Remedial Measures

Zhenisbek Assylbekov

Department of Mathematics

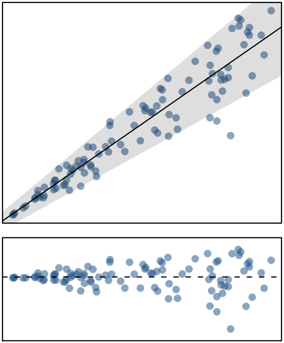
Regression Analysis

Weighted Least Squares

Ridge Regression

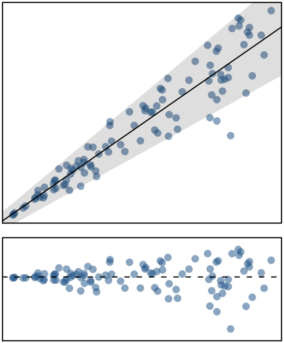
Robust Regression

Non-homogeneous variance



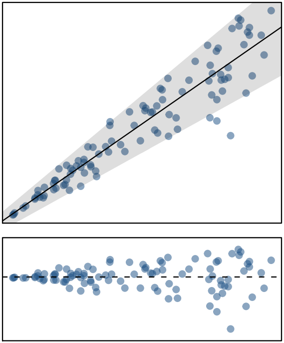
- Chapters 3 and 6 discuss transformations of $x_1, \ldots, x_k$ and/or Y.

Non-homogeneous variance



- Chapters 3 and 6 discuss transformations of $x_1, \ldots, x_k$ and/or Y.
- More advanced remedy: **weighted least squares** (WLS) regression.

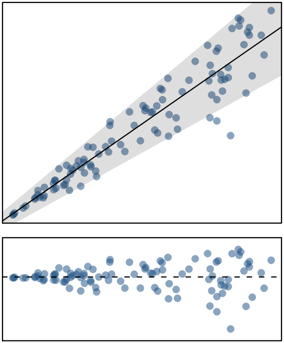
Non-homogeneous variance



- Chapters 3 and 6 discuss transformations of $x_1, \ldots, x_k$ and/or Y.
- More advanced remedy: **weighted least squares** (WLS) regression.
- Model is as before

$$Y_i = \beta_0 0 + \beta_1 \cdot x_{i1} + \ldots + \beta_k \cdot x_{ik} + \epsilon_i,$$

Non-homogeneous variance



- Chapters 3 and 6 discuss transformations of $x_1, \ldots, x_k$ and/or Y.
- More advanced remedy: **weighted least squares** (WLS) regression.
- Model is as before

$$Y_i = \beta_0 0 + \beta_1 \cdot x_{i1} + \ldots + \beta_k \cdot x_{ik} + \epsilon_i,$$

- Except that

$$\epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_i^2)$$

Weighted least squares

- We have $\mathrm{Var}[Y_i] = \sigma_i^2$.

Weighted least squares

- We have $\mathrm{Var}[Y_i] = \sigma_i^2$.
- Idea: Give observations with higher variance less weight in the regression fitting.

Weighted least squares

- We have $\mathrm{Var}[Y_i] = \sigma_i^2$.
- Idea: Give observations with higher variance less weight in the regression fitting.
- Let $\omega_i = \frac{1}{\sigma_i^2}$. WLS solves

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \omega_i [Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2$$

Weighted least squares

- We have $\mathrm{Var}[Y_i] = \sigma_i^2$.
- Idea: Give observations with higher variance less weight in the regression fitting.
- Let $\omega_i = \frac{1}{\sigma_i^2}$. WLS solves

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \omega_i [Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2$$

- Or in matrix notation:

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

where $\boldsymbol{\Omega} = \mathrm{diag}[\omega_1, \ldots, \omega_n] \in \mathbb{R}^{n \times n}$.

Solving WLS

Let us rewrite the WLS objective as:

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Solving WLS

Let us rewrite the WLS objective as:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{1/2} \boldsymbol{\Omega}^{1/2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

Solving WLS

Let us rewrite the WLS objective as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{1/2} \boldsymbol{\Omega}^{1/2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta})
\end{aligned}$$

Solving WLS

Let us rewrite the WLS objective as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{1/2} \boldsymbol{\Omega}^{1/2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}\|^2 \qquad \text{(This is OLS!)}
\end{aligned}$$

Solving WLS

Let us rewrite the WLS objective as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{1/2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}\|^2 \qquad \text{(This is OLS!)}
\end{aligned}$$

Hence

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \left((\boldsymbol{\Omega}^{1/2}\mathbf{X})^\top (\boldsymbol{\Omega}^{1/2}\mathbf{X}) \right)^{-1} \left(\boldsymbol{\Omega}^{1/2}\mathbf{X} \right)^\top \boldsymbol{\Omega}^{1/2}\mathbf{Y}$$

Solving WLS

Let us rewrite the WLS objective as:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{1/2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Omega}^{1/2}\mathbf{Y} - \boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}\|^2 \qquad \text{(This is OLS!)}
\end{aligned}
$$

Hence

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{WLS}} &= \left((\boldsymbol{\Omega}^{1/2}\mathbf{X})^\top(\boldsymbol{\Omega}^{1/2}\mathbf{X})\right)^{-1}\left(\boldsymbol{\Omega}^{1/2}\mathbf{X}\right)^\top \boldsymbol{\Omega}^{1/2}\mathbf{Y} \\
&= (\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{Y}
\end{aligned}
$$

But σ_i's are unknown!

- However, $\sigma_1, \ldots, \sigma_n$ are usually unknown!

But σ_i's are unknown!

- However, $\sigma_1, \ldots, \sigma_n$ are usually unknown!
- Notice, that in OLS

$$\mathrm{E}[e_i^2] = \mathrm{Var}[e_i] + (\mathrm{E}[e_i])^2 = \mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2(1 - h_{ii})$$

But σ_i's are unknown!

- However, $\sigma_1, \ldots, \sigma_n$ are usually unknown!
- Notice, that in OLS

$$\mathrm{E}[e_i^2] = \mathrm{Var}[e_i] + (\mathrm{E}[e_i])^2 = \mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2(1 - h_{ii})$$

- So e_i^2 in OLS estimates σ^2 and $|e_i|$ estimates σ if $h_{ii} \approx 0$.

But σ_i's are unknown!

- However, $\sigma_1, \ldots, \sigma_n$ are usually unknown!
- Notice, that in OLS

$$\mathrm{E}[e_i^2] = \mathrm{Var}[e_i] + (\mathrm{E}[e_i])^2 = \mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2(1 - h_{ii})$$

- So e_i^2 in OLS estimates σ^2 and $|e_i|$ estimates σ if $h_{ii} \approx 0$.
- Look at plots of $|e_i|$ from an OLS fit against x_i's and $\hat{Y}_i$'s to see how σ changes with predictors or fitted values.

But σ_i's are unknown!

- However, $\sigma_1, \ldots, \sigma_n$ are usually unknown!
- Notice, that in OLS

$$\mathrm{E}[e_i^2] = \mathrm{Var}[e_i] + (\mathrm{E}[e_i])^2 = \mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2(1 - h_{ii})$$

- So e_i^2 in OLS estimates σ^2 and $|e_i|$ estimates σ if $h_{ii} \approx 0$.
- Look at plots of $|e_i|$ from an OLS fit against x_i's and $\hat{Y}_i$'s to see how σ changes with predictors or fitted values.
- For example, if $|ei|$ increases linearly with $\hat{Y}_i$, then we'll fit

$$|e_i| = \alpha_0 + \alpha_1 \cdot x_{i1} + \ldots + \alpha_k \cdot x_{ik} + \delta_i$$

and obtain the fitted values $\widehat{|e_i|}$.

But σ_i's are unknown!

- However, $\sigma_1, \ldots, \sigma_n$ are usually unknown!
- Notice, that in OLS

$$\mathrm{E}[e_i^2] = \mathrm{Var}[e_i] + (\mathrm{E}[e_i])^2 = \mathrm{Var}[Y_i - \hat{Y}_i] = \sigma^2(1 - h_{ii})$$

- So e_i^2 in OLS estimates σ^2 and $|e_i|$ estimates σ if $h_{ii} \approx 0$.
- Look at plots of $|e_i|$ from an OLS fit against x_i's and $\hat{Y}_i$'s to see how σ changes with predictors or fitted values.
- For example, if $|ei|$ increases linearly with $\hat{Y}_i$, then we'll fit

$$|e_i| = \alpha_0 + \alpha_1 \cdot x_{i1} + \ldots + \alpha_k \cdot x_{ik} + \delta_i$$

  and obtain the fitted values $\widehat{|e_i|}$.
- Or, e.g., if $|\epsilon_i|$ increases linearly w.r.t. x_{i4} only, then we'll fit $|e_i| = \alpha_0 + \alpha_1 \cdot x_{i4} + \delta_i$

Putting it all together

1. Regress Y against predictor variable(s) as usual (OLS), and obtain $e_1, \ldots, e_n$ & $\hat{Y}_1, \ldots, \hat{Y}_n$.

Putting it all together

1. Regress Y against predictor variable(s) as usual (OLS), and obtain $e_1, \ldots, e_n$ & $\hat{Y}_1, \ldots, \hat{Y}_n$.
2. Regress $|e_i|$ against (all or some) predictors $x_1, \ldots, x_k$ or fitted values $\hat{Y}$.

Putting it all together

1. Regress Y against predictor variable(s) as usual (OLS), and obtain $e_1, \dots, e_n$ & $\hat{Y}_1, \dots, \hat{Y}_n$.
2. Regress $|e_i|$ against (all or some) predictors $x_1, \dots, x_k$ or fitted values $\hat{Y}$.
3. Let $\widehat{|e_i|}$ be the fitted values for the regression in 2.

Putting it all together

1. Regress Y against predictor variable(s) as usual (OLS), and obtain $e_1, \ldots, e_n$ & $\hat{Y}_1, \ldots, \hat{Y}_n$.
2. Regress $|e_i|$ against (all or some) predictors $x_1, \ldots, x_k$ or fitted values $\hat{Y}$.
3. Let $\widehat{|e_i|}$ be the fitted values for the regression in 2.
4. Define $\omega_i = 1/\widehat{|e_i|}^2$ and feed them into `lm` command using the `weights` parameter.

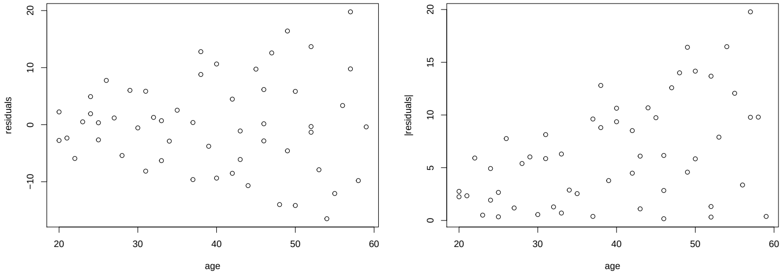
Example: diastolic blood pressure vs age

We are interested in studying the relationship between diastolic blood pressure and age among healthy adult women 20 to 60 years old.

Example: diastolic blood pressure vs age

We are interested in studying the relationship between diastolic blood pressure and age among healthy adult women 20 to 60 years old.

Fitting an OLS $\text{dbp}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \epsilon_i$ gives:



OLS vs WLS

https://github.com/zh3nis/MATH440/blob/main/chp11/dbp.R

```
> summary(ols)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.15693    3.99367  14.061  < 2e-16 ***
age          0.58003    0.09695   5.983 2.05e-07 ***
---
Residual standard error: 8.146 on 52 degrees of freedom
Multiple R-squared:  0.4077,Adjusted R-squared:  0.3963
```

OLS vs WLS

https://github.com/zh3nis/MATH440/blob/main/chp11/dbp.R

```
> summary(ols)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.15693    3.99367  14.061  < 2e-16 ***
age          0.58003    0.09695   5.983 2.05e-07 ***
---
Residual standard error: 8.146 on 52 degrees of freedom
Multiple R-squared:  0.4077,Adjusted R-squared:  0.3963

> summary(wls)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.56577    2.52092  22.042  < 2e-16 ***
age          0.59634    0.07924   7.526 7.19e-10 ***
---
Residual standard error: 1.213 on 52 degrees of freedom
Multiple R-squared:  0.5214,Adjusted R-squared:  0.5122
```

Comments

- $\mathrm{s}[b_1]$ reduced from 0.097 (OLS) to 0.079 (WLS)

Comments

- $\mathrm{s}[b_1]$ reduced from 0.097 (OLS) to 0.079 (WLS)
- R^2 is no longer interpreted the same way in terms of amount of total variability explained by model.

Comments

- $\mathrm{s}[b_1]$ reduced from 0.097 (OLS) to 0.079 (WLS)
- R^2 is no longer interpreted the same way in terms of amount of total variability explained by model.
- In WLS, standard inferences about coefficients may not be valid for small sample sizes when weights are estimated from the data.

Comments

- $\mathrm{s}[b_1]$ reduced from 0.097 (OLS) to 0.079 (WLS)
- R^2 is no longer interpreted the same way in terms of amount of total variability explained by model.
- In WLS, standard inferences about coefficients may not be valid for small sample sizes when weights are estimated from the data.
- If MSE of the WLS regression is near 1, then our estimation of the σ_i function is okay. Here it's 1.21.

Weighted Least Squares

Ridge Regression

Robust Regression

Ridge Regression

If some predictors are collinear, then columns of $\mathbf{X}$ become linearly dependent, and $\mathbf{X}$ loses its rank.

Ridge Regression

If some predictors are collinear, then columns of $\mathbf{X}$ become linearly dependent, and $\mathbf{X}$ loses its rank.

Exercise: Show that for $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $n \geq d$

$$\mathrm{rank}(\mathbf{X}) < d \quad \Rightarrow \quad \nexists\, (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Ridge Regression

If some predictors are collinear, then columns of $\mathbf{X}$ become linearly dependent, and $\mathbf{X}$ loses its rank.

Exercise: Show that for $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $n \geq d$

$$\operatorname{rank}(\mathbf{X}) < d \quad \Rightarrow \quad \nexists \, (\mathbf{X}^\top \mathbf{X})^{-1}.$$

This is bad for OLS, because $\mathbf{X}^\top \mathbf{X}$ will not be invertible.

Simple solution: add penalty term into the cost function:

$$Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2,$$

where λ is a *hyperparameter*, to be chosen through a criterion like PRESS or training/validation approach.

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$Q(\boldsymbol{\beta}) =$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^\top \mathbf{Y}$$

Proof.

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \end{aligned}$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \\ \nabla_{\boldsymbol{\beta}} Q &= -\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \end{aligned}$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \\ \nabla_{\boldsymbol{\beta}} Q &= -\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \\ &= 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^\top\mathbf{Y} \end{aligned}$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$\begin{aligned}
Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \\
\nabla_{\boldsymbol{\beta}} Q &= -\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \\
&= 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^\top\mathbf{Y} = \mathbf{0}
\end{aligned}$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \\ \nabla_{\boldsymbol{\beta}} Q &= -\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \\ &= 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^\top\mathbf{Y} = \mathbf{0} \end{aligned}$$

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y} \quad \Rightarrow$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$\begin{aligned}
Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \\
\nabla_{\boldsymbol{\beta}} Q &= -\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \\
&= 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^\top\mathbf{Y} = \mathbf{0}
\end{aligned}$$

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y} \quad \Rightarrow \quad \boxed{\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}}$$

$\square$

Solving Ridge Regression

Theorem. The function $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ reaches its minimum at

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$$

Proof.

$$\begin{aligned}
Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&= \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{I}\boldsymbol{\beta} \\
\nabla_{\boldsymbol{\beta}} Q &= -\mathbf{X}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + 2\lambda\mathbf{I}\boldsymbol{\beta} \\
&= 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^\top\mathbf{Y} = \mathbf{0}
\end{aligned}$$

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y} \quad \Rightarrow \quad \boxed{\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}}$$

$\square$

Exercise. Show that $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$ is *always* invertible for $\lambda > 0$.

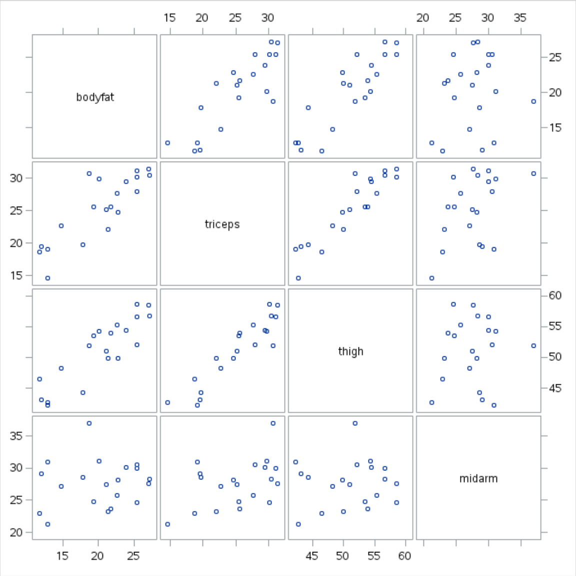
Chapter 7 example: Body fat

$n = 20$ healthy females 25–34 years old.

- x_1 = triceps skinfold thickness (mm)
- x_2 = thigh circumference (cm)
- x_3 = midarm circumference (cm)
- Y = body fat (%)

Obtaining Y_i, the percent of the body that is purly fat, requires immersing a person in water. Want to develop model based on simple body measurements that avoids people getting wet.

Scatterplot



Correlation coefficients

Pearson Correlation Coefficients, N = 20 Prob > |r| under H0: Rho=0			
	triceps	**thigh**	**midarm**
triceps	1.00000	0.92384 <.0001	0.45778 0.0424
thigh	0.92384 <.0001	1.00000	0.08467 0.7227
midarm	0.45778 0.0424	0.08467 0.7227	1.00000

Correlation coefficients

Pearson Correlation Coefficients, N = 20 Prob > |r| under H0: Rho=0			
	triceps	**thigh**	**midarm**
triceps	1.00000	0.92384 <.0001	0.45778 0.0424
thigh	0.92384 <.0001	1.00000	0.08467 0.7227
midarm	0.45778 0.0424	0.08467 0.7227	1.00000

There is high correlation among the predictors.

Correlation coefficients

Pearson Correlation Coefficients, N = 20 Prob > |r| under H0: Rho=0			
	triceps	thigh	midarm
triceps	1.00000	0.92384 <.0001	0.45778 0.0424
thigh	0.92384 <.0001	1.00000	0.08467 0.7227
midarm	0.45778 0.0424	0.08467 0.7227	1.00000

There is high correlation among the predictors. For example $r = 0.92$ for triceps and thigh. These two variables are *essentially carrying the same information*.

Correlation coefficients

Pearson Correlation Coefficients, N = 20 Prob > |r| under H0: Rho=0			
	triceps	**thigh**	**midarm**
triceps	1.00000	0.92384 <.0001	0.45778 0.0424
thigh	0.92384 <.0001	1.00000	0.08467 0.7227
midarm	0.45778 0.0424	0.08467 0.7227	1.00000

There is high correlation among the predictors. For example $r = 0.92$ for triceps and thigh. These two variables are *essentially carrying the same information*. Maybe only one or the other is really needed.

Effects of multicolinearity

```
lm(formula = bodyfat ~ triceps + thigh + midarm,
   data = bodyfat_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
triceps        4.334      3.016   1.437    0.170
thigh         -2.857      2.582  -1.106    0.285
midarm        -2.186      1.595  -1.370    0.190
```

Effects of multicolinearity

```
lm(formula = bodyfat ~ triceps + thigh + midarm,
   data = bodyfat_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
triceps        4.334      3.016   1.437    0.170
thigh         -2.857      2.582  -1.106    0.285
midarm        -2.186      1.595  -1.370    0.190
```

- Two of the three regression effects are *negative*.

Effects of multicolinearity

```
lm(formula = bodyfat ~ triceps + thigh + midarm,
   data = bodyfat_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
triceps        4.334      3.016   1.437    0.170
thigh         -2.857      2.582  -1.106    0.285
midarm        -2.186      1.595  -1.370    0.190
```

- Two of the three regression effects are *negative*.
- Holding midarm and triceps constant, increasing the thigh circumference *decreases* bodyfat.

Effects of multicolinearity

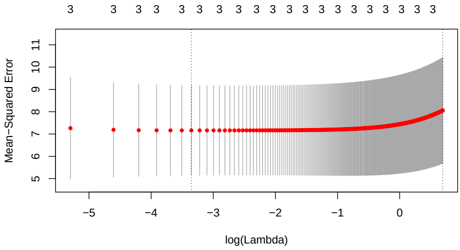
```
lm(formula = bodyfat ~ triceps + thigh + midarm,
   data = bodyfat_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
triceps        4.334      3.016   1.437    0.170
thigh         -2.857      2.582  -1.106    0.285
midarm        -2.186      1.595  -1.370    0.190
```

- Two of the three regression effects are *negative*.
- Holding midarm and triceps constant, increasing the thigh circumference *decreases* bodyfat.
- This may not make sense!

OLS vs Ridge on bodyfat data

https://github.com/zh3nis/MATH440/blob/main/chp11/ridge.R




```
> mean(ols$residuals^2)
[1] 4.920244

> mean((y-ridge_yhat)^2)
[1] 5.340952
```

```
> coef(ols)
(Intercept)     triceps       thigh      midarm
 117.084695    4.334092   -2.856848   -2.186060

> t(coef(ridge))
   (Intercept)   triceps     thigh     midarm
s0   0.6555753 0.8074414 0.1588796 -0.3266744
```

Weighted Least Squares

Ridge Regression

Robust Regression

Back to outliers

- Leverages h_{ii} and deleted residuals t_i useful for finding outlying $\mathbf{x}_i$ and Y_i cases.

Back to outliers

- Leverages h_{ii} and deleted residuals t_i useful for finding outlying $\mathbf{x}_i$ and Y_i cases.
- Cook's D_i and DFFIT_i indicate which cases are highly influencing the fit of the model.

Back to outliers

- Leverages h_{ii} and deleted residuals t_i useful for finding outlying $\mathbf{x}_i$ and Y_i cases.
- Cook's D_i and DFFIT_i indicate which cases are highly influencing the fit of the model.
- What to do with influential and/or outlying cases? Are they transcription errors or somehow (un)representative of the target population?

Back to outliers

- Leverages h_{ii} and deleted residuals t_i useful for finding outlying $\mathbf{x}_i$ and Y_i cases.
- Cook's D_i and DFFIT_i indicate which cases are highly influencing the fit of the model.
- What to do with influential and/or outlying cases? Are they transcription errors or somehow (un)representative of the target population?
- Outliers are often interesting in their own right and can help in building a better model.

Back to outliers

- Leverages h_{ii} and deleted residuals t_i useful for finding outlying $\mathbf{x}_i$ and Y_i cases.
- Cook's D_i and DFFIT_i indicate which cases are highly influencing the fit of the model.
- What to do with influential and/or outlying cases? Are they transcription errors or somehow (un)representative of the target population?
- Outliers are often interesting in their own right and can help in building a better model.
- **Robust regression** weakens the effect of outlying cases on estimation to provide a better fit to the majority of cases.

Back to outliers

- Leverages h_{ii} and deleted residuals t_i useful for finding outlying $\mathbf{x}_i$ and Y_i cases.
- Cook's D_i and DFFIT_i indicate which cases are highly influencing the fit of the model.
- What to do with influential and/or outlying cases? Are they transcription errors or somehow (un)representative of the target population?
- Outliers are often interesting in their own right and can help in building a better model.
- **Robust regression** weakens the effect of outlying cases on estimation to provide a better fit to the majority of cases.
- Useful in situations when there's no time for "influence diagnostics" or a more careful analysis.

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

where $\rho(\cdot)$ is some function.

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

where $\rho(\cdot)$ is some function.
- $\rho(u) = u^2$ gives OLS

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

  where $\rho(\cdot)$ is some function.
- $\rho(u) = u^2$ gives OLS
- $\rho(u) = |u|$ gives **least absolute residual** (LAR) regression

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

  where $\rho(\cdot)$ is some function.
- $\rho(u) = u^2$ gives OLS
- $\rho(u) = |u|$ gives **least absolute residual** (LAR) regression
- Huber's method is a compromise between OLS and LAR.

M-estimation

- Robust regression is effective when the error distribution is not normal, but heavy-tailed.
- **M-estimation** is a general class of estimation methods.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

  where $\rho(\cdot)$ is some function.
- $\rho(u) = u^2$ gives OLS
- $\rho(u) = |u|$ gives **least absolute residual** (LAR) regression
- Huber's method is a compromise between OLS and LAR. It looks like u^2 for u around zero, and like $|u|$ for u further away from zero.

Iteratively reweighted least squares (IRLS)

Outlying residuals are (iteratively) given less weight in the estimation process.

Algorithm 1: IRLS

1 **Input**: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$

2 Fit OLS. Let $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$

3 Initialize weights: $\omega_i \leftarrow \frac{1}{e_i^2}$, $\boldsymbol{\Omega} = \text{diag}[\omega_1, \ldots, \omega_n]$

4 Fit WLS: $\hat{\boldsymbol{\beta}}_{\text{WLS}} \leftarrow (\mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{Y}$

5 Estimate: $\hat{\sigma} \leftarrow \text{median}_i \left\{ \frac{|y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{\text{WLS}}|}{\Phi^{-1}(0.75)} \right\}$

6 Update weights: $\omega_i \leftarrow w\left(\frac{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{\text{WLS}}}{\hat{\sigma}}\right)$, where

$$w(u) = \begin{cases} 1, & |u| < 1.345 \\ \frac{1.345}{|u|}, & |u| > 1.345 \end{cases}$$

7 Repeat steps 4–6 until $\hat{\sigma}$ and $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ stabilize.

8 **Output**: $\hat{\boldsymbol{\beta}}_{\text{WLS}}$

IRLS example

https://github.com/zh3nis/MATH440/blob/main/chp11/irls.R

```
require(foreign)
require(MASS)

cdata <- read.dta(
    "https://stats.idre.ucla.edu/stat/data/crime.dta")

plot(crime ~ poverty, data=cdata)
ols <- lm(crime ~ poverty, data = cdata)
abline(ols)

irls <- rlm(crime ~ poverty, data=cdata)
abline(irls, col='blue')

summary(ols)
summary(irls)
```

IRLS example