EMSE 4765: DATA ANALYSIS

For Engineers and Scientists

Session 11: Multiple Regression, Residual Diagnostics, Outlier Detection

Version: 3/30/2021

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Lecture Notes by: J. René van Dorp[1]

www.seas.gwu.edu/~dorpjr

[1] Department of Engineering Management and Systems Egineering, School of Engineering and Applied Science, The George Washington University, 800 22nd Street, N.W., Suite 2800, Washington D.C. 20052. E-mail: dorpjr@gwu.edu.

ID	NAME	DESCRIPTION
Y	PRICE	Sale price in $000 per acre
X_1	COUNTY	Santa Mateo = 0, Santa Clara = 1
X_2	SIZE	Size of the property in Acres
X_3	ELEVATION	Average elevation in feet above sea level
X_4	SEWER	Distance (in feet) to nearest sewer connection
X_5	DATE	Date of sale counting backward from current time (in months)
X_6	FLOOD	Subject to flooding by tidal action = 1; otherwise = 0
X_7	DISTANCE	Distance in miles from Leslie property (in almost all case, this toward San Francisco

Leslie Salt Property (1968): 246.8 Acres, right on San Francisco Bay, Exactly at sea level (but diked in).

Mountan View, CA, began legal proceedings to acquire this land.

- **It was upto the courts to determine a fair market value**. **Hence, it was decided to build a regression model.**

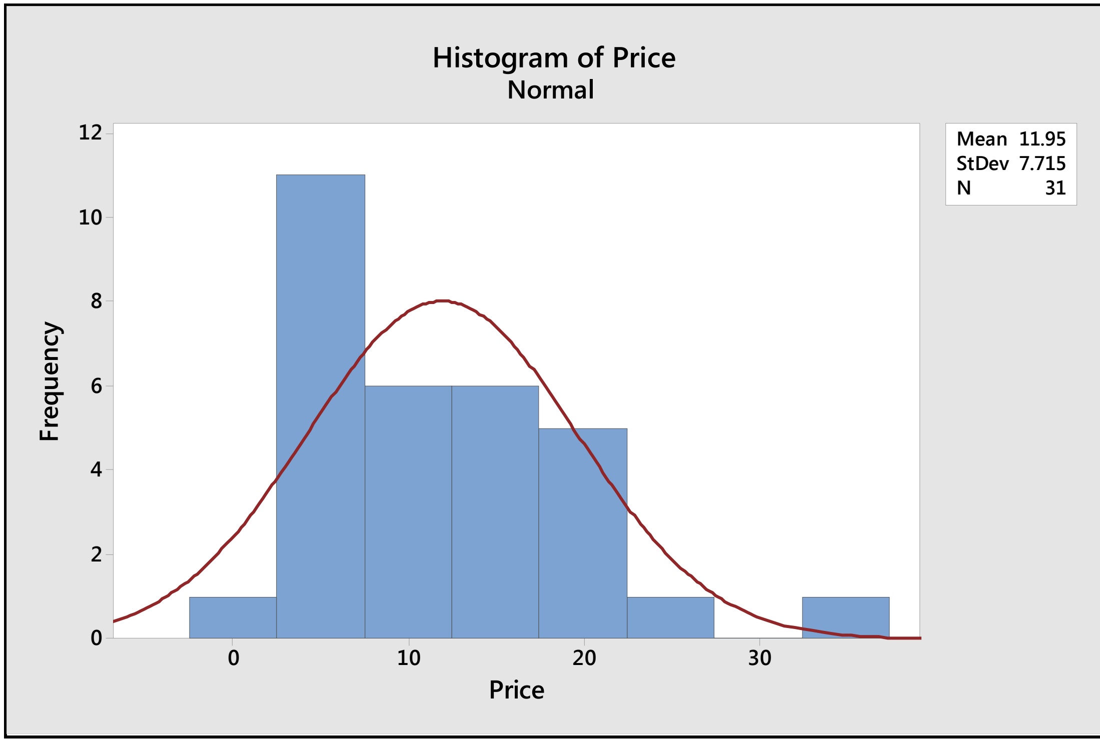
- Data collection: 31 bayland properties, that were sold during the previous 10 years.

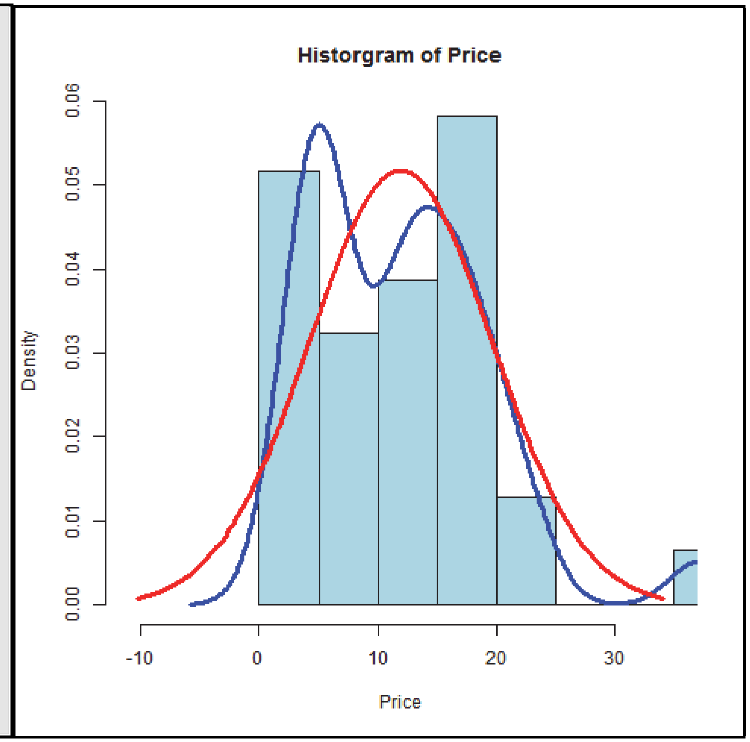
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Data	Price	County	Size	Elevation	Sewer	Date	Flood	Distance
3	**1.70**	0	16.10	0	2640	-98	1	10.30
4	**5.00**	0	**1695.20**	1	3500	-93	0	14.00
6	**3.30**	1	**6.90**	2	10000	-86	0	0.00
26	**37.20**	0	**15.00**	5	0	-39	0	7.20

- **Observe great variability in sizes** of these properties. One could expect that variability increases with the size of the property. Hence it was decided to focus on **the price per acre Y.**

FIRST STEP: PLOT A HISTOGRAM OF THE DEPENDENT VARIABE.

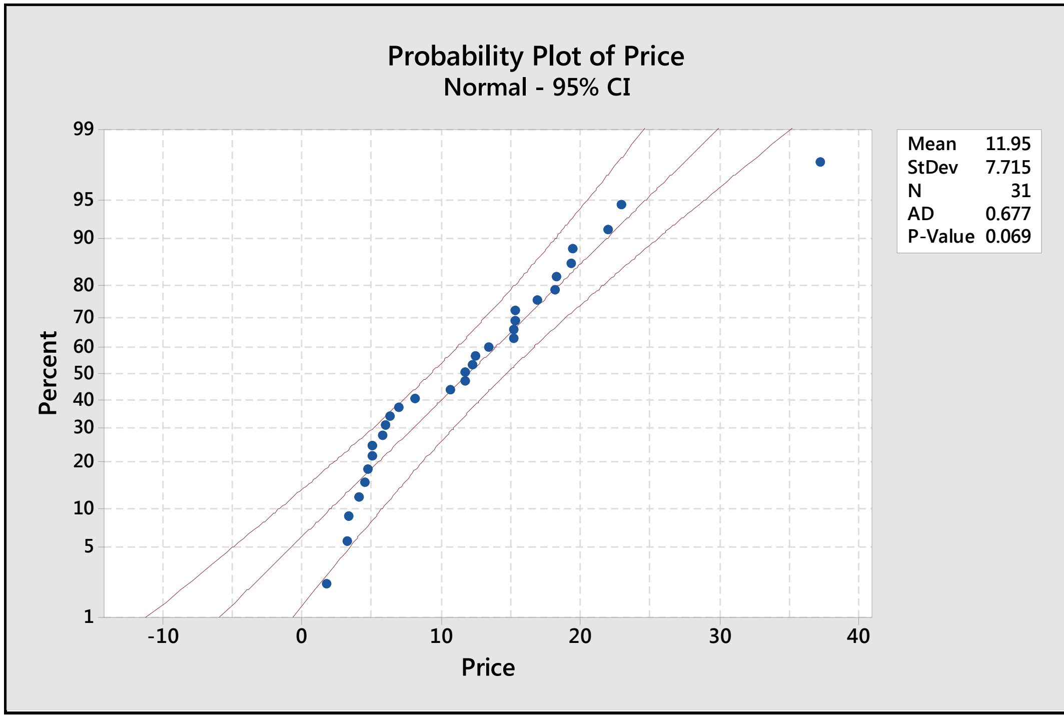
MINITAB AND R-HISTOGRAMS WITH NORMAL DISTRIBUTION FIT

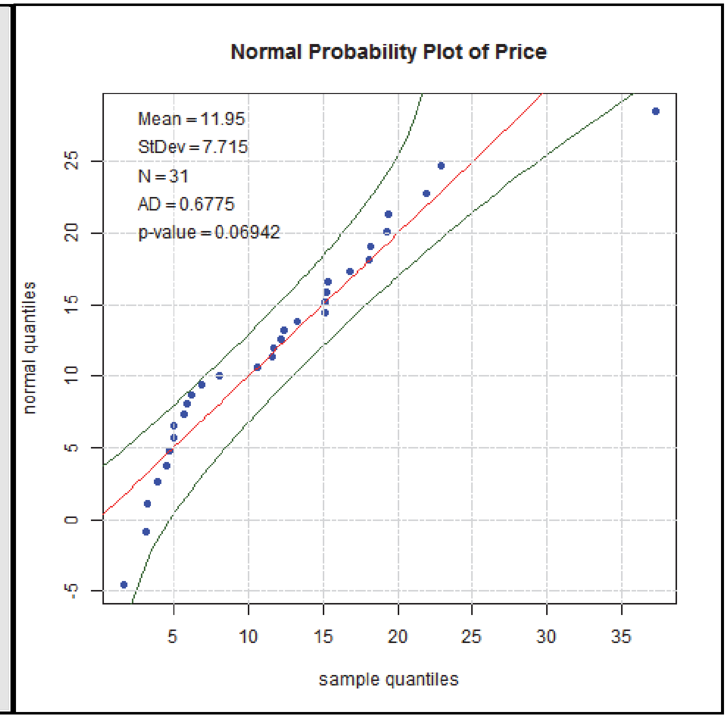




Observation: Minitab histogram appears skewed towards the left which could be problematic in the regression analysis. **A histogram of the dependent variable that is bell-shaped is desirable, but not a requirement! Normal distribution fit allows for negative price values which is problematic too!**

- Recall, the dependent variable Y is a **linear sum of multiple explanatory variables** $x_1, \ldots, x_p$ and an error term. Hence, a "bell-shaped" histrogram for the dependent variable is **a good starting point for regression analysis.**

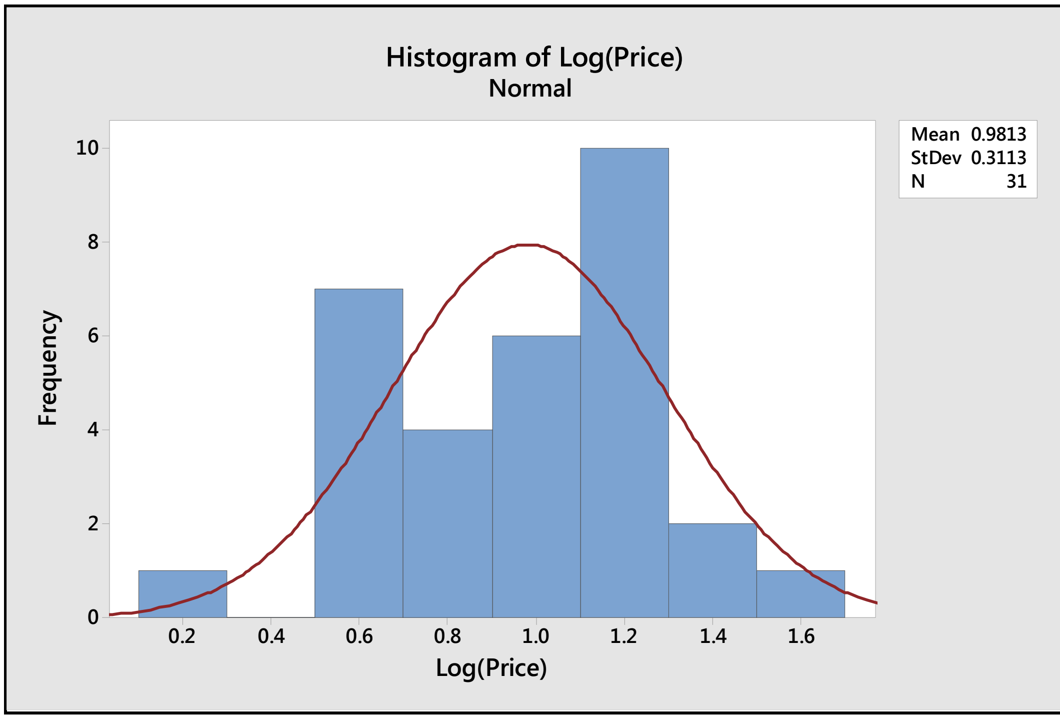


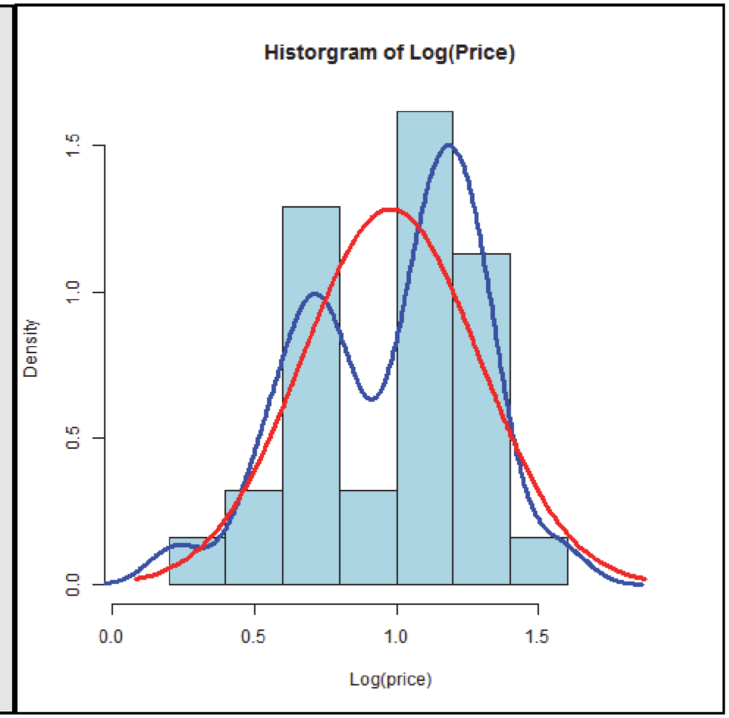


Normal Probability Plots indicate non-normality of the dependent variable Y.

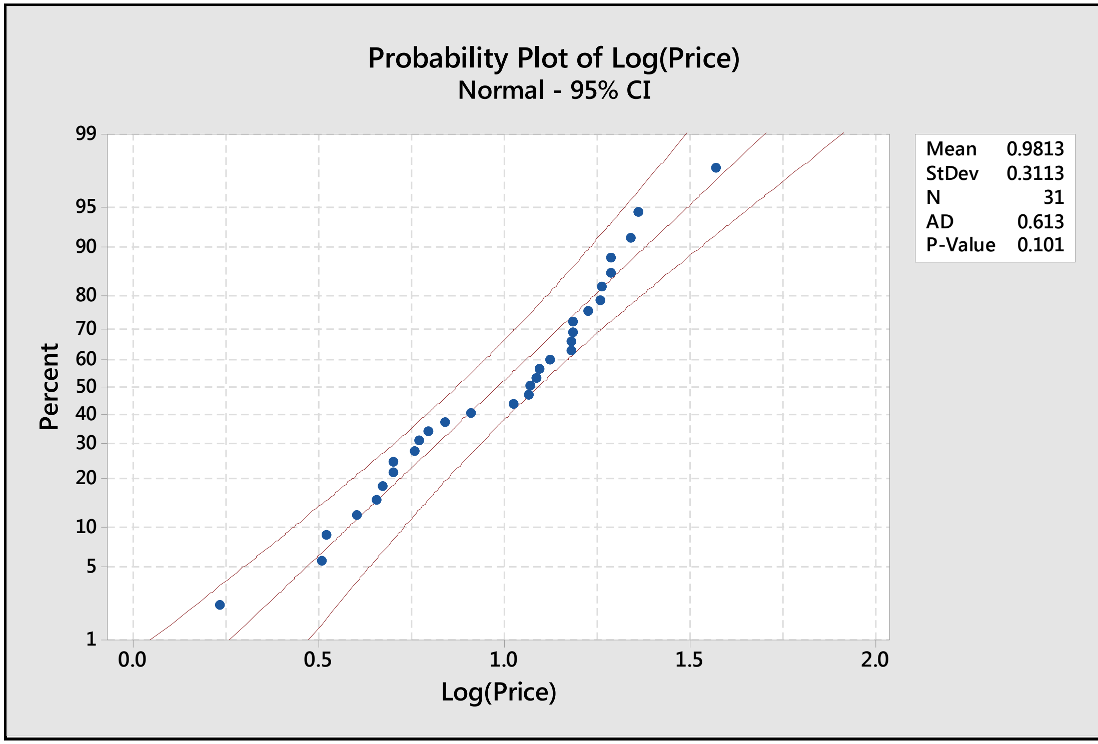
- In the case of **asymmetry in the dependent variable Y**, one attempts to transform this variable Y (a trial and error exercise) until one achieves symmetry. Also, $Z = Log(Y) \in (-\infty, \infty) \Leftrightarrow Y = 10^Z \in (0, \infty)$

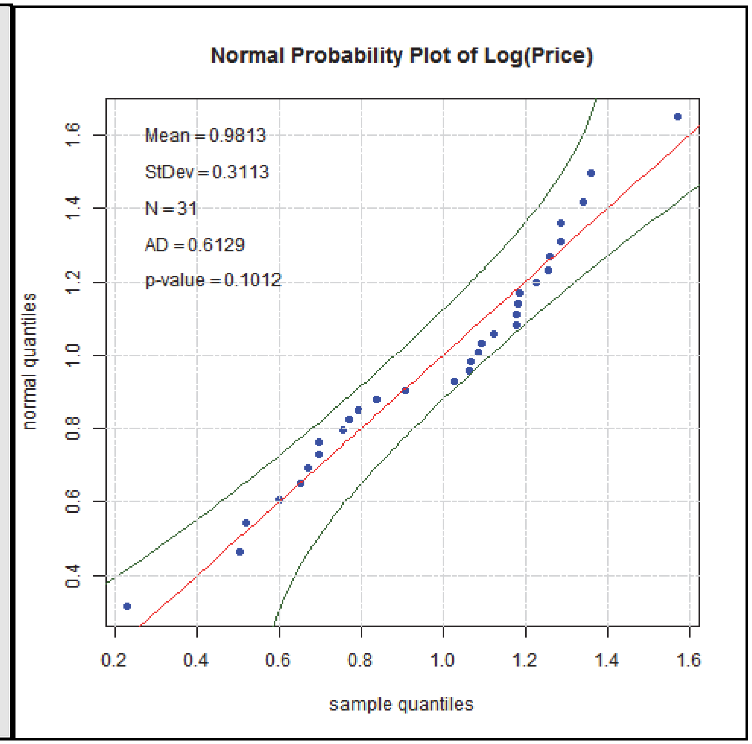
MINITAB HISTOGRAM R HISTOGRAM





Do these histograms reflect more symmetry? Does they look more bell-shaped?





Less deviation from normality in LOG(PRICE) plot than in the PRICE plot (although at the center we have larger deviations). **Perhaps equally important, when LOG(Price) is negative, Price is still positive valued!**

Modeling Decision: LOG(PRICE) becomes the dependent variable.

HOW DO WE SELECT INITIAL SET OF EXPLANATORY VARIABLES?

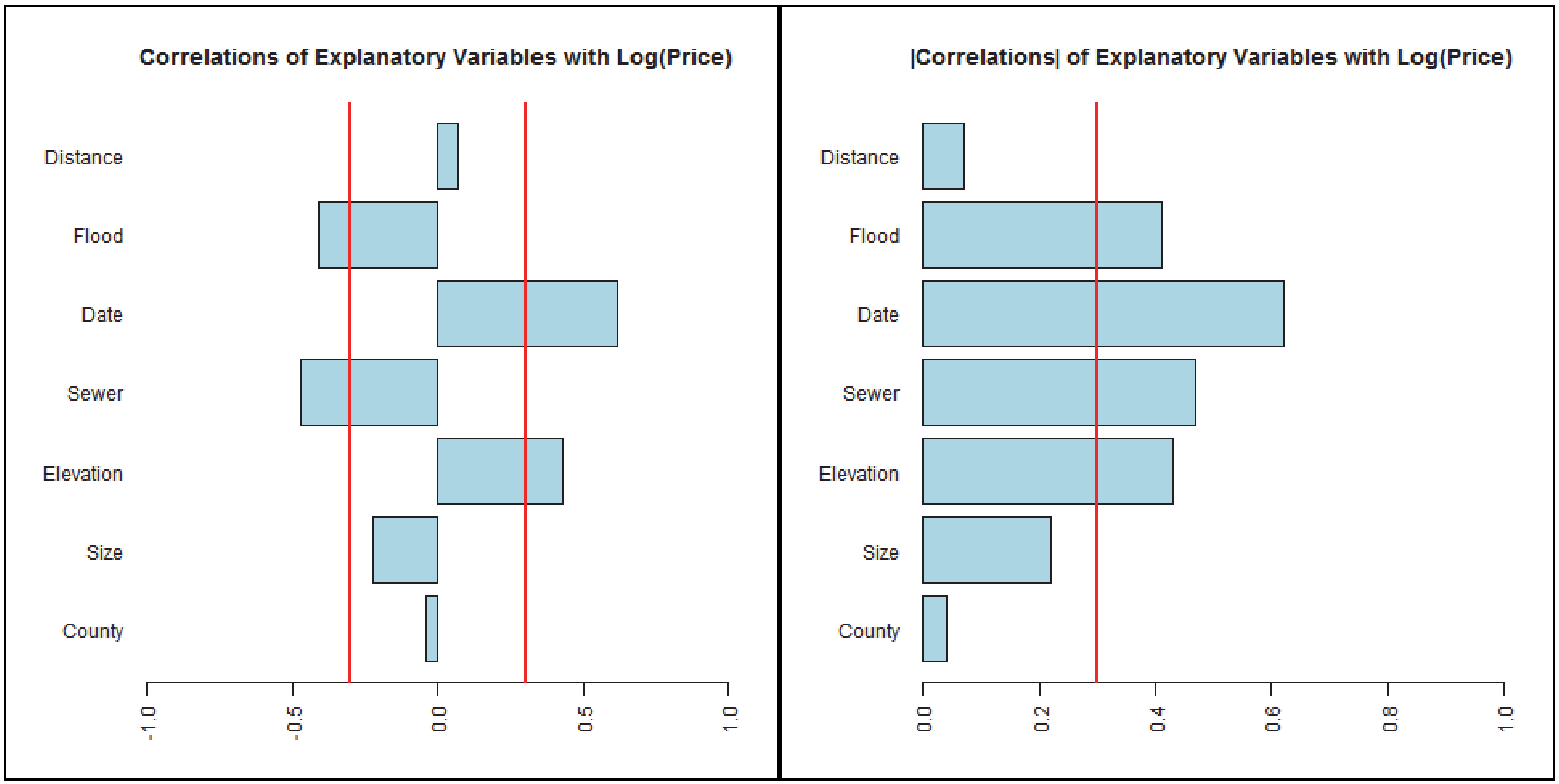
	Log(Price)	*County*	*Size*	*Elevation*	*Sewer*	*Date*	*Flood*	*Distance*
Log(Price)	1							
County	-0.044161	1						
Size	-0.22024	-0.339441	1					
Elevation	0.433356	0.475173	-0.209456	1				
Sewer	-0.467591	-0.050044	0.053381	-0.359408	1			
Date	0.62016	-0.369839	-0.349463	-0.056509	-0.151495	1		
Flood	-0.407298	-0.551804	0.108902	-0.373081	-0.113055	0.015361	1	
Distance	0.065871	-0.742204	0.556946	-0.36246	-0.158654	0.044383	0.423308	1

- **Recall correlation is a measure of linear dependence**, thus it is a good idea to obtain a feel for the data by **studying the correlations between the dependent and the explanatory variables**.

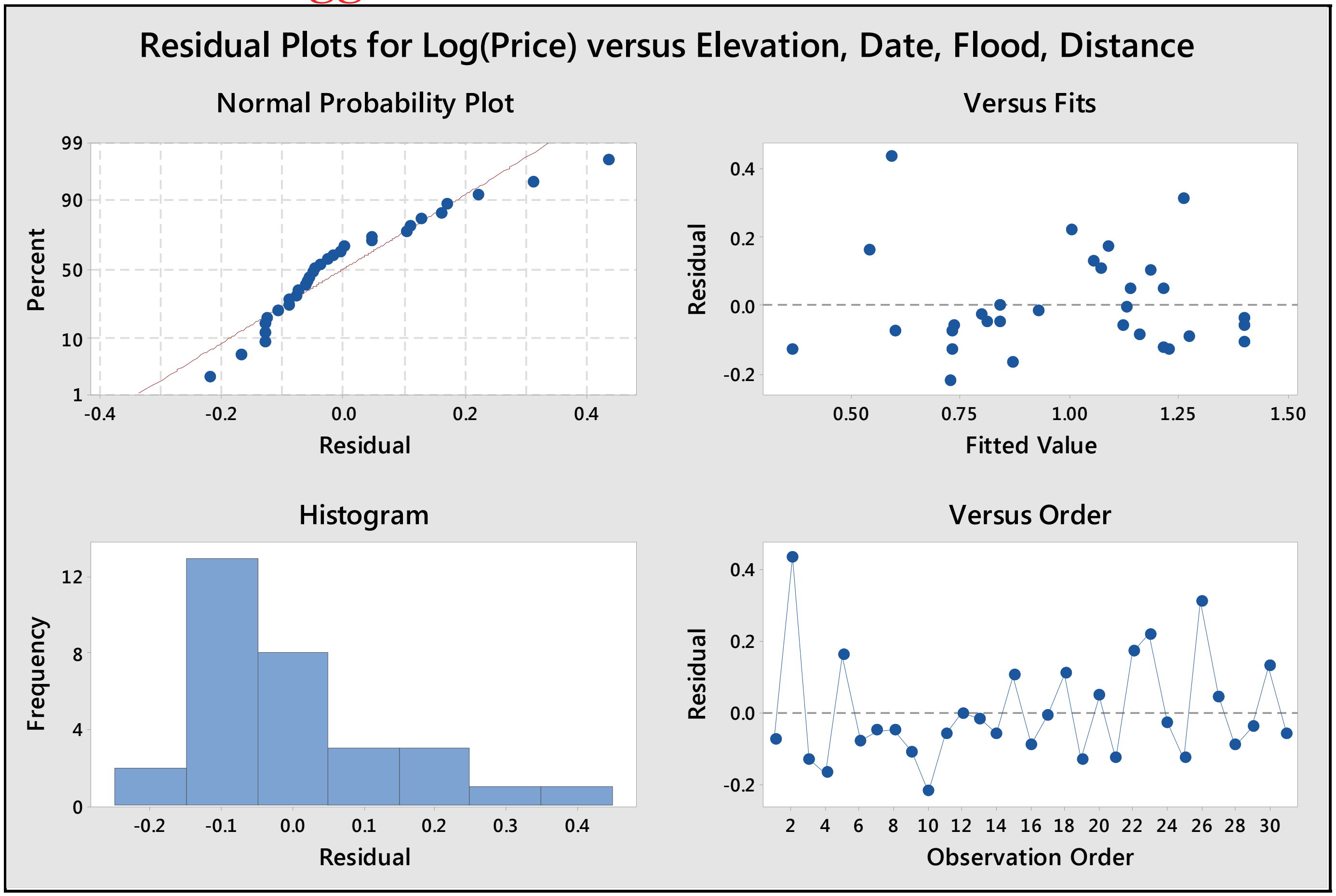
Based on the correlation matrix we select as initial explanatory variables:
ELEVATION, SEWER, DATE, FLOOD

The book, however, suggests the variables
ELEVATION, DATE, FLOOD, DISTANCE

- Coloring in correlation matrix above produced in Ms Excel using **conditional formatting**. Instead one could also use bar-charts in Excel, Minitab or R.

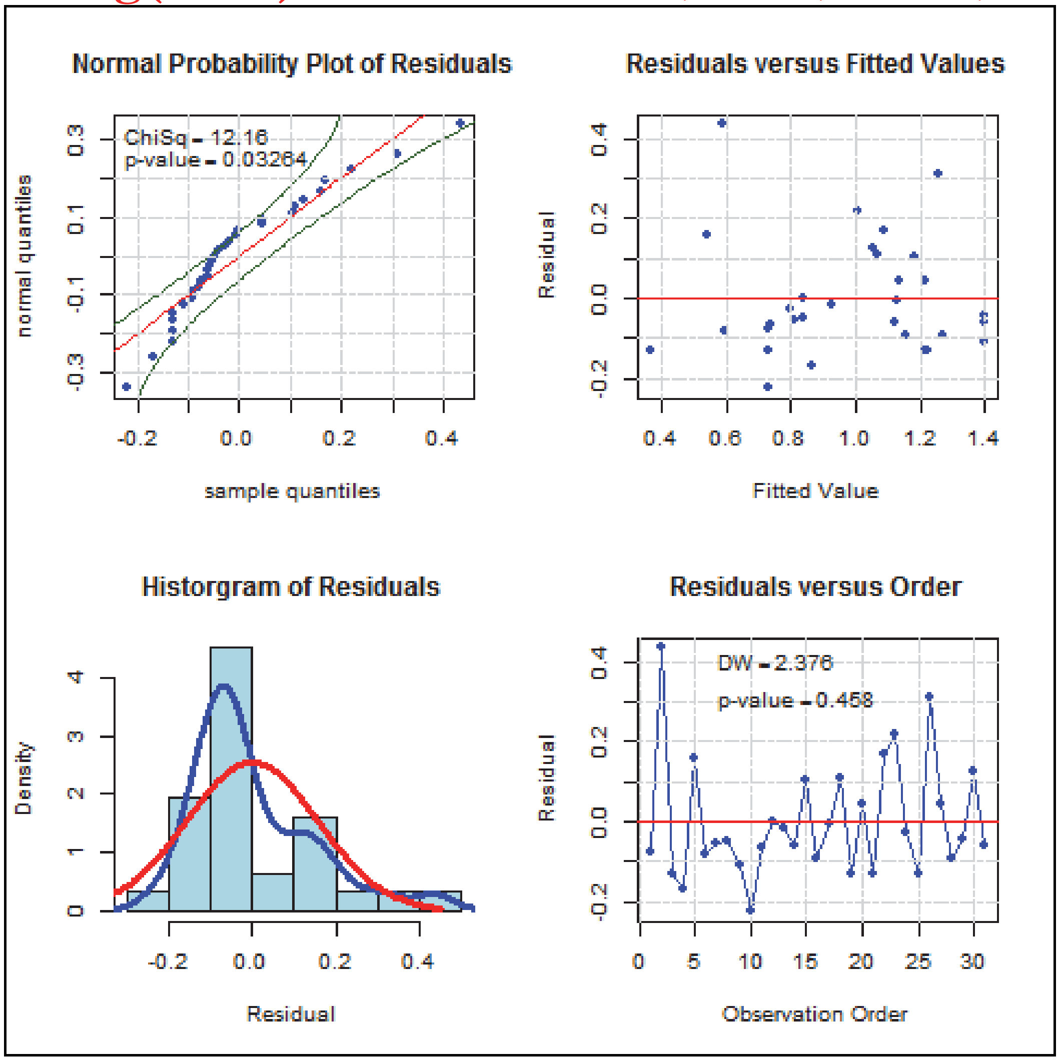


Book-Suggestion: $R^2 \approx 78.1\%$, DW-Statistic ≈ 2.37

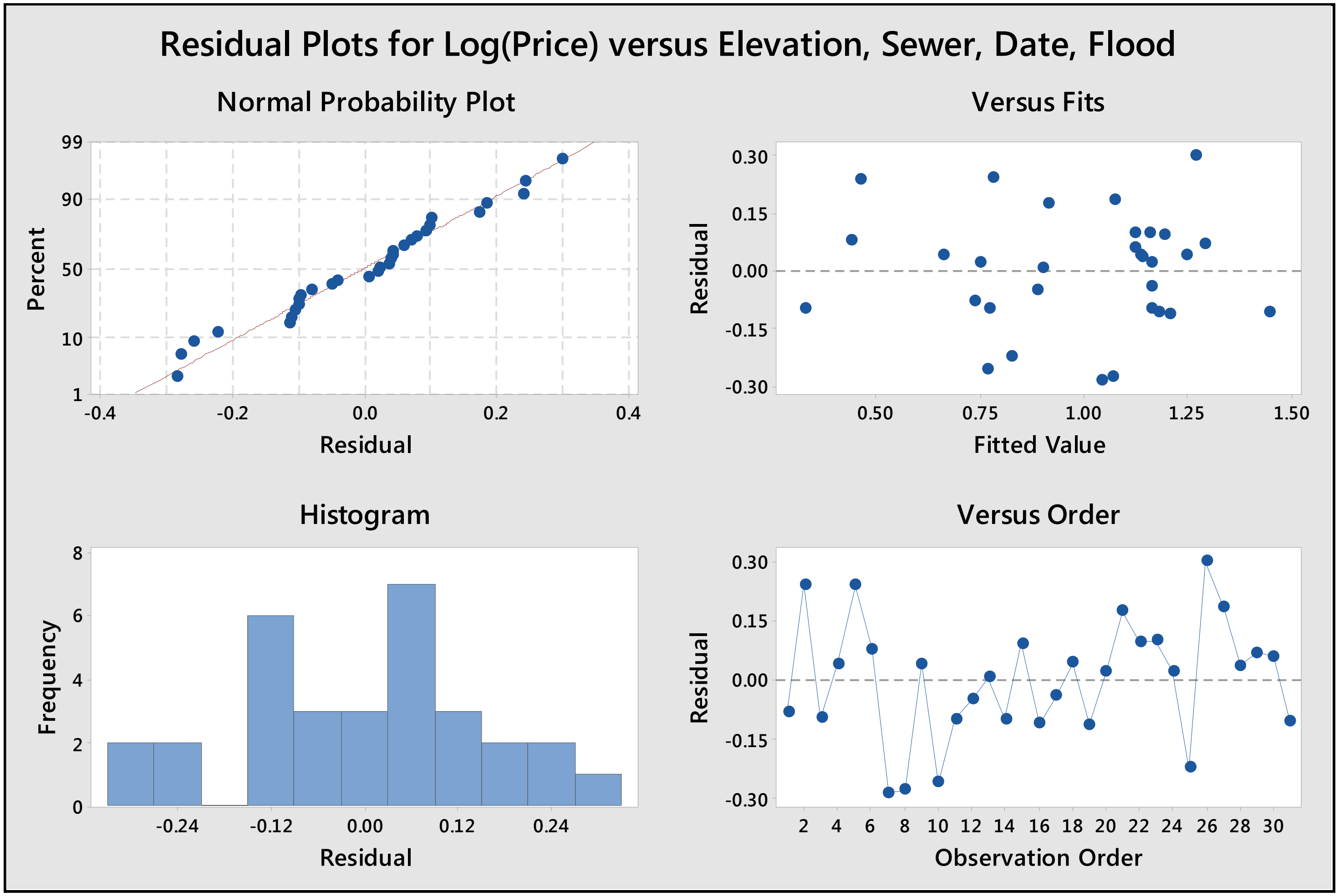


- **Durbin-Watson statistic tests auto-correlation** from residual observation to residual observation. $DW \approx 2(1-r)$ where r is the one-step autocorrelation amongst the residual observations. **Preferably $DW \in (1.5, 2.5)$.**

Residual Plot Log(Price) versus Elevation, Date, Flood, Distance in *R*

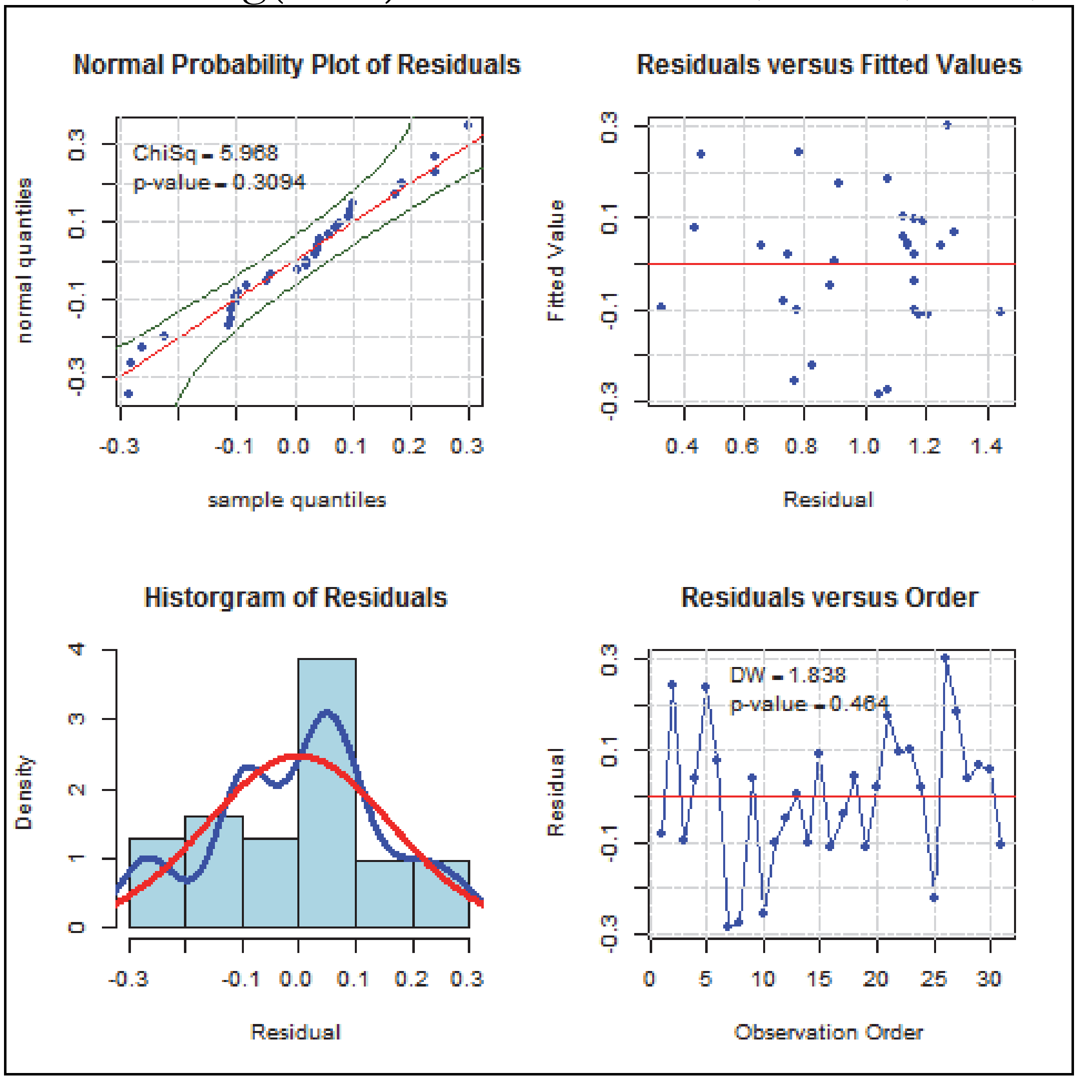


$$R^2 \approx 76.8\%, \text{ DW-Statistic} \approx 1.84$$

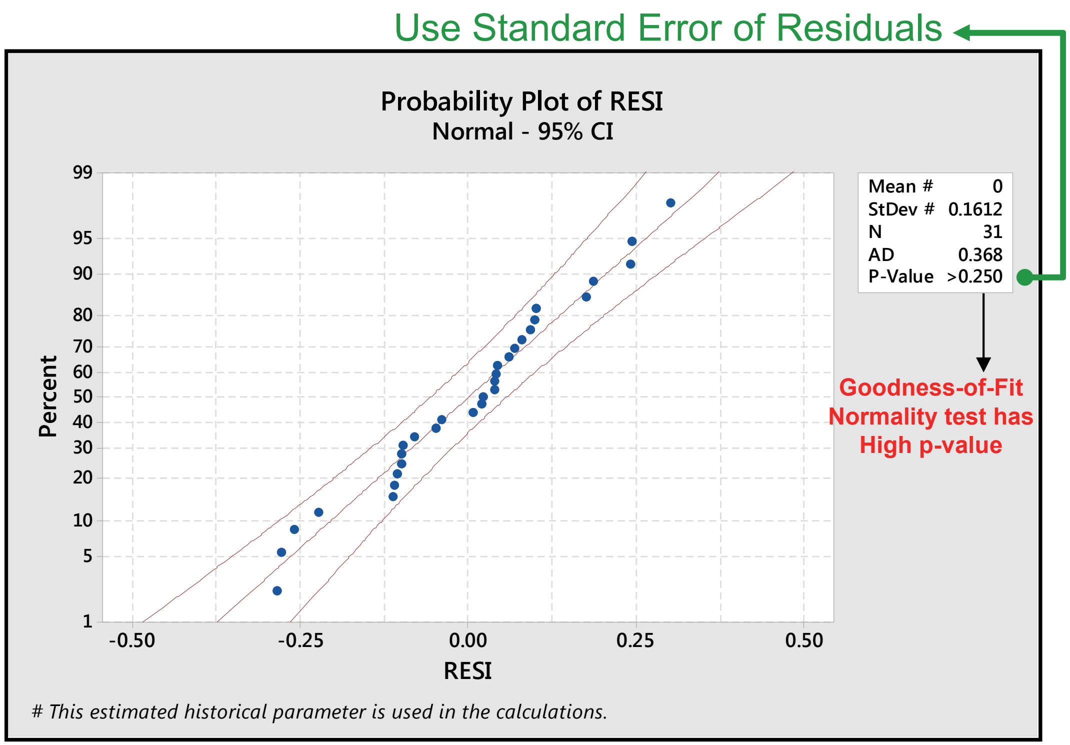


- The residual analysis (at least a first glance) seems to support more a normality assumption of the residuals and a lesser deviation in auto correlation from 2.

Residual Plot Log(Price) versus Elevation, Sewer, Date, Flood



Residual Analysis in Minitab by storing the residuals and creating a probability plot



- Despite the lower R^2-value of 76.8% against the 78.1% the model behavior of ELEVATION, SEWER, DATE, FLOOD is preferred over the model behavior of ELEVATION, DATE, FLOOD, DISTANCE because of the residual analysis.

We will continue to use: ELEVATION, SEWER, DATE, FLOOD

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.876204631
R Square	0.767734555
Adjusted R Square	0.73200141
Standard Error	0.161156071
Observations	31

$$F = \frac{\sum_i (\hat{y}_i - \overline{y})^2 / p}{\sum_i (\hat{y}_i - y_i)^2 / (n - p - 1)}$$

When model fits well F-value will be high

$$H_0 : b_1 = b_2 = \ldots = b_p = 0$$

Reject

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	2.231994794	0.557998698	21.48522185	6.28672E-08
Residual	26	0.675253263	0.025971279		
Total	30	2.907248057			

Low P-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.489073451	0.091484916	16.27671007	3.75383E-15	1.301023513	1.677123388
Elevation	0.014109922	0.008164115	1.728285526	0.095797643	-0.002671657	0.030891502
Sewer	-4.40859E-05	1.3516E-05	-3.261766947	0.003089944	-7.18684E-05	-1.63035E-05
Date	0.007411666	0.001224892	6.050871004	2.15959E-06	0.004893864	0.009929469
Flood	-0.318349977	0.088676005	-3.590035162	0.001348477	-0.500626116	-0.136073838

Output generated by MicroSoft EXCEL

Although the **F-Statistic is statistically significant** it is **still possible** that some **of the individual parameters are equal to zero**.

Recall: $var(\widehat{\boldsymbol{b}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, ν_{kk}: k-th diagonal element of $var(\widehat{\boldsymbol{b}})$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.876204631
R Square	0.767734555
Adjusted R Square	0.73200141
Standard Error	0.161156071
Observations	31

$$t = \frac{\hat{b}_k - 0}{\sqrt{\nu_{kk}}}$$ T- distribution with (n-p-1) degrees of freedom

$$H_0 : b_k = 0$$

Reject for all coefficients

Low P-values

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	2.231994794	0.557998698	21.48522185	6.28672E-08
Residual	26	0.675253263	0.025971279		
Total	30	2.907248057			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.489073451	0.091484916	16.27671007	3.75383E-15	1.301023513	1.677123388
Elevation	0.014109922	0.008164115	1.728285526	0.095797643	-0.002671657	0.030891502
Sewer	-4.40859E-05	1.3516E-05	-3.261766947	0.003089944	-7.18684E-05	-1.63035E-05
Date	0.007411666	0.001224892	6.050871004	2.15959E-06	0.004893864	0.009929469
Flood	-0.318349977	0.088676005	-3.590035162	0.001348477	-0.500626116	-0.136073838

Regression Analysis: Log(Price) versus Elevation, Sewer, Date, Flood

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	2.2320	0.55800	21.49	0.000
Error	26	0.6753	0.02597		
Total	30	2.9072			

Model Summary

S	R-sq	R-sq(adj)
0.161156	76.77%	73.20%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	1.4891	0.0915	16.28	0.000
Elevation	0.01411	0.00816	1.73	0.096
Sewer	-0.000044	0.000014	-3.26	0.003
Date	0.00741	0.00122	6.05	0.000
Flood	-0.3183	0.0887	-3.59	0.001

Regression Equation

Log(Price) = 1.4891 + 0.01411 Elevation - 0.000044 Sewer + 0.00741 Date - 0.3183 Flood

- **Multicolinearity:** In the Leslie case study data, we were able to identify the effect of FLOOD on the LOG(PRICE) **despite a negative correlation between (ELEVATION, SEWER) and (ELEVATION, FLOOD)**.

	Log(Price)	*County*	*Size*	*Elevation*	*Sewer*	*Date*	*Flood*	*Distance*
Log(Price)	1							
County	-0.044161	1						
Size	-0.22024	-0.339441	1					
Elevation	0.433356	0.475173	-0.209456	1				
Sewer	-0.467591	-0.050044	0.053381	-0.359408	1			
Date	0.62016	-0.369839	-0.349463	-0.056509	-0.151495	1		
Flood	-0.407298	-0.551804	0.108902	-0.373081	-0.113055	0.015361	1	
Distance	0.065871	-0.742204	0.556946	-0.36246	-0.158654	0.044383	0.423308	1

- Hence, regression is robust to some correlation between the explanatory variables. **However, too much correlation however can cause instability in the regression coefficients.**

- Recall that: $\widehat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$**.** Too high multicolinearity in $\boldsymbol{X}^T\boldsymbol{X}$ will result in **a matrix determinant** of $|\boldsymbol{X}^T\boldsymbol{X}|$ close to zero.

- The inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ has elements that are proportional to $1/|\boldsymbol{X}^T\boldsymbol{X}|$. Hence, **small changes in the explanatory data** will therefore result in **large changes of regression coefficients** (instability), which **makes interpretation of the effect of the explanatory variables difficult** (or next to impossible).

- **There are multiple ways to detect collinearity.** MINITAB provides a **variance inflation factor (VIF)** for each regression parameter. If the VIF < 1, there is no multicollinearity but if the VIF is > 1, predictors may be correlated.

- Montgomery and Peck suggest that **if the VIF is between 5 - 10, the regression coefficients are poorly estimated.**

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.161156	76.77%	73.20%	67.24%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.4891	0.0915	16.28	0.000	
Elevation	0.01411	0.00816	1.73	0.096	1.46
Sewer	-0.000044	0.000014	-3.26	0.003	1.30
Date	0.00741	0.00122	6.05	0.000	1.04
Flood	-0.3183	0.0887	-3.59	0.001	1.27

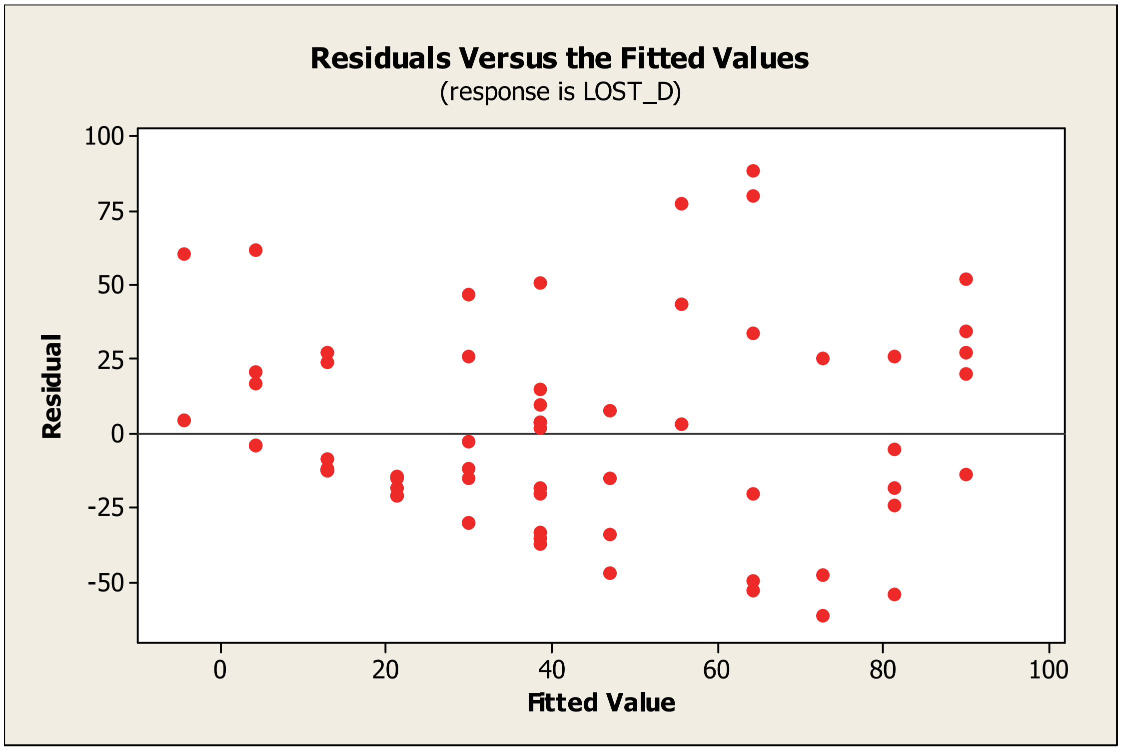
Regression Equation

Log(Price) = 1.4891 + 0.01411 Elevation - 0.000044 Sewer + 0.00741 Date - 0.3183 Flood

Output generated by MINITAB

- **Possible solution to multicollinearity: Eliminate explanatory variables** from the model (especially if deleting them has little effect on R^2).

- **Heteroscedasticity** means that the residuals **do not have a constant variance**. That is, **some relationship can be observed** between the residuals and the dependent variable and **the explanatory variables**.



Possible solutions to the problem of heteroscedasticity are:

1. variable transformations
2. weighted least squares regression (not part of this class).

- The Leslie Salt data is **time-series data**. In **the case of time-series data** one major concern may be **the dependence from one observation in one year to the next year**. This is called **auto-correlation.** It can be detected by evaluating **the Durbin-Watson Statistic amongst residuals**:

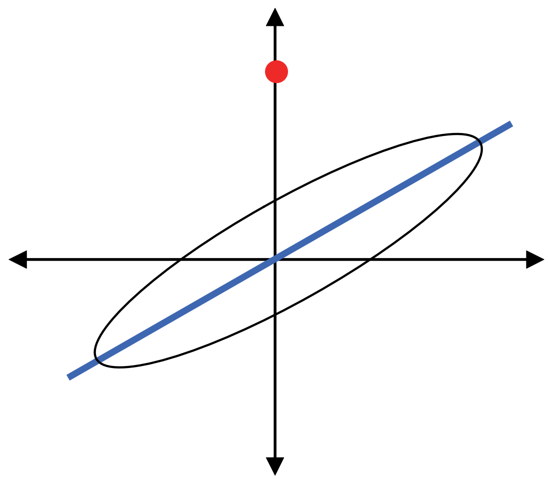
$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{\sum_{i=2}^{n} e_i^2 - 2\sum_{i=2}^{n} e_i e_{i-1} + \sum_{i=1}^{n-1} e_{i-1}^2}{(n-p-1)\sigma^2}$$

- **Independence residuals** $\Rightarrow$ **no autocorrelation** $\Rightarrow \sum_{i=2}^{n} e_i e_{i-1} = 0 \Rightarrow$

$$DW = \frac{\sum_{i=2}^{n} e_i^2 + \sum_{i=1}^{n-1} e_{i-1}^2}{(n-p-1)\sigma^2} \approx \frac{2 \times (n-p-2)\sigma^2}{(n-p-1)\sigma^2} \approx 2$$

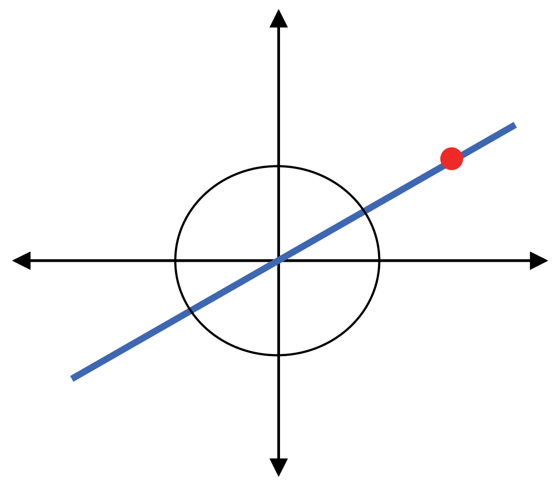
Thus large deviations of DW from 2.0 indicate **the presence of auto-correlation amongst the residuals, which contradicts independence.**

Large value of
dependent variable Y



Outlier easy to detect
From residuals

Large value of
independent variable X

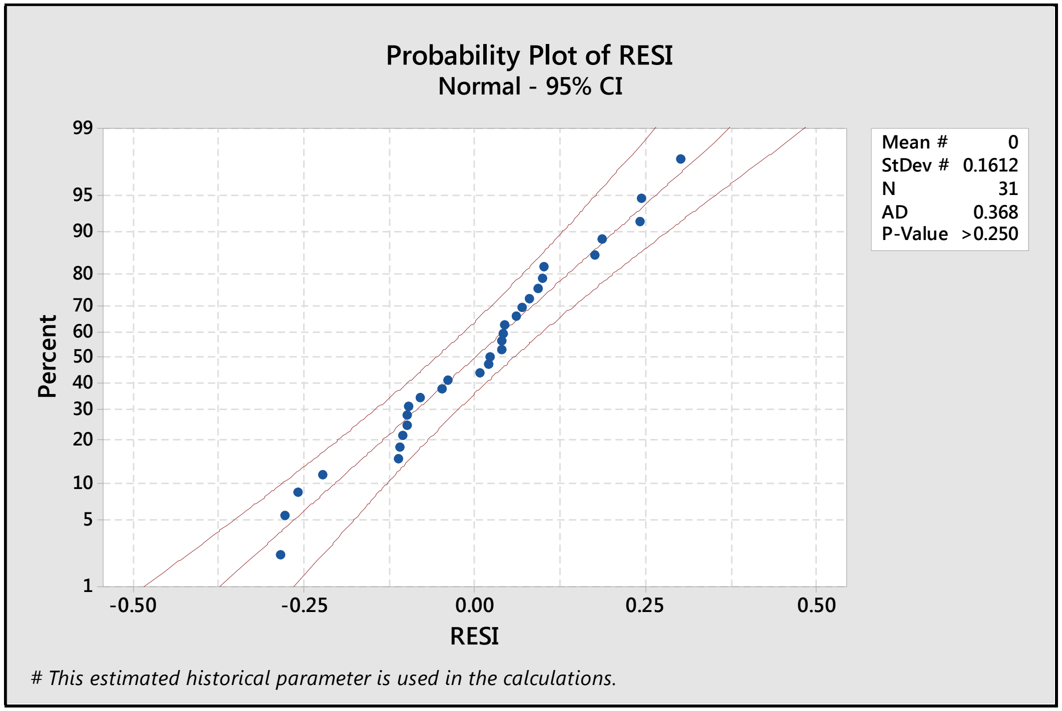


Outlier difficult to detect
From residuals

Conclusion:

Behavior of residuals does not determine **all influential observations**!

- **Outliers may be visually observed** from residual probability plots when they fall outside the confidence boundaries.



- Another method for **determining influential observations** is to calculate **studentized residuals (called deleted-*t* residuals in Minitab).**

$$e_i^* = \frac{e_i}{s(i)\sqrt{1-h_{ii}}} \sim T_{n-p-2},$$

$s(i)$: **Standard deviation of residuals** when **omitting** observation i
(hence, studentized residuals have one degree of freedom less)

h_{ii} : The i-th diagonal element of the matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \Leftrightarrow \widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{b}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Hence, the matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ determines the fitted values $\widehat{y}$.

- To evaluate **the combined effect of a single observation on all regression coefficients** MINITAB calcutates a DFIT coefficient per observation. (Similar to the DFBETAS in the book). **Observations of |DFIT| values greater than $2\sqrt{(p+1)/n}$ are considered large** and these observations should be examined for accuracy, where **p is the number of explanatory variables** and **n is the number of observations**.

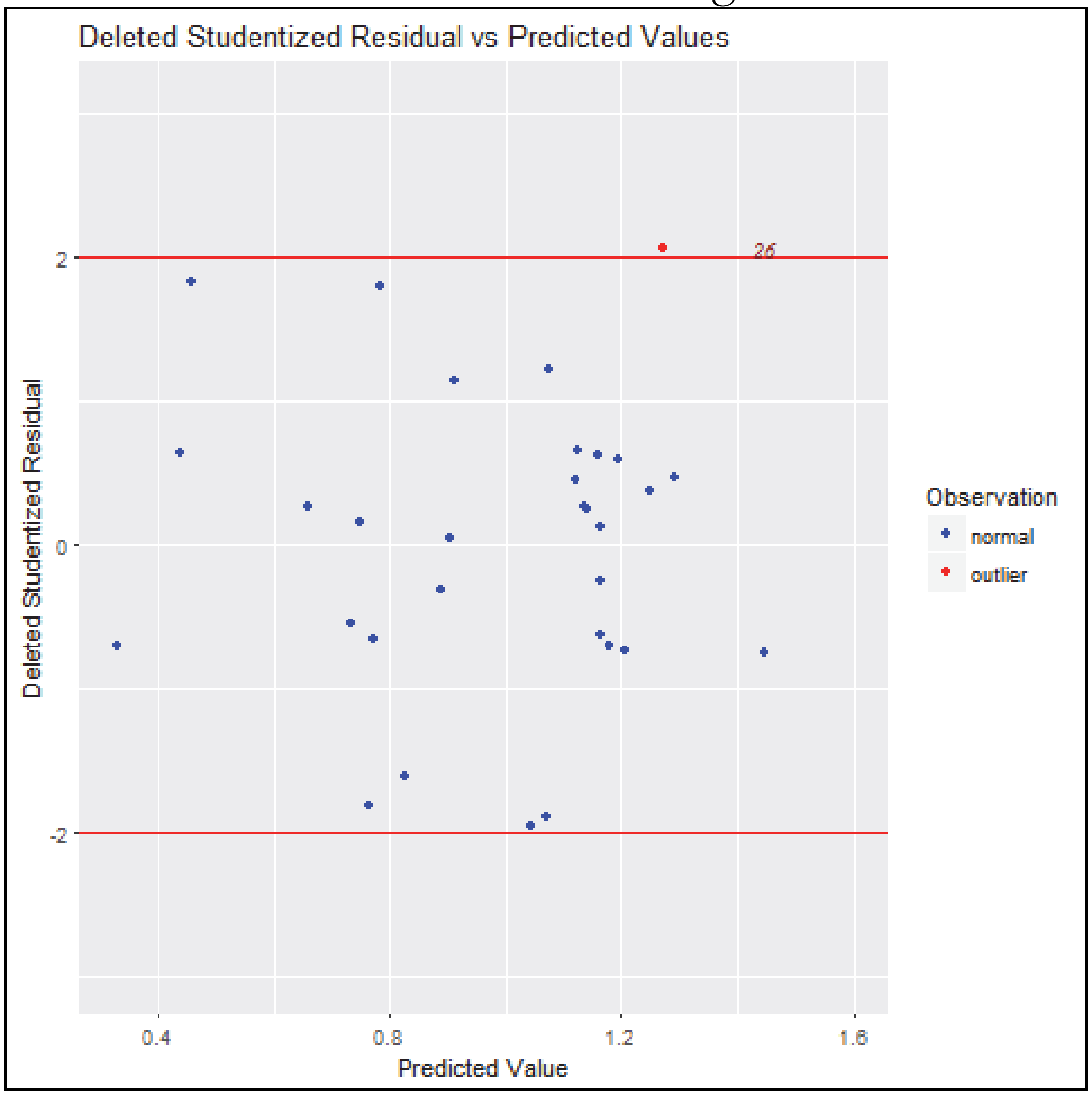
$$e_2^* \approx 1.79, \mathrm{DFIT}_2 \approx 0.96 > 2\sqrt{5/31} \approx 0.80 \Rightarrow \text{Should be checked.}$$

$$e_5^* \approx 1.8296, \mathrm{DFIT}_5 \approx 1.11 > 2\sqrt{5/31} \approx 0.80 \Rightarrow \text{Should be checked.}$$

Deleted Student Residuals and Dfit values generated by MINITAB

Data	TRES1	DFIT1		p	4	
1	-0.56	-0.27537		n	31	
2	1.79	0.968737		DFIT THRESHOLD	**0.803219**	
3	-0.71	-0.46051				
4	0.26	0.108643		α	5%	
5	1.82	1.111812		TRES1 Threshold	1.708141	
6	0.63	0.519808				
7	-1.96	-0.63989				
8	-1.90	-0.59833				
9	0.37	0.403163				
10	-1.81	-0.726				
11	-0.66	-0.26599				
12	-0.31	-0.08664				
13	0.04	0.013538				
14	-0.64	-0.16958				
15	0.59	0.184402				

Deleted Student Residuals Plot generated in R



DFIT Value Plot generated in *R*