Supplementary exercise 2.57 of IPS7e

Data: Particulate pollution in a city and a rural location close to the city in a direction so that prevailing winds blow from the rural location to the city. Interest is in predicting city pollution levels from rural levels. Readings are available every 6 days over a 6-month period, although there are many missing values (stated to usually be due to equipment failure).

The data are clearly sampled as two responses, and we will assume that the missing values occur randomly without any relation to the values that could have been obtained. This is an important assumption when the data have so many missing values; for example, it would be a serious problem if missing values occurred predominantly at high pollution levels.

Model: The statistical model of interest is a linear regression model:

$$\text{City}_i = \beta_0 + \beta_1 \text{Rural}_i + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{26}$ are assumed i.i.d. from $\text{N}(0, \sigma)$.

Regression Analysis: City versus Rural

The regression equation is
City = - 2.580 + 1.094 Rural

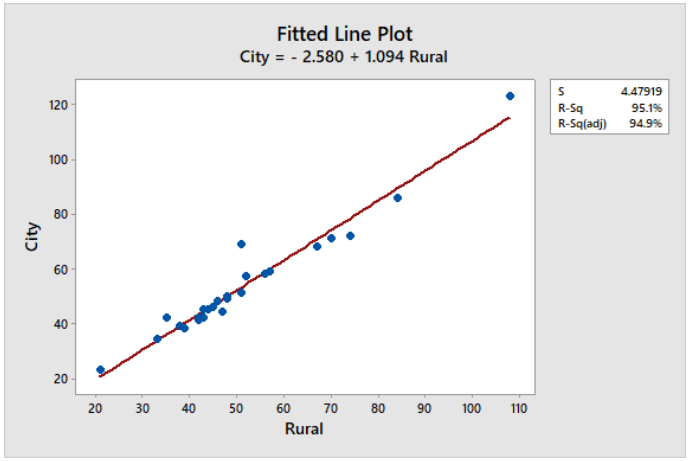
Model Summary

S	R-sq	R-sq(adj)
4.47919	95.11%	94.91%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	9371.10	9371.10	467.08	0.000
Error	24	481.52	20.06		
Total	25	9852.62			

Fitted Line: City versus Rural



(a) The fitted line plot shows a strong positive and apparently quite linear relation between the two pollution levels. There is one point somewhat off the fitted line. Nevertheless, the regression seems to be useful for prediction across the range of rural pollution levels. The largest rural pollution value (108 for day 15, i.e. row no. 15 in the dataset) is substantially larger than all other values, and we should be cautious with predictions for high rural pollution levels, exceeding 90 (say, when determined visually this value is somewhat arbitrary). The estimated regression line is stated above:

$$\text{City} = -2.580 + 1.094 \cdot \text{Rural}$$

The listing also gives $R^2 = 95.1\%$, a very high value. The linear regression accounts for 95% of the variation in city pollution values.

For the following questions, we reestimate the model in the Regression menu with residual plots and subsequently perform the prediction.

```
MTB > Name C3 "SRES" C4 "HI" C5 "COOK".
MTB > Regress;
SUBC>   Response 'City';
SUBC>   Nodefault;
SUBC>   Continuous 'Rural';
```

```
SUBC>   Terms Rural;
SUBC>   Constant;
SUBC>   Unstandardized;
SUBC>   Rtype 2;
SUBC>  GFOURPACK;
SUBC>   Gvariable 'Rural';
SUBC>   Tmethod;
SUBC>   Tanova;
SUBC>   Tsummary;
SUBC>   Tcoefficients;
SUBC>   Tequation;
SUBC>   TDiagnostics 0;
SUBC>   Sresiduals 'SRES';
SUBC>   Hi 'HI';
SUBC>   Cookd 'COOK'.
MTB > Predict 'City';
SUBC>   Nodefault;
SUBC>   KPredictors 88;
SUBC>   TEquation;
SUBC>   TPrediction.
```

Regression Analysis: City versus Rural

Method

Rows unused 10

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	9371.1	9371.10	467.08	0.000
Rural	1	9371.1	9371.10	467.08	0.000
Error	24	481.5	20.06		
Lack-of-Fit	19	312.5	16.45	0.49	0.884
Pure Error	5	169.0	33.80		
Total	25	9852.6			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.47919	95.11%	94.91%	93.25%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2.58	2.73	-0.95	0.354	
Rural	1.0935	0.0506	21.61	0.000	1.00

Regression Equation

City = -2.58 + 1.0935 Rural

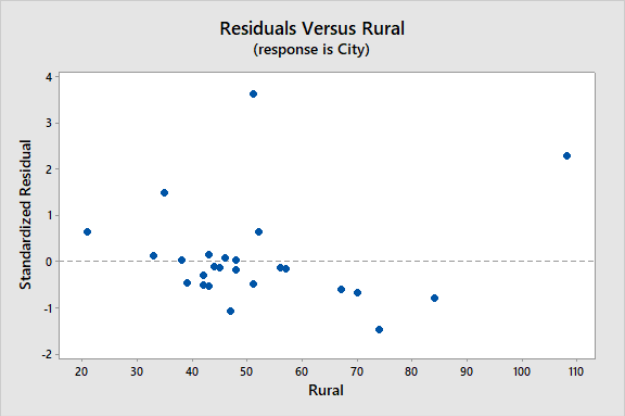
Fits and Diagnostics for Unusual Observations

Obs	City	Fit	Resid	Std Resid	
15	123.00	115.52	7.48	2.26	R X
26	69.00	53.19	15.81	3.60	R

R Large residual
X Unusual X

Residual Plots for City

Residuals from City vs Rural



Prediction for City

Regression Equation

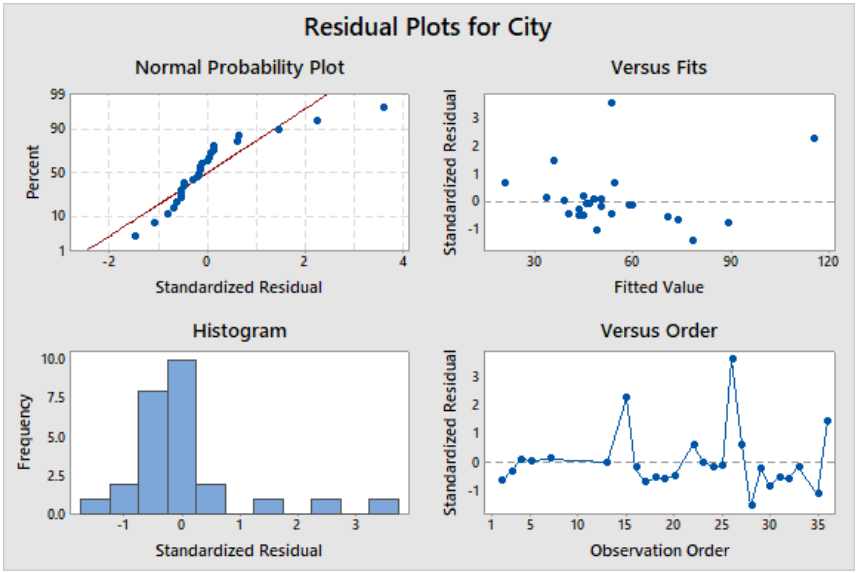
City = -2.58 + 1.0935 Rural

Settings

Variable	Setting
Rural	88

Prediction

Fit	SE Fit	95% CI	95% PI
93.6485	2.06618	(89.3841, 97.9128)	(83.4677, 103.829)



(b) The above analysis included plots of the standardized residuals (preferable over the raw residuals) against the fitted values, the observed rural pollution values and the order of observations.

The first two of these plots (which only differ in the scale of the x-axes) show that most residuals are around and below zero, with a few scattered points further away from zero. The two largest standardized residuals are listed in the table of unusual observations: 2.26 for day 15, and 3.60 for day 26. Among the remaining points there seems to a negative association between residuals and fitted values; the points appear to be essentially situated around a line with negative slope. This would represent a serious concern for the validity of model assumptions because there should be no obvious patterns in the plot. In this case there is a quite simple explanation: the extreme point to the right (day 15) is so influential that the entire regression line is dragged upwards to adapt to this point. The net result is that the line does not fit so well with the other points. Try yourself to refit the model without the day 15 observation, and compare the results! The plot of residuals against days also suggests that there may be increasing variation over time. That would also be a violation of model assumptions because the model assumes the same variance (homogeneity) for all observations.

(c) It is not straightforward to visually assess the most influential observation. The two obvious candidates are the points already discussed above with extreme residuals, although a very influential point does not necessarily have an extreme residual. The table of influential observations indicated by "X" that day 15 is potentially very influential; this is because the x-value is substantially larger than any other observation. Further diagnostics (beyond the scope of the course) are needed to determine exactly how influential each of these observations are.

The Minitab analysis stored values of leverage (`HI`) and Cook's D (`COOK`). The leverage measures potential influence of an observation, and the indicated X is based on leverage. So we know the leverage is largest for day 15. The Cook's D statistic measures actual influence, and it is seen that the value for this statistic is by far largest for day 15. We have therefore established that the most extreme residual was not for the most influential observation. The explanation of this is that the value for day 26 is central among the x-values, so that the poor fit for this observation does not strongly affect the regression line.

(d) The predicted value for $x = 88$ is $\hat{y} = 93.65$. This can either be seen in the Minitab listing or be computed manually from the estimated regression line. The Minitab listing gives a 95% prediction interval of $(83.5, 103.8)$.

(e) The residual plots included a normal plot and a histogram for the standardized residuals. The normal plot is far from a straight line, and the deviations from the line appear to be associated with more than the two points discussed above. The histogram shows that the distribution of the standardized residuals is right-skewed. Overall, the residual plots show that the model assumptions are not really satisfied for this analysis. This is mostly due to the influence of the two extreme points discussed. They affect the model fit quite strongly because all other points are much closer to the line.