HUDM 5123 - Linear Models and Experimental Design
Notes 02 - Regression Diagnostics

1 Introduction

Let index i denote the study participant, $i = 1, 2, \ldots, n$. For outcome variable Y_i, predictors, $X_{i1}, X_{i2}, \ldots, X_{ip}$, and error term ϵ_i, the model is as follows.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

where β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_p$ are multiple regression slopes. As we saw last class, the four assumptions required for valid inferences about regression coefficients in multiple linear regression are related to the error term.

1. **Linearity.** $E[\epsilon_i] = E[\epsilon | X_{i1} = x_{i1}, X_{i2} = x_{i2}, \ldots, X_{ip} = x_{ip}] = 0$ for each $i \in 1, 2, \ldots, n$.

2. **Constant variance.** $\text{Var}(\epsilon_i) = \text{Var}(\epsilon | X_{i1} = x_{i1}, X_{i2} = x_{i2}, \ldots, X_{ip} = x_{ip}) = \sigma^2_\epsilon$. In other words, at all values of the predictors, the variance of the error, ϵ_i, is the same, σ^2_ϵ.

3. **Normality.** $\epsilon_i \sim N(0, \sigma^2_\epsilon)$. In other words, the conditional distribution of the residuals, ϵ, given the predictors, is normal. That is $\epsilon_i | X_{i1} = x_{i1}, X_{i2} = x_{i2}, \ldots, X_{ip} = x_{ip} \sim N(0, \sigma^2_\epsilon$.

4. **Independence.** $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for each $i \neq j$. The errors for each unit are assumed to be independent.

The four assumptions may be written concisely as follows:

$$\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2_\epsilon\right),$$

which is read as "the error term for the ith unit is independent and identically distributed (iid) with a normal distribution with mean zero and variance σ^2_ϵ."

An additional condition that is not theoretically required but is essential for estimation of coefficients and their standard errors is that there is no multicollinearity. Recall from last class that the solution to least squares estimation of the regression coefficients is given by the normal equations:

X: design matrix

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}.$$

Note that the solution requires taking the inverse of the matrix $\mathbf{X}^\mathbf{T}\mathbf{X}$. A well-known theorem from linear algebra states that a square matrix of the form $\mathbf{X}^\mathbf{T}\mathbf{X}$ is invertible if, and only if, the columns of the matrix $\mathbf{X}$ are *linearly independent*. This means that no predictor variable can be expressed as a linear combination of the remaining variables in the model. This could be violated if, for example, there were three predictors, X_1, X_2, X_3, in the model that were individual item responses ($0 =$ incorrect; $1 =$ correct) and a predictor, X_4, that was the total sum score of those three items was also included. Then $X_4 = \frac{1}{3} * X_1 + \frac{1}{3} * X_2 + \frac{1}{3} * X_3$. Thus, while no multicollinearity is not a required theoretical assumption for multiple regression, it is a practical consideration that is required for stable estimation of beta coefficients.

Finally, the presence of outliers can lead to drastic changes in regression parameter estimates if the outliers are *influential*. Thus, while the presence of outliers doesn't necessarily violate any of the assumptions required for statistical inference, it warrants further investigation.

2 Unusual and Influential Data

Following Fox, chapter 11, we will discuss leverage, discrepancy, and influence. Definitions:

- A point's *leverage* is the extent to which it has the potential to be influential on the regression slopes based on its predictor values. Leverage has nothing to do with the outcome value of the point; only with its predictor value(s).

- A point's *discrepancy* is the extent to which a point's outcome value falls outside of the expected cloud of outcome values, given its predictor values. A point with high discrepancy has an unusual outcome value for its predictor profile.

- A point's *influence on the regression slope coefficients* is determined, roughly speaking, by its leverage and discrepancy as in the following heuristic:

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}$$

That is, points with high leverage have the *potential* to be influential, but only are if they are coupled with high discrepancy. Likewise, points with high discrepancy have the *potential* to be influential, but only if they are coupled with high leverage.

https://en.wikipedia.org/wiki/Leverage_(statistics)

2.1 Measures of Leverage, Discrepancy, and Influence

https://learnche.org/pid/least-squares-modelling/outliers-discrepancy-leverage-and-influence-of-the-observations

The *hat-value* is a measure of leverage in OLS regression. In simple regression, for the ith point, h_i measures the distance from that point to the mean of the predictor variable. In multiple regression, for the ith point, h_i measures the scaled distance from the centroid (point of means) of the predictor variables, taking into account the variances and covariances of the predictors; it is very closely related to the Mahalanobis distance. The higher the hat-value, the more leverage the point has.

To detect outliers (i.e., points with high discrepancy), we need a quantitative measure of how unusual the Y value of a point is based on where its predictors are situated. While we might consider using the residual value for each point, high leverage values tend to have small residuals because they pull the regression prediction surface in their direction. A solution is to delete the ith observation, so that it will not impact the prediction surface, and then calculate the value of the standardized residual (residual divided by its standard error). These quantities are referred to as *studentized residuals* because they follow Student's t-distribution. The larger the studentized residual, the more discrepant (outlying) the point.

Cook's distance is a measure of the influence of a point. It may be though of as a measure that takes the product of a point's discrepancy and its influence.

https://en.wikipedia.org/wiki/Cook%27s_distance

2.2 An Example: Simple Linear Regression

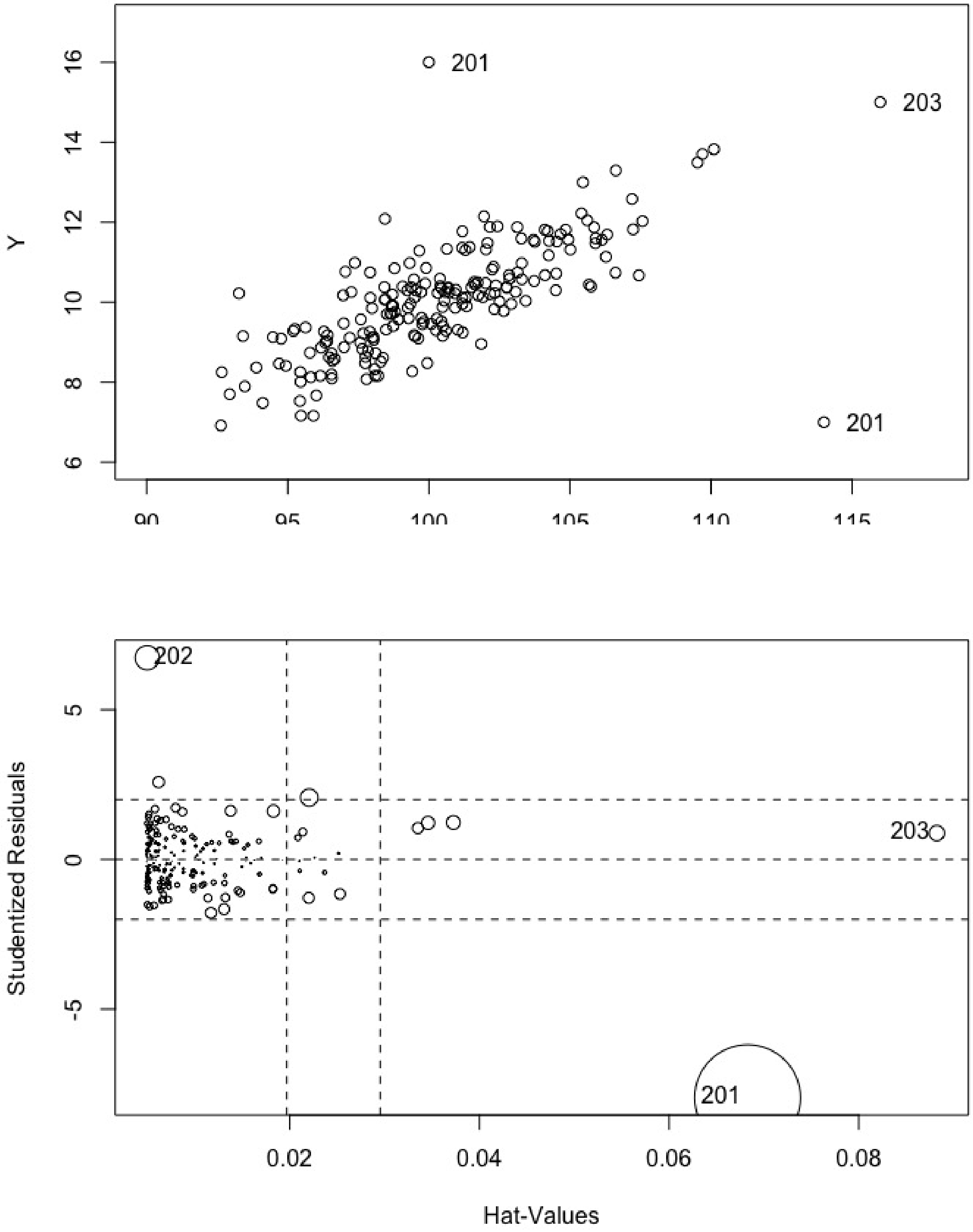

Figure 1: The diagnostic plot (lower panel) displays hat-values on the horizontal axis, studentized residuals on the vertical axis, and the area of each circle represents Cook's distance; vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the Studentized-residual scale; the plot was produced by a the `influencePlot()` function in package **car**

2.3 Should Unusual Data Be Thrown Out?

While it is important to detect unusual data, especially if it is influential, it should not be automatically thrown out.

- One of the most important reasons to identify unusual data is to identify data entry mistakes, impossible values, and the like. An example that comes to mind involves a data set I worked with where I identified one weight value that was impossibly low given the height and age for the subject. The weight was supposed to have been reported in pounds but I suspected this person might have entered his weight in kilograms instead. When I made the conversion, the weight fell in line with expectations. In other cases, I have seen people haphazardly enter things like 999 in order to answer a survey question without actually answering it. Clearly, it is important to detect these kinds of aberrations and treat them as what they are: missing data.

- In other cases, the presence of outliers might motivate you to think about whether the model you specified is correct. Might there be other important predictor variables that were left out of the model that should be included to help explain the outlying data? Or perhaps a different functional form is called for (more on this in subsequent classes).

- It is also important to note that the impact of a single unusual point on the estimated regression slopes will decrease as the total sample size increases. Thus, we might justifiably be a little less concerned about the impact of influential points when working with very large samples.

3 Diagnosing Violations of Assumptions about the Error Term: Non-Normality, Nonconstant Error Variance, and Nonlinearity

3.1 Non-Normality

For a particular population, as sample sizes get larger, the estimates of regression coefficients and their standard errors may be estimated with smaller and smaller error even when the assumption of normally distributed residuals is violated. So, why should we care about the assumption? First, because good behavior is not guaranteed in small samples. Second, because the *efficiency* of least squares estimation is not guaranteed when normality is violated. An estimation procedure is said to be *efficient* if, as sample size goes up, it converges to the population parameters as fast or faster than any competing estimators. If normality is violated, this property is no longer guaranteed.

In particular, error distributions that are heavily skewed can be problematic because (a) they generate outliers and (b) the mean is not a very meaningful measure of the center of a highly skewed distribution. In those cases, we might try to transform the data to produce a symmetric error distribution before analyzing with OLS regression.

3.1.1 Density Plots and Histograms

One straightforward way of assessing the normality assumption is to plot the density of the studentized residuals and compare it with the normal distribution. With large samples, histograms with many bins are useful. In small to moderately sized samples, smooth kernel density plots are useful.

https://en.wikipedia.org/wiki/Studentized_residual

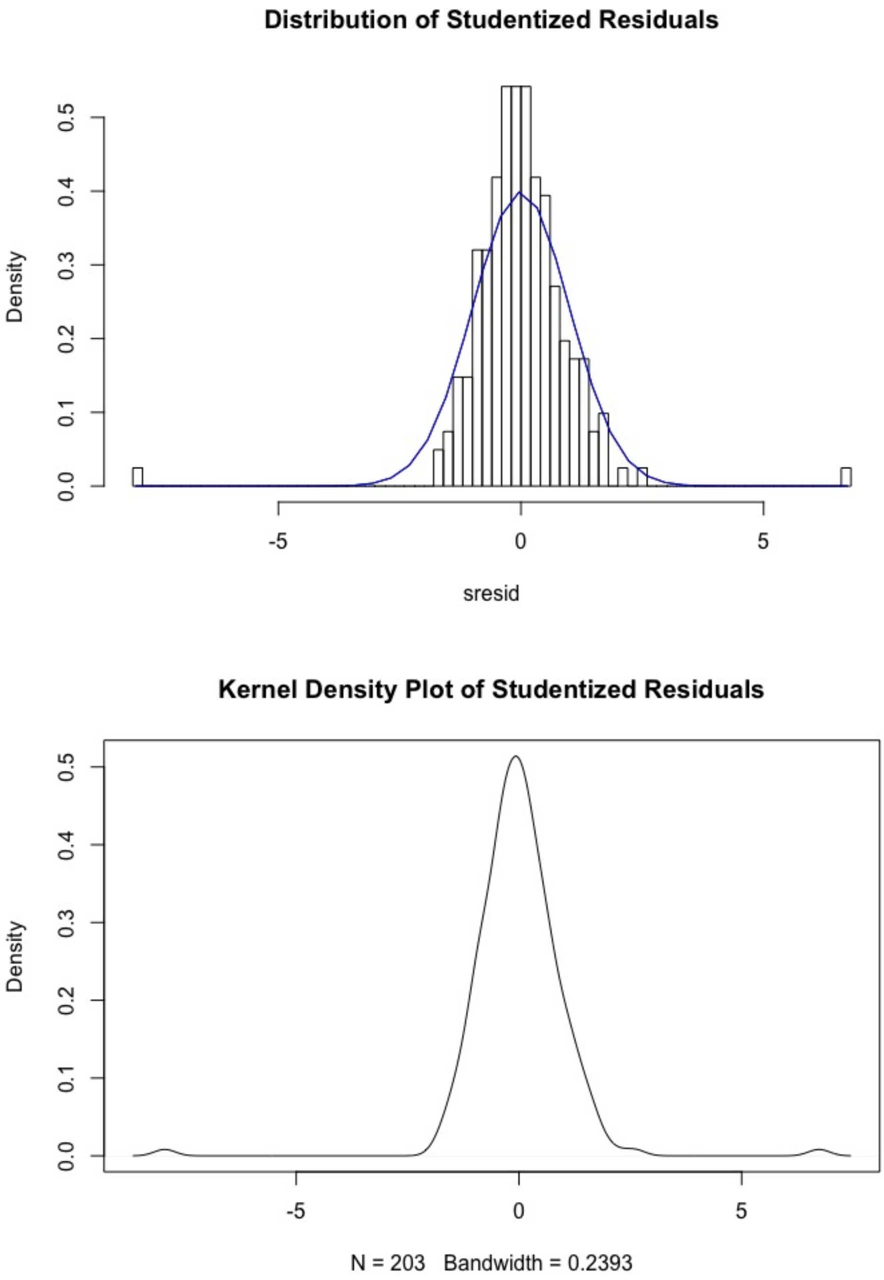

Figure 2: Histogram of studentized residuals with normal curve overlay (upper panel); kernel density plot of studentized residuals (lower panel)

3.1.2 Quantile-Quantile Plots

A quantile-quantile plot (or QQ plot) is a graphical method for comparing two distributions. In theory, the studentized residuals from a linear model fit to data that have normally distributed residuals should follow a t distribution with $n - k - 2$ degrees of freedom. The quantile-quantile plot compares the (ordered) studentized residuals to the theoretical t distribution. To the extent that the points correspond, they will lie along a 45 degree line. To the extent that the residuals differ from the theoretical expectation, they will fall off of the line.

The QQ plots generated by package **car** (stands for "**c**ompanion to **a**pplied **r**egression") also include a 95% confidence band based on simulated sampling. Those points that fall outside of the 95% confidence band may be interpreted as particularly unlikely under the assumption that normally distributed errors holds true. Thus, if we see a lot of points outside the confidence band, we might question the validity of the assumption.

Furthermore, if we see systematic diversion from the 45 degree line, we might suspect a particular type of diversion from normality. Examine the uppermost panel in the plot below, for example, and you will see that the distribution of the raw data is skewed to the right. This shows up in the QQ plot as tails that rise above the 95% confidence band on both ends; left skew is the opposite: tails that fall below. Similar patterns can be observed with heavy and light-tailed data.

Chapter 4 of the Fox book is entitled "Transforming Data" and has myriad strategies for how to transform both outcome and, in some cases, predictor variables to reduce problems such as non-normality. We won't spend much time covering these strategies explicitly in class, but if you encounter a data set for which normality is violated in problematic way, consider looking over the detailed instructions in chapter 4 for advice on how to transform your way out of it.

3.2 Non-constant Error Variance

Similar to the normality assumption, coefficients will still be consistently estimable if the assumption of constant error variance (also called *homoskedasticity*) is violated, but they will no longer be most efficient. Residual plots are most useful for detecting problems related to non-constant error variance (also called *heteroskedasticity*). It is not uncommon to see error variance grow as the average value of the outcome gets larger. While we could plot residuals against the observed outcome to examine this, the plot would be tilted because of built in correlation between the two. Instead, it is useful to plot the residuals against the predicted outcome values because they are uncorrelated by design.

What should we do if we detect non-constant error variance? One option is to use heterogeneity-robust standard errors (also called Huber-White standard errors; see Fox, p. 275 for more details).

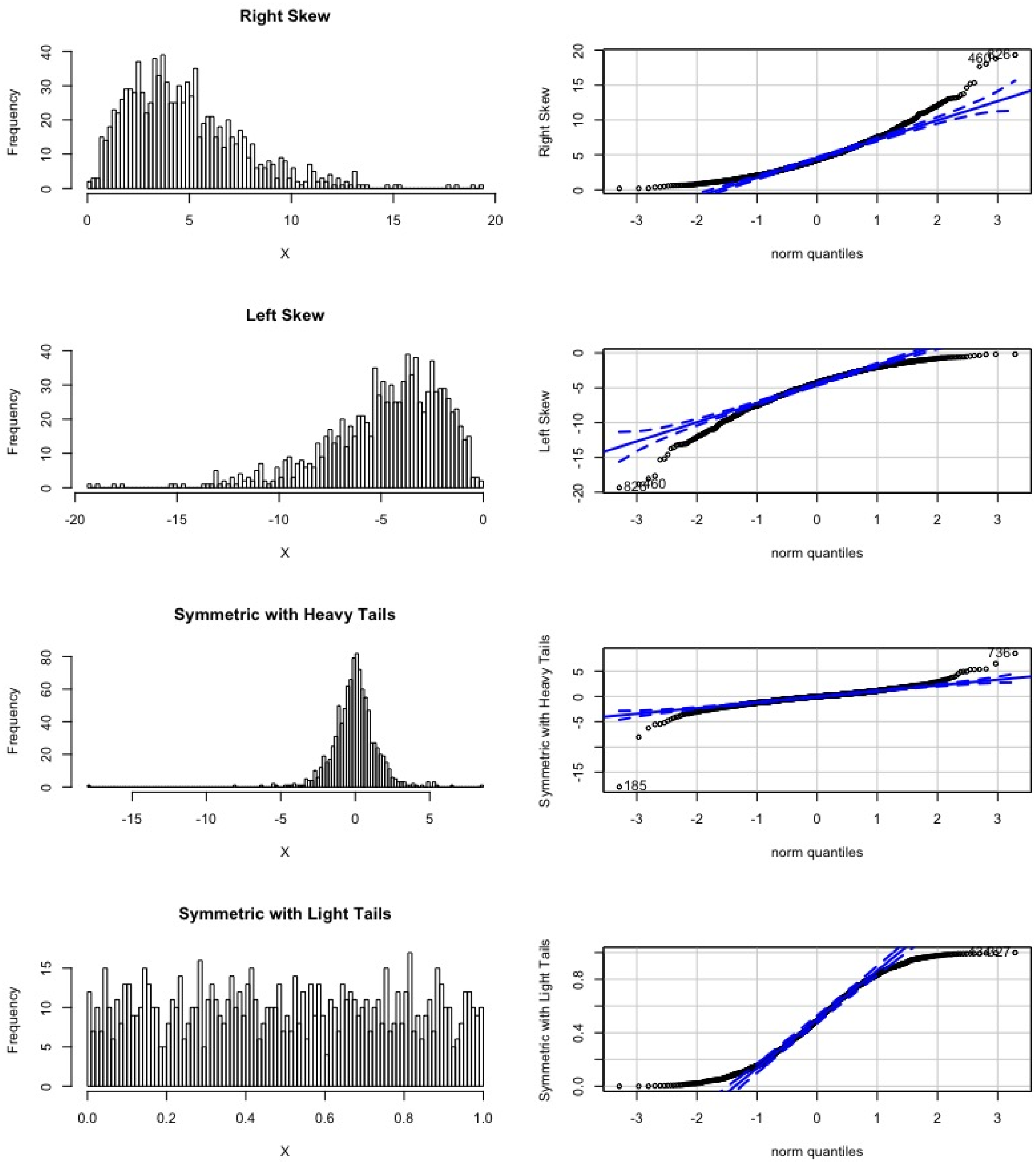

Figure 3: Examples of QQ plots with data generated from four different non-normal distributions

3.2.1 Plots of Residuals vs Fitted Values

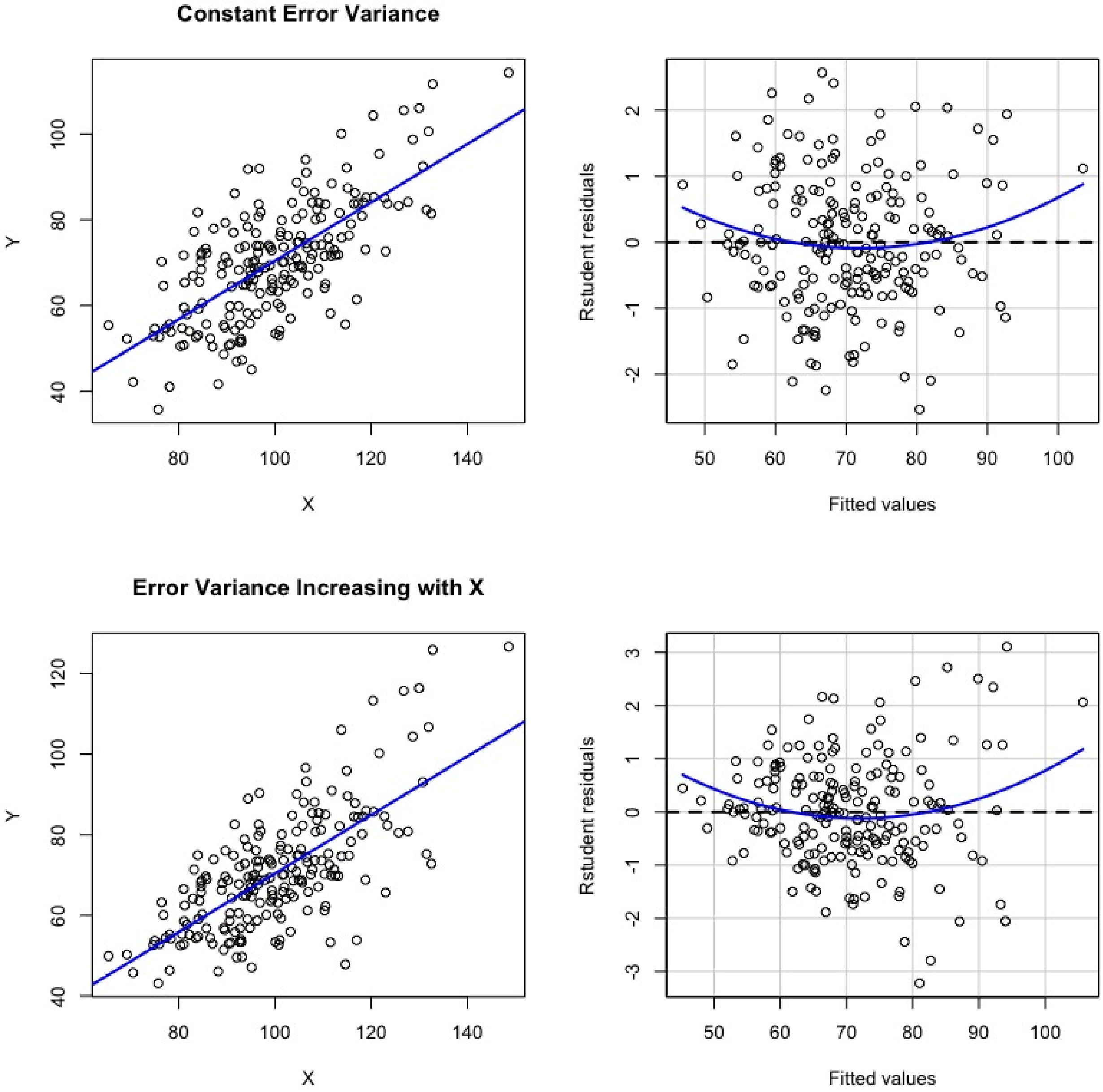

Figure 4: Examples of two data sets and their residual vs fitted values plots

3.3 Detecting Non-Linearity

Non-linearity is present when, by definition, the average value of the error term is not zero over some portion of the predictor space. Two-dimensional plots of each predictor and the outcome are a useful starting point for examining the linearity assumption.

However, we are ultimately interested in the partial relationships between each predictor and the outcome, after conditioning on the other predictors in the model. For this

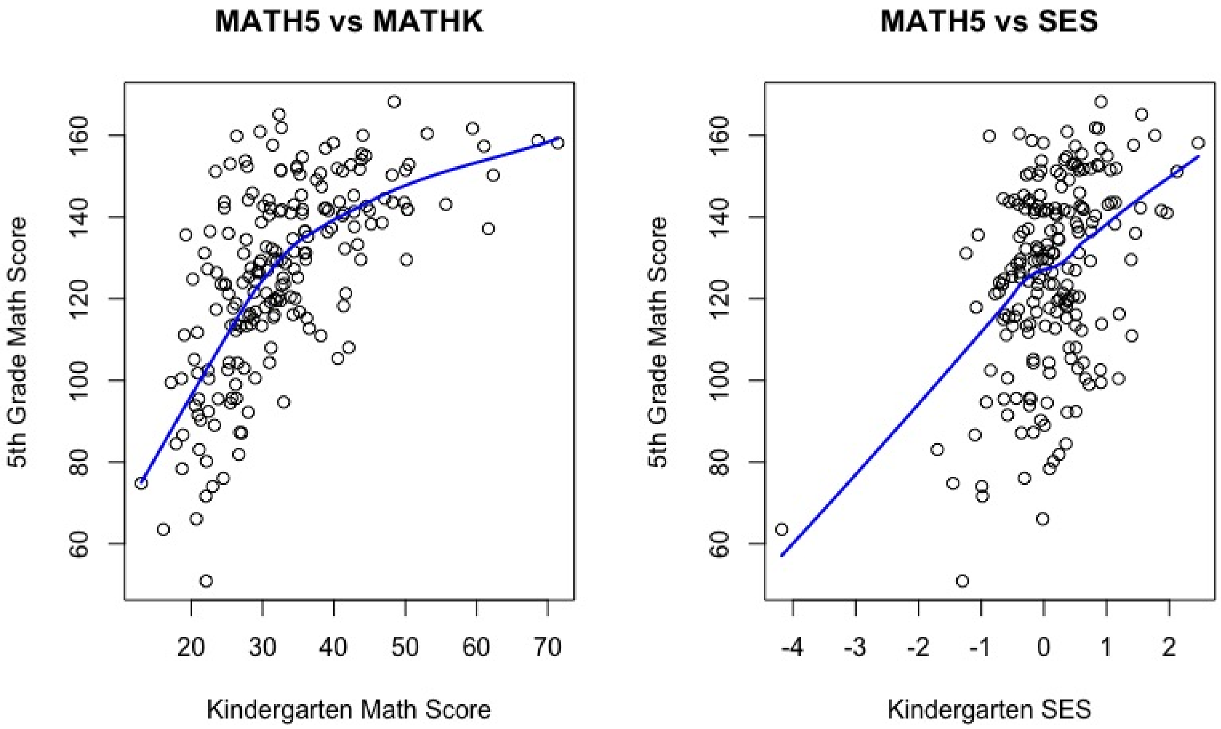

Figure 5: Scatterplots showing the two-dimensional marginal relationship between predictors MATHK and SES and the outcome MATH5

reason, residual plots are more useful for detecting nonlinearity in multiple regression. However, residual plots cannot distinguish between monotone and non-monotone nonlinearity. *Component-plus-residual plots* are an effective alternative because they simultaneously display (a) trends in the residuals and (b) overal linear trend in the data. The component-plus-residual plots for the predictors MATHK and SES are shown in the plots below.

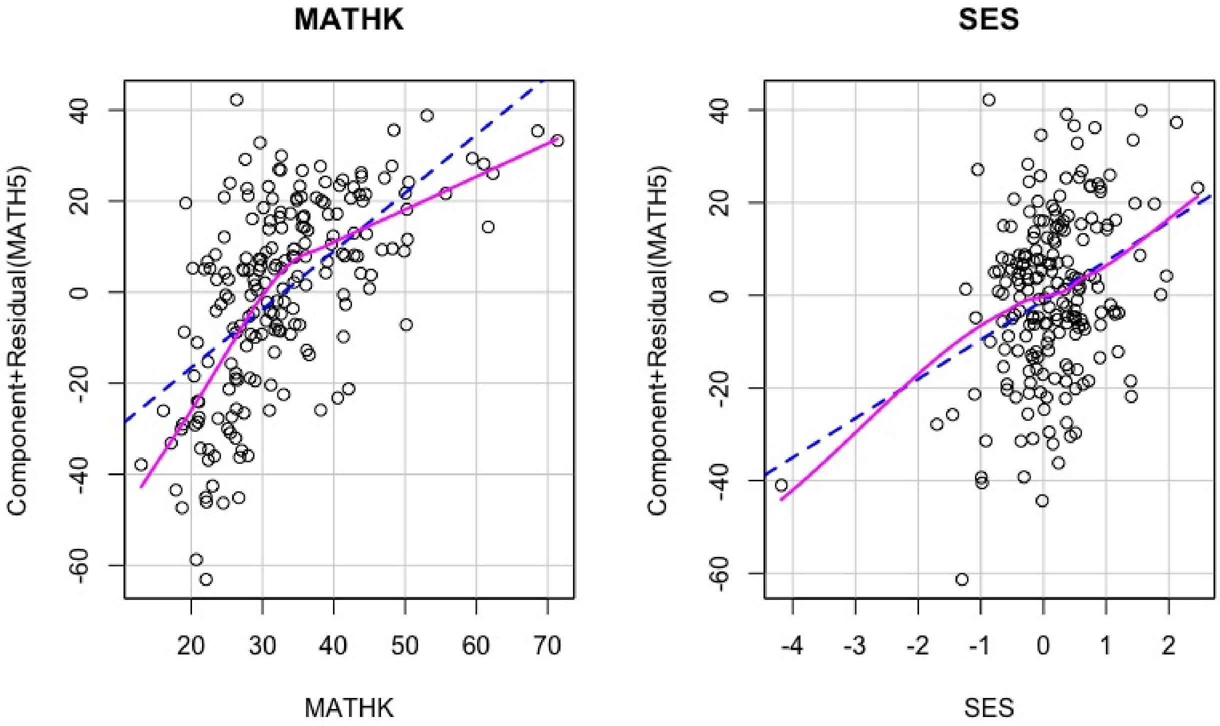

Figure 6: Component-plus-residual plots for predictors MATHK and SES with non-parametric regression overlays

The component-plus-residual plot for MATHK shows a monotonically increasing relationship with downward curvature. One solution would be to add a quadratic term to the model for MATHK to try to pick up on the curved relationship between MATHK and MATH5.

4 Detecting Multicollinearity

Perfect multicollinearity among predictor variables will render cause OLS slopes to be inestimable. However, you can also get into trouble running a multiple regression if there is *approximate* (as opposed to perfect) multicollinearity. Consider the sampling variance of the estimated slope coefficient $\hat{\beta}_j$ in multiple regression.

$$\text{var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j,)^2}$$

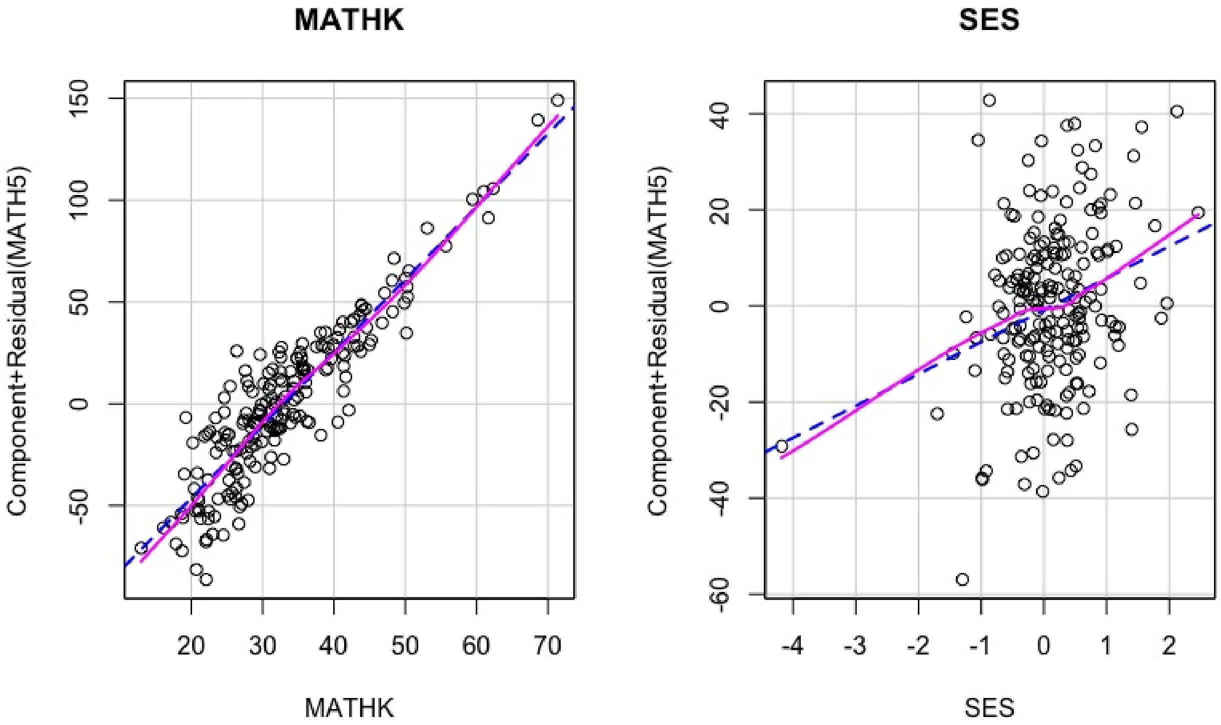

Figure 7: Component-plus-residual plots for predictors MATHK and SES with nonparametric regression overlays after adding a quadratic term to the model for MATHK

where R_j^2 is the squared multiple correlation from the regression of X_j on all the other predictors. The first term is called the *variance inflation factor*, or VIF. The VIF for a predictor X_j ranges from a minimum possible value of 1 when the other predictors have no explanatory power in predicting X_j (i.e., $R_j^2 = 0$), to a maximum possible value of positive infinity when the other predictors can perfectly predict X_j (i.e., $R_j^2 = 1$). The variance inflation factor is strong, therefore, when the predictor X_j is strongly correlated with the other predictors in the model. Note that a very large VIF will make the sampling variance of the slope coefficient β_j very large. Thus, the problem with approximate collinearity is that it can inflate the variance of slope coefficients, which can be problematic for estimation.

Consider the following reconfiguration of the sampling variance.

$$\text{var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$$

$$\text{var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\epsilon^2}{\frac{n-1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$$

$$\text{var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\epsilon^2}{(n-1)s_{X_j}^2}$$

Beyond the VIF, we can see the other sources that contribute to the magnitude of the sampling variance of a regression coefficient. Large error variance, σ_ϵ^2, small sample size n, and small variance for the predictor X_j all cause the sampling variance to be larger.

4.1 Variance Inflation Factor

It is not sufficient to examine the correlation matrix for predictors when looking for multicollinearity because multicollinearity may occur across more than two variables. Recall the example above for which three indicators were summed up to create a fourth. None of the indicators would be perfectly correlated with their sum, but all three combine to perfectly predict the sum.

The variance inflation factor, defined above, is a useful measure of the extent to which the variability in a given predictor can be explained by the remaining predictors.

$$\text{VIF} = \frac{1}{1-R_j^2},$$

where where R_j^2 is the squared multiple correlation from the regression of X_j on all the other predictors. If all the variability in X_j is perfectly predicted by the remaining predictors in the model, then R_j^2 will be 1 and the VIF will be infinitely large. If a lot (but not all) of the variability of X_j is explained by other predictors in the model, then R_j^2 will be close to 1 and the VIF will be large, but not infinite. If, on the other hand, the other predictors in the model explain none of the variability in X_j, the VIF will be 1, the lowest value possible.

4.2 Diagnosing Multicollinearity with VIFs

There are various rules of thumb for how large a VIF needs to be to cause concern that including that predictor in the model will unduly inflate standard errors of regression coefficients. A multiple R^2 of .35 is considered “large” by Cohen’s (1988) guidelines. The corresponding VIF is 1/(1 - .35) = 1.54.

R_j^2	VIF
.30	1.43
.40	1.67
.50	2.00
.60	2.50
.70	3.33
.80	5.00
.90	10.00
.95	20.00
.99	100.00

If you search you will find a number of rules of thumb for how large a VIF is too large, but there is no absolute authority on the issue because the rules are all subjective. Most will agree that a VIF of 10 is too large, but between 1 and 10 opinions vary, with some going as low as 2.5. As Fox points out, there is no "quick fix" for multicollinearity. If you encounter it with your data, one option is to drop an offending variable and respecify the model. If the variable is not expendable, then you may need to consider how to combine variables in the model to minimize potential for multicollinearity.