y). The engineers conducting this experiment wish this minimize this measure.

a. Perform a thorough analysis of the results including residual plots.

Solution



From the normality plot, we can see that the residuals roughly conform to the normal distribution.



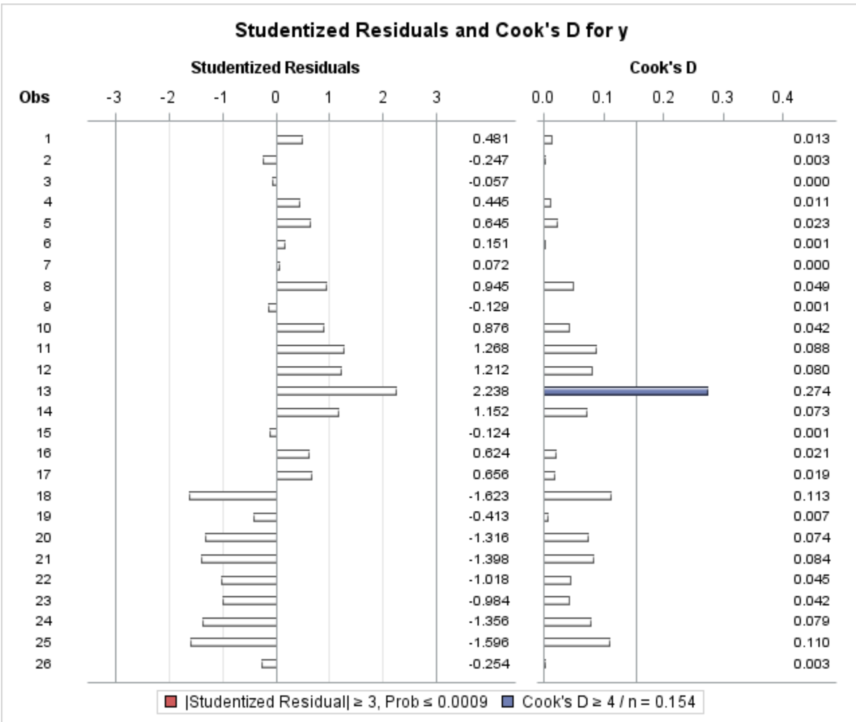
From the residuals plot versus predicted values, however, the residuals does not distribute equably, which indicates that there seems to be some slight problem with variance.

```
proc import datafile='E:\Applied Regression Analysis\Data Sets\data-prob-4-20.xls' out = file_data dbms=XLS replace;
    getnames=yes;
```

```
run;
data file_data;
    set file_data;
    rename acid_temp=x1 acid_conc=x2 water_temp=x3 sulf_conc=x4 amt_bl_=x5;
run;
proc print data=file_data;
run;
proc reg data=file_data;
    model y=x1 x2 x3 x4 x5/r;
    output out=residual_data residual=res student=stu_res press=press_res rstudent=rstu_res h=h;
    plot r.*p.;
run;
proc univariate data=residual_data normal plot;
    var res;
run;
```

b. Perform the appropriate test for lack of fit.

Solution



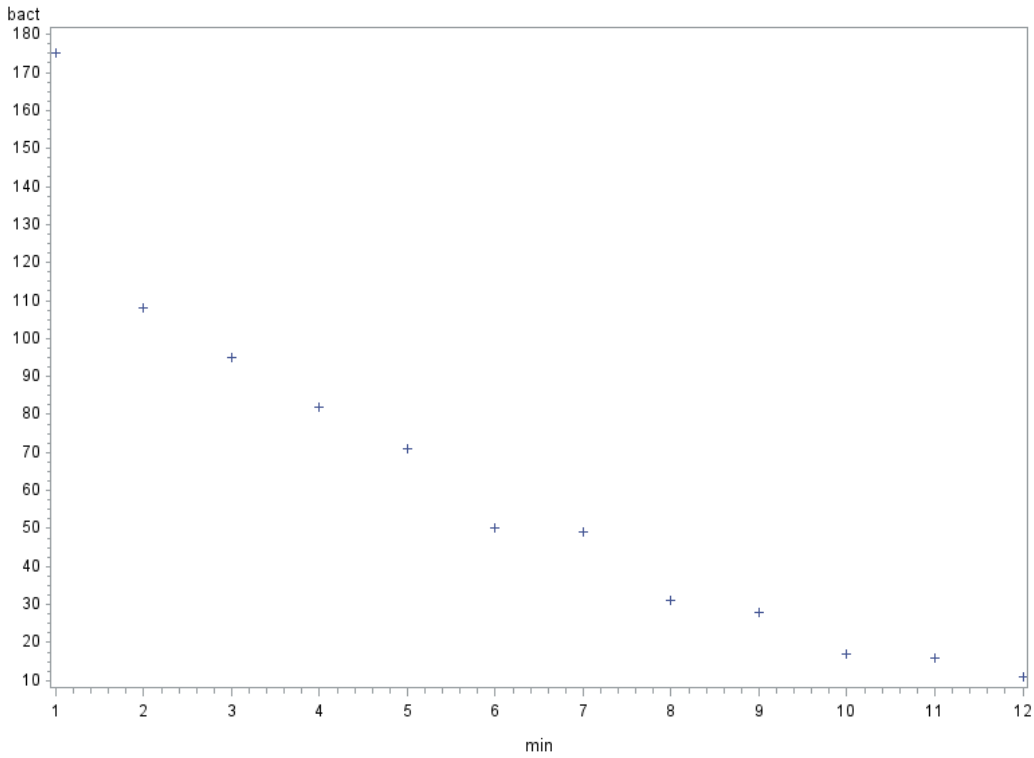
From the figure above, we can know that there is an influential point in the fitted model. Thus, the model can be further modified.

5.3

The data present the average number of surviving bacteria in a canned food product and the minutes of exposure to 300°F heat.

a. Plot a scatter diagram. Does it seem likely that a straight-line model will be adequate?

Solution

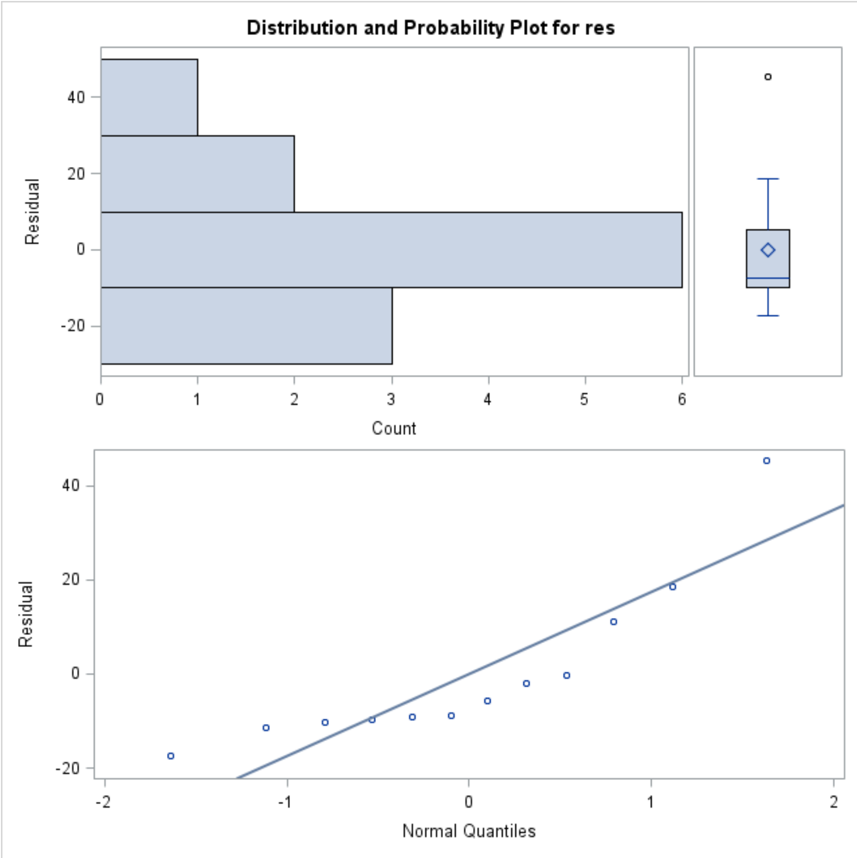


From the figure above we can see that these points seem to display a linear relationship except the first point.

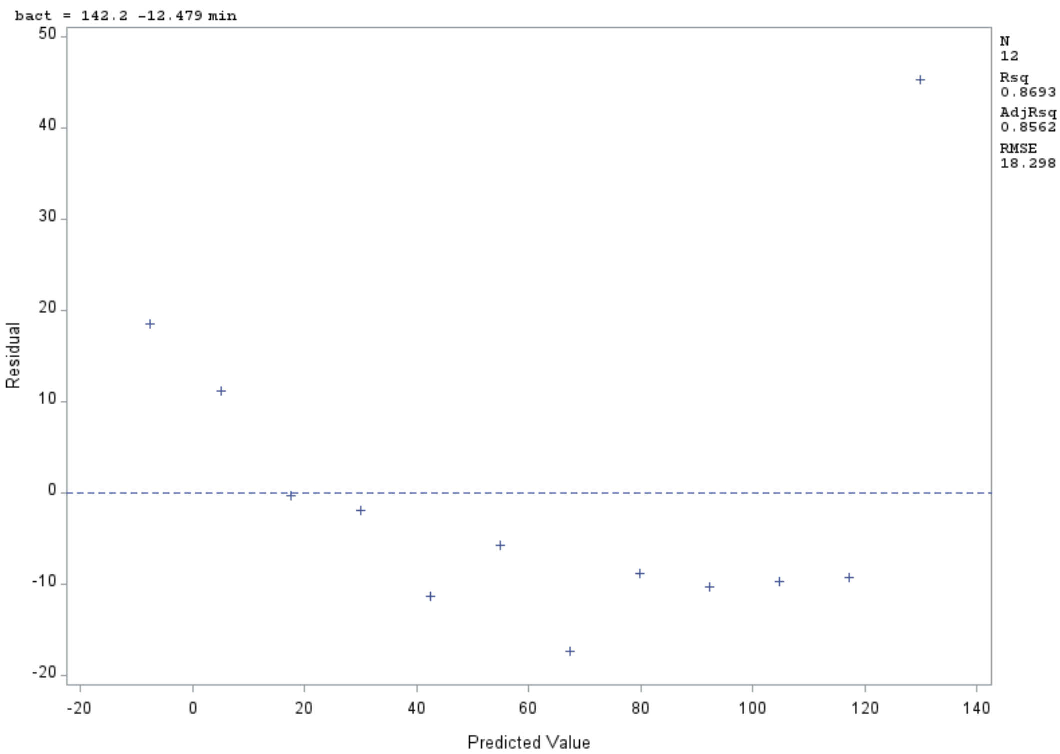
```
proc import datafile='E:\Applied Regression Analysis\Data Sets\data-prob-5-3.xls' out = file_data dbms=XLS replace;
    getnames=yes;
run;
proc gplot data=file_data;
    plot bact*min;
run;
```

b. Fit the straight-line model. Compute the summary statistics and the residual plots. What are your conclusions regarding model adequacy?

Solution



From the normality plot, there seems to be some problems with this model, since the residuals does not fit the normal distribution well.



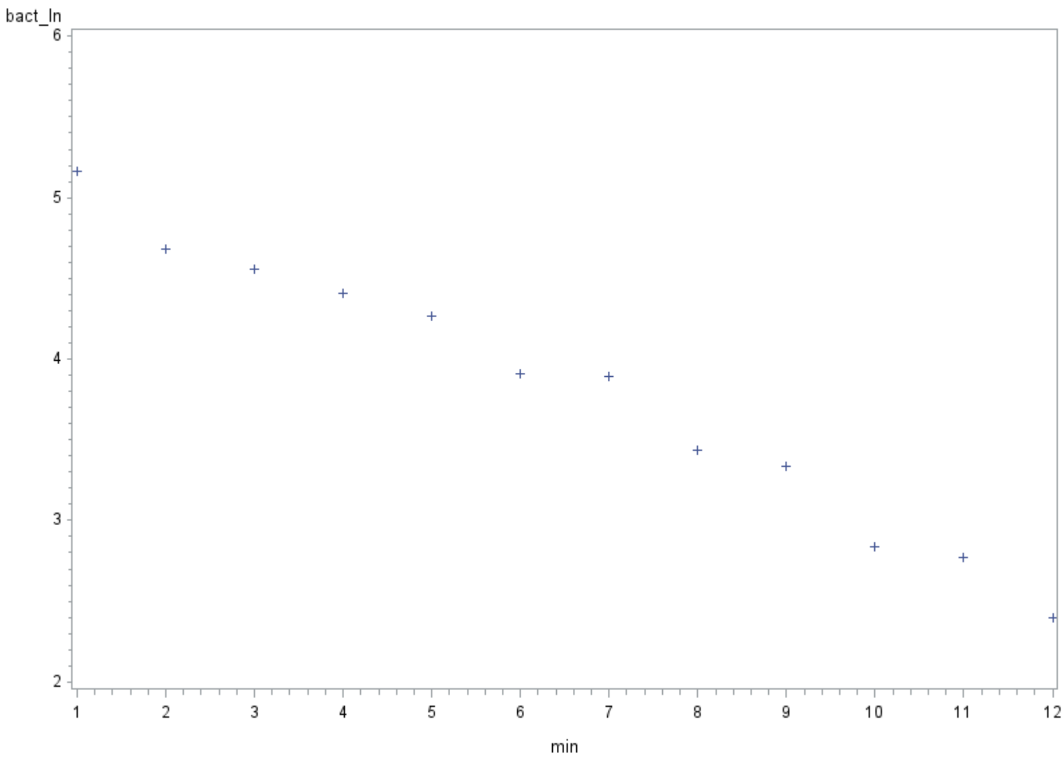
From the scatter plot for residuals, we can clearly see that the residuals distribute unequally, which means residuals does not conform to the previous assumption.

```
proc reg data=file_data;
    model bact=min/r;
    output out=residual_data residual=res student=stu_res press=press_res rstudent=rstu_res h=h;
    plot r.*p.;
run;
proc univariate data=residual_data normal plot;
    var res;
run;
```

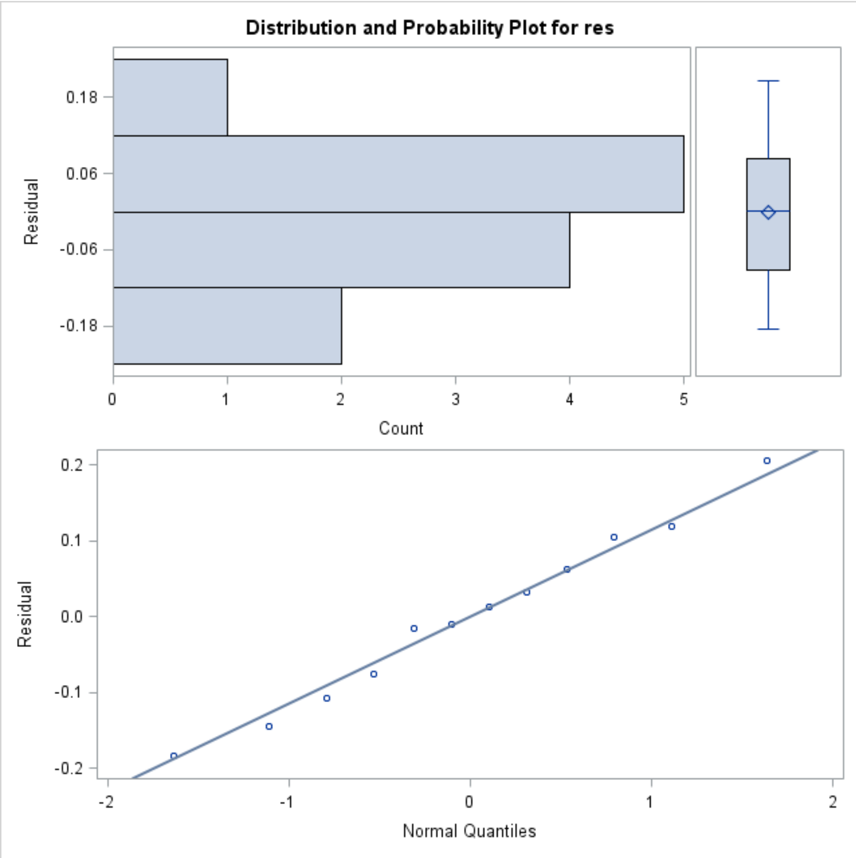
c. Identify an appropriate transformed model for these data. Fit this model to the data and conduct the usual tests of model adequacy.

Solution

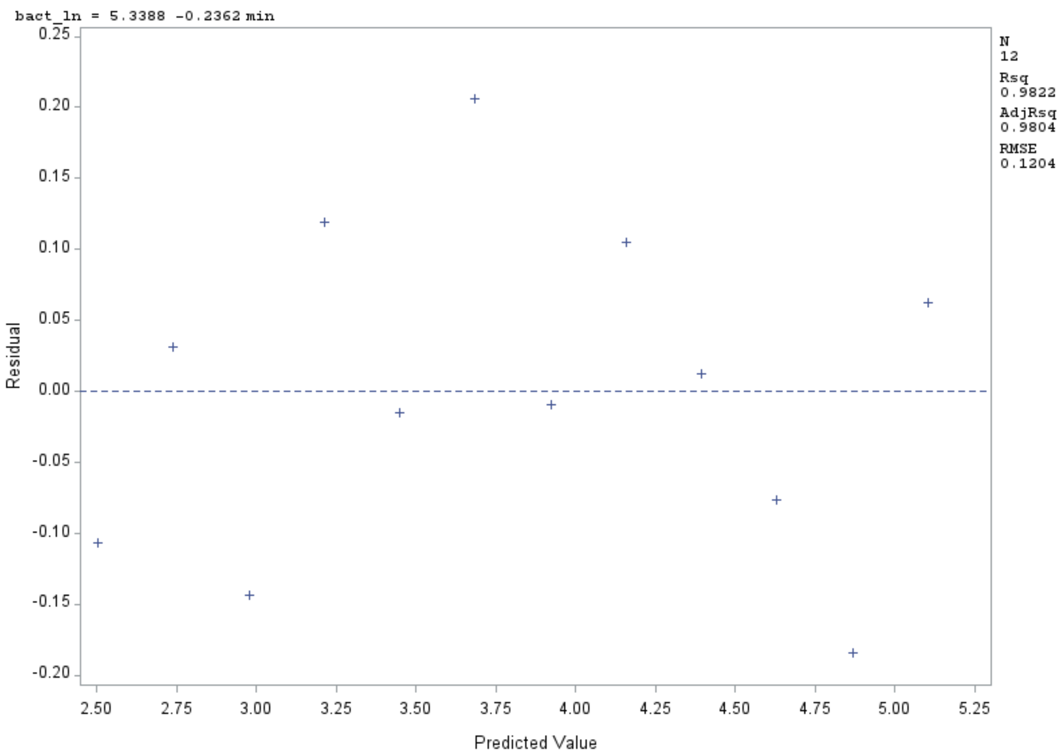
From the first figure in this problem, we can transform the response *bact* to $\ln(bact)$.



After the transformation, we can clearly know that the variables show a significant linear relationship seen from the scatter plot.



In addition, the distribution of residuals adequately fit the normal distribution.



The residuals also distribute adequately.

In conclusion, the model modified after transformation is better.

```
data file_data;
    set file_data;
    bact_ln = log(bact);
```

```
run;
proc print data=file_data;
run;
proc gplot data=file_data;
    plot bact_ln*min;
run;
proc reg data=file_data;
    model bact_ln=min/r;
    output out=residual_data residual=res student=stu_res press=press_res rstudent=rstu_res h=h;
    plot r.*p.;
run;
proc univariate data=residual_data normal plot;
    var res;
run;
```

5.14

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the variance of ε_i is proportional to x_i^2, that is, $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$.

a. Suppose that we use the transformations $y' = y/x$ and $x' = 1/x$. Is this a variance-stabilizing transformation?

Solution

$$\mathrm{Var}(y_i') = \frac{1}{x_i^2} \times \sigma^2 x_i^2 = \sigma^2$$

Thus, this is a variance-stabilizing transformation.

b. What are the relationships between the parameters in the original and transformed models?

Solution

$$x' = \frac{1}{x}$$

c. Suppose we use the method of weighted least squares with $w_i = 1/x_i^2$. Is this equivalent to the transformation introduced in part a?

Solution

Original model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and transformed model after the transformation of the methods of weighted least squares is $y_i' = \beta_1 + \beta_0 x_i + \varepsilon_i'$, where $y_i' = y\sqrt{w_i} = y_i/x_i$ and $\varepsilon_i' = \varepsilon w_i = \varepsilon_i/x_i^2 = \sigma^2$. It is equivalent to the transformation introduced in part a.

5.17

Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$. Assume that $\mathbf{V}$ is known but not σ^2. Show that

$$\left(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right)/(n-p)$$

is an unbiased estimate of σ^2.

Solution

$$\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{y}'\left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\right]\mathbf{y} = \mathbf{y}'\mathbf{A}\mathbf{y}$$

$$E(\mathbf{y}'\mathbf{A}\mathbf{y}) = \sigma^2\,\operatorname{trace}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = (n-p)\sigma^2$$