

Chapter 1 Simple Linear Regression (part 5)

1 Diagnostics for regression model

For the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

the assumptions are

- (L) Linearity of Regression function: $E(Y_i) = \beta_0 + \beta_1 X_i$ or equivalently $E\varepsilon_i = 0$
- (N) Normality of Error Terms: $\varepsilon_i \sim N(0, \sigma^2)$
- (I) Independent/ uncorrelated Error Terms: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, if $i \neq j$
- (E) Equal (constant) Error Variance: $\text{Var}\{\varepsilon_i\} = \sigma^2$

We should always check fitted models to make sure that these assumptions have not been violated. If some assumptions, then (1) the efficiency of estimators cannot be guaranteed; (2) some statistical inference (e.g. the test) is no longer correct; (3) the conclusions may not be correct.

- if nonlinearity exists, the parameter estimates are NOT valid. Solution: One can include polynomial terms to improve the fitting.
- if the variance is unequal, parameter estimates are valid, but the confidence intervals are misleading. Solution: We can consider the weighted Least squares estimation
- Outliers¹ far from the pattern of the rest of the Xs may affect the line. Solution: We can remove the outliers to improve the estimation

¹an outlier is an observation that is numerically distant from the rest of the data.

- dependent observations may reduce the efficiency of the parameters estimator, but again parameter estimates are valid. Solution: (to be discussed later)
- distribution of the random errors are not normally distributed: parameters estimates are still correct, but the confidence intervals are misleading. Solutions:
 - Including additional predictors sometimes solves this problem
 - Another solution is to transform Y
 - * $\ln(Y)$ or \sqrt{Y} draws in data skewed to high values
 - * $1/Y$ or $1/\sqrt{Y}$ draws in data skewed to low values
 - * use transformed Y instead of original Y
 - * interpret parameters according to transformed Y!

However, the violation and departures from the underlying assumptions cannot be detected using any of the summary statistics we've examined so far such as the t or F statistics or R^2 . In fact, tests based on these statistics may lead to incorrect inference since they are based on many of the assumptions above.

2 How do you check the assumptions?

In general ... plot your data!

- Simply plotting the data can be one of the most powerful model checking techniques
- From a simple plot of Y on X that includes the fitted regression line, we can check:

(1) linearity; (2) normality; (3) equal (constant) variance; (4) outliers, etc.

2.1 Linear relationship

- Is the model correct?
 - Is this the right line?
 - Are there outliers for which the model may be wrong?

- Assess with graphs: plot Y against X and the fitted line/model
 - If there are repeated observation on the each X_i : Note that the observations with the same X_i forms a sub-sample. We can estimate its mean. If the sub-sample mean is obviously different from the regression line, then the linear relationship may be violated.
 - If there is no repeated observation on each X_i , we can slice of the region of X and take each slice as from one sub-sample.

2.2 Independent observations/individuals

- Are all the individuals surveyed independent from one another?
- Cannot be assessed graphically or statistically
- Must know how the data were collected, for example, when the data are collected over time. (we shall discuss this later in Chapter 2)

2.3 Normally distributed errors

- At every value of X , the observed points should follow a roughly normal distribution centered at the fitted value of Y .
- Assess with residual plots

2.4 Equal variance

- At every value of X , the observed points should follow a roughly normal distribution with the same variance across all X s
- Assess with residual plots

2.5 4 types of assumption violations; see figure 1

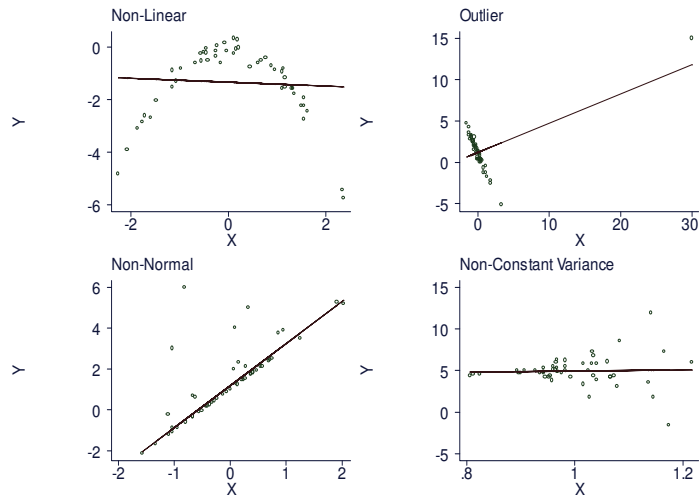


Figure 1:

3 Residual Analysis:

3.1 Standardized residuals

- Since our model states: $\varepsilon_i \sim N(0, \sigma^2)$, the standardized residuals,

$$\frac{e_i - 0}{\hat{\sigma}} \quad \text{where } \hat{\sigma} = \text{MSE}$$

should follow (roughly) a standard normal distribution (more exactly t -distribution)

3.2 Residual Analysis

If the model fits the data well, we expect:

- A histogram of the standardized residuals should look normal.
- Check for asymmetry and outliers.
- A plot of the residuals vs. X should look like a random scatter (no systematic relationship)
- A plot of the residuals vs. \hat{Y}_i (the fitted values) should also look like a random scatter.

3.3 Residual Analysis: Example plots

Example: Relationship between health status and pollution in 20 geographic areas; see figure 2

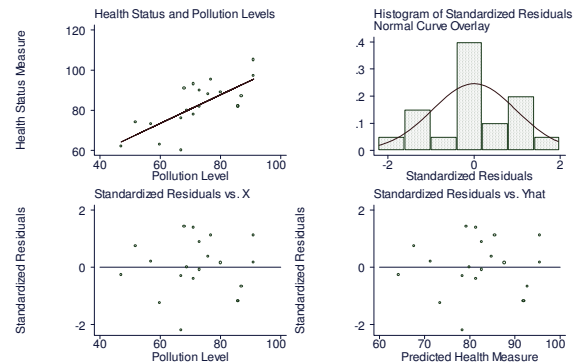


Figure 2:

Conclusion for the example

- Regression scatterplot looks good
- Standard Residuals appear fairly normally distributed
- Standard Residuals vs X appear randomly scattered (i.e. no apparent patterns & no extreme outliers)
- Standardized Residuals vs predicted values appear randomly scattered (i.e. no apparent patterns & no extreme outliers)

Example: Nonlinear Example plots; see figure 3

Conclusion for the example

- Regression scatterplot shows non-linear relationship
- Standardized Residuals dont look normally distributed
- Standardized Residuals vs X shows nonlinear relationship
- Standardized Residuals vs predicted values shows non-linear relationship

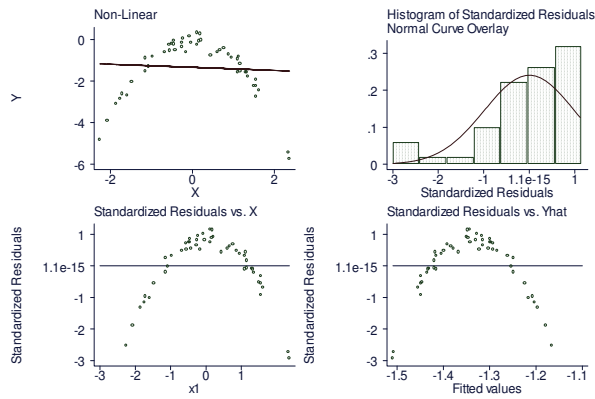


Figure 3:

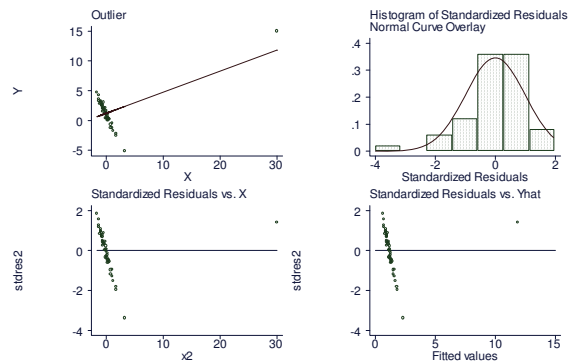


Figure 4:

Example: Outlier Example plots; see figure 4

Conclusion for the example

- Regression scatterplot shows outlier
- Standardized Residuals look normal but large residual present
- Standardized Residuals vs X shows a pattern & the outlier
- Standardized Residuals vs Y shows a pattern & the outlier

Example: Non-Normal Example plots; see figure 5

Conclusion for the example

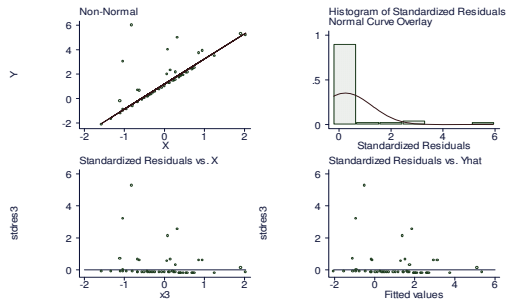


Figure 5:

- Regression scatterplot shows non-even spread
- Standardized Residuals dont look normally distributed
- Standardized Residuals vs X shows noneven spread
- Standardized Residuals vs predicted values shows non-even spread

Example: Unequal Variance Example plots; see figure 6

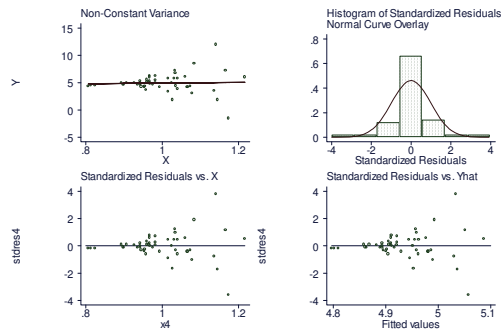


Figure 6:

Conclusion for the example

- Regression scatterplot shows increasing variability
- Standardized Residuals do look normally distributed
- Standardized Residuals vs X shows increasing variability

- Standardized Residuals vs predicted values shows increasing variability

Example 3.1 ((Data), (R code)) bus transit map example

Y - increase of daily bus ridership; X - number of free distributed bus transit maps

conclusion: the linearity assumption cannot be accepted; see figure 7

R code

```
busmap=read.table(file="busmap.dat")
reg=lm(busmap[,1] ~ busmap[,2]) # or reg=lm(busmap$V1 ~ busmap$V2)
plot(busmap[,2], busmap[,1], xlab="maps distributed",
      ylab="increase in ridership",
      xlim = c(80, 240), ylim=c(0, 8) )
lines(busmap[,2], reg$fitted)
plot(busmap[,2], reg$residuals, xlab="maps distributed",
      ylab="fitted residuals",
      xlim = c(80, 240), ylim=c(-1, 1) )
lines(c(80, 240), c(0, 0))
```

Plot of Y against X and Plot residuals against X

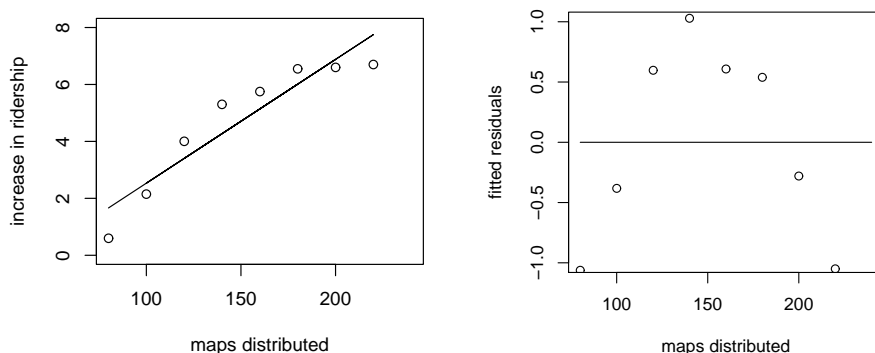


Figure 7:

Example 3.2 Wage Data; see figure 8

$$\hat{Y}_i = 8.38 + 0.04 \times Experience_i$$

Conclusion

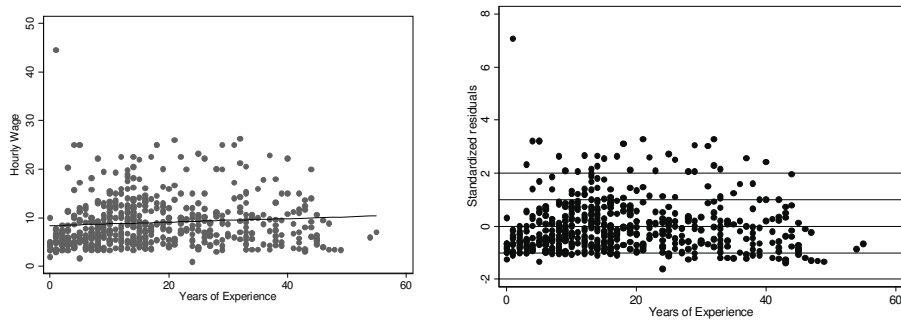


Figure 8:

- There seems to be a residual pattern for people with very little experience. (could add polynomials to allow for non-linearity)
- There is one outlier. (should check data entry and information for that person, or remove it from the data)
- The variance is unequal.
- The random errors are not normally distributed. here is the reason: consider

$$\text{Standardized residual} = \frac{e_i}{\sqrt{MSE}}$$

- *Standardized residual* > 2 standard deviations from 0 should happen only 5% of the time
- *Standardized residual* > 3 standard deviations from 0 should happen only 1% of the time
- The data were collected by randomly sampling workers, so independence can be assumed.