

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.

Special Models

- Residual plots (in SLR or MLR) may provide evidence that the linear trend does not hold or that the error variance is not constant.
- Special models can explicitly account for these situations.

- **The polynomial regression model allows for a curvilinear relationship between Y and the predictor(s).**

- **A multiplicative model (see Sec. 8.6.2) is useful when the error variance increases for larger values of Y .**

- **Note this model is equivalent to the linearized form:**

- **In practice, both the polynomial regression model and the multiplicative model can be fit by transforming variables and using MLR software.**

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 effect of each individual variable may be masked with multicollinearity.**

Common Problems Caused by Multicollinearity

(1) large standard errors for estimated regression coefficients → leads to concluding individual variables are not significant even though overall model may be highly 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) \rightarrow

Rules of thumb:

- **VIF = 1 $\rightarrow X_j$ not involved in any multicollinearity**
- **VIF > 10 $\rightarrow X_j$ 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.**

Variable Selection

- 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 many independent variables:
- A model with few 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.

Goals:

- (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 parsimonious model.

Achieving the Goals

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

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

- $C(p) \gg p + 1 \rightarrow$

- $C(p) \ll p + 1 \rightarrow$

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 \rightarrow 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 simpler model is usually better.

Problems with Automatic Variable Selection

(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 using the data to suggest hypotheses – can alter Type I error rates.

Note: If we initially have a large number (≥ 20) of independent variables, finding the “best” model can be time-consuming.

- Pages 434-435 discuss initial screening methods (stepwise methods) to eliminate some variables quickly.

Detecting Outliers and Influential Points

- **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:
 $|\text{studentized residual}| > 2.5 \rightarrow \text{possible outlier}$
- **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 high-leverage points.**

Picture:

- **Hard to “visualize” leverage when there are several 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, \dots, X_m values.

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

Influence Diagnostics

- **Question:** How much would the regression line change if we estimated it after removing a particular observation?
- If the regression line would change greatly, that point is an influence point.
- DFFITS, for each observation, measures the difference between:
 - * the predicted value from the regression estimated with that observation included
 - and
 - * the predicted value from the regression estimated with that observation removed.
- Any observation with a $|\text{DFFITS}|$ 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.**

Rain example:

Outlier detection:

Hat diagonals:

DFFITS: