8 Residual analysis: checking model assumptions

We move on to discuss some practical aspects of regression analysis. Our first topic will be residual analysis: the general goal is to check if the modeling assumptions are adequate (correct). To remind, the general linear model is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \mathrm{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

We break this down into three separate assumptions:

1. *Linearity*: $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0} \Longleftrightarrow \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ (make sure you can prove this equivalence).
2. *Homoscedastic* (equal-variance) errors: $\mathrm{Var}\,(\epsilon_i) \equiv \sigma^2$ is the same for all $i = 1, \ldots, n$
3. *Uncorrelated* errors: $\mathrm{Cov}\,(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

In the normal model, also:

4. *Normality* of errors: $\boldsymbol{\epsilon}$ is multivariate normal $\Rightarrow$ $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ (combined with 2,3).

We will first present some basic tools for detecting "substantial" departures from assumptions 1-4. If such departures are detected, we will discuss some possible fixes. The main assumptions are linearity (Assumption 1) and homoscedasticity (Assumption 2); we will want to check these first.

Let us start intuitively with the case of simple linear regression (i.e., single predictor, $p = 1$). In that case we can look at a scatterplot of the data and try to visually detect violations of the linearity and homoscedasticity assumptions. The plot below shows 4 different simulated datasets. The fitted (LS) regression line is shown in blue in all four panels. In the top left panel the data was generated from a linear model with equal error variances. We see that there is roughly the number of points lying above and below the LS line throughout the range of the x values, i.e., if we take a small $\Delta > 0$, then the points in the dataset whose X values are in $(x - \Delta, x + \Delta)$, are roughly equally scattered above and below the LS line; this indicates good agreement with the linearity assumption. Furthermore, the *spread* of the points around the LS line is roughly constant throughout the range of the x values, i.e., the variance in Y for points in $(x - \Delta, x + \Delta)$ is more or less constant in the location x; this indicates good agreement with the homoscedasticity (equal-variances) assumption.

In the top right panel there is strong violation of the linearity assumption: the points show a clear trend in their location about the LS line: for x values with small absolute value, almost all points are below the LS line, and the trend is opposite for x values with large absolute value. However, there is still good agreement with the homoscedasticity assumption: the dashed black line represents a nonparametric regression fit; you can think of this as a smoothed version of an estimator that bins the points on the x axis, and takes the average of the Y values in each bin separately–which we considered as a *nonparametric* (or "general") regression method, as opposed to a *linear* regression method. The points indeed seem to have a constant spread (throughout the range of the x values) around this dotted line.

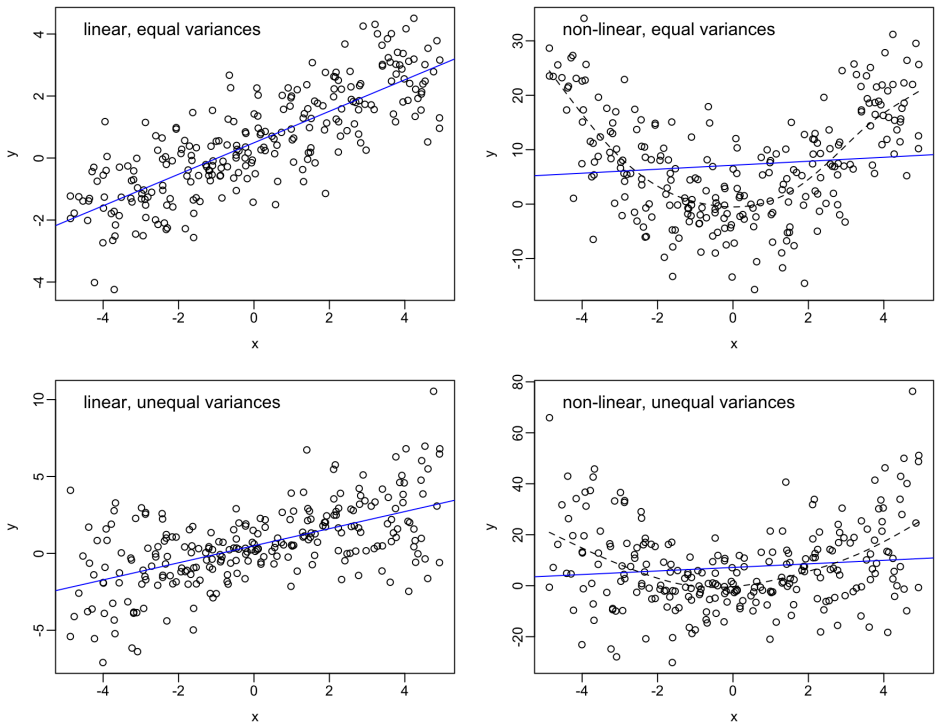


Figure 11: Violations of linearity and/or homoscedasticity

In the bottom right panel the data is simulated from a linear model, but the errors have unequal variances: this is shown clearly in the "fanning out" of the points as we move away from $x = 0$ in either direction. Lastly, in the bottom right panel both linearity and homoscedasticity are violated. The Normality assumption is harder to check by just visually examining the scatterplot; we will need some device to help us check normality. But even if we stick to the linearity and homoscedasticity assumptions, the situation is more complicated when dealing with *multiple* regression, because we cannot really visualize a scatterplot for the case of more than two predictors (this would require a plot in more than 3 dimensions). The workaround wil be to consider plots of the *residuals* instead of scatterplots. In fact, residual plots are usually more convenient to inspect than scatterplots also in the simple regression case. Maybe the most standard residual plot is e_i against the fitted values $\hat{Y}_i$: under the linearity assumption, this plot should have roughly the same number of points above and below zero (the X axis) throughout the range of $\hat{Y}_i$, i.e., in each interval $(x - \Delta, x + \Delta)$ (some refer to this informally as a "patternless cloud of points"). We now give a rough explanation why this is expected. Recall that the residuals are

$$e_i = Y_i - \hat{Y}_i$$

which in vector form can be written

$$\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{Y},$$

where $\boldsymbol{Q} := \boldsymbol{P}_{\mathrm{Im}^{\perp}(X)}$. Under the general linear model, we have

$$\mathbb{E}(\boldsymbol{e}) = \mathbb{E}(\boldsymbol{Q}Y) = \boldsymbol{Q}\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

(by the linearity assumption). So, under the general linear model, the residuals have zero means, $\mathbb{E}(e_i) = 0$, regardless of the value of $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$. Thus, in particular, the e_is should have mean zero regardless of the

value of $\sum \beta_j x_{ij} = \mathbb{E}[Y_i]$, because this is a function of $\boldsymbol{x}_i$. When the sample size n is much larger than the number of predictors p, and under suitable assumptions (basically, if we assume (X_i, Y_i) are i.i.d. pairs), the fitted value $\hat{Y}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$ will be a consistent (in a proper sense) estimator of its mean $\mathbb{E}\left[\hat{Y}_i\right] = \mathbb{E}[Y_i] = \sum \beta_j x_{ij}$, hence, in this case we can roughly treat $\hat{Y}_i \approx \mathbb{E}[Y_i] = \sum \beta_j x_{ij}$ (this is of course imprecise). This means we expect the residuals to be approximately zero-mean, $\mathbb{E}(e_i) = 0$ regardless of the value of $\hat{Y}_i$.

As for the variances of the residuals, note that

$$\operatorname{cov}(\boldsymbol{e}) = \operatorname{cov}(\boldsymbol{Q}) = \boldsymbol{Q} \operatorname{cov}(\boldsymbol{Y}) \boldsymbol{Q}^\top = \sigma^2 \boldsymbol{Q}$$

so even under the linear model the residuals are generally not uncorrelated (hence not independent), and not even homoscedastic (i.e., $\operatorname{Var}(\epsilon_i) = Q_{ii}$ depends on i); compare this with ϵ_i, which have equal variances and are independent. Still, when the sample size n is large we argued that $\hat{Y}_i \approx \mathbb{E}[Y_i] = \sum \beta_j x_{ij}$, so its variance is approximately zero. In that case

$$\operatorname{Var}(e_i) = \operatorname{Var}\left(Y_i - \hat{Y}_i\right) \approx \operatorname{Var}(Y_i) = \operatorname{Var}(\epsilon_i) \equiv \sigma^2$$

so we can also expect the spread of the e_i 's about the x axis (the horizontal linear at $y = 0$) to be roughly constant regardless of the value of $\hat{Y}_i$. To summarize, if the linearity assumption holds then in a small X-window we expects approx. equal number of points above and below the X axis, no matter the location of the window; and, regardless of whether the linearity assumption seems appropriate, if the equal-variance assumption holds, we expect the points to have roughly the same spread (empirical variance) as we vary the location of the window.

The discussion above can be made even more intuitive by saying that, when n is large, we expect "$e_i \approx \epsilon_i$" and "$\hat{Y}_i \approx \sum \beta_j x_{ij}$ " in some sense, but a plot of ϵ_i vs. $\sum \beta_j x_{ij}$ should be throughout centered about zero, and have constant variance. In other words, we expect no pattern at all. The figure below shows the corresponding plots for e_i vs. $\hat{Y}_i$ for the 4 scenarios in Figure 11.

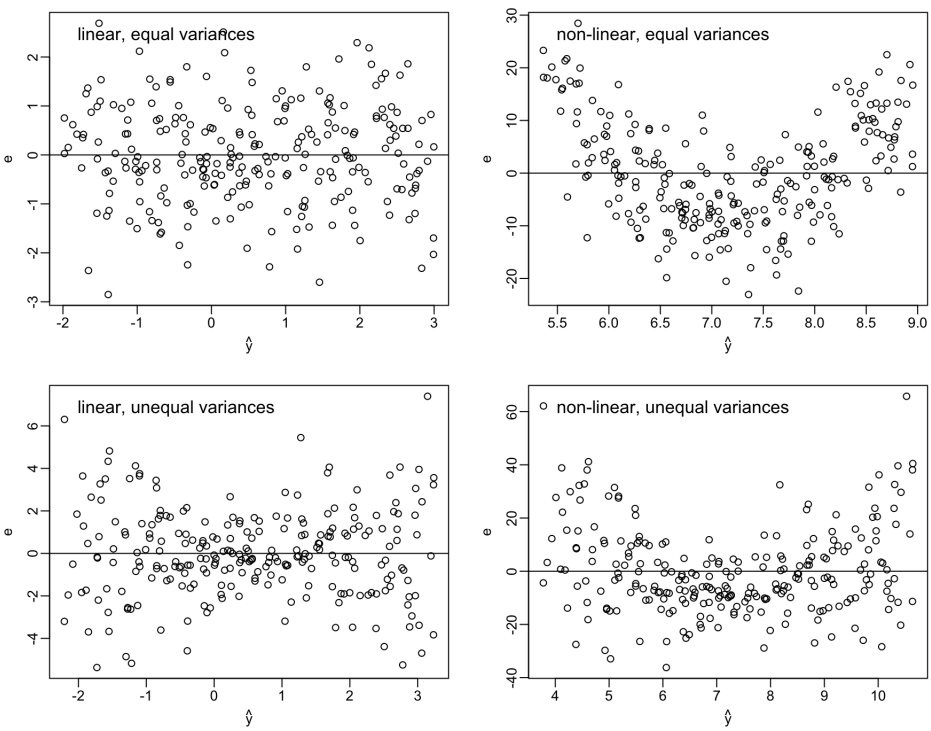


Figure 12: Residuals e_i vs. fitted values $\hat{Y}_i$

Other residual plots that can be used is the residuals against an individual predictor, e_i vs. X_{ij} (for $j = 1, ..., p$).

Checking normality. To check the normality assumption, we need a better tool because this is a more subtle assumption and harder to see visually by just eyeballing the residual plots. A preliminary check can be done using simply a histogram of the residuals e_i. Recall that, to form a histogram for a sample $W_1, \ldots, W_n$, we fix a grid $w_1, \ldots, w_K$, and let $n_k =$ number of observations falling in the bin $(w_{k-1}, w_k]$, and $h_k = w_k - w_{k-1}$ be the width of the k th bin, $k = 1, \ldots, K$. Then we plot a rectangle whose base is the bin $(w_{k-1}, w_k]$ and whose height is

$$f_k := \frac{1}{n} \cdot \frac{n_k}{h_k}.$$

The resulting histogram is an estimate for the density of W_i. The following command in R forms a histogram with bins of equal width: hist(r, breaks = 'scott', freq = FALSE).

Thus, under the normal model, the histogram of the residuals e_i should approximate a $\mathcal{N}\left(0, \sigma^2\right)$ density:

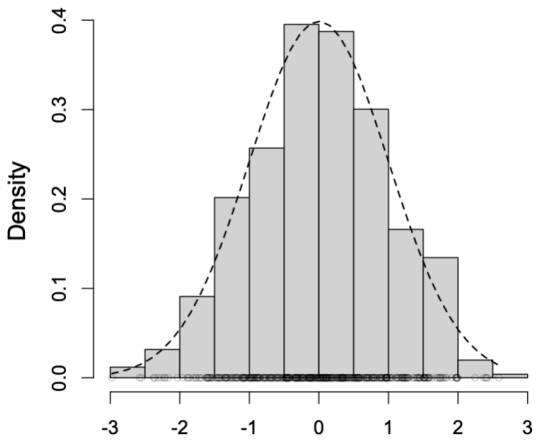


Figure 13: Histogram of residuals e_i

We could add to the plot the density of a Normal r.v. with mean and variance equal to the average and standard error of the W sample, respectively, as plotted in the Figure (dashed line). But a more convenient and precise way is to use a Quantile-Quantile plot (Q-Q plot).

To introduce the Q-Q plot, consider first CDF's F and G corresponding to any two distributions, and consider the plot

$$y = F^{-1}(p) \quad \text{vs.} \quad x = G^{-1}(p), \quad p \in (0,1)$$

If the two distributions are the same, i.e., $F(t) = G(t)$ for all t, then of course we expect the plot to form the line $y = x$. In a Q-Q plot, G is taken to be a theoretical reference distribution, and F is taken to be an empirical CDF computed from the sample. Note that, if F is an empirical CDF, then F^{-1} gives the order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Of course, in this case we never expect the points to align exactly on a straight line, but we do expect this approximately if the sample is iid draws from G.

Thus, to check normality of the errors, we will basically want to plot

$$e_{(i)} \text{ vs. } \quad \Phi^{-1}\left(\frac{i}{n}\right), \qquad i = 1, \ldots, n$$

where $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$ are the order statistics of the residuals, and Φ^{-1} is the inverse CDF of a normal distribution with proper mean and variance. However, if F, G in (11.2) are both normal distributions (each with its own-but possibly different-mean and variance), it is a simple exercise to show that the graph (11.2) will still be a straight line, though not necessarily with slope 1 and intercept 0 . In other words, in (11.3) we can take Φ^{-1} to be a standard normal $\mathcal{N}(0,1)$, and compare to a general straight line instead of the identity line $y = x$. Recall that, under the normal model and assuming n is large, we expect e_i to be approximately $\mathcal{N}\left(0, \sigma^2\right)$ and approximately independent. In this case, we expect (11.3) to lie approximately on a straight line. The basic command in R is

```
qqnorm(resid(fm1))  # generate Normal Q-Q plot
qqline(resid(fm1))  # add reference line
```

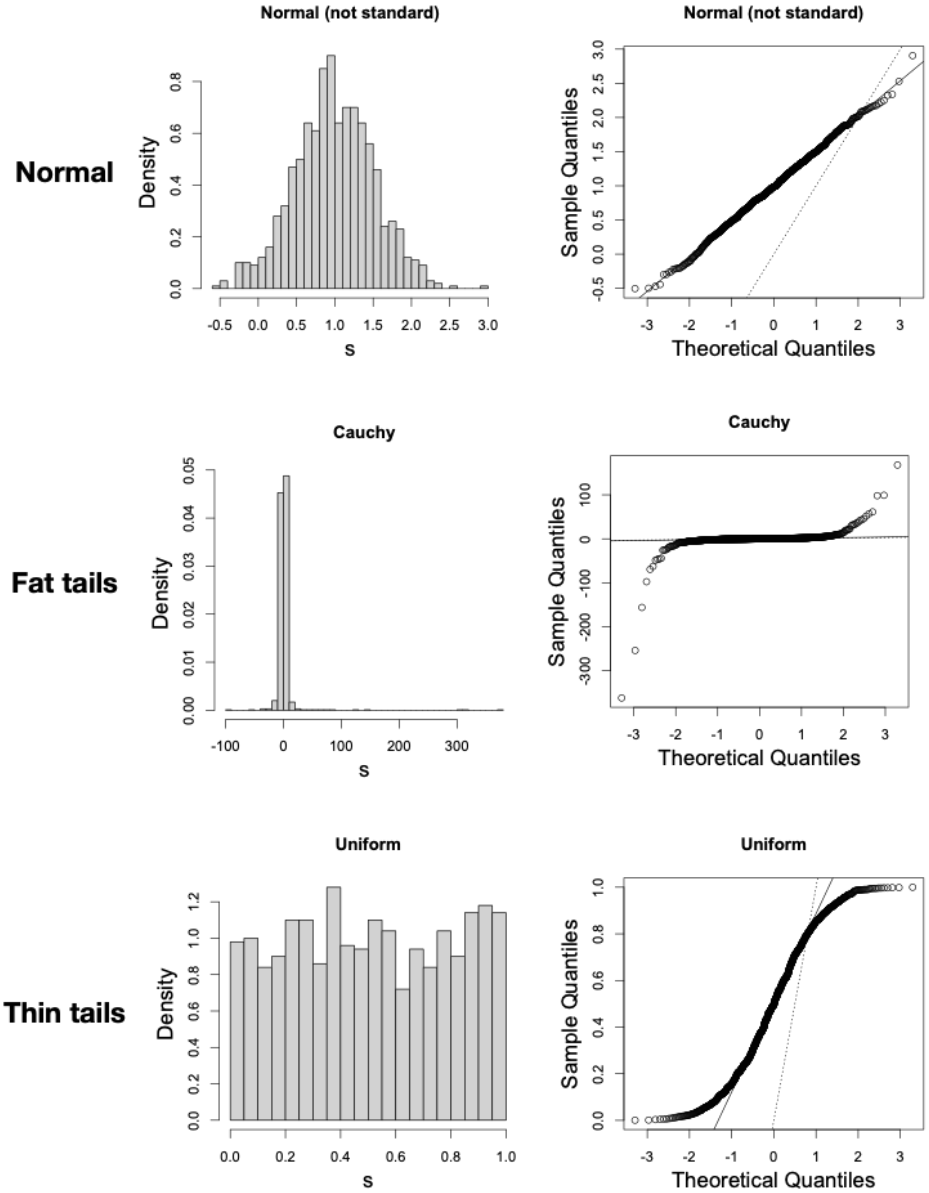


Figure 14: QQ-plot

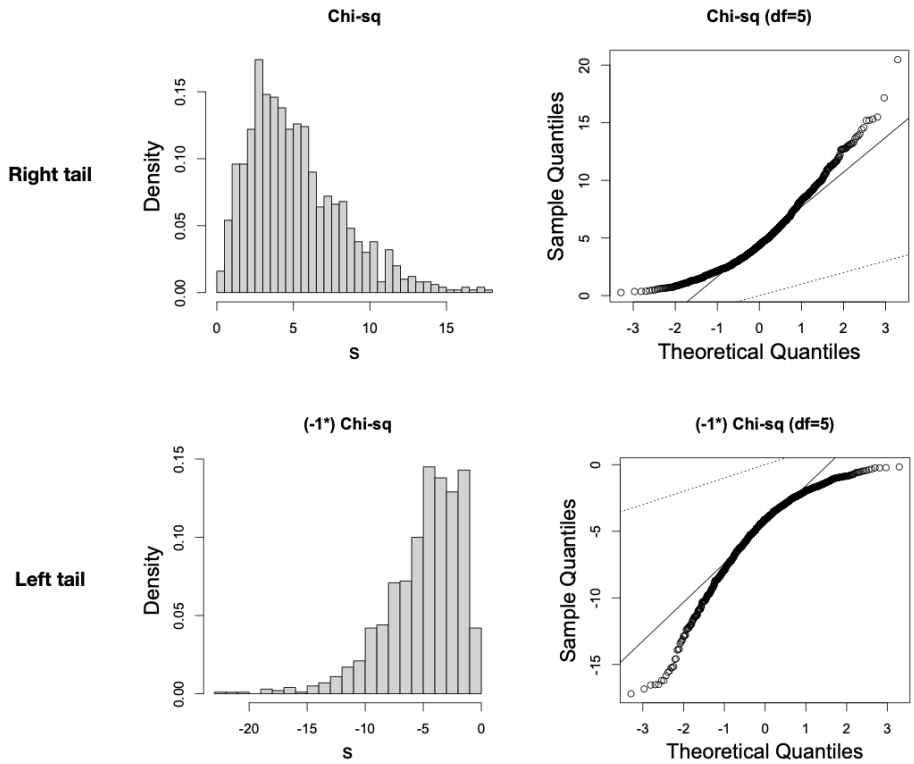


Figure 15: QQ-plot (continued)

Outliers. An outlier is a sample point $(\boldsymbol{X}_i, Y_i)$ that shows clear disagreement with the fitted model for the dataset. To identify outliers, one may compute $\hat{Y}_{(i)}$ = fitted value for X_i when the i th observation is removed from the dataset. If the difference $Y_i - \hat{Y}_{(i)}$ is large, the i th observation is suspect as an outlier. Note: this does not necessarily imply that the residual from the fit to the entire dataset (including the i th point) is small!

A formal measure for identifying candidate outliers is the leverage of a point, defined as

$$\text{leverage}_i := \text{Cov}\left(\hat{Y}_i, Y_i\right) = [\text{cov}(\hat{\boldsymbol{Y}}, \boldsymbol{Y})]_{ii} = [\boldsymbol{P}_X]_{ii} = \left[\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right]_{ii}$$

Note that this depends on the matrix $\boldsymbol{X}$ only, i.e., only on the explanatory variables. Observations with high leverage have the potential of being outliers. Intuitively, if the covariance between $\hat{Y}_i$ and Y_i is large (in absolute value), this means that Y_i has large impact on the fitted value, i.e., a changing the value of Y_i will change the (entire) LS fit considerably. Figure 16 gives intuition for why it makes sense that this quantity depends (only) on the X valus.

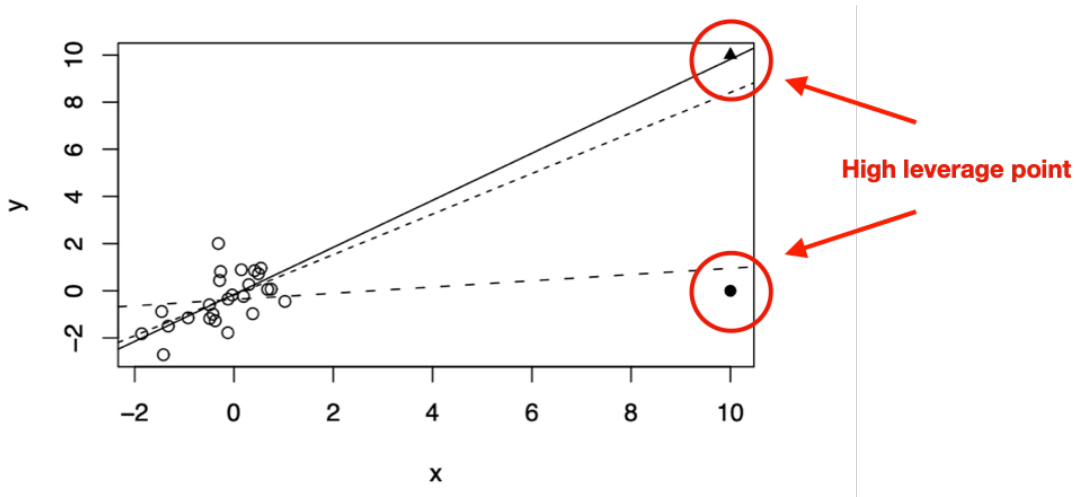
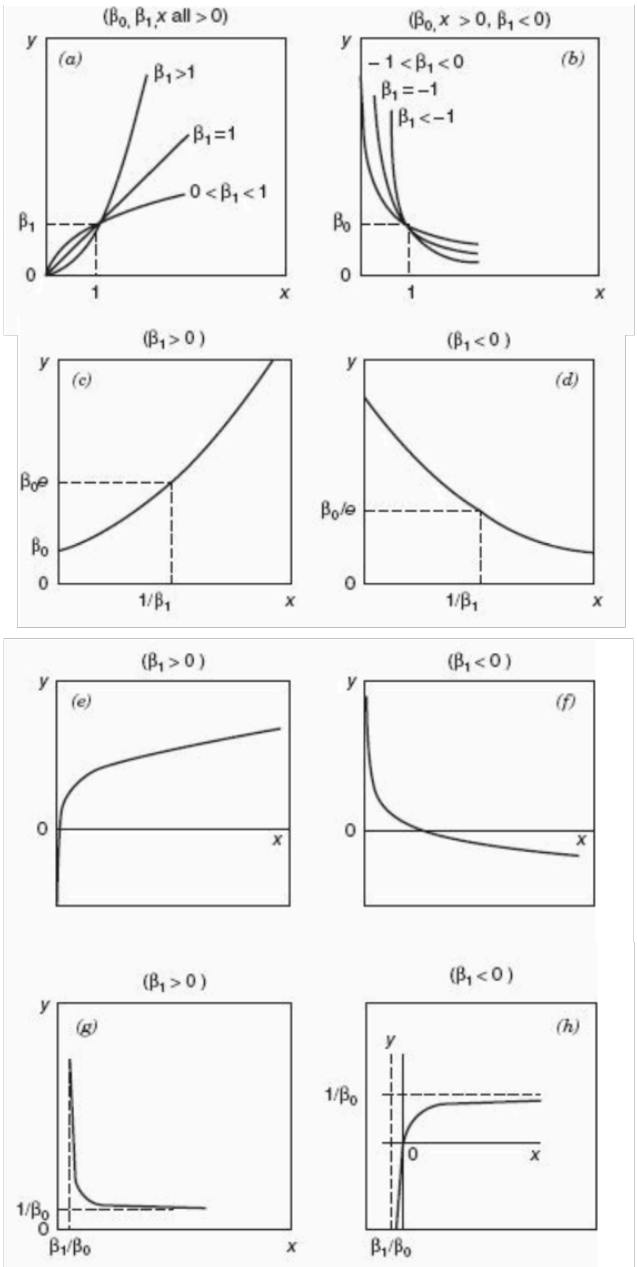


Figure 16: Illustration of leverage

Strategies for fixing violations of the model assumptions. We will now see a few options for fixes that can help to correct the situation when there are clear violations of the basic model assumptions.

Violations of the linearity assumption. The linearity assumption says that the expectation of Y_i is linear in $\boldsymbol{X}_i$, that is, $\mathbb{E}Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$. There are situations where the linear model is inadequate for regressing the original Y_i s on the original X_i s, but will become appropriate one applying a *transformation* to Y_i or X_i or both.

Examples. Here are some examples that can be used as guidelines for suggesting a "correcting" transformation.



$$y = \beta_0 x^{\beta_1} \longrightarrow \tilde{y} = \log(y), \tilde{x} = \log(x)$$
$$\tilde{y} = \log\left(\beta_0\right) + \beta_1 \tilde{x}$$

$$y = \beta_0 e^{\beta_1 x} \longrightarrow \tilde{y} = \ln(y)$$
$$\tilde{y} = \ln\left(\beta_0\right) + \beta_1 x$$

$$y = \beta_0 + \beta_1 \log(x) \quad \longrightarrow \quad \tilde{x} = \log(x)$$
$$\tilde{y} = \beta_0 + \beta_1 \tilde{x}$$

$$y = \frac{x}{\beta_0 x - \beta_1} \quad \longrightarrow \quad \tilde{y} = \frac{1}{y}, \tilde{x} = \frac{1}{x}$$
$$\tilde{y} = \beta_0 - \beta_1 \tilde{x}$$

Figure 17: Transformations

Violations of the equal-variance (homoscedasticity) assumption. There are cases where the assumption of equal variances for the errors ϵ_i is not appropriate. For example, this would be the case when the distribution of the outcome variable Y is such that the variance $\mathrm{Var}[Y]$ is functionally related to $\mathbb{E}[Y]$.

Examples.

- $Y \sim \text{Pois}(\lambda) : \mathbb{E}[Y] = \lambda, \quad \text{Var}(Y) = \lambda$

- $Y \sim \text{Binom}(n, p)/n : \mathbb{E}[Y] = p, \quad \text{Var}(Y) = p(1-p)/n$

Hence, if, say, $Y_i \sim \text{Pois}(\lambda_i)$ with the mean $\lambda_i = \mathbb{E}[Y_i]$ being a function of X_i, then $\text{Var}[\epsilon_i] = \lambda_i$ will not be the same for all $i = 1, \ldots, n$.

In such cases, it may be possible to apply a variance stabilizing transformation to Y in order to recover again a situation with equal-variance errors.

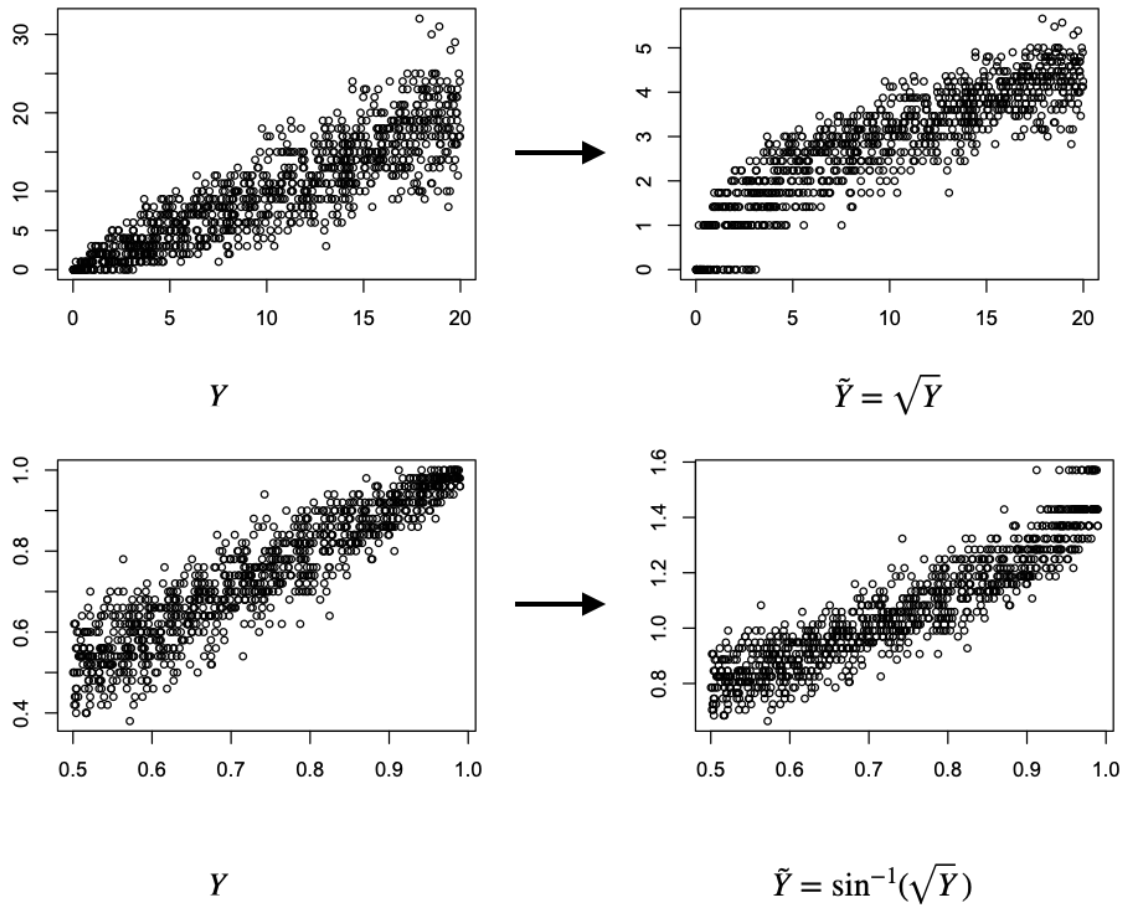


Figure 18: Variance-stabilizing transformations

How to obtain the correct transformation for variance stabilizing? Suppose that we know that

$$\sigma^2 = \phi(\mu_Y)$$

for some known function $\phi(\cdot)$. For example, in the binomial case,

$$\mu_Y = np, \quad \sigma_Y = [np(1-p)]^{1/2} = [\mu_Y(1 - \mu_Y/n)]^{1/2}.$$

If we now define $Z = f(Y)$, and use the so-called delta method to write

$$Z = f(Y) \approx f(\mu_Y) + f'(\mu_Y)(Y - \mu_Y) \Longrightarrow \sigma_Z^2 \approx [f'(\mu_Y)]^2 \sigma_Y^2 = [f'(\mu_Y)]^2 \phi^2(\mu_Y),$$

then to stabilize the variance would mean approximately to have

$$f'(\mu_Y)\phi(\mu_Y) \equiv \sigma_Z$$

implying

$$f(y) = \sigma_Z \int \frac{1}{\phi(y)} dy$$

Thus, for example, in the Binomial case,

$$f(y) = \sigma_Z \int [y(1 - y/n)]^{-1/2} dy = \sigma_Z \cdot 2n^{1/2} \arcsin\left[(y/n)^{1/2}\right].$$