Weighted Least Squares

GR 5205 / GU 4205
Section 3

Columbia University
Xiaofei Shi

Least Squares

- When we use ordinary least squares to estimate linear regression, we minimize the mean squared error;

$$\|y - x\beta\|^2 = \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2$$

Other Least Squares

- When we use ordinary least squares to estimate linear regression, we minimize the mean squared error;

$$\|y - x\beta\|^2 = \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2$$

- Suppose we change the task a little bit:

$$\sum_{i=1}^{n} w_i \left(y_i - x_i^\top \beta\right)^2 = (y - x\beta)^\top w\,(y - x\beta)$$

Other Least Squares

- When we use ordinary least squares to estimate linear regression, we minimize the mean squared error;

$$\|y - x\beta\|^2 = \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2$$

- Suppose we change the task a little bit:

$$\sum_{i=1}^{n} w_i \left(y_i - x_i^\top \beta\right)^2 = (y - x\beta)^\top w\,(y - x\beta)$$

Different weights are assigned to different data points to illustrate that they are not of the same importance!

Weighted least squares

Weighted least squares estimator:

$$\widehat{\beta} = \left(x^\top w x\right)^{-1} x^\top w Y$$

- Relationship: $Y_i = x_i^\top \beta + \epsilon_i$
- Assumptions: $\mathbb{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma_i^2.$
- when $w = \sigma^{-2} I_n$, we recover the OLS case.

Recall our linear regression models:

Recall what we have for simple linear regression model:

- Relationship: $Y = x\beta + \epsilon$;
- Assumptions: $\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \sigma^2 I_n$.

Recall our linear regression models:

Recall what we have for simple linear regression model:

- Relationship: $Y = x\beta + \epsilon$;
- Assumptions: $\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \sigma^2 I_n$.

If we change the model assumptions a little bit:

- Assumptions: $\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \Sigma$.

Recall our linear regression models:

Ordinary Least Square (OLS)

Recall what we have for simple linear regression model:

- Relationship: $Y = x\beta + \epsilon$;
- Assumptions: $\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \sigma^2 I_n$.

Weighted Least Square (WLS)

If we change the model assumptions a little bit:

- Assumptions: $\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \Sigma$.

Weighted Least Squares

- Focusing accuracy: we may care very strongly about predicting the response for certain values of the input — ones we expect to see often again, ones where mistakes are especially costly or embarrassing or painful, etc. — than others. If we give the points near that region big weights, and points elsewhere smaller weights, the regression will be pulled towards matching the data in that region.

- Discounting imprecision.
  - homoskedastic: constant variance
  - heteroskedastic: magnitude of noise is not constant

- Try a different model for the data.

Heteroskedasticity

Assume we know the value of $\sigma_1, \ldots, \sigma_n$

- Relationship: $Y_i = x_i^\top \beta + \epsilon_i$
- Assumptions: $\mathbb{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma_i^2.$

Heteroskedasticity

Assume we know the value of $\sigma_1, \ldots, \sigma_n$

- Relationship: $Y_i = x_i^\top \beta + \epsilon_i$
- Assumptions: $\mathbb{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma_i^2.$

How to estimate the model parameters?

In other words, how to choose the weights for different data points?

Heteroskedasticity

Assume we know the value of $\sigma_1, \ldots, \sigma_n$

Heteroskedasticity

Assume we know the value of $\sigma_1, \ldots, \sigma_n$

with the weight matrix w,

- $\widehat{\beta} = \left(x^\top w x\right)^{-1} x^\top w Y.$
- $\mathbb{E}[\widehat{\beta}] = \beta, \quad \mathrm{Var}[\widehat{\beta}] = \left(x^\top w x\right)^{-1} x^\top w \Sigma w x \left(x^\top w x\right)^{-1}.$

The Gauss-Markov Theorem

$\widehat{\beta} = \left(x^\top \Sigma^{-1} x\right)^{-1} x^\top \Sigma^{-1} Y$ has the least variance among all possible linear, unbiased estimators of the regression coefficients.

- When $\Sigma = \sigma^2 I_n$, the Gauss-Markov's theorem yields the optimality of OLS.
- The unbiased constraint is powerful. In fact, biased estimator will have even smaller variance.

The Gauss-Markov Theorem

- If all the noise variances are equal, then we've proved the optimality of OLS.
- The theorem doesn't rule out linear, biased estimators with smaller variance. As a trivial example, 0Y is linear and has variance 0, but is (generally) very biased.
- The theorem also doesn't rule out non-linear unbiased estimators of smaller variance. Or indeed non-linear biased estimators of even smaller variance.
- The proof actually doesn't require the variance matrix to be diagonal.

How to find the variance and weights

Multiple measurements. The easiest case is when our measurements of the response are actually averages over individual measurements, each with some variance σ^2. If some Y_i are based on averaging more individual measurements than others, there will be heteroskedasticity. The variance of the average of n_i uncorrelated measurements will be σ^2/n_i, so in this situation we could take $w_i \propto n_i$.

Binomial counts Suppose our response variable is a count, derived from a binomial distribution, i.e., $Y_i \sim \text{Binom}(n_i, p_i)$. We would usually model p_i as a function of the predictor variables — at this level of statistical knowledge, a linear function. This would imply that Y_i had expectation $n_i p_i$, and variance $n_i p_i (1 - p_i)$. We would be well-advised to use this formula for the variance, rather than pretending that all observations had equal variance.