Statistics 203: Introduction to Regression and Analysis of Variance

Robust methods

Jonathan Taylor

Today's class

- Weighted regression.
- Robust methods.
- Robust regression.

Heteroskedasticity

- In our standard model, we have assumed that

$$\varepsilon \sim N(0, \sigma^2 I).$$

  That is, that the errors are independent and have the same variance (*homoskedastic*).
- We have discussed graphical checks for non-constant variance (*heteroskedasticity*) but not "remedies" for heteroskedasticity.
- Suppose that

$$\varepsilon \sim N(0, \sigma^2 D)$$

  for some known diagonal matrix D.
- Where does D come from? Suppose that we see that variance increases like $f(X_j)$, then we might choose $D_i = f(X_{ij})$.
- What is the "maximum likelihood" thing to do?

MLE for one sample problem

- Consider the simpler problem

$$Y_i \sim N(\mu, \sigma^2 D_i)$$

with σ^2 and D's known.

-

$$-2 \log L(\mu|Y, \sigma) = \sum_{i=1}^n \frac{(Y_i - \mu)^2}{\sigma^2 D_i} + n \log(2\pi\sigma^2) + \sum_{i=1}^n \log(D_i)$$

- Differentiating

$$-2 \sum_{i=1}^n \frac{(Y_i - \widehat{\mu})}{\sigma^2 D_i} = 0$$

implying

$$\widehat{\mu} = \sum_{i=1}^n \frac{Y_i}{D_i} \Big/ \sum_{i=1}^n \frac{1}{D_i}.$$

- Observations are weighted inversely proportional to their variance.

Weighted least squares

- $$\begin{aligned} -2\log L(\beta,\sigma|Y,D) &= \sum_{i=1}^n \frac{(Y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j X_{ij})^2}{\sigma^2 D_i} \\ &= \frac{1}{\sigma^2}(Y - X\beta)^t D^{-1}(Y - X\beta) \\ &= \frac{1}{\sigma^2}(Y - X\beta)^t W(Y - X\beta). \end{aligned}$$

  with $W = D^{-1}$.
- Normal equations:

  $$-2X^tW(Y - X\widehat{\beta}_W) = 0$$

  or,

  $$\widehat{\beta}_W = (X^tWX)^{-1}X^tWY.$$
- Distribution of $\widehat{\beta}_W$

  $$\widehat{\beta}_W \sim N(\beta, \sigma^2(X^tWX)^{-1}).$$

Estimating σ^2

- What are the right residuals?
- If we knew β exactly

$$Y_i - \beta_0 - \sum_{i=1}^{p-1} X_{ij}\beta_j \sim N(0, \sigma^2/W_i).$$

- Suggests that the natural residual is

$$e_{W,i} = \sqrt{W_i}\frac{Y_i - \widehat{Y}_{W,i}}{=}\sqrt{W_i}e_i$$

where

$$\widehat{Y}_W = (X^tWX)^{-1}X^tWY.$$

- Estimate of σ^2

$$\widehat{\sigma}^2_W = \frac{1}{n-p}\sum_{i=1} e^2_{W,i} = \frac{1}{n-p}\sum_{i=1} w_i e_i^2 \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}$$

Weighted regression example

Robust methods

- We also discussed outlier detection but no specifi c remedies.
- One alternative is to discard potential outliers – not always a good idea.
- Outliers can really mess up the sample mean, but have relatively effect on the sample median.
- Could also “downweight” outliers: basis of robust techniques.
- Another “cause” of outliers may be that the data is not really normally distributed.

Example

- Suppose that we have a sample $(Y_i)_{1 \leq i \leq n}$ from
$$f(y|\mu, \sigma) = \frac{1}{2\sigma} e^{-|y-\mu|/\sigma}.$$
This has heavier tails than the normal distribution.
- MLE for μ
$$\widehat{\mu} = \text{argmin}_{\mu} \sum_{i=1}^{n} |Y_i - \mu|.$$
- It can be shown that $\widehat{\mu}$ is the sample median (exercise in STATS116).
- Take home message: if errors are not really normally distributed then least squares is not MLE and the MLE downweights large residuals relative to least squares.

M-estimators

- Depending on the error distribution of ε (assuming i.i.d.) we get different optimization problems.
- Suppose that

$$f(y|\mu, s) \propto e^{-\rho((y-\mu)/s)}.$$

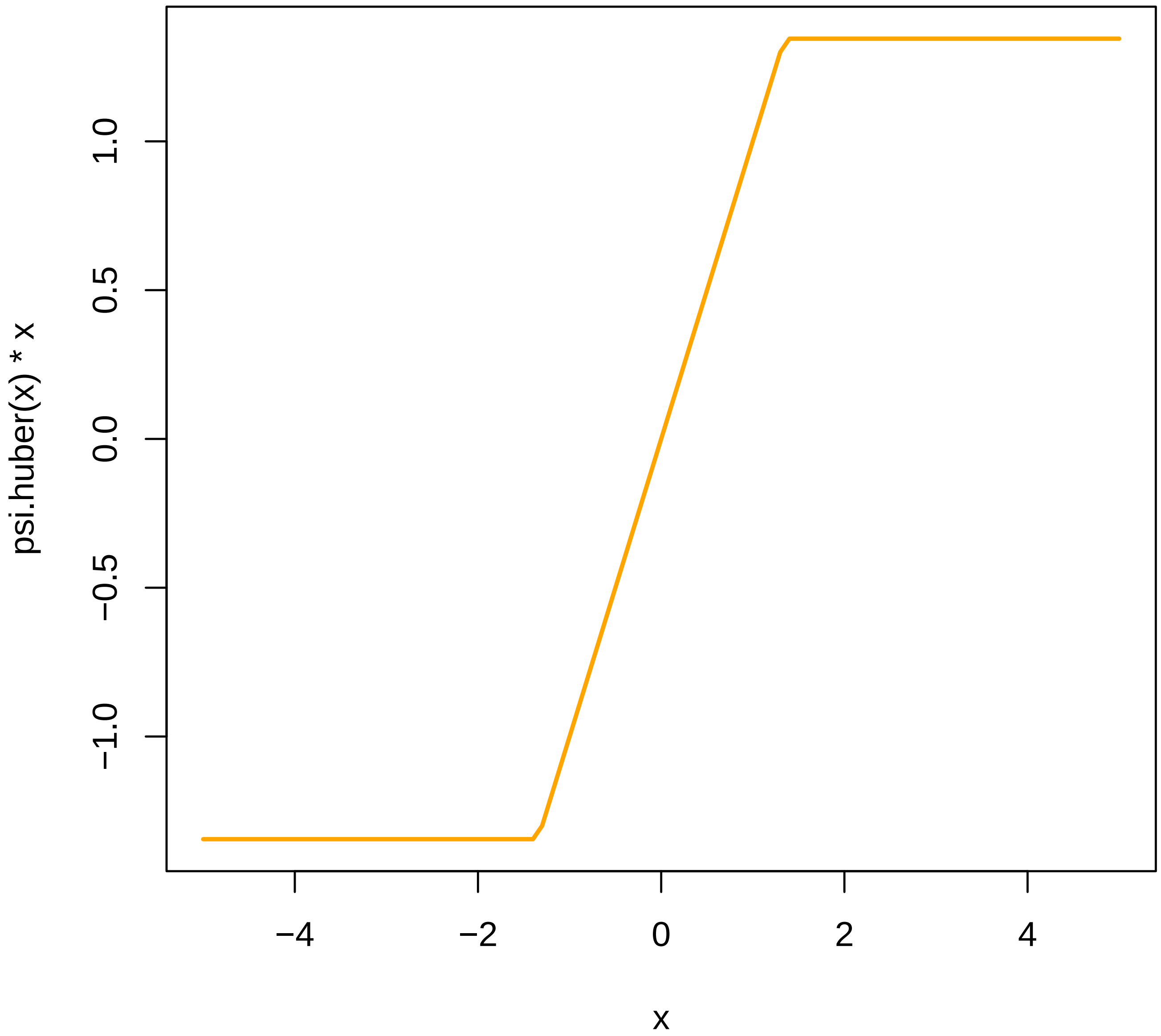
- MLE:

$$\widehat{\mu} = \text{argmin}_{\mu} \sum_{i=1}^{n} \rho\left(\frac{Y_i - \mu}{s}\right).$$

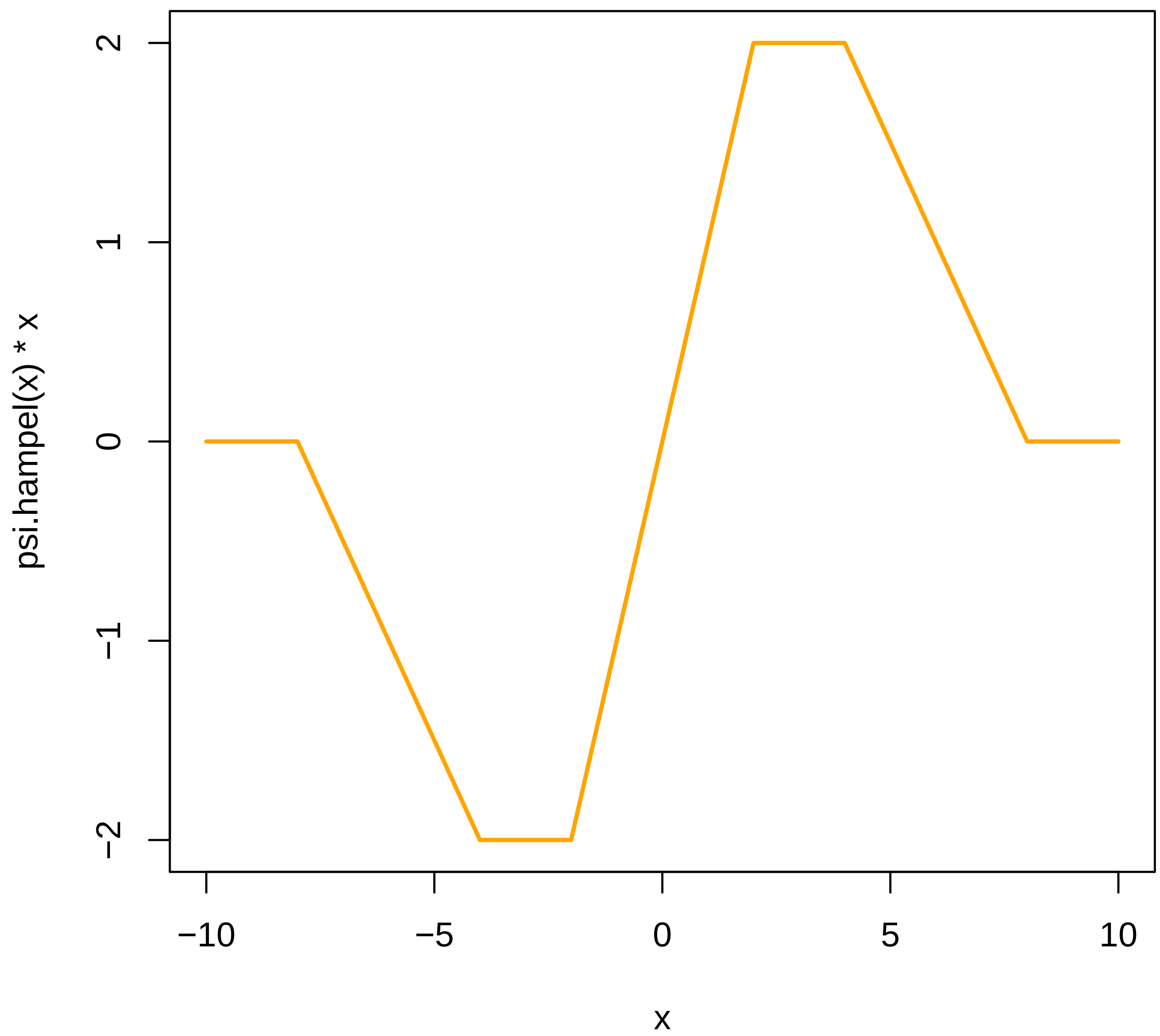
- For every ρ, we get a different estimator: an M-estimator. Let $\psi = \rho'$.
- Generalizes easily to regression problem

$$\widehat{\beta} = \text{argmin}_{\beta} \sum_{i=1}^{n} \rho\left(\frac{Y_i - \beta_0 - \sum_{j=1}^{p-1} X_{i,j}\beta_j}{s}\right).$$

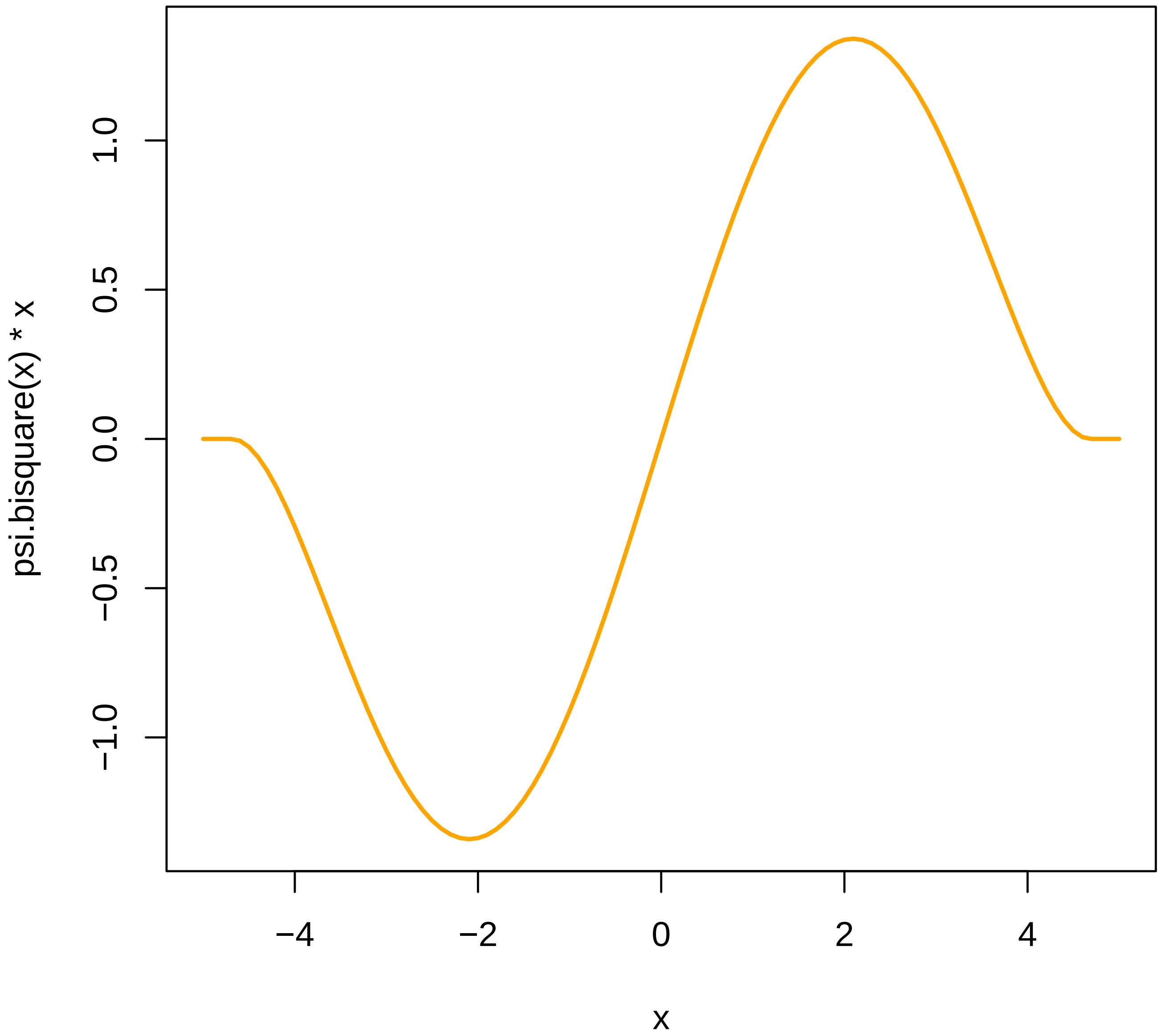
Huber's ψ



Hampel's ψ



Tukey's ψ



Solving for $\widehat{\beta}$

- Assume for now that s is known in the M-estimator.
- Normal equations:

$$\sum_{i=1}^{n} X_{ij}\psi\left(\frac{Y_i - \sum_{j=0}^{p-1} X_{ij}\widehat{\beta}_j}{s}\right) = 0, \qquad 0 \leq j \leq p-1$$

where $\psi = \rho'$.

- Set

$$W_i = \frac{\psi\left(\frac{Y_i - \sum_{j=0}^{p-1} X_{ij}\widehat{\beta}_j}{s}\right)}{Y_i - \sum_{j=0}^{p-1} X_{ij}\widehat{\beta}_j}.$$

- Then

$$\sum_{i=1}^{n} X_{ij}W_i(Y_i - \sum_{j=0}^{p-1} X_{ij}\widehat{\beta}_j) = 0, \qquad 0 \leq j \leq p-1.$$

Or

$$\widehat{\beta}X^tWX - X^tWY = 0.$$

Iteratively reweighted least squares (IRLS)

- Although we are trying to estimate β above, given an initial estimate $\widehat{\beta}^0$ we can compute initial weights

$$W_i^0 = \psi\left(\frac{Y_i - \sum_{j=0}^{p-1} X_{ij}\widehat{\beta}_j^0}{s}\right) \Big/ \left(Y_i - \sum_{j=0}^{p-1} X_{ij}\widehat{\beta}_j^0\right)$$

ans solve for $\widehat{\beta}^1$

$$\widehat{\beta}^1 = (X^tW^0X)^{-1}X^tW^0Y.$$

- Now we can recompute weights, and reestimate $\widehat{\beta}$.
- In general, given weights W^j we can solve

$$\widehat{\beta}^{j+1} = (X^tW^jX)^{-1}X^tW^jY.$$

- This is very similar to a Newton-Raphson technique.
- Used over and over again in statistics.

Robust estimate of scale

- In general s needs to be estimated. A popular estimate

$$s = MAD(e_1, \dots, e_n)/0.6745.$$

where

$$MAD(e_1, \dots, e_n) = \text{Median}|e_i - \text{Median}(e_1, \dots, e_n)|.$$

- Scale can be estimated at each stage of IRLS procedure based on current residuals, or based on some "very resistant" fit. (LMS or LTS below)
- The constant 0.6745 is chosen so that s is asymptotically unbiased for σ if the e_i's are $N(0, \sigma^2)$.

Other resistant fitting methods

- If ψ is not redescending (i.e. does not return to 0) then robust regression is still susceptible to outliers.
- Least median of squares (LMS)

$$\widehat{\beta} = \underset{\beta}{\text{argmin}}\, \mathrm{Median}(Y_i - \sum_{j=0}^{p-1} X_{ij}\beta_j)^2$$

- Least trimmed squares (LTS): fi x some $q < n$

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{q} (Y_i - \sum_{j=0}^{p-1} X_{ij}\beta_j)^2_{(i)}.$$

Here the $\cdot_{(i)}$ represents “order statistic”.

Why not always use robust regression?

- Inference: seems reasonable to treat estimate covariance of $\widehat{\beta}$ as

$$\widehat{\sigma}^2(X^t\widehat{W}X)^{-1}, \qquad \widehat{\sigma}^2 = \frac{\sum_{i=1}^n \widehat{W}_i e_i^2}{n-p}.$$

  What about degrees of freedom?
- R uses a different estimate.
- Effi ciency: this robustness comes at a cost. Even asymptotically, confi dence intervals for robust estimates are wider than least squares (if least squares model is applicable).
- *Asymptotic Relative Efficiency*:

$$ARE(\widehat{\beta}_{LS}, \widehat{\beta}_{robust}) = \lim_{n\to\infty} \frac{\text{Var}(\widehat{\beta}_{LS})}{\text{Var}(\widehat{\beta}_{robust})} < 1.$$

- Biggest advantage: can have higher breakdown. Median has 50% breakdown, sample mean has 0%.