HUDM 5123 - Linear Models and Experimental Design Notes 01 - Simple and Multiple Linear Regression and Assumptions

1 Introduction

Linear regression refers to a method of modeling the relationship between one or more explanatory variables and a single outcome variable. The explanatory variables in regression may also be referred to as predictors, covariates, or as X variables, and the outcome variable may also be referred to as a dependent variable or Y variable. The difference between simple and multiple linear regression is that simple refers to the case where there is only one explanatory variable; multiple linear regression refers to the case where there are at least two explanatory variables. It is also possible to simultaneously model linear relationships between one or more predictors and more than one dependent variable. In that case, linear regression is referred to as multivariate. Note the distinction between multiple regression and multivariate regression: multiple means one or more predictors and a single outcome; multivariate means one or more predictors and more than one outcome. We will focus on multiple, not multivariate, regression in this course.

1.1 Notation

I will generally use Greek letters for population parameters, such as β_0 for the intercept in a linear regression model. I will use letters from the English alphabet for observed variables. For example, $Y_1, Y_2, \ldots, Y_n$ may be used to represent the values of the observed outcome. Boldface lowercase letters from the English alphabet will be used for column vectors. For example,

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}.$$

I will use boldface uppercase letters from the English alphabet for matrices. For example,

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nn} \end{bmatrix}.$$

Hats will be used to represent estimated values of parameters. For example, $\hat{\beta}_1$ represents the estimated value of the population parameter β_1. The sample mean of a vector of n numbers, say $Y_1, Y_2, \ldots, Y_n$, will be expressed with an overbar, and is defined as follows.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

The sample variance of a vector of n numbers, say $Y_1, Y_2, \ldots, Y_n$, will be expressed as s_Y^2, and is defined as follows.

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 .$$

The sample covariance between two vectors of n numbers, say $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, will be expressed as s_{XY}, and is defined as follows.

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left[\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right] .$$

The sample correlation between two vectors will be expressed as r_{XY} and defined as follows.

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} .$$

2 Simple Linear Regression

A linear regression model is referred to as "simple" if there is only one predictor variable. Let Y represent 5th grad math achievement score, and let X represent kindergarten math achievement score. A linear model fit to a sample of n student data points to predict 5th grade math based on kindergarten math has the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i ,$$

where $i = 1, \ldots, n$.

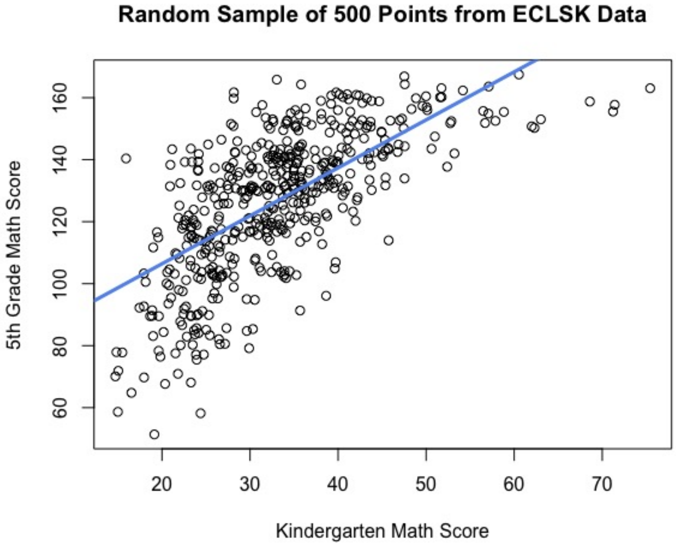


Figure 1: Scatterplot of kindergarten and 5th grade math scores for a random sample of 500 cases from ECLSK data with simple linear regression fit

The estimates for the regression coefficients in Figure 1 are $\hat{\beta}_0 = 75.5$ and $\hat{\beta}_1 = 1.5$. The prediction equation is obtained by substituting the estimates in to the original linear model.

$$\hat{Y}_i = 75.5 + 1.5X_i$$
$$\widehat{\text{math5}}_i = 75.5 + 1.5\text{mathK}_i$$

Where do those estimates come from? Why choose 75.5 and 1.5? Why not another combination of numbers? For example, what if I choose $\hat{\beta}_0 = 50.5$ and $\hat{\beta}_1 = 1.5$, or $\hat{\beta}_0 = 120.0$ and $\hat{\beta}_1 = -2$? See Figure 2 for graphical plots that show these lines with the ECLSK data. The answer has to do with how we frame the problem. What is our mathematical goal when fitting a straight line to these data? We want to get the best fitting line possible. One way to define best fitting is in reference to the residuals. For each point of data, the residual for a given linear model fit to the data is defined as the difference between the observed data point and the predicted data point. That is,

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$
$$\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

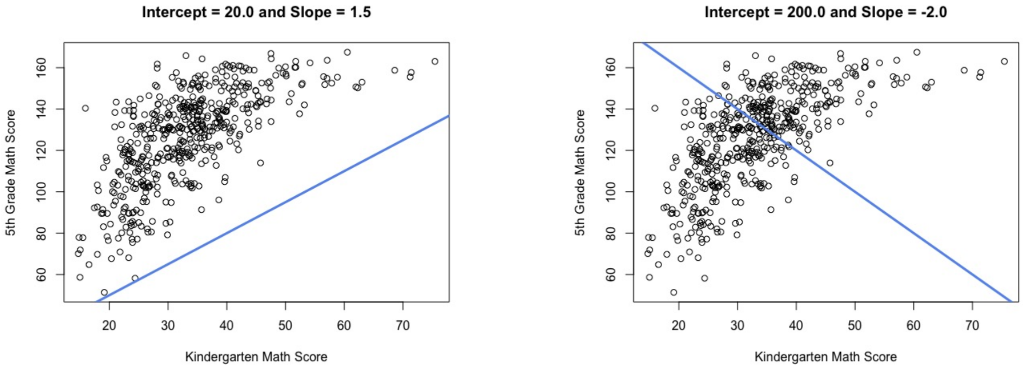
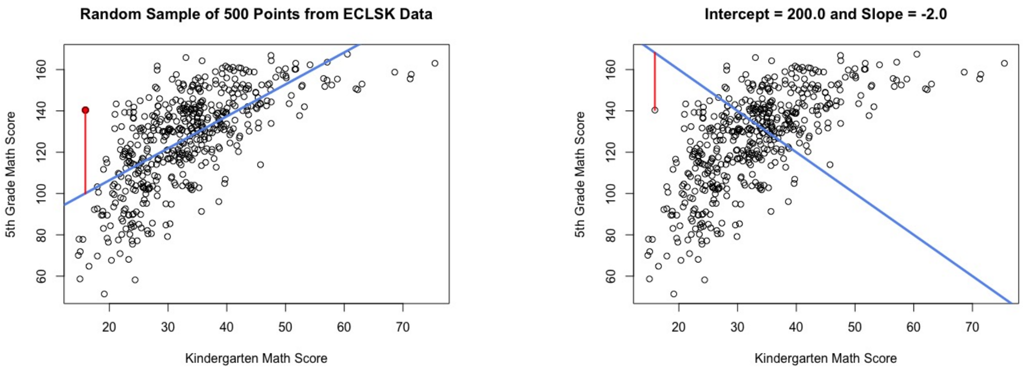


Figure 2: Scatterplots of kindergarten and 5th grade math scores for a random sample of 500 cases from ECLSK data with alternative linear fits

Notice the outlying point at approximate coordinates (15, 140), highlighted in the left panel of Figure 3. Using the line of best fit, this student's residual is 140 - 99 = 41 points. Using the third solution, with $\hat{\beta}_0 = 120.0$ and $\hat{\beta}_1 = -2$, the predicted value is about 168, which gives a residual of -28. Notice that, even though the line of best fit provides a much better fit to the data overall, it yields a larger residual for case 23 than the arbitrarily selected line with intercept 200 and slope -2. To judge fit for the entire sample as a whole, we need a quantitative measure that summarizes fit for all data points. The residual sum of squares (RSS), also called the sum of squared errors or sum of squared residuals is the approach most commonly used.

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$$



Random Sample of 500 Points from ECLSK Data

Intercept = 200.0 and Slope = -2.0

Figure 3: Residuals for case number 23 based on the best fit (left panel) and an arbitrarily selected fit (right panel)

Note that RSS is a function of $\hat{\beta}_0$ and $\hat{\beta}_1$. One way to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize RSS is to take partial derivatives of RSS with respect to both $\hat{\beta}_1$ and $\hat{\beta}_1$ and set them equal to zero, which gives the normal equations.

$$\frac{\partial \text{RSS}}{\partial \beta_0} = -2\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] = 0$$

$$\frac{\partial \text{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] = 0$$

Solving the normal equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.$$

2.1 A Note on Interpretation

The estimated regression coefficients in a simple linear regression are interpretable as follows. The intercept, $\hat{\beta}_0$, is the predicted outcome value for a unit with a score of 0 on the predictor variable. The slope, $\hat{\beta}_1$, is the change in the predicted value of the outcome associated with a one unit increase in the predictor.

2.2 Simple Correlation

To define correlation, we first have to define how we measure variability in the outcome variable, Y_i. The first way to measure variability in the outcome variable is based on its mean, $\bar{Y}$. For the ECLSK data, the mean of the 5th grade math score is $\bar{Y} = 127$. The sum of squared deviations from the mean of the outcome variable, also referred to as sum of squares total (SST) or total sum of squares (TSS), can be expressed as follows:

$$\text{TSS} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 .$$

The residual sum of squares, defined above, measures the variability of the outcome around the regression line:

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 .$$

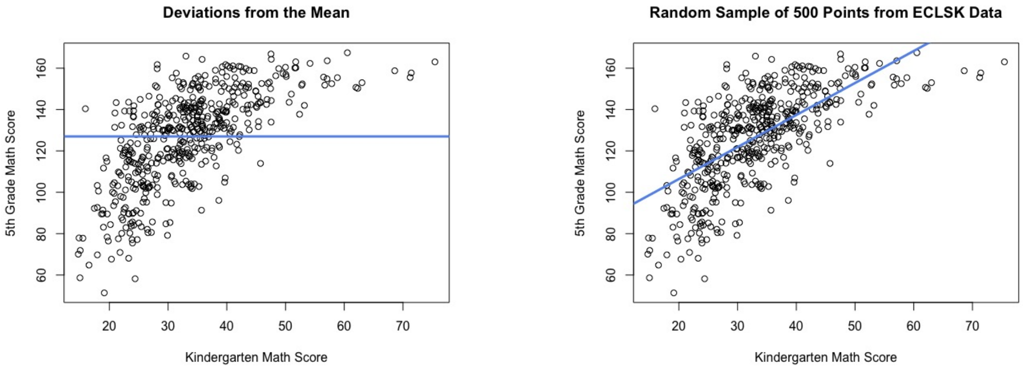


Figure 4: Total sum of squares are determined using the mean of the outcome, $\bar{Y}$, shown as a horizontal line in the left panel; residual sum of squares are determined using the regression line, $\hat{Y}$, shown in the right panel

The regression sum of squares (RegSS) is defined as the difference, TSS - RSS, and represents the amount of improvement in model fit based on using the best fitting regression line instead of the mean. The square of the correlation coefficient, also called the correlation ratio, is defined as follows.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = \frac{\text{RegSS}}{\text{TSS}},$$

and may be interpreted as the proportional reduction in squared error due to the linear regression. If the linear regression offers no improvement over the mean, $RSS = TSS$, and $R^2 = 0$. If the linear regression fits all the data perfectly (i.e., all the points fall on a straight line), then $RSS = 0$, and $R^2 = 1$. To find the value of the correlation coefficient, r, take the positive square root of R^2 if $\hat{\beta}_1 \geq 0$ and the negative square root of R^2 if $\hat{\beta}_1 < 0$.

2.3 Residual Standard Error

The formulas given above for $\hat{\beta}_0$ and $\hat{\beta}_1$ permit us to estimate the coefficients using data to plot the line of best fit and use it to make predictions. We may also be interested in the variance, or standard deviation, of the residuals. The residual variance is defined using the degrees of freedom for the model.

degrees of freedom = sample size (n) - number of parameters estimated in the model

For simple linear regression, there are two parameters estimated, β_0 and β_1, and the sample size is n. Thus, the degrees of freedom for simple linear regression is $n-2$. The *residual variance* may be estimated using the sum of squared residuals and the degrees of freedom as follows:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

The *residual standard error* is simply the square root of the residual variance.

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}.$$

3 Multiple Linear Regression

Suppose we wish to predict 5th grad math using more than one predictor so that, in addition to kindergarten math score, our new model will also use each student's socioeconomic status as measured in kindergarten. Then our model would have the following form.

$$\text{MATH5}_i = \beta_0 + \beta_1\text{MATHK}_i + \beta_2\text{SES}_i + \epsilon_i.$$

With two predictor variables, the multiple linear regression model fits a hyperplane to the data. The least squares estimates give the following prediction equation:

$$\widehat{\text{MATH5}}_i = 79.80 + 1.39\text{MATHK}_i + 5.27\text{SES}_i + \epsilon_i.$$

See Figure 5 for a representation of the prediction plane fit to the ECLSK data.

In general, the multiple linear regression model with p covariates can be expressed as follows.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \cdots + \beta_p X_i + \epsilon_i.$$

R^2 and estimated residual standard error may be calculated the same way as for simple linear regression. To estimate coefficients (betas) the normal equations may be created by taking partial derivatives with respect to the betas. The normal equations for multiple regression are more easily expressed using matrix notation. The multiple regression model in matrix notation is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \quad \mathbf{X}_{n\times p+1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}; \quad \boldsymbol{\beta}_{p+1\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}; \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The residual sum of squares:

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

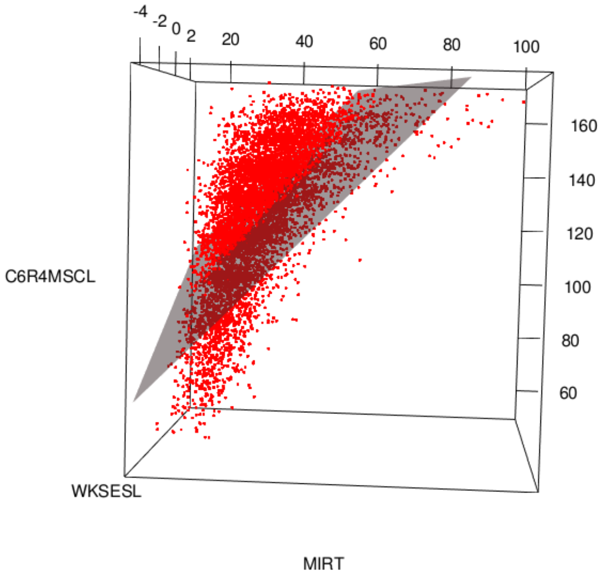


Figure 5: Multiple linear regression of 5th grade math score on kindergarten math score and SES

The normal equations may be shown to be as follows.

$$\mathbf{X^TX}\boldsymbol{\beta} = \mathbf{X^Ty}$$
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X^TX}\right)^{-1}\mathbf{X^Ty}$$

3.1 A Note on Interpretation

The estimated regression coefficients in a multiple linear regression are interpretable as follows. The intercept, $\hat{\beta}_0$, is the predicted outcome value for a unit with a score of 0 on all the predictor variables. The slope, $\hat{\beta}_1$, is the change in the predicted value of the outcome associated with a one unit increase in the predictor *while holding constant the values of the other predictors in the model*. The slope, $\hat{\beta}_2$, is the change in the predicted value of the outcome associated with a one unit increase in the predictor *while holding constant the values of the other predictors in the model.* And so on.

3.2 Assumptions for Statistical Inference

1. **Linearity.** $E[\epsilon_i] = E[\epsilon|X_{i1} = x_{i1}, X_{i2} = x_{i2}, \ldots, X_{ip} = x_{ip}] = 0$ for each $i \in 1, 2, \ldots, n$. This implies that the mean of the outcome really does follow the linear functional form specified by the model. This would be violated, for example, if the true relationship between the mean of the outcome and the covariates is curvilinear but we model it only as linear.

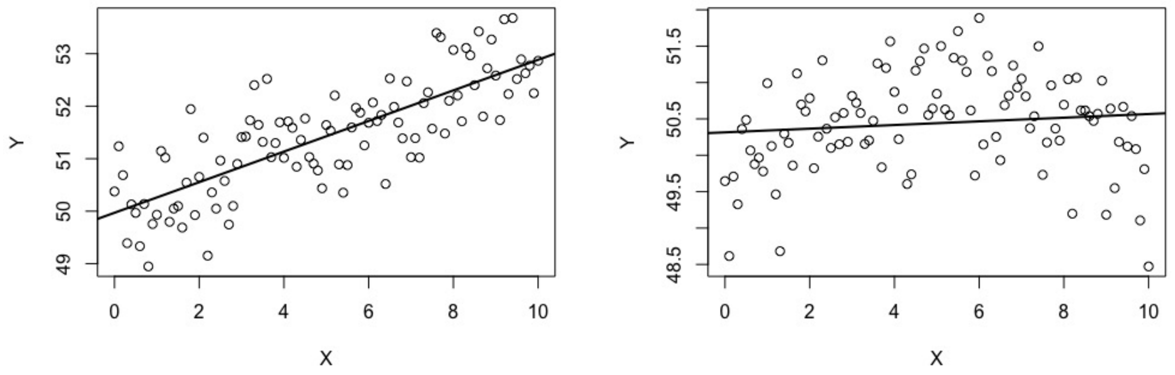


Figure 6: In the left panel, the linearity assumption is satisfied; in the right it is not. In the left panel, notice that the average value of the residuals is equal to the predicted value based on the regression line. In the right panel this is not the case; notice, for example, for $X_i = 0$, the mean of the errors (observed - predicted) is negative, while for $X_i = 5$, the mean of the errors is positive. This violates the assumption that the mean of the error term should be zero for any value or combination of values of predictors (i.e., linearity).

2. **Constant variance.** $\text{Var}(Y_i) = \text{Var}(Y|X_{i1} = x_{i1}, X_{i2} = x_{i2}, \ldots, X_{ip} = x_{ip}) = \sigma_\epsilon^2$. In other words, at all values of the predictors, the variance of the outcome Y is the same, σ_ϵ^2.

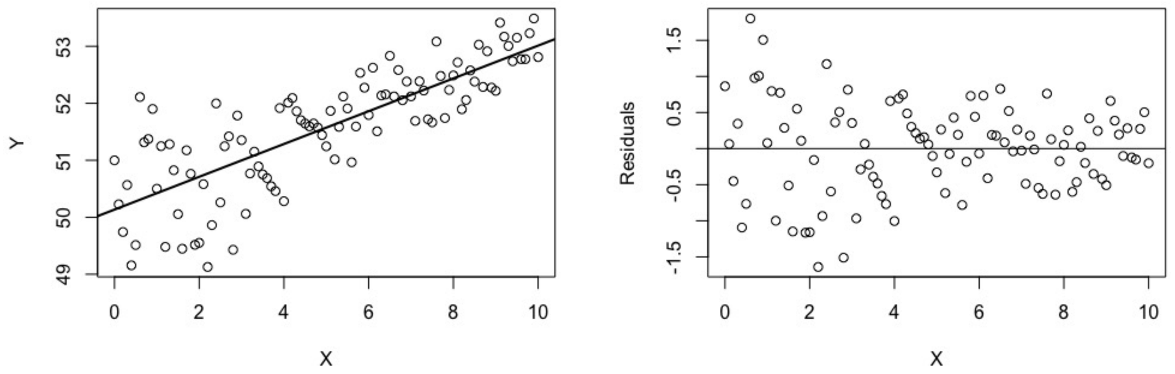


Figure 7: In the left panel, note that the variance of the outcome, Y, is larger for values of X closer to zero. This violates the assumption of constant variance. The residuals are plotted in the right panel. Note that the residuals are more variable for smaller values of X.

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

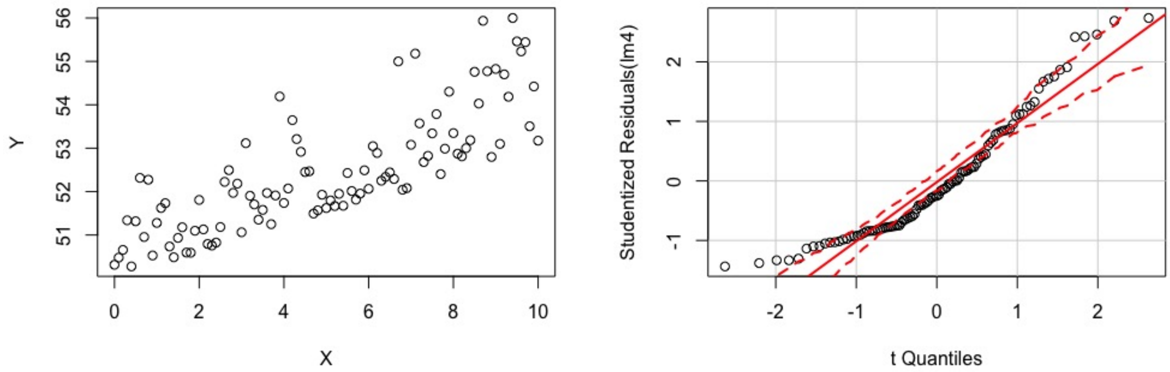


Figure 8: In the left panel, the errors are skewed upward. This violates the assumption of normality. In the right panel, a quantile-quantile (qq) plot is used to detect deviation from normality.

4. **Independence.** $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for each $i \neq j$. The errors for each unit are assumed to be independent. This assumption is hard to check with data. Instead, as we have discussed, it is typically assessed based on what is known about how the data were generated. Some examples of designs that would invalidate the assumption:

   - longitudinal or repeated measures data where certain units are assessed more than once,
   - nested data structures where units are clustered such as, for example, students in schools or patients in hospitals.