Multiple Regression
Leverage, outliers, and influence measures

Leverage

- **Leverage - a measure of "outlier" in "x-direction" in SLR;**

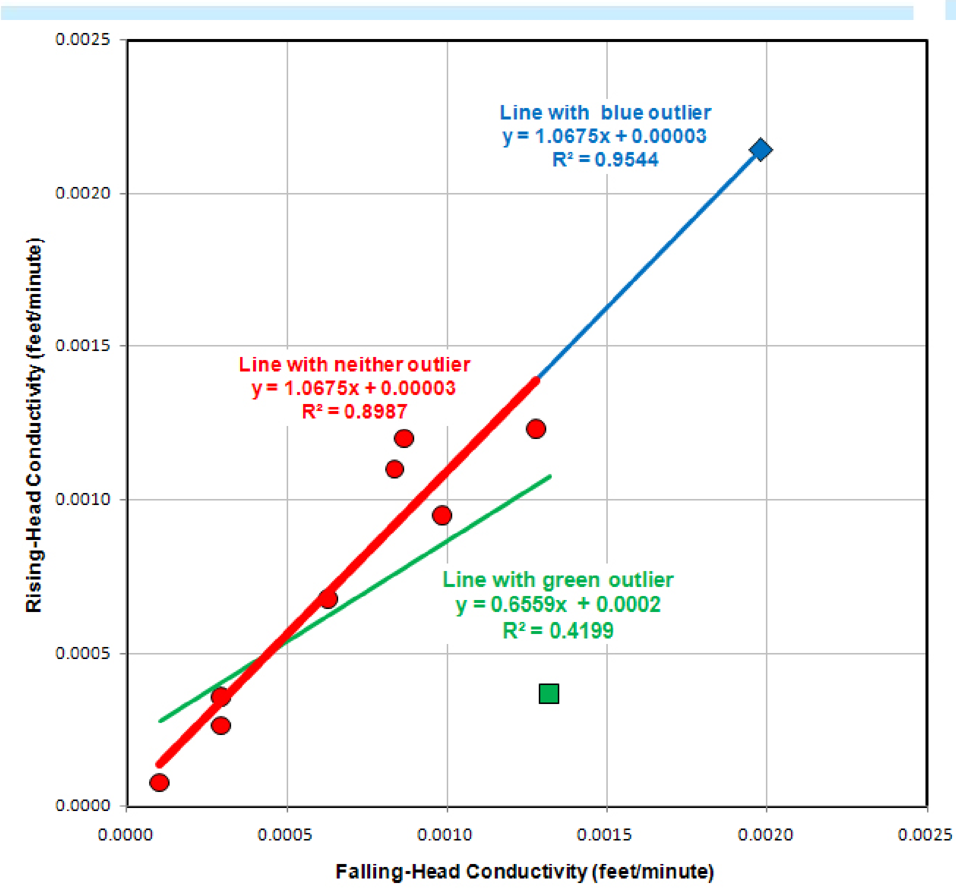
  - a measure of distance of a point (x_1, x_2, …xn) from the center of the data given by $(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n},)$ in MLR.

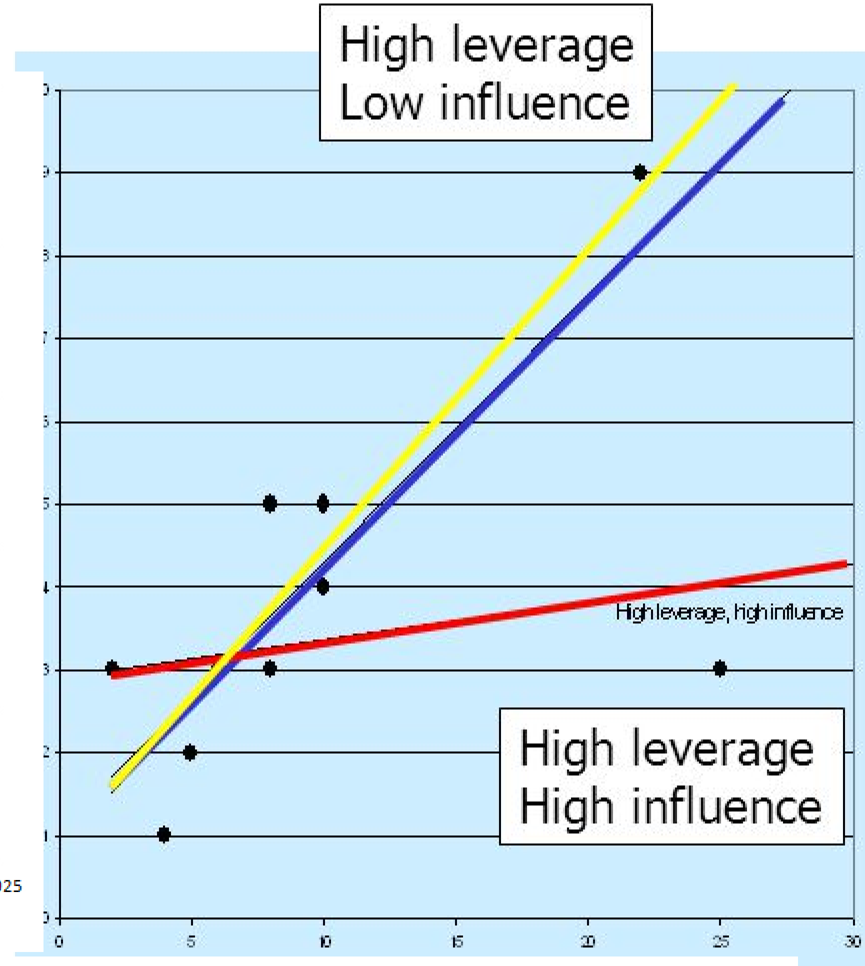
- **For SLR: leverage of x_i is:** $$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{SSx}.$$

- **For MLR there is a more complex formula for H_i=leverage of predictors of ith observation. MINITAB computes Hi's.**

- **High leverage point has hi > 3k/n, where k=#of predictors, and n=number of observations.**

3. Leverage vs. Influence





High influence: when removal of an outlier changes regression line.

High leverage is not enough for influence!

Outliers in Y direction: standardized residuals

- Residual $e_i = Y_i - \widehat{Y}_i$.
- **Standard error** of a residual: $s\sqrt{1 - h_i}$
- **Standardized residual**: $e_{si} = e_{i_{s\sqrt{1-h_i}}}$
- **OUTLIERS**: observations with $|e_{si}| > 2\ or > 3\ (\text{extreme outlier})|$
- Why? Standardized residuals should have standard normal distribution, so |values|>2 are rare, and |values|>3 are extremely rare.
- Also: Many outliers may suggest not normal distribution! However, plots are better to check that.

Outliers in Y direction: studentized residuals (TRESIDs)

- **TRESID$_i$ = $\frac{e_{(i)}}{SD_{e_{(i)}}}$, where $e_{(i)}$ is the ith prediction residual, i.e.**
- $e_{(i)} = Y_i - \widehat{Y}_i$, **where $\widehat{Y}_i$ is computed using regression model estimated with ith observation deleted, and $SD_{e_{(i)}}$ is the standard deviation of $e_{(i)}$.**
- **TRESID~ $t_{(n-k-1)}$ if the model holds with normal errors.**
- **Large TRESIDi suggests that ith observation may hold an outlier in y-direction, so could be influential.**

Measures of influence
COOK's D and DFFITS

- **COOK's D is a very popular measure of influence.**
- **Observation I has COOK's D** $D_i = \frac{1}{k}\left(\frac{h_i}{1 - h_i}\right)(\text{standardized residual}_i)^2$
- D_i **combines leverage of predictors for observation I with a measure of "outlier" in the "Y-direction" of observation i.**
- **Observation I has high influence if** $D_i > F_{k+1,n-k,0.1}$
- **For** $n > 30$**, critical values for** D_i **are about 1.6 to 2.**

DFFITS – related to studentized residuals

- **DFFITS of observation i is** $\mathrm{DFFITS_i} = \mathrm{TRESID_i}\sqrt{\frac{h_i}{1-h_i}}$
- **Observation I is considered to have high influence if** $|\mathrm{DFFITS_i}| > 2\sqrt{k/n}$

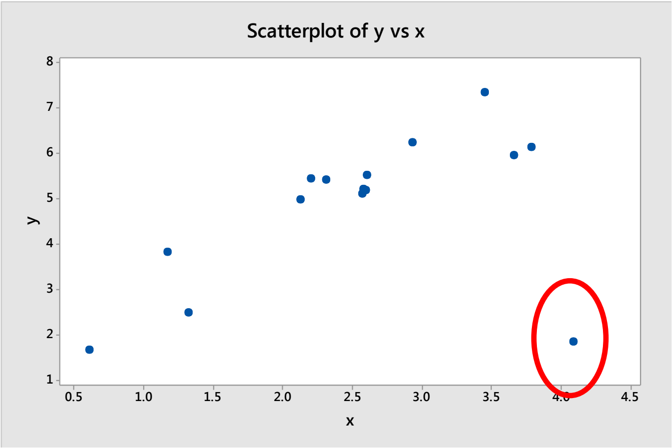
SUMMARY

- **To identify outliers in Y-direction use standardized or studentized residuals.**
- **To identify influential observations with use Cook's D or DFFITS.**
- **To identify observations with unusual x use leverage statistics** h_i.
- **Often influence can be detected by high leverage and outlier in Y direction.**

If an influential point is detected…

- **Check it out for possible error.**
- **If error detected, correct it if possible, or delete the observation.**
- **If no error consider other models that may fit the point better or use procedures robust to outliers.**

Example: data set **influenceclass16.MTW**



Observation 15 is far from the rest of the data. Is it influential?

Need to find the thresholds for the measures of influence: Cook's D, DFFITS, and TRESID, and compare them to these measures computed for observation 15.

```
Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    2.74     1.10     2.49    0.027
x          0.819    0.406     2.02    0.065  1.00

Regression Equation
y = 2.74 + 0.819 x

Fits and Diagnostics for Unusual Observations

Obs        y     Fit   Resid  Std Resid
 15    1.833   6.095  -4.262      -3.26  R
R  Large residual
```

Example: data set influenceclass16.MTW

obs	y	x	FITS	SRES	TRES	HI	COOK	DFIT
1	4.97468	2.13042	4.48819	0.33742	0.32561	0.078740	0.004865	0.095193
2	1.66211	0.60897	3.24196	-1.29328	-1.33112	0.338663	0.428251	-0.952552
15	1.83260	4.09203	6.09496	-3.26301	-7.36915	0.243801	1.71634	-4.18424

Thresholds for the measures of influence:

For Cook's D: $F_{2,\ 14,\ 0.05}$ = 2.73, D15=1.716

For DFFITS: $2\sqrt{1/15} = $ 0.516, DFFITS15 = -4.184

For TRESID: t_{13}, 0.05= 1.771, TRESID15 = -7.37.

It looks like obs 15 can be influential. W will compute regression eqn without obs 15 to see this.

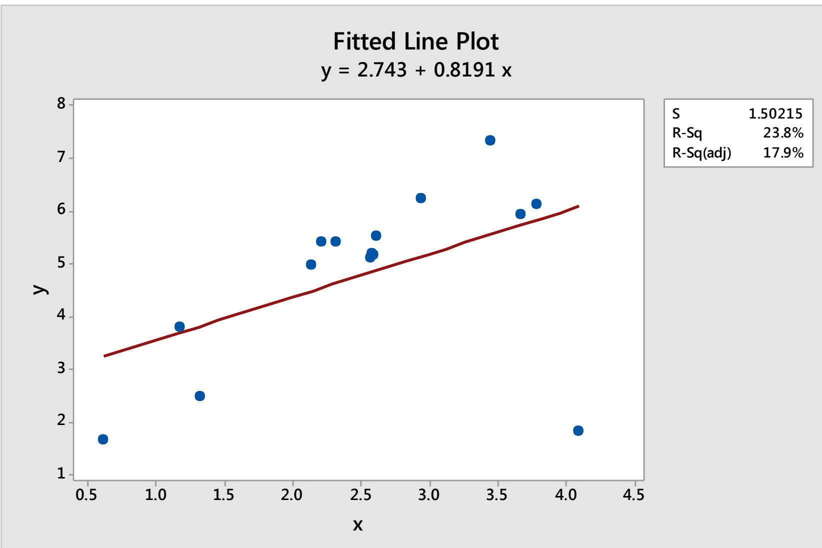
Example: data set influenceclass16.MTW

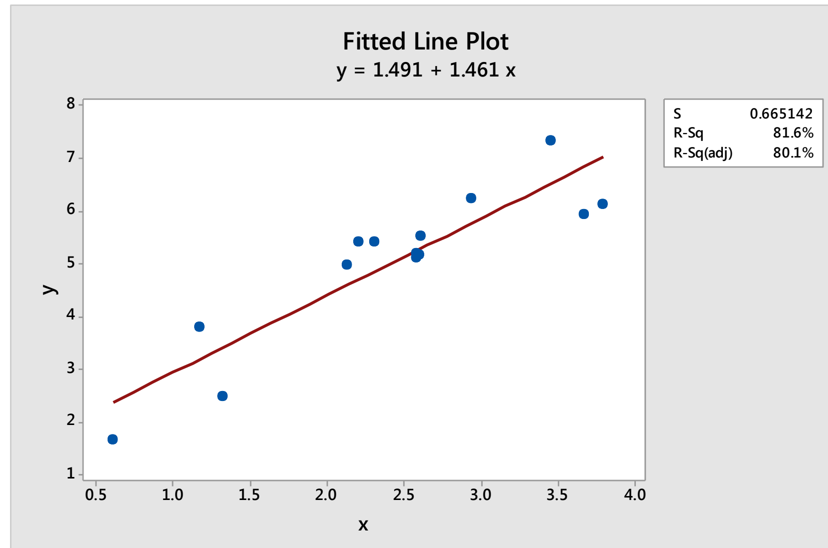
Regression Equation
(with observation 15)

y = 2.74 + 0.819 x

Regression Equation
(without observation 15)

y = 1.491 + 1.461 x





Looks like observation 15 was quite influential. After we removed if from the data, we got a quite different regression line.

Multicollinearity

- Multicollinearity means that at least one predictor is closely related to one (or more) other predictors.

- Consequences:
  - Problems with stat inference, unreliable regression equation, etc.
  - Unstable regression equation.
  - Stepwise procedures may produce different (contradictory) models.

- Diagnostics: Compute correlation coefficients between all pairs of predictors. If the correlation is large, we may have a problem with collinearity.

- Solutions/remedies:

  - Eliminate some predictors in severe cases,
  - Use principal components in less severe cases (outside the scope of this course)

Measure of Multicollinearity: Variance Inflation factors (VIF)

- **Variance Inflation Factor for predictor j is:**

$$VIF_j = \frac{1}{1 - R_j^2}, \text{where } R_j^2 \text{ is the multiple Rsquared}$$

from the prediction of jth predisctor on all other predictors.

If multicollinearity is present, then for some j we have $R_j^2 \approx 1$, so then VIF_j would be very large.

If no multicollinearity, then all $R_j^2 \approx 0$, so then VIF_j would be close to 1.

Serious problem with variable j if $R_j^2 \approx 0.9$, then $VIF_j > 10$.

Multicollinearity example: multicollclass data

Model with 4 explanatory variables:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.08	1.33	3.06	0.012	
x1	-4.64	3.47	-1.34	0.211	77.55
x2	9.73	6.98	1.39	0.194	256.69
x3	4.70	5.26	0.89	0.393	920.42
x4	0.17	1.20	0.14	0.892	342.03

Regression Equation

y = 4.08 - 4.64 x1 + 9.73 x2 + 4.70 x3 + 0.17 x4

Model with 2 explanatory variables:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.56	1.32	2.70	0.019	
x1	0.719	0.413	1.74	0.107	1.05
x2	-0.506	0.457	-1.11	0.290	1.05

Regression Equation

y = 3.56 + 0.719 x1 - 0.506 x2