C. Common approach for Robust Regression:

1. Minimize difference $\sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j|$.

2. Note the absence of $^2$, which would be present in regular least squares regression.

3. Equivalent to minimizing $\sum_j e_j^+ + e_j^-$, for $e_j^+ \geq 0$, $e_j^- \geq 0$,
   $e_j^+ - e_j^- = Y_j - \beta_1 - \beta_2 X_j \forall j$.
   a. This minimization is an example of linear programming.
   b. Solution is well-known, but computationally intensive.
      i. Harder than taking derivative, since function to be optimized is non-differentiable.

4. This solution is called $L^1$ regression, after the power on $|Y_j - \beta_1 - \beta_2 X_j|$.

5. Solution is also called quantile regression.
   a. If $\beta_2 = 0$, best $\beta_1$ is median.
   b. Just as median is not always uniquely defined, these estimates are not necessarily uniquely defined.
   c. Can replace $\sum_j e_j^+ + e_j^-$ by $\tau \sum_j e_j^+ + (1 - \tau) \sum_j e_j^-$.
      i. Runs regression line through $1 - \tau$ quantile of $Y|X$.

6. At minimum $\hat{\beta}$, $\sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j|$ should not decrease as
β moves away from \( \hat{\beta} \).

a. Since \( \sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j| \) is piecewise linear, function \( S(\beta) \) giving “derivative” should be piecewise constant, with jumps.

b. \( \hat{\beta} \) should satisfy \( S(\hat{\beta}) \approx 0 \)
   
i. Estimate either sets the function to \( 0 \), or is a point where it jumps across \( 0 \).

c. \( S(\beta) \) may be expressed as a rank statistic

d. \( H_0 : \beta = \beta^o \) may be tested by comparing \( S(\beta^o) \) to \( 0 \)
   
i. Either exactly or asymptotically.

e. Test can be inverted to give confidence sets for \( \beta \). [Mark A R]
   
[Mark A sas]

MMY: 3

XIV. Robustness

A. Breakdown tolerance

1. How many bad values in a data set does it take to completely mess up an estimator?

2. Definition:

   a. Let \( \epsilon_n^* = \inf \{ m | \sup \{ \hat{\theta}(y, x) | y \in \mathbb{R}^m \} = \infty \}/n \).
b. Suppose that there exists a an integer such that

i. \( \hat{\theta} \in (X_{(a+1)}, X_{(n-a)}) \)

ii. For fixed \( x_{(a+2)}, \ldots, x_{(n)} \), if \( x_{(a+1)} \to -\infty \), then \( \hat{\theta} \to -\infty \).

iii. For fixed \( x_{(1)}, \ldots, x_{(n-a-1)} \), if \( x_{(n-a)} \to +\infty \), then \( \hat{\theta} \to +\infty \).

iv. Then define the tolerance as \( \tau_n = a^*/n \), where \( a^* \) is the smallest such integer.

v. The asymptotic tolerance \( \tau \) is the limit of this value.

3. Examples:

a. Mean: \( \tau_n = 0 \), \( \tau = 0 \).

b. Median: \( \tau_n = ((n - 1)/2)/n \) if \( n \) odd, \( \tau_n = (n/2 - 1)/n \), \( \tau = 1/2 \).

c. \( \alpha \)-Trimmed mean (throw out \( \lfloor \alpha n \rfloor \) highest and lowest variables and take average of rest): \( \tau_n = \lfloor \alpha n \rfloor /n \), \( \tau = \alpha \).

d. Midrange \( \tau_n = \tau = 0 \).

4. When parameter value is bounded (ex., binomial parameter in \([0, 1]\), can replace \( \pm \infty \) by endpoints.

B. Sensitivity Curve
1. Suppose we have an estimator \( \hat{\theta} \) adaptable to various sample sizes.

2. Calculate it for a sample \( x_1, \ldots, x_n : \hat{\theta}(x_1, \ldots, x_n) \).

3. Consider an additional observation \( x_0 \).

4. Sensitivity curve is the effect of adding \( x_0 : \hat{\theta}(x_0, x_1, \ldots, x_n) - \hat{\theta}(x_1, \ldots, x_n) \).

5. Examples:
   a. For mean, is linear with slope \( 1/(n + 1) \)
   b. For median,
      i. \( n \) even, is linear with slope 1 for \( x_0 \in [x(n/2), x(n/2)+1] \), and flat on \([x(n/2), x(n/2)+1]^c\).
      ii. \( n \) even, is linear with slope \( 1/2 \) for \( x_0 \in [x((n-1)/2), x((n+1)/2)+1] \), and flat on \([x((n-1)/2), x((n+1)/2)+1]^c\).
   c. For trimmed mean, has slope \( 1/(m + 1) \) when \( x_0 \) is in the range of untrimmed observations, and flat outside that, with \( m = \) the number of untrimmed observations.

6. Curve depends on sample size, and so to get a curve independent of sample size often multiply by \( n - 1 \) to get standardized curve.
Lecture 13

7. Curve depends on data, but basic shape does not.
   a. Curve bounded indicates that an outlier stops having an increasing effect at some point,
   b. unbounded curve indicates that the effect of an outlier keeps increasing as its magnitude increases.  

C. Way to get a more robust estimator:
   1. Model: \( Y_i = \beta^T x_i + \epsilon_i \), \( \epsilon_i \) iid, \( \text{Var}[\epsilon_i] = \sigma^2 \) for \( \sigma^2 \) known.
   2. Recall: Least squares estimator minimizes \( \sum_{i=1}^{n}(Y_i - \beta^T x_i)^2 \)
      a. Solves: Least squares estimator \( \sum_{i=1}^{n} 2(Y_i - \beta^T x_i)x_i = 0 \)
   3. Downweight large residuals: \( \sum_{i=1}^{n} \psi((Y_i - \beta^T x_i)/\sigma)\sigma x_i = 0 \)
   4. Can express as: \( \sum_{i=1}^{n} w_i(Y_i - \beta^T x_i)x_i = 0 \) for \( w_i = w(r) = \psi(r)/r \), \( r \) is standardized residual.
   5. Common choices for \( w(r) \)
      a. Huber: \( w(r) = \begin{cases} 1 & \text{if } |r| < c \\ c/|r| & \text{if } |r| \geq c \end{cases} \).
      b. Bisquare: \( w(r) = \begin{cases} (1 - (r.c)^2)^2 & \text{if } |r| < c \\ 0 & \text{if } |r| \geq c \end{cases} \).
      c. Median: \( w(r) = \begin{cases} 1/c & \text{if } r = 0 \\ 1/|r| & \text{if } r \neq 0 \end{cases} \).
   D. In more realistic setting with \( \text{Var}[\epsilon_i] \) to be estimated, estimating equations are modified a bit to avoid opportunity to get better
estimate by shrinking $\sigma$ to zero. [Mark Császár]