Remedies for Violations of Error Assumptions

● **Assumptions about ε the same as in SLR.**

● **Residual plots will again help find violations in MLR.**

● **Transformations may help fix violations (trickier in MLR).**

● **Using** $Y^* = \log(Y)$ **or** $Y^* = \sqrt{Y}$ **can stabilize non-constant error variance or fix non-normality violation.**

Example (Surgical Data):

● **Again, drawback is that model may be less interpretable.**

Multicollinearity

● **If several independent variables measure similar phenomena, their sample values may be strongly correlated.**

● **This is known as multicollinearity.**

Example: Predicting javelin throw length based on ability in bench press (X_1), military press (X_2), curl (X_3), chest circumference (X_4).

● Natural association among the independent variables.

● Including many similar independent variables in the model is easy to do, and may actually improve prediction.

● Problem: The <u>effect of each individual variable</u> may be masked with multicollinearity.

<u>Common Problems Caused by Multicollinearity</u>

(1) large standard errors for estimated regression coefficients → leads to concluding individual variables are <u>not</u> significant even though overall model may be <u>highly</u> significant.

(2) Signs of estimated regression coefficients seem "opposite" of intuition (idea of "holding all other X's constant" doesn't make sense).

● A common measure to detect multicollinearity is the Variance Inflation Factor (VIF).
● For an independent variable X_j, its VIF is:

High R_j^2 (near 1) →

Rules of thumb:

- **VIF = 1 → X_j not involved in any multicollinearity**
- **VIF > 10 → X_j involved in severe multicollinearity**

- **In practice we obtain VIF values from computer.**

Example:

Remedies for Multicollinearity

(1) Drop one or more variables from model

(2) Rescale variables (often to account for trends over time like population increases)

(3) More advanced: Principal components regression, Ridge regression.

- **Important note: Multicollinearity does not typically harm the predictive ability of a model.**

<u>Variable Selection</u>

● Often a very large number of possible independent variables are considered in a study.

● Which ones are really worth including in the model?

● A model with <u>many</u> independent variables:

● A model with <u>few</u> independent variables:

● If there are *m* independent variables under consideration, how many possible subsets of variables do we have?

● Computer procedures can help search among many possible models.

<u>Goals</u>:
(1) Choose a model that yields accurate (i.e., unbiased) estimates and predictions
(2) Choose a model that explains much of the variation in *Y*.
(3) Choose a <u>parsimonious</u> model.

Achieving the Goals

(1) Mallows' C(p) statistic measures the bias in the bias under consideration, relative to the full model.

● For a model having *p* independent variables, we would want C(p) to be near *p* + 1.

● C(p) >> *p* + 1 →

● C(p) << *p* + 1 →

Formula:

Note: If $MSE_p \approx MSE_{full}$, then:

(2) Normally R^2 tells us the proportion of variation in *Y* that the model explains.
● But R^2 always increases when new variables are added to a model → inappropriate to compare models with a different number of independent variables using R^2.

Better: Adjusted R^2 (R_a^2), which penalizes models having more variables.

R_a^2 =

Compare R^2 and R_a^2:

● Choosing the model with the maximum R_a^2 is equivalent to choosing the model with the minimum MSE.

● The "best" model is usually a compromise between the choice using the C(p) criterion and the choice using the R_a^2 criterion.

● All else being equal, a <u>simpler</u> model is usually better.

<u>Problems with Automatic Variable Selection</u>

(1) If we really are really interested in the (partial) effect of some independent variable X_j on Y, we may need to include X_j even if it's not in the "best" subset.

(2) Using the data to choose the "best" model and then examining P-values amounts to <u>using the data to suggest hypotheses</u> – can alter Type I error rates.

<u>Note</u>: If we initially have a large number ($\geq$ 20) of independent variables, finding the "best" model can be time-consuming.

● Pages 387-388 discuss initial screening methods (stepwise methods) to eliminate some variables quickly.

<u>Detecting Outliers and Influential Points</u>

● Outliers: Observations that do not fit the general pattern of points.

● With MLR, cannot see outliers using a simple scatter plot of the data. Why?

● Examining residuals still helps find outliers.
Rule of thumb:
$|\text{studentized residual}| > 2.5 \rightarrow$ possible outlier

● Outlying data values that occur near the extremes of the range of *X* values often greatly influence the position of the least-squares line.

● These points are called <u>high-leverage</u> points.

Picture:

● Hard to "visualize" leverage when there are <u>several</u> independent variables.

● In MLR, the $n \times n$ "hat" matrix is:

● For each observation, the corresponding diagonal element of the hat matrix measures how similar that observation is to the others, in terms of its $X_1, X_2, \ldots, X_m$ values.

<u>Rule of thumb</u>: If the i-th hat diagonal is greater than $2(m + 1) / n$, then the i-th observation is a high-leverage point.

<u>Influence Diagnostics</u>

● <u>Question</u>: How much would the regression line change if we estimated it <u>after removing</u> a particular observation?

● If the regression line would change greatly, that point is an <u>influence point</u>.

● DFFITS, for each observation, measures the difference between:
*** the predicted value from the regression estimated <u>with that observation included</u>**
and
*** the predicted value from the regression estimated <u>with that observation removed</u>.**

● Any observation with a |DFFITS| greater than:
is considered an influence point.

● **What to do if we have influence points?**

(1) Verify the data point is recorded correctly.
(2) Fit the regression line with and without the point(s). Do the substantive conclusions about the regression change?
(3) Ask: Does the observation reflect a fluke or a truly important event?

● **Automatically deleting outliers and influence points can be bad practice.**

<u>Rain example</u>:

Outlier detection:

Hat diagonals:

DFFITS: