Robust Statistics

KJM

November 24, 2024

Contents

1 Motivation

In statistics and mathematical modeling in general, in addition to the sample that we observe, any inference or conclusion one may reach is facilitated by a host of assumptions. These may include the distributional form from which we sample, independence, randomness, and so on. This is a necessary step, not designed to produce a model absolutely resembling reality, but rather to produce something that is reasonable with only small error, as well as mathematically convenient. Any conclusions we draw, therefore, should consequently be reasonable and small in error.

However, countless examples reveal that common statistical procedures are remarkably sensitive to small departures from some set of common assumptions. Most notably, the sample mean can be highly variable under slight departures from normality. In reality, distributions can be skewed, they can have heavy tails, and random samples often have outliers, leading to a plethora of practical problems even if these deviations are minor in scale. For example, the standard error of the sample mean is inflated substantially by heavy tails, resulting in the probability coverage of standard methods for computing confidence intervals differing greatly from the nominal value, and the usual sample variance can give a distorted view of the amount of dispersion. Thus, these small deviations from normality give a false sense of the amount of spread of the data, and the central or "average" value around which most the population are distributed.

Consequently, it is desirable to find estimators that do not vary so greatly due to small deviations from our assumptions, since this can lead to a false sense of the distribution we are working with, as well as any inferences we make from it. That is, we want to use techniques that are robust. Robustness of a given statistical procedure signifies insensitivity to small deviations from the assumptions upon which the procedure was constructed.

The scenario below is a classic example, due to Tukey, demonstrating how a small deviation from normality can greatly affect the well-known sample standard deviation. It turns out that the mean absolute deviation, to be defined below, is a more robust measure of scale (an indication of how spread out the values in a distribution are), and how it can greatly outperform the standard deviation in the face of outliers.

Imagine some proportion of observations, $1-\epsilon$, is generated by a normal model and the remaining proportion by some unknown mechanism, in our case another normal with greater variance. For example, we are taking repeated measurements of something: 95% of the time, we take measurements correctly but the remaining 5% there is some error like an apparatus failure. Suppose the distribution for the correctly measured data has distribution $G = \mathcal{N}(\mu, \sigma^2)$ and erroneous data $H = \mathcal{N}(\mu, 9\sigma^2)$. Our data will thus follow the distribution F, where:

$$F = (1-\epsilon)G + \epsilon H$$

So, in our experiment the data we collect follows a distribution F, but in our model we assume that we are sampling from a distribution resembling normality. Now consider two measures of scale, mean absolution deviation (MAD), d_n, and root mean square deviation (RMSD), s_n:

$$d_n = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|, \qquad s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

The asymptotic relative efficiency (ARE) is a statistical measure used to compare the efficiency of two estimators when drawing from a particular distribution as the sample size n grows large. Here, the ARE is given when sampling from the mixed normal, dependent on ϵ:

$$\text{ARE}(\epsilon) = \lim_{n\to\infty}\frac{Var(s_n)/(\mathbb{E}(s_n)^2)}{Var(d_n)/(\mathbb{E}(d_n)^2)} = \frac{\frac{1}{4}\left[\frac{3(1+80\epsilon)}{(1+8\epsilon)^2}\right]-1}{\frac{\pi(1+8\epsilon)}{2(1+2\epsilon)^2}-1}$$

Intuitively, we take the ratio of the variances of the two estimators as n grows large. But since the MAD and RMSD are estimates of different measures of scale, their raw variances alone wouldn't provide a fair comparison. By using the squared expectation in the denominator for each, the variance is scaled relative to the central tendency of each estimator. In the case above, if $\text{ARE}(\epsilon) > 1$ then we conclude the RMSD is less efficient than MAD.

Assuming absolute normality ($\epsilon = 0$), we get an ARE of roughly 0.88. That is, RMSD is 12% more efficient than MAD. Different values of ϵ yield the following:

ϵ	ARE(ϵ)
0	0.876
0.001	0.948
0.01	1.439
0.05	2.035
0.10	1.903
0.5	1.017
1.0	0.876

Table 1: Table of ϵ and its corresponding ARE(ϵ)

Therefore, if ϵ is only equal to 0.01 (there is a 1% chance of taking erroneous measurements), the 12% advantage of the RMSD has vanished and been replaced by a 44% advantage for MAD. Namely, an experiment with a very small error rate disproportionately increases the variance in RMSD compared to MAD, thus rendering it suspect as an ideal measure of scale.

This egregious inflation of the variance of the RMSD indeed has dire consequences upon inferences one may make from the data when assuming normality. Set μ, σ, and ϵ to 0, 1, and 0.1 respectively. The probability distribution functions of G and F are plotted below. The standard normal has variance 1 but the mixed normal has a variance of (...). As a result, if sampling is from the mixed normal, the length of the standard confidence interval for the population mean, μ, will be over (...) times longer than it would be when sampling from the standard normal distribution instead.

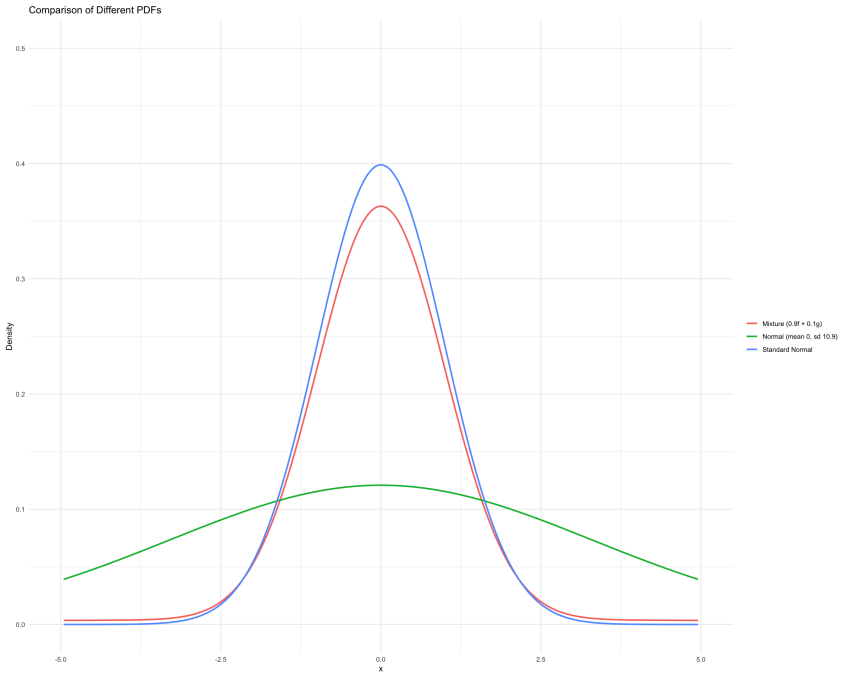

Figure 1: Tukey's Mixed Normal. The diagram below is a plot of the probability distribution function (pdf) of X, in blue, and the mixed normal, in red.

From the image, it seems that the normal distribution provides a good approximation of the contaminated normal. Only the tails of the mixed normal are slightly above the tails of the normal. But this small proportion of the population has had an inordinately large effect on the variance of the distribution and the RMSD, thus underscoring the need to find alternative measures of a distribution that do not result in wildly different outcomes as a result of a slight shift from our assumptions.

One might question, however, the necessity of robust methods. If we were in the scenario above, could we not apply some rule for outlier rejection and then proceed to use standard techniques? Only one reason which is in the negative is due to masking. This is when the presence of an outlier prevents

an outlier rejection test from identifying another outlier. For instance, the presence of an outlier can inflate the standard error of the sample mean, increasing the width of a given confidence interval that could envelop other outliers.

Robust methods, therefore, are a collection of new and improved methods that provide a more accurate and more nuanced understanding of data. under general conditions they can have substantially higher power compared to more traditional techniques. Robust methods should strike a balance between the following three traits:

i **Efficiency**: Given the assumed model we are sampling from, the statistical procedure should have low variance

ii **Stability**: Small deviations from the model assumptions should only impair model performance to a small extent

iii **Breakdown**: Large outliers should not cause a drastic change in the performance of the model.

2 Measuring Robustness

A measure is called robust if slight changes in the distribution from which it is sampled results in a relatively small effect on its value. As demonstrated earlier, the sample mean and standard deviation are not robust. There are three different tools that are utilized to asses whether a given measure has the desired robustness properties.

2.1 Qualitative Robustness

Often, our measure of location can be viewed as functionals of their empirical distributions or distribution functions. For example, the sample and population means from a continuous distribution can be found with respect to the functional T:

$$T(F) = \mathbb{E}_f(X) = \int_{-\infty}^{\infty} xf(x)dx, \qquad T(\hat{F}) = \mathbb{E}_{\hat{f}}(X) = \sum_{i=1}^{n} x_i \hat{f}(x_i)$$

Say we are sampling from some true underlying distribution H, which we have modeled as G. Say that G is "close" to H - in the sense of some metric. Then, it should be the case that our measure of location when sampling from G, $T(G)$ is also "close" to that same measure had we sampled from the true distribution, $T(H)$.

Another intuitive idea behind robustness is that if we have some sample $(x_1, \dots, x_n)$ and another sample $(y_1, \dots, y_n)$, where the difference between them is either small differences between observation i in each, that is $x_i \approx y_i$, or large changes between x_i and y_i for only some i due to gross errors or blunders but otherwise $x_i = y_i$. Then we should have that the empirical distributions of each are "close" and therefore so are the estimates we obtain.

More formally, a measure, T, is said to be qualitatively robust at H if T is continuous at H according to the metric d. That is,

$$\forall\, \epsilon > 0\ \exists\, \delta > 0\ :\ d(H, G) < \delta\ \Rightarrow\ |T(H) - T(G)| < \epsilon$$

2.2 Infinitesimal Robustness

Given our functional T, we want to be able to measure the sensitivity and amount of change of an estimator due to outliers. More specifically, we want to measure the rate of change of the estimator T when a small amount of contamination is added. This is done using the influence function (IF), defined below:

$$IC(x, F, T) = \lim_{\epsilon \to 0^+} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}$$

Intuitively, we are measuring the rate of change of T when an infinitesimal amount of contamination has been added at point x. As the value of the IF varies with x, we say that F is infinitesimally robust if the IF is bounded.

2.3 Quantitative Robustness

Let Θ be the set of values that a parameter θ can take. For our measure of location or scale, this is the interval $[-\infty, \infty]$ or $[0, \infty]$ respectively. Additionally, consider $F_{x,\epsilon} = (1-\epsilon)F + \epsilon\delta_x$.

The breakdown point of the estimator T at F, denoted by $\epsilon^*(T, F)$ is the largest $\epsilon^* \in (0, 1)$ such that:

$$\forall \epsilon < \epsilon^*\ \forall G\ \exists K \subset \Theta\ :\ T((1-\epsilon)F + \epsilon G) \in K$$

Where K is closed and bounded. The breakdown point is the smallest fraction of bad observations that may cause an estimator to take on arbitrarily large values.

3 M-Estimation

3.1 M-Estimates of Location

Let $x_1, \ldots, x_n$ be an i.i.d sample from some distribution F An M-estimator, $\hat{\theta}$, of some population parameter, θ, is one that is found by minimizing a criterion function, $M_n(\theta)$, expressed in terms of a some loss function $\rho(x_i; \theta)$, given by

$$M_n(\theta) := \sum_{i=1}^{n} \rho(x_i; \theta)$$

Often, we can solve this implicitly by partially differentiating and solving the set of equations that are set to zero.

$$\Psi_n(\theta) := \sum_{i=1}^{n} \psi(x_i; \hat{\theta}) = 0,$$

Where $\partial\rho(x;\theta)/\partial\theta = \psi(x;\theta)$. Note that if ρ is smooth and concave in θ, then the estimates we obtain are equivalent. Sometimes, we will encounter ρ that are not concave in θ and this can lead to multiple local minima. Methods to find the global minima will be addressed in a later section, but for now it is sufficient to note that all minima we find follow the same properties.

In our case, we are mainly concerned with measures of location, μ. A measure of distribution is a statistical value or function that helps summarize and describe the characteristics of a probability distribution. Measures of location and scale are two examples of such measures that characterize a given distribution.

Let X be a random variable with distribution F, and $\theta(X)$ some measure of F. $\theta(X)$ is said to be a measure of location if it satisfies the following conditions. For any given constants a and b:

i $\theta(X + b) = \theta(X) + b$ (location equivariance)

ii $\theta(-X) = -\theta(X)$

iii $X \geq 0 \Rightarrow \theta(X) \geq 0$

iv $\theta(aX) = a\theta(X)$ (scale equivariance)

Also referred to as a measure of central tendency, these provide a central or "average" value around which the data points are distributed. Examples include the sample mean. This can be shown using standard properties of expectation. When searching for a measure of location, we are looking for some value that is in some sense close to all of our data points. We aim to find the estimate $\hat{\mu}$ that minimizes the total error across all data points by summing up the individual errors, measured by the function ρ. So in this setup, we want to find $\hat{\mu}$ such that:

$$\sum_{i=1}^{n} \rho(x_i - \hat{\mu}) = \min!, \qquad \sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0.$$

To evaluate the performance of M-estimators, it is necessary to calculate their distributions. However, for finite samples there is no close-form or analytical formula that describes the exact sampling distribution of the M-estimator, and so must be approximated. It can be shown that M-estimators are asymptotically normally distributed:

$$\hat{\mu} \rightsquigarrow \mathcal{N}(\mu_0, \frac{\nu}{n})$$

Where:

- μ_0 is the asymptotic value of $\hat{\mu}$, where:

$$\mathbb{E}_F[\psi(X - \mu_0)] = 0$$

- ν is the asymptotic variance of $\hat{\mu}$, given by:

$$\nu = \frac{\mathbb{E}_F[\psi(X - \mu_0)^2]}{(\mathbb{E}_F[\psi'(X - \mu_0)])^2}$$

Note that it is often extremely difficult or impossible to analytically solve for μ_0 and ν. We will have to defer to numerical techniques that will be discussed later.

3.1.1 Maximum Likelihood Estimation

It should be noted that maximum likelihood estimation (MLE) is a special case of M-estimation. Consider the log-likelihood below.

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \log\ f_0(x_i - \theta)$$

The maximum likelihood estimate ($\hat{\theta}_{\text{MLE}}$) is then given by the value of θ that maximises the log-likelihood. If we set $\rho = -\log\ f_0$, then this is equivalent to:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(x_i - \theta)$$

Which is the same form of an M-estimator. It can, in fact, be shown that $\hat{\theta}_{\text{MLE}}$ is the "optimal" estimate, in that it attains the lowest possible asymptotic variance among a "reasonable" class of estimators, given F. But maximum likelihood estimation is not in general robust, and in reality we are not truly sampling from distribution F but rather something that is "close" to it.

We can further see that the sample mean and median are special cases of M-estimation. Consider the following case where $\rho(x) = x^2$ and thus $\psi(x) = 2x$, with observed sample x_i for $i = 1, 2, \dots, n$. We want to find $\hat{\theta}$ minimising (...):

$$\sum_{i=1}^{n} \psi(x_i - \hat{\theta}) = \sum_{i=1}^{n} (x_i - \hat{\theta}) = 0$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{\mathbf{x}}$$

Thus we obtain the sample mean. Now consider setting $\rho(x) = |x|$ and $\psi(x) = \text{sgn}(x)$. We then find $\hat{\theta}$ by the same procedure:

$$\sum_{i=1}^{n} \psi(x_i - \hat{\theta}) = \sum_{i=1}^{n} \text{sgn}(x_i - \hat{\theta}) = 0$$

In this situation, we want to pick $\hat{\theta}$ such that the number of x_i where $x_i - \hat{\theta}$ is negative is equal to the number of x_i so that $x_i - \hat{\theta}$ is positive. So $\hat{\theta}$ is greater than half the values and less than the other half, namely a sample median. Thus it has been shown that the sample mean and median are indeed M-estimates.

3.1.2 Scale Equivaruance

Note that in its current form, our M-estimate of location is note scale-equivariant. Indeed, say we scale our data points $x_1, \dots, x_n$ by some constant $a \neq 0$. It cannot be guaranteed $\psi(ax_i - a\mu) = a\psi(x_i - \mu)$ and therefore it is not in general true that $a\Sigma_i \psi(x_i - \mu) = \Sigma_i \psi(ax_i - a\mu)$. Therefore they are not minimized by them same value $\hat{\mu}$

The way around this is to standardize what is inside by dividing by some measure of scale, to be defined later. This way ..... Scaling our data so we have $ax_1, \dots, ax_n$. Then call our measure of location for this data is $\sigma_a = a\sigma$ and:

$$\sum_{i=1}^{n} \psi\left(\frac{ax_i - a\mu}{\sigma_a}\right) = \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right)$$

Namely, our estimate is the same.

3.2 Picking the ρ-function

We want an estimator that is "nearly optimal" when sampling from exactly F and also from distributions "close" to F. The function ρ, or its derivative, ψ, can be chosen in such a way to provide the estimator with such desirable properties.

Consider the mean, with ρ-function $\rho(x) = x^2$, which grows rapidly as $|x|$ increases. This excessive growth makes the mean highly sensitive to outliers because large residuals exert disproportionate influence. The median, on the other hand, is extremely resistant to large outliers as ρ increases far more slowly, thus not influencing the estimate we obtain as greatly. However, the ρ-function here is too insensitive to small residual values. In a typical dataset, most residuals are relatively small and arise from random variability around the true model. Therefore, these small residuals provide the bulk of the information about the central tendency of the data or the true parameter value. Not penalizing these smaller residuals appropriately can lead to inefficiency of our estimate due to the random nature of how our data was generated.

This suggests that we should seek a ρ-function that does not grow excessively for large residuals, which would lead to overly penalizing due to the presence outliers, while still assigning sufficiently large values to small residuals to ensure efficiency. Below are examples of two such rho-functions.

3.2.1 Huber functions

Huber functions blend elements of the mean and median. For $k > 0$, define.

$$\psi_k(r) = \begin{cases} r & \text{if } |r| \le k, \\ k \cdot \operatorname{sgn}(r) & \text{if } |r| > k. \end{cases}, \qquad \rho_k(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le k, \\ k|r| - \frac{1}{2}k^2 & \text{if } |r| > k. \end{cases}$$

For $k = 0$, define $\rho_0(r) = |r|$ and $\psi_0 = \operatorname{sgn}(r)$. It can be seen that ρ is quadratic in some central region, but only increases linearly to infinity. Also note that k plays the role of a sort of tuning parameter. As $k \to 0$, ρ becomes that for the median, whereas when $k \to \infty$ then ρ becomes that for the mean. This suggests for small k, the Huber function may become more robust but less efficient, and for large k we lose robustness but gain efficiency, when sampling from a Normal. We therefore seek a value of k that strikes a balance between the two.

The table below demonstrates the how the asymptotic variance of the Huber M-estimate varies with differing values of k, as well as with differing levels of contamination in the mixed normal distribution. We take $G = \mathcal{N}(0, 1)$, $H = \mathcal{N}(0, 10)$, and sample from $F = (1 - \epsilon)G + \epsilon H$

k	$\epsilon = 0$	$\epsilon = 0.5$	$\epsilon = 0.10$
0	1.571	1.722	1.897
0.7	1.187	1.332	1.501
1.0	1.107	1.263	1.443
1.4	1.047	1.227	1.439
1.7	1.023	1.233	1.479
2.0	1.010	1.259	1.550
∞	1.000	5.950	10.900

Table 2: Approximate asymptotic variances of Huber M-estimate for various values of k and ϵ.

3.2.2 Bisquare functions

Bisquare functions give rise to a broader range of M-estimates called "re-descending M-estimates" which shall be discussed later. It is defined as follows.

$$\psi_k(r) = \begin{cases} r\left(1 - \left(\frac{r}{k}\right)^2\right)^2 & \text{if } |r| \le k, \\ 0 & \text{if } |r| > k. \end{cases}, \qquad \rho_k(r) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{r}{k}\right)^2\right)^3\right] & \text{if } |r| \le k, \\ \frac{k^2}{6} & \text{if } |r| > k. \end{cases}$$

Note that the ρ-function is steep in some central region and flat in the outer region. In other words, after the residual passes a certain threshold, it contributes no further to the loss function and dramatically down-weights the influence of outliers. The steep central region, meanwhile, ensures that smaller residuals are penalized effectively.

Also note that k plays the same role here as before. For larger k we allow a greater central region and a greater threshold that we must pass before large residuals are down-weighted.

An intuitive view

In most cases that will be discussed, it will be the case that $\psi(0) = 0$ and $\psi'(0)$ exists. Define

$$W(x) = \begin{cases} \frac{\psi(x)}{x} & \text{if } x \neq 0, \\ \psi'(0) & \text{if } x = 0. \end{cases}$$

So we can re-write:

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = \sum_{i=1}^{n} W(x_i - \hat{\mu})(x_i - \hat{\mu}) = 0.$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}, \quad \text{where } w_i = W(x_i - \hat{\mu})$$

So we can re-write our M-estimate as a weighted average of our observations. These weights depend on $\hat{\theta}$ itself, but gives intuition on how we calculate the M-estimate. This representation will further be useful when we try to numerically compute our estimate $\hat{\mu}$.

In the cases we consider, ρ will be of an order less than quadratic as $|x|$ grows large, and therefore ψ will be less than linear this region. This was part of our desirable properties when picking ρ described earlier. With this in mind, Therefore, $W(x)$ is decreasing for large $|x|$. So any observations far from our estimate will have very small weight.

Consider, then, the weight functions for the Huber (left) and Bisquare (right) estimates

$$w(r) = \begin{cases} 1 & \text{if } |r| \leq c, \\ \frac{c}{|r|} & \text{if } |r| > c. \end{cases}, \qquad w(r) = \begin{cases} \left(1 - \left(\frac{r}{c}\right)^2\right)^2 & \text{if } |r| \leq c, \\ 0 & \text{if } |r| > c. \end{cases}$$

From figure 2, ...

3.3 M-Estimates of Scale

A non-negative function, $\sigma(X)$ is said to be a measure of scale if for any constants $a > 0$ and b:

i $\sigma(aX) = a\sigma(X)$ (scale equivariace)

ii $\sigma(X + b) = \sigma(X)$ (location invariance)

iii $\sigma(-X) = \sigma(X)$ (sign invariance)

Notice from properties (i) and (iii) that if $a < 0$ then $\sigma(aX) = \sigma(-aX) = -a\sigma(X) = |a|\sigma(X)$.

3.4 ...

Given our normal distribution $\mathcal{N}(\mu, \sigma^2)$, say we know σ. We obtain our estimate $\hat{\mu}$ by:

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(x_i - \mu)$$

And we obtain the limiting distribution for $\hat{\mu} \rightsquigarrow \mathcal{N}(\mu_0, \frac{\nu}{n})$, but now we get a new expression for ν:

$$\nu = \sigma^2 \frac{\mathbb{E}_F[\psi((X - \mu_0)/\sigma)^2]}{(\mathbb{E}_F[\psi'((X - \mu_0)/\sigma)])^2}$$

In reality, we often do not know σ and therefore must estimate it. This may appear to complicate things, as we will now have to work with the sampling distribution of $\hat{sigma}$ to determine our asymptotic distribution for $\hat{\mu}$. Fortunately, it can be shown that if $\hat{\sigma} \xrightarrow{\mathbb{P}} \sigma$, then our limiting distribution is that above.

It should also be noted that we seek a $\hat{\sigma}$ that is robust. It is clear that it directly affects the performance and robustness of the location estimator itself.

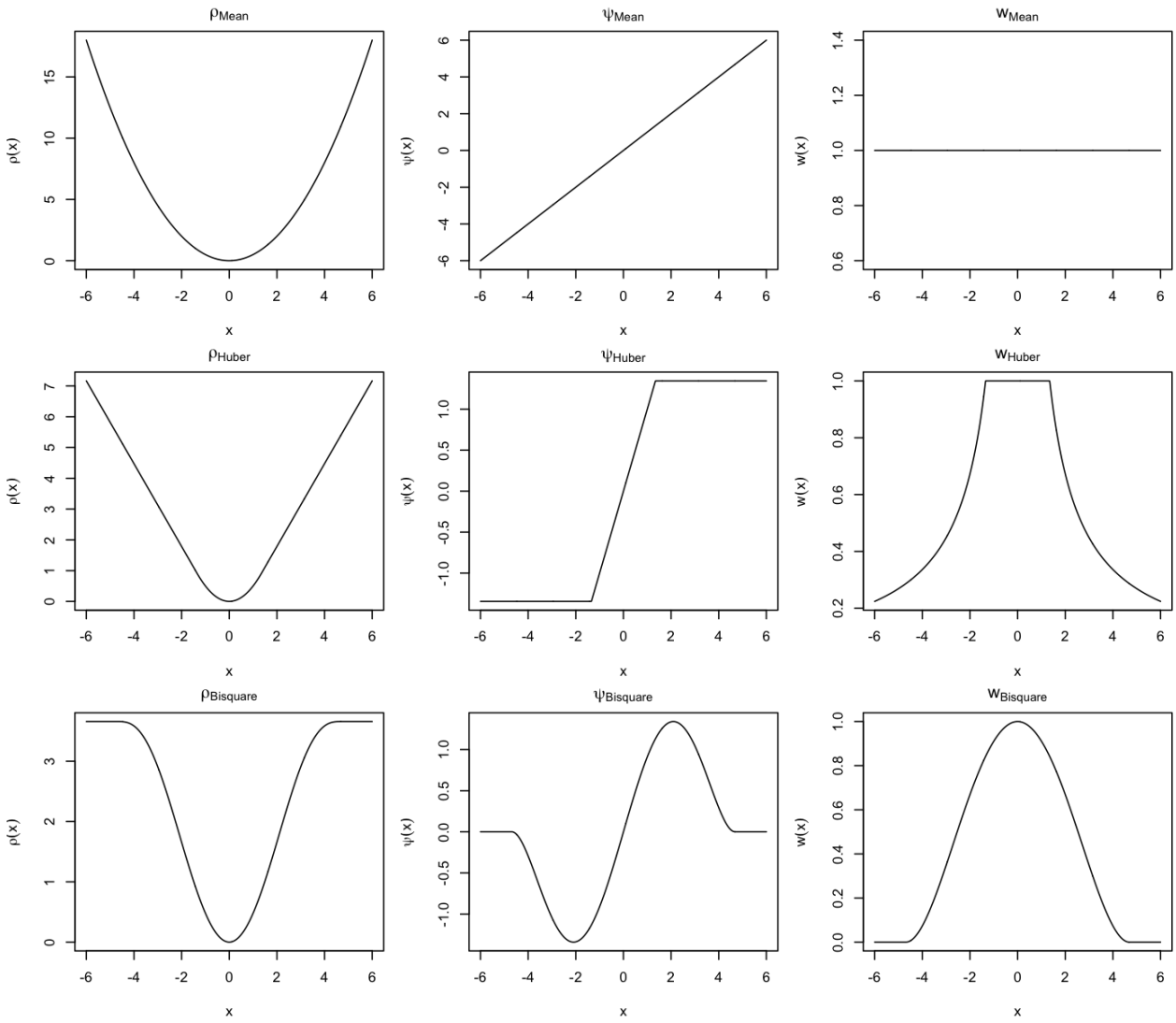

Figure 2: The image consists of a 3x3 grid of plots, illustrating the ρ, ψ, and weight functions for three types of ρ-functions: the Mean (row 1), Huber ($k = 1.345$) (row 2), and Bisquare ($k = 4.685$) (row 3), with column 1 the ρ function, column 2 the ψ function, and column 3 the weight function.

3.5 Numerical Computation

More often than not, it is extremely difficult or impossible to find closed-form expressions to calculating an M-estimate. This can be due to the complex form of our ρ and ψ-functions, high dimensionality and complicated supports, large data sample, or solving non-linear simultaneous equations. This necessitates numerical methods to find approximate solutions, and here we will explore a technique called iterative re-weighting to achieve this goal.

The algorithm is as follows:

1. Compute $\hat{\sigma} = \text{MADN}(x)$ and $\hat{\mu}_0 = \text{Med}(x)$
2. For $k = 0, 1, 2, , , ,$, compute the weights:

$$w_{k,i} = W\left(\frac{x_i - \hat{\theta}_k}{\hat{\sigma}}\right), \text{for} i = 1, \dots, n$$

   and then compute $\hat{\mu}_{k+1}$, given by:

$$\frac{\Sigma_{i=1}^{n} w_{k,i} x_i}{\Sigma_{i=1}^{n} w_{k,i}}$$

3. Stop when $|\hat{\mu}_{k+1} - \hat{\mu}_k| < \epsilon\hat{\sigma}$

If $W(x)$ is bounded and non-increasing for $x > 0$, then the sequence converges to the global minimum.

4 Robust Regression

(not considering issue of whether actually appropriate model or not)

4.1 Notation and Assumptions

Assume that p unknown parameters $\beta_1, \dots, \beta_p$ are to be estimated from n observations $y_1, \dots, y_n$ to which they are linearly related by:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

Which can be represented by matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where we define:

- $\mathbf{Y^T} = (y_1, \dots, y_n)$ is the vector of responses, each y_i called the response variable,
- $\mathbf{X}$ is the design matrix, each x_{ij} called a predictor variable,
- $\boldsymbol{\beta^T} = (\beta_1, \dots, \beta_p)$ is the p-dimensional parameter vector, each β_i fixed and unknown coefficients, called parameters,
- $\boldsymbol{\epsilon^T} = (\epsilon_1, \dots, \epsilon_n)$ is the vector of errors, each ϵ_i called an error variable.

We also make the following assumptions:

i **Linearity**: $\mathbb{E}(\epsilon_i) = 0$

ii **Homoscedasticity**: $\mathrm{Var}(\epsilon_i) = \sigma^2$

iii **Independence**: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$

iv **Normality**: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Consequently, $\mathbf{y}$ is an n-dimensional normally distributed random vector. Specifically, $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the n-dimensional identity matrix.

4.2 Standard Linear Regression

Commonly, one estimates the parameters of a linear regression model using the method of least squares. In our framework, it can be demonstrated that least-squares estimation is equivalent to maximum likelihood estimation. As we will explore robust regression through the framework of M-estimation, it is instructive to re-frame standard linear regression through the lens of maximum likelihood, thus providing a natural bridge to understanding how robust methods extend and generalize these ideas to handle violations of key assumptions.

We obtain our estimates by finding $\beta_1, \dots, \beta_p$ that minimizes the set of equations:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} \right)^2$$

Equivalently, we can find our estimate $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ by differentiating with respect to $\boldsymbol{\beta}$ and setting to zero:

$$\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \right) \mathbf{x}_i = 0$$

It will be assumed that $\mathbf{X}$ is of full rank, so we can therefore write:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Following the assumptions outlined in the previous section, the sampling distribution of the vector parameter is given by:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

We are often interested in a linear combination of the parameters. For example, we may want to just know the distribution of a singular $\hat{\beta}_j$, or the expected value of our response given a set of predictors. This can be achieved through the linear combination $\mathbf{c}^T\hat{\boldsymbol{\beta}}$, where $\mathbf{c}$ is a p-dimensional column vector with a value of 1 in component j and 0 everywhere in the former case, otherwise component j is the value of the j^{th} predictor. This linear form has the following sampling distribution:

$$\mathbf{c}^T\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{c}^T\boldsymbol{\beta}, \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c})$$

Of course, σ^2 is unknown and must be estimated. The maximum likelihood estimate gives $\frac{1}{n}\Sigma_i\hat{\epsilon_i}$, but this is biased. By a simple re-scaling, we can make it unbiased to obtain our unbiased estimate:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2$$

It then follows that $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ follows a t distribution with $n-p$ degrees of freedom:

$$\frac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-p}$$

4.3 Regression M-Estimation

As established in Chapter 2, maximum likelihood estimation, and therefore in our case least squares estimation, is a special case of M-estimation. In the context of M-estimation, $\hat{\boldsymbol{\beta}}$ is that which minimises:

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)$$

Where $\hat{\sigma}$ is an scale estimate required to make $\hat{\boldsymbol{\beta}}$ scale equivariant. This estimate is obtained by fitting an $L1$ regression to our sample and then obtaining an estimate from the median of the non-negative residuals (then normalise by dividing by 0.675 so it is unbiased).Differentiating we obtain $\hat{\boldsymbol{\beta}}$ as the solution to the implicit equation:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\mathbf{x}_i = 0$$

We can see that in the case of least squares, we take $\rho(x) = x^2$, and $\hat{\sigma}$ becomes a constant outside the summation sign and so is equivalent to minimizing the sum of squares of the residuals.

In order to perform inference or prediction, we need the distributional form of the parameter vector. In most scenarios, σ is unknown and therefore we must estimate it and then proceed with this sampling distribution. However, it can be shown that if $\hat{\sigma} \xrightarrow{\mathbb{P}} \sigma$ then for large n the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by:

$$\hat{\boldsymbol{\beta}} \rightsquigarrow \mathcal{N}_p(\boldsymbol{\beta}, v(\mathbf{X}^T\mathbf{X})^{-1})$$

Where:

$$v = \sigma^2\frac{\mathbb{E}[\psi(\epsilon/\sigma)^2]}{(\mathbb{E}[\psi'(\epsilon/\sigma)])^2}$$

And it is clear that we must estimate v:

$$\hat{v} = \hat{\sigma}^2\frac{\text{ave}(\psi(\hat{\epsilon}_i/\hat{\sigma})^2)}{(\text{ave}[\psi'(\hat{\epsilon}_i/\hat{\sigma})])^2}\frac{n}{n-p}, \quad \text{ave}(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

For large n, we thus assume $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \hat{v}(\mathbf{X}^T\mathbf{X})^{-1})$

So it can be seen that our sampling distribution for $\hat{\boldsymbol{\beta}}$ is the same as the case for ordinary least squares, bar our estimate for the standard deviation.

- How it adapts concepts from linear regression:
    i Replacing least squares with alternative criteria (e.g., Huber, Tukey loss).
    ii Weighted least squares or iterative methods.
- Practical applications and algorithms (e.g., IRLS, MM-estimation).
- Examples or case studies comparing linear and robust regression outcomes (highlight trade-offs in robustness, efficiency, and computational complexity).

1. **Core Adjustments** Explain how robust regression modifies least squares to handle outliers and assumption violations. Link this back to M-estimation and the ψ-functions from the Fundamentals chapter.
2. **Methods** Iteratively reweighted least squares (IRLS) as an intuitive algorithm. Examples of robust methods (Huber regression, Tukey's bisquare).
3. **Comparison** Use small simulations or examples (e.g., synthetic datasets with and without outliers) to show the advantages of robust regression over OLS.

Linear Regression

Before delving into the intricacies of robust regression, it is essential to first address the principles underpinning standard linear regression. With this foundation, we can then examine how robust regression adapts the approach of linear regression to accommodate potential deviations in our assumptions, such as the presence of outliers. In what follows, the key assumptions of linear regression will be introduced, as well as the method of least squares for estimating model parameters, and the approaches used for statistical inference and prediction, which will serve as a baseline for developing robust methods.

References

[1] Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed., Wiley.

[2] Wilcox, R. R. (2021). *Introduction to Robust Estimation and Hypothesis Testing*. 5th ed., Academic Press.

[3] Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley.

[4] Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.