6.S085 Statistics for Research Projects	**IAP 2014**

Lecture 4: January 24

Lecturer: Ramesh Sridharan and George Chen — *Notes by: William Li*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

4.1 More Regression

If testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$:

- construct 95% CI
- ask if 0 is in the CI
- if 0 not in the CI, then p-value is ¡ 0.05

4.2 Residual Analysis

Residual for ith point: $\hat{y} - y_i$

Intuitively, $\hat{y} - y_i$ versus $\hat{y}$ should look like noise if the model is good

$$\begin{aligned}
\hat{\epsilon} &= y - \hat{y} && (4.1)\\
&= y - X\hat{\beta} && (4.2)\\
&= y - [X(X^TX)^{-1}X^T]y \qquad = (I-H)y && (4.3)\\
&= (I-H)X\beta + (I-H)\epsilon && (4.4)\\
&= (I-H)\epsilon && (4.5)
\end{aligned}$$

Standardized residuals: $\frac{-\hat{\epsilon}_i}{\sqrt{1-H_{ii}}}$

Standarized residuals will have variance σ^2

Model says that the standardized residuals should each have variance σ^2

Fact: Under the model, $\hat{y} - y_i$ is uncorrelated with $\hat{y}_i$

4.3 Outliers

Informally:

Outlier: point far away from the rest of the points

Leverage: How far point is from the rest of the points along the x axis

Influential Point: point (typically with high leverage) that substantially affects the estimated slope β_1

Leverage for ith point is defined H_{ii}

In 1D case: $H_{ii} = \frac{x_i^2}{\sum x_j^2}$

4.3.1 Influence

Measured using Cook's distance:

For the ith point:

$D_i = \frac{1}{p \cdot \text{MSE}} \frac{H_{ii}}{(1-H_{ii})^2} \hat{\epsilon}^2$

Higher leverage means higher influence ($\frac{H_{ii}}{(1-H_{ii})^2}$)

Higher fitting error means higher influence ($\hat{\epsilon}^2$)

4.4 Robust Regression

Can we be resilient to to outliers without manually removing them beforehand?

Recall: We are trying to minimize a cost function:

$\min_\beta \sum_{i=1}^{n} (Y_i - X_i\beta)^2$

$(Y_i - X_i\beta)^2$ (squared "loss")

As the squared loss gets bigger, the loss function will become enormous, so the

Squared loss: $\rho(r) = r^2$

4.4.1 Least absolute deviation (LAD)

$\rho(r) = |r|$

$\min_\beta \sum_{i=1}^{n} (Y_i - X_i\beta)$

Large deviations don't hurt us as much

However, not stable: a change in x may dramatically impact β

4.4.2 Huber Loss

Close to the origin: squared loss

Further away: grow linearly

4.4.3 Bisquare

When your points are really far away, may not even consider them

4.4.4 Probabilistic Interpretations

Different cost functions correspond to different distributions used (for ϵ)

$\epsilon \sim N(0, \sigma^2)$ corresponds to squared loss (98% of the probability mass is within 3 standard deviations)

$\epsilon \sim$ Laplace corresponds to LAD (more probability mass further out)

4.4.5 RANSAC

Previous approaches we've talked about so far solve optimization problems

RANSAC (Random Sample Consensus) – widely used in practice

Randomly choose two points, fit a line, find and count inliers

- for iteration t=1...T:
  - choose random subset of points, I_t as "inliers"
  - fit model to points I_t
  - find all points that are within some α of the model and add to I_t
  - (Optional:) Refit model to points I_t
  - Compute score for model
- Choose model with the highest score

4.5 Sparse Regression

Imagine p (the number of predictors) is huge (e.g. $p \approx 10^6$)

$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$

Want to figure out some small subset of predictors relevant to predicting y (want most $\hat{\beta}_k$'s to be 0

Why?

4.6 Logistic Regression