Advanced Remedial Measures

Previously, we have addressed violations of the model assumptions by

- Transforming Y (e.g., Box-Cox, natural log)
- Transforming one or more X variables
- Adding terms to the model (polynomials, additional predictors, interactions)
- Variable selection (to minimize multicollinearity)

Sometimes these tools fail...

Advanced remedial measures

- Weighted least squares (WLS)
- Ridge regression
- Robust regression
- Nonparametric methods: Bootstrapping

Unequal Error variances Remedial Measures---Weighted Least Squares (WLS)

WLS

Our model assumes that the errors are *iid* with constant variance, σ^2

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

But what if each subpopulation (i.e., each unique combination of X values) has its own, potentially unique error variance instead?

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

Where the diagonal matrix D reflects that the variance could be non-consistent.

$$\sigma^2(\varepsilon) = \sigma^2 D = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Define a diagonal weight matrix $W, \ \ such\ that\ w_i = 1/\sigma_i^2$

$$W = \begin{bmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n^2 \end{bmatrix} \qquad \sigma^2(\varepsilon) = \sigma^2 D = \mathrm{W}^{-1} \qquad W^{1/2} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n \end{bmatrix}$$

The weighted matrix W can be used to create a (weighted) data with constant variance

Multiple $W^{1/2}$ to $Y = X\beta + \varepsilon$, we obtain $\boldsymbol{W^{1/2}Y} = \boldsymbol{W^{1/2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{W^{1/2}\varepsilon}$

This becomes $\boldsymbol{Y}_{\mathbf{w}} = \mathbf{X}_{\mathbf{w}}\boldsymbol{\beta} + \boldsymbol{\varepsilon_w}$ where

$\boldsymbol{Y}_{\mathbf{w}} = \boldsymbol{W^{1/2}Y}$

$\boldsymbol{X}_{\mathbf{w}} = \boldsymbol{W^{1/2}X}$

$\boldsymbol{\varepsilon}_{\mathbf{w}} = \boldsymbol{W^{1/2}\varepsilon}$

$$\boldsymbol{E(\varepsilon_w) = E(W^{1/2}\varepsilon) = W^{1/2}\, E(\varepsilon) = 0}$$

$$\boldsymbol{\sigma^2(\varepsilon_w) = \sigma^2(W^{1/2}\varepsilon) = W^{1/2}\sigma^2(\varepsilon)W^{1/2} = W^{1/2}W^{-1}W^{1/2} = I}$$

$$b_w = (X_w'X_w)^{-1}X_w'Y_w = (X'WX)^{-1}X'WY$$

$$s^2\{b_w\} = \mathrm{MSE_w}(X'WX)^{-1} = \frac{\Sigma w_i(Y-\hat{Y})^2}{n-p}(X'WX)^{-1}$$

If all weights are equal, w_i is identically equal to a constant, and WLS reduces to OLS.

Advantage

Valid inference in presence of non-constant variance (heteroscedasticity).

Disadvantage

$$\sigma^2(\varepsilon_w) = \sigma^2(W^{1/2}\varepsilon) = W^{1/2}\sigma^2(\varepsilon)W^{1/2} = W^{1/2}W^{-1}W^{1/2} = I$$

MSE_w is close to 1 in a good WLS model. Therefore, MSE_w could be used in model diagnosis but it has no clear contextual interpretation and cannot be used to compare models.

Next, we need to estimate the variance matrix D, or W^{-1}

Method 1: use replicated observations at each X_i to estimate each σ_i^2, which may require new data.

Method 2: regress the residual |e| with a MLR function on X variables.

$$|e| = U_0 + U_1 X_1 + .. + U_{p-1} X_{p-1}$$

since $\mathrm{D} = \frac{1}{\mathrm{W}} = \sigma^2(\varepsilon) = E(\varepsilon^2) - [E(\varepsilon)]^2 = E(\varepsilon^2)$,
and $E(\varepsilon^2)$ is estimated by $|e|^2 = \left(\widehat{U}_0 + \widehat{U}_1 X_1 + +\widehat{U}_{p-1} X_{p-1}\right)^2$

Then, W is estimated by $1/\ |e|^2$

The process could take several interactions with a weights added in the MLR model until the estimates become stable. This process is also known as the Interactively Reweighted Least Squares (IRLS)

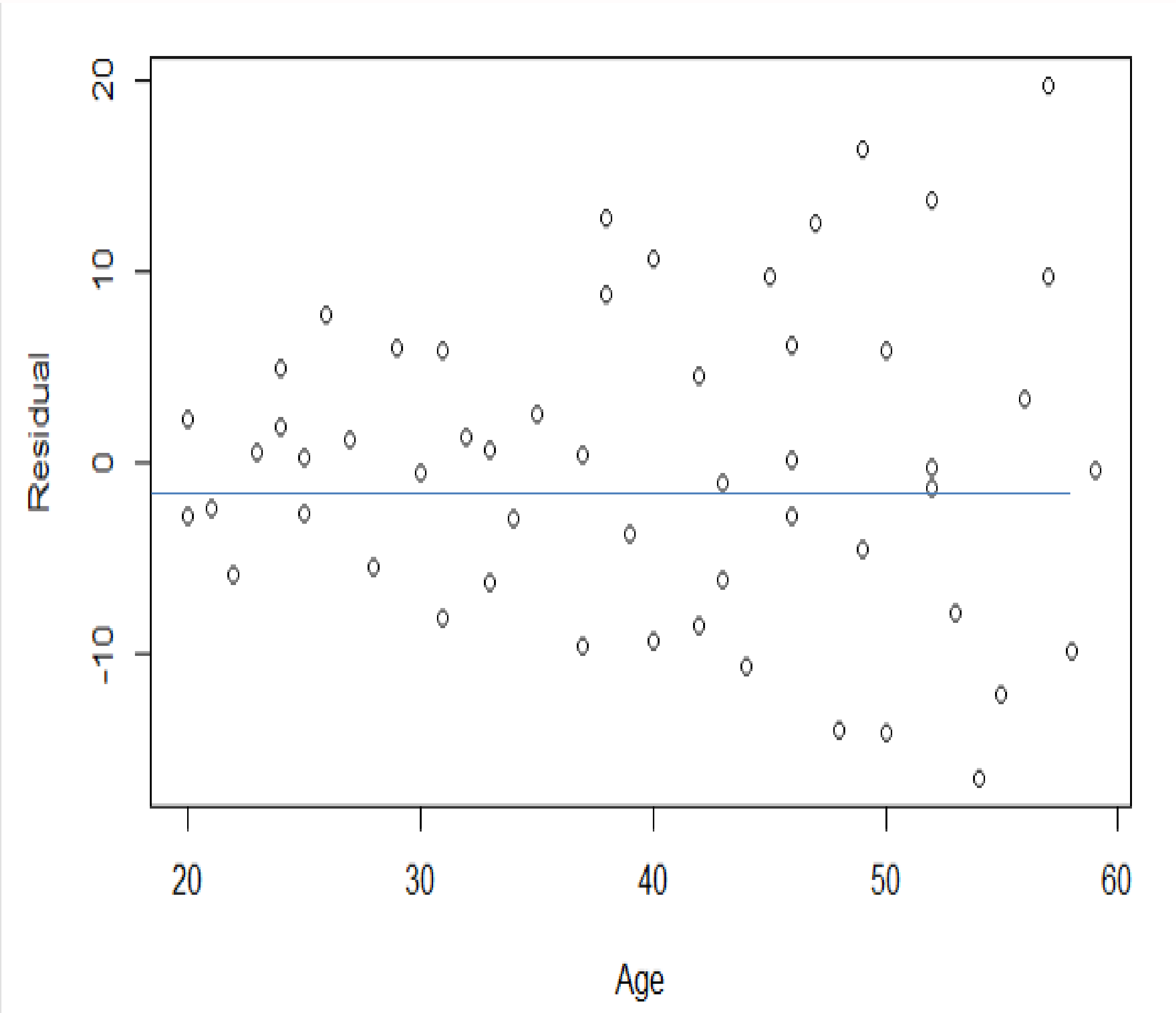
Although this method does not require new data, it does assume that residuals can be predicted with a MLR function on the predictors.

Example: Modeling blood pressure as a function of age

A heath researcher who is interested in studying the relationship between diastolic blood pressure and age among healthy adult women 2- to 60 years old, collected data on 54 subjects.

- Y is diastolic blood pressure
- X is age in years
- $n = 54$ healthy adult women aged 20 to 60 years old

Diagnostic plots detecting unequal error variance



The algorithm run down

```
pres.mod<-lm(bp~age, pres)
wts1<-1/fitted(lm(abs(residuals(pres.mod))~age, pres))^2
pres.mod2<-lm(bp~age, weight=wts1, data=pres)
```

1. Fit $Y \sim X\beta + \varepsilon$ by unweighted LS
2. Save the residuals e_i
3. Fit the model $|e_i| \sim U_i' X + \phi_i$
4. Use the fitted values from Step 3 to calculate weights

$$W_i = \frac{1}{\widehat{e_i^2}}$$

5. Use the estimated weights to fit $Y = X\beta + \varepsilon$ by WLS
6. (If necessary) Repeat Steps 2–5 until the values of ***b*** stabilize (typically, 1–3 iterations).

Unweighted linear model (OLS) vs. Weighted linear model (WLS)

Unweighted linear model (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
F-statistic: 35.79 on 1 and 52 DF,  p-value: 2.05e-07
```

Weighted linear model (WLS)

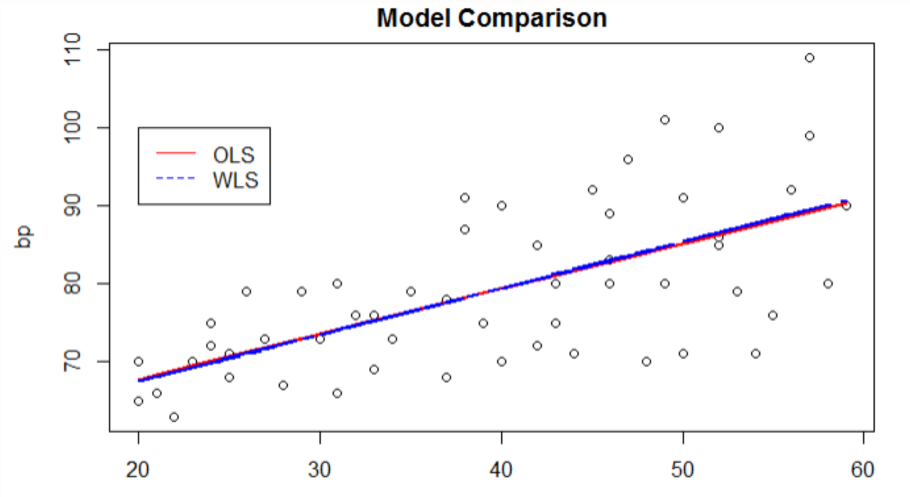
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.56577    2.52092  22.042  < 2e-16 ***
age          0.59634    0.07924   7.526 7.19e-10 ***

Residual standard error: 1.213 on 52 degrees of freedom
Multiple R-squared:  0.5214,    Adjusted R-squared:  0.5122
F-statistic: 56.64 on 1 and 52 DF,  p-value: 7.187e-10
```

- Comparing to the OLS, in the WLS,
  - the standard error of the coefficient is smaller, and
  - the Multiple R-square and the F-statistic are larger

  probably because the heteroscedasticity in the errors is accounted for by the chosen weighting scheme.
- We cannot compare the residual standard error (1.213 in WLS vs 8.146 in OLS) because the residuals have been altered and not comparable.

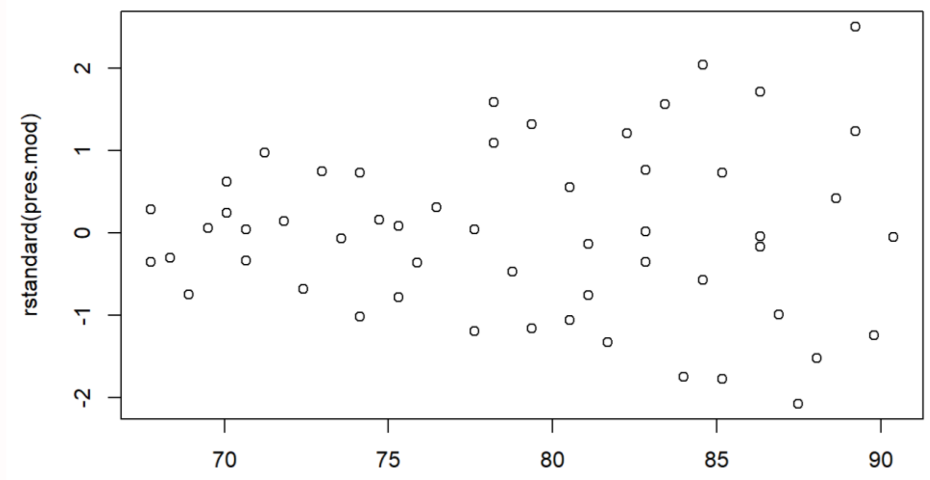
OLS and WLS Comparison via the Standardized Residual Plot



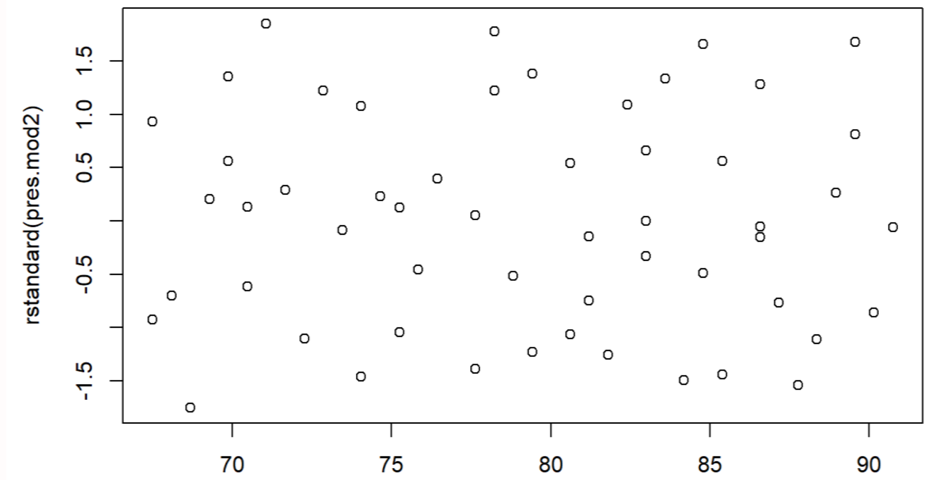
- In the scatter plot, the OLS and WLS do not show much difference.

```
plot(fitted(pres.mod), rstandard(pres.mod))
plot(fitted(pres.mod2), rstandard(pres.mod2))
```

- Rstandard residuals are standardized residuals, which means that they are scaled by the estimated standard deviation of the residuals.
This makes them more appropriate for assessing homogeneity of variance, as they consider the potential differences in variances at different levels of the predictor variables.
- From the plot of the (standardized residual, fitted value), we can see that the Rstandard residuals are constant across the range of the fitted values in the WLS.
- In WLS, the weights used to estimate the regression coefficients also affect the estimated standard deviation of the error term, which can lead to a change in the magnitude of the rstandard residual.



OLS model



WLS model

OLS and WLS Comparison via the Studentized Breusch-Pagan test

```
library(lmtest)
bptest(pres.mod)
bptest(pres.mod2)
```

```
        studentized Breusch-Pagan test

data:  pres.mod
BP = 12.541, df = 1, p-value = 0.0003981


        studentized Breusch-Pagan test

data:  pres.mod2
BP = 0.43608, df = 1, p-value = 0.509
```

- The null hypothesis of the BP test is homoscedasticity,
so, a significant p-value in the OLS model would suggest that the data exhibit heteroscedasticity. But WLS model doesn't has the issue.

Confidence inference for coefficient in the weighted linear model (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 
F-statistic: 56.64 on 1 and 52 DF,  p-value: 7.187e-10
```

$$b_{w1} \pm t(0.975;52)SE\{b_{w_1}\} = 0.59634 \pm 2.007(0.07924) = (0.437, \quad 0.755)$$

- (IMPORTANT) The T and F method here still based on the assumption that the random error follows Normal with constant variance! We could consider **Bootstrapping** for a more precise evaluation.

Multicollinearity Remedial Measures---Ridge regression

Multicollinearity and Ridge Regression

Previous approaches for dealing with serious multicollinearity include,

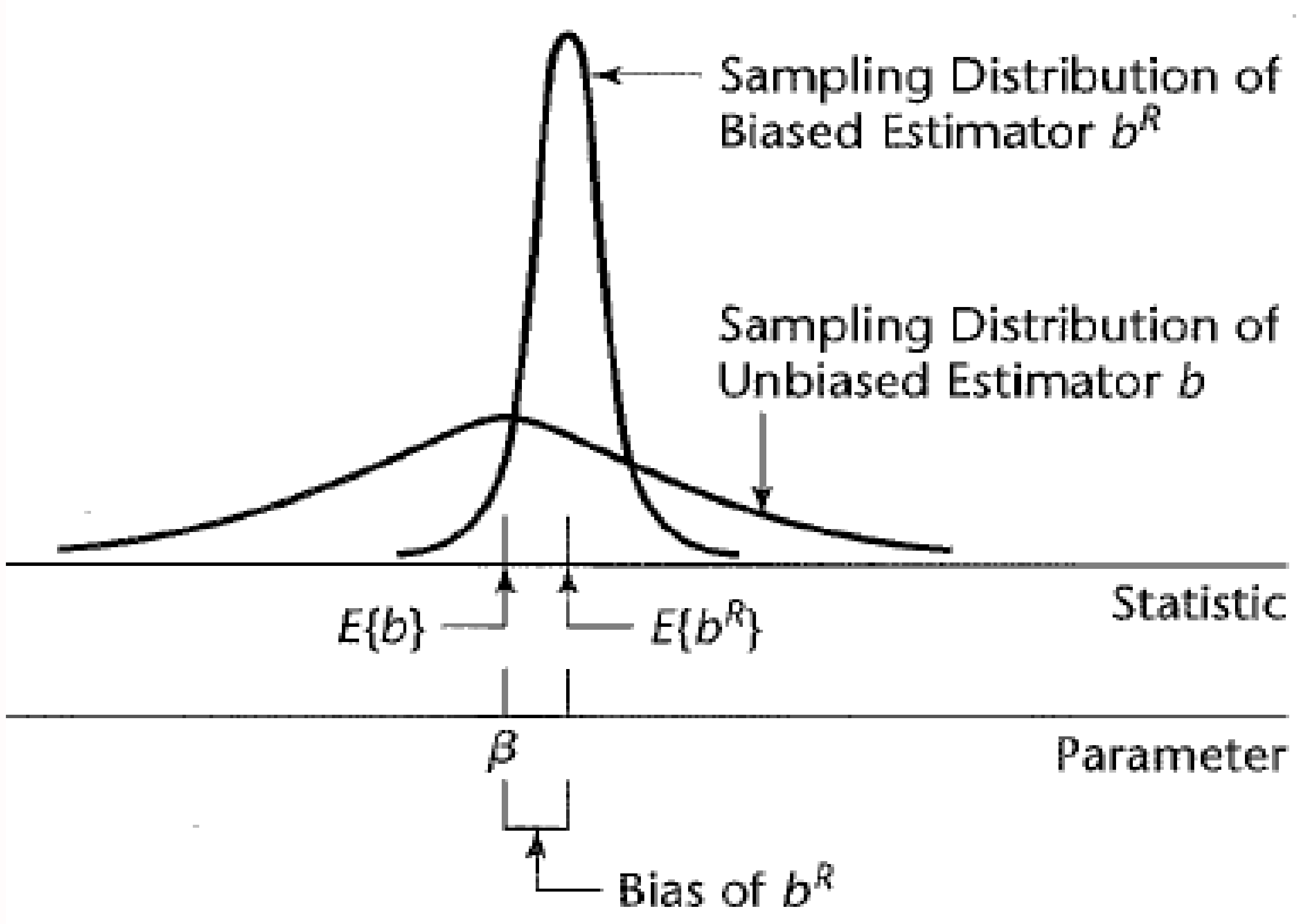
- Keeping the collinear predictors and restricting prediction to similarly collinear cases
- Centering of predictors in polynomial regression
- Model selection (drop some of the predictors)

Some additional possibilities include:

- Add new data points that break the pattern of collinearity (this can be difficult)
- Use of supplementary data that come from other contexts
- Principle components analysis (PCA) to create one or more

  composite variables that combine the collinear predictors
- Biased (a.k.a. “shrinkage”) estimation methods such as *ridge regression*

Ridge Regression



$$E\{b^R\} \neq \beta$$
$$E\{b\} = \beta$$

$$But\ s\{b^R\} < s\{b\}$$

Then on average, estimates based on the biased estimator, b^R, Will be closer to the true parameter β, than those based on The unbiased estimator, b.

Two equivalent formulations of ridge regression

Ridge regression shrinks estimators by adding a size penalty: $\lambda \Sigma_{j=1}^{p} \beta_j^2$

Penalized Residual Sum of Squares:

$$b^R = \arg_\beta \min\{\Sigma_{\mathrm{i=1}}^{\mathrm{n}} \left(Y_i - \left(\beta_0 + \Sigma_{j=1}^{p} X_{ij}\beta_j\right)\right)^2 + \lambda \Sigma_{j=1}^{p} \beta_j^2\}$$

Or in matrix form: $b^R = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$

- λ controls the amount of bias (shrinkage) of the parameter estimates.
- Large $\lambda \rightarrow$ greater shrinkage (toward zero), the less variable of the coefficients.
- A commonly used method to determine λ is the *ridge trace, which simultaneously trace the b^R with different λ.*
- The value of VIF also tend to reduce as λ $(also\ denoted\ by$ k or c) is increased.
- You will choose the smallest value when the regression confidents become stable in the ridge trace and the VIFs become sufficiently small.

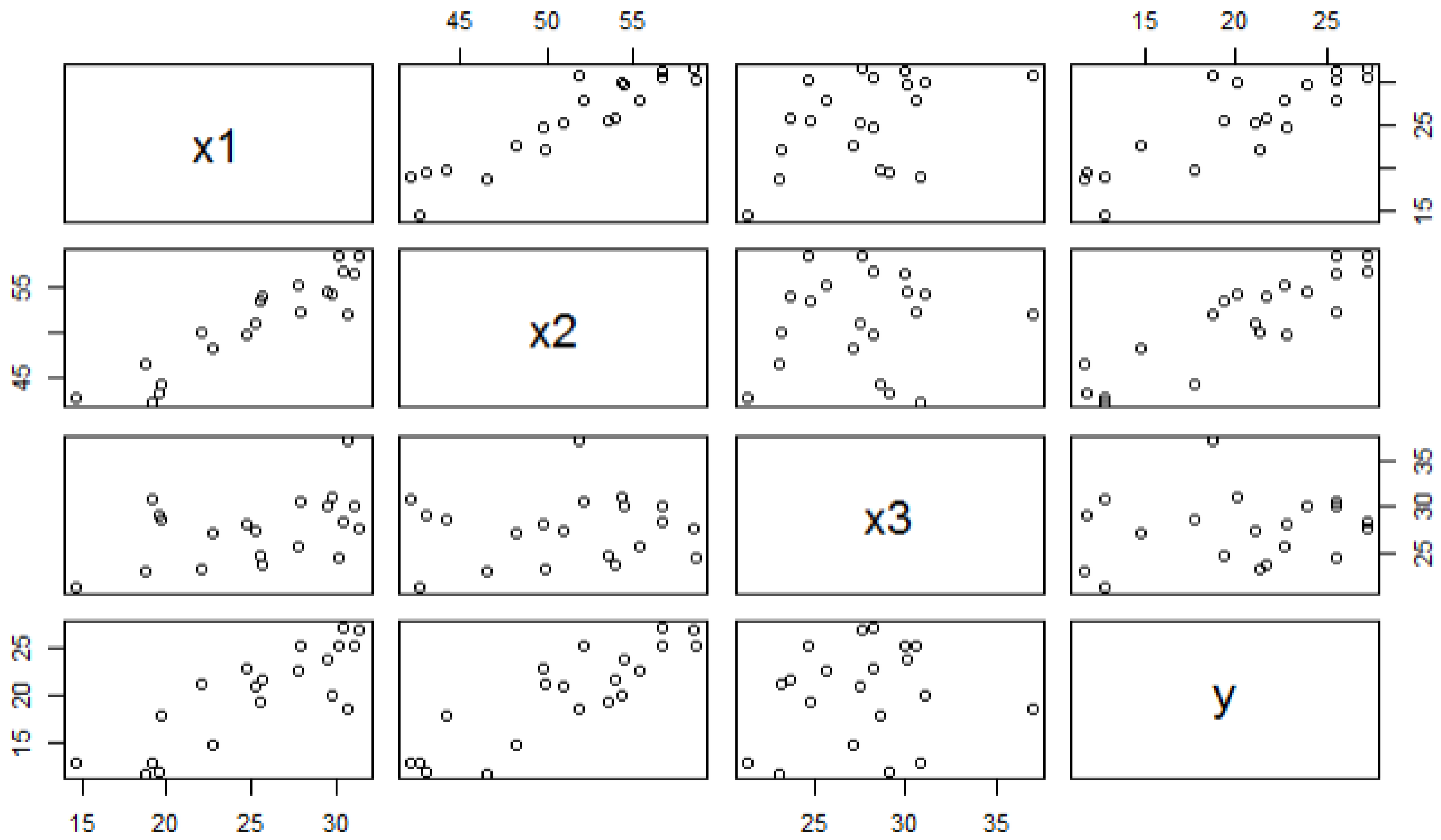
Choosing a value of λ

This is a judgment call. Select the smallest value of λ for which:

- Variance inflation factors are close enough to 1.
- Estimated coefficients are stable (trace lines approximately horizontal).
- Either R^2 or $\hat{\sigma}$ $(RMSE)$ are changing slowly.
- From cross validation

The body fat example

- 20 healthy female subjects ages 25-34
- Y is fraction body fat
- X_1 is triceps skin fold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference

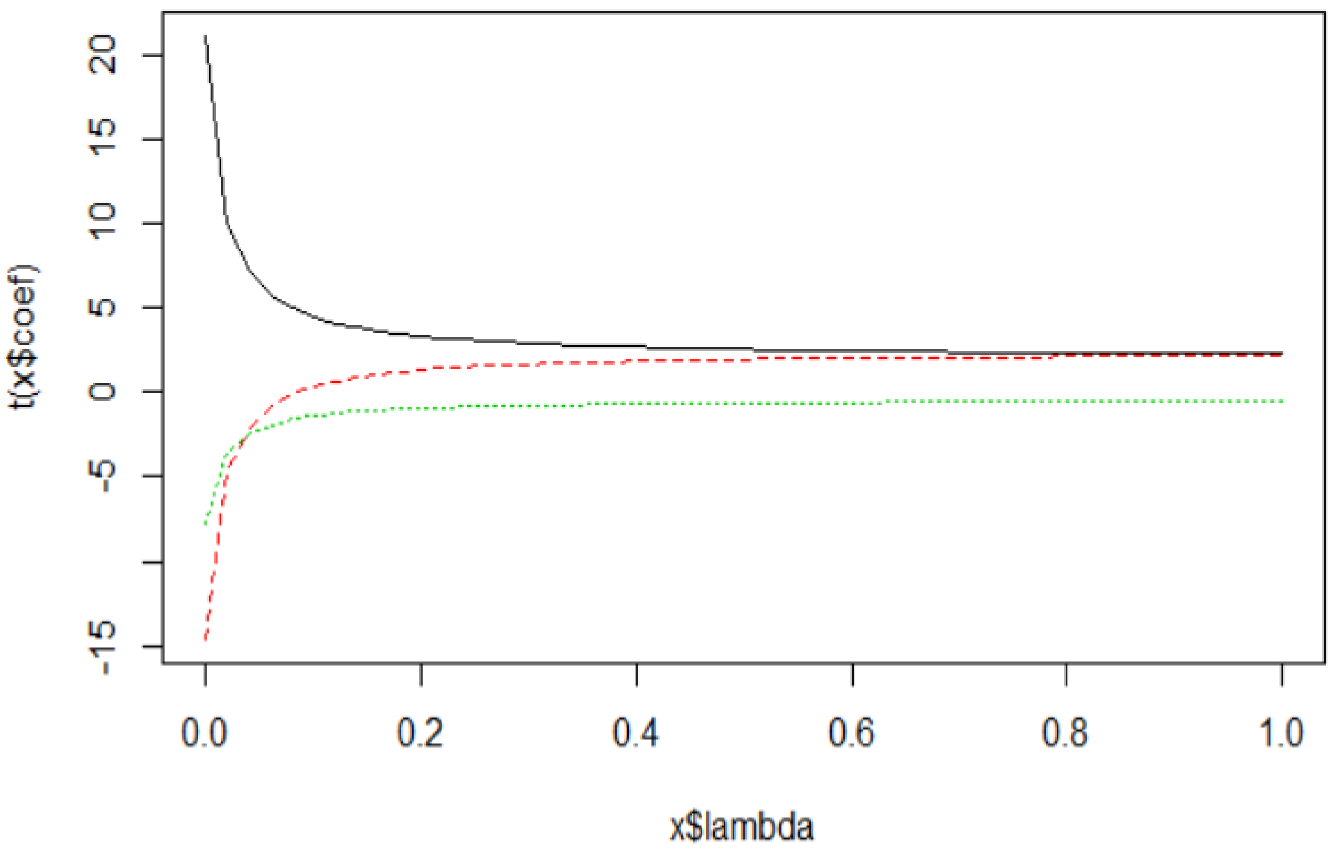


$VIF_1 = 708.84 \qquad VIF_2 = 564.34 \qquad VIF_3 = 104.61$

The selection of λ is subjective

```
library(MASS)
mod1<-lm.ridge(y~x1+x2+x3, data=bodyfat, lambda=seq(0, 1, 0.02))
plot(mod1)
select(mod1)
```

Ridge trace plot

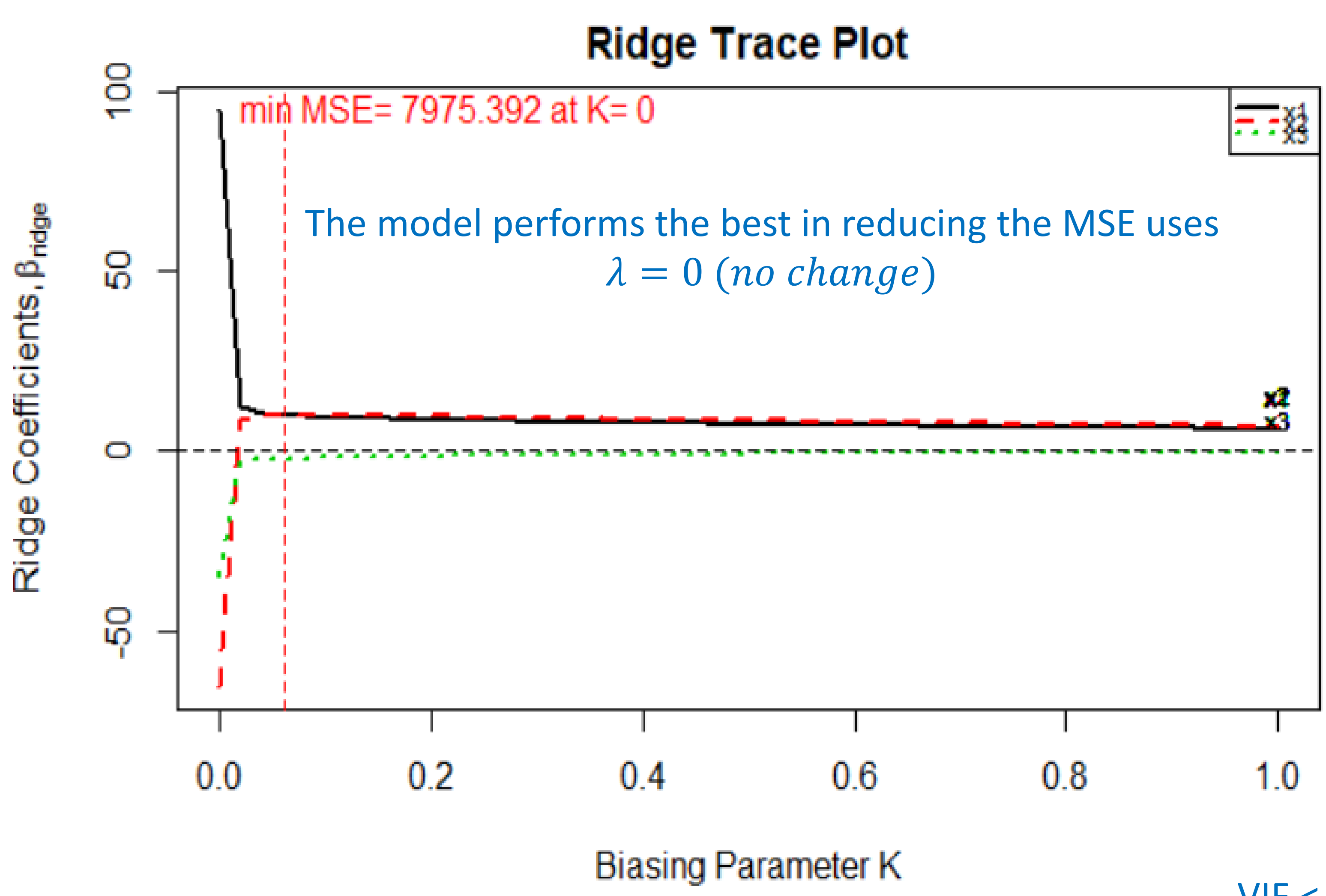



```
modified HKB estimator is 0.008505093
modified L-W estimator is 0.3098511
smallest value of GCV  at 0.02
```

The model that performs the best in the cross-validation uses $\lambda = 0.02$

The selection of λ is subjective

```r
library(lmridge)
mod2<-lmridge(y~x1+x2+x3, data=as.data.frame(bodyfat), K=seq(0,1, 0.02))
plot(mod2)
vif(mod2)
```



VIF

The model performs the best in reducing the multicollinearity uses $\lambda = 0.02$

	x1	x2	x3
k=0	708.84291	564.34339	104.60601
k=0.02	1.10255	1.08054	1.01051
k=0.04	0.45279	0.55529	0.88140
k=0.06	0.32437	0.44543	0.83060
k=0.08	0.27615	0.39984	0.79339
k=0.1	0.25155	0.37347	0.76137
k=0.12	0.23639	0.35499	0.73232
k=0.14	0.22577	0.34047	0.70540
k=0.16	0.21762	0.32825	0.68022
k=0.18	0.21096	0.31751	0.65652
k=0.2	0.20525	0.30782	0.63415

VIF < 1 because VIF is closely related to $1/(1+\lambda^2)$ in the algorithm

Model summary for different λ

```
summary(lmridge(y~x1+x2+x3, data=as.data.frame(bodyfat), K=seq(0,1, 0.02)))
```

```
Coefficients: for Ridge parameter K= 0
          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept 117.0847     1914.1817   3412.1592       0.5610   0.5826
x1          4.3341       94.8988     64.0559       1.4815   0.1579
x2         -2.8569      -65.1851     57.1552      -1.1405   0.2709
x3         -2.1861      -34.7530     24.6072      -1.4123   0.1770

Ridge Summary
      R2    adj-R2  DF ridge          F       AIC       BIC
 0.80140   0.77800   3.00001   22.86042  37.86718 100.76904
Ridge minimum MSE= 7975.392 at K= 0
```

```
Coefficients: for Ridge parameter K= 0.02
          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept  -7.4034     -633.1991    161.1205      -3.9300   0.0011 **
x1          0.5554       12.1599      2.5781       4.7167   0.0002 ***
x2          0.3681        8.4000      2.5522       3.2913   0.0043 **
x3         -0.1916       -3.0464      2.4681      -1.2343   0.2339
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
     R2   adj-R2 DF ridge        F      AIC      BIC
 0.76340  0.73560  2.00448 21.95136 37.75478 99.66535
Ridge minimum MSE= 7975.392 at K= 0
P-value for F-test ( 2.00448 , 17.93165 ) = 1.500203e-05
```

```
Coefficients: for Ridge parameter K= 1
          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept  -2.2485     -486.8614     71.6147      -6.7983   <2e-16 ***
x1          0.2844        6.2268      0.9584       6.4969   <2e-16 ***
x2          0.3025        6.9013      1.0795       6.3930   <2e-16 ***
x3         -0.0083       -0.1322      1.3966      -0.0947   0.9256
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
      R2    adj-R2  DF ridge          F       AIC       BIC
 0.33290   0.25440   1.15722   15.42241  43.60628 104.67321
Ridge minimum MSE= 7975.392 at K= 0
P-value for F-test ( 1.15722 , 18.37261 ) = 0.0006391673
```

Statistical Inference from the Ridge Regression at $\lambda = 0.02$

```
Coefficients: for Ridge parameter K= 0.02
           Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept   -7.4034     -633.1991    161.1205      -3.9300   0.0011 **
x1           0.5554       12.1599      2.5781       4.7167   0.0002 ***
x2           0.3681        8.4000      2.5522       3.2913   0.0043 **
x3          -0.1916       -3.0464      2.4681      -1.2343   0.2339
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge Summary
      R2    adj-R2  DF ridge         F       AIC       BIC
 0.76340   0.73560   2.00448  21.95136 37.75478 99.66535
Ridge minimum MSE= 7975.392 at K= 0
P-value for F-test ( 2.00448 , 17.93165 ) = 1.500203e-05
```

- The Estimate (Sc) column shows that the coefficient estimate is scaled to reduce the impact of the predictor's unit (i.e., Km vs m)
- The $\Pr(> |t|)$ gives a pvalue for the significant marginal Effect test for X_i.
  - E.g., X_3 has little marginal effect for a model with $X_1 and\ X_2$
- The Confidence interval

$$Estimate\ (SC) \pm t\left(1 - \frac{\alpha}{2}, n - df.ridge\right) StdErr(SC)$$

Could be used to estimate the linear impact of the predictor

  - E.g., a 95% CI for X3's impact is estimated by

$$-3.0464 \pm t(0.975, 20 - 2.00448)2.4681 = (-8.25, 2.16)$$

- (IMPORTANT) The T and F method here still based on the assumption that the random error follows Normal with constant variance!

We could consider **Bootstrapping** for a more precise evaluation.

Remedial measures for influential cases—Robust regression

Robust regression

Tools that have been used to detect outliers and influential points.

- Hat matrix, studentized deleted residuals
- DFFITS, Cook’s distance, and DEBETTAS measures.
- LS method is particularly susceptible to outliers and influential cases.

Outlying and influential case may lead to the finding of model inadequacies.

- Missing interaction, missing important predictors or choice of an incorrect functional form

In OLS, using least square errors is not robust

- Outliers are heavily weighted

An alternative to discarding outlying cases that is less severe is to dampen the influence of these cases.

Iteratively reweighted least squares (IRLS) robust regression

1. Choose a ***weight function*** for weighting the case
2. Obtain the ***starting weights*** for all cases.
3. Use the starting weights in weighted least squares and ***obtain the residuals*** from the fitted Regression function.
4. Use the residuals in step 3 to obtain revised weights.
5. Continue the iterations until convergence, which can be judged by whether
   - The weights change relatively little, or
   - The residuals change relatively little, or
   - The estimated regression coefficients change relatively little, or
   - The fitted values change relatively little

$$W = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

Where the w_i is computed differently for outliers or non-outliers.

Weight function (Huber estimator and Bisquare estimator)

Many weight functions have been proposed for dampening the influence of outlying cases. Two widely used weight functions are the Huber and Tukey's Bisquare weight functions

Huber: $w = \begin{cases} 1 & |u| \leq 1.345 \\ \frac{1.345}{|u|} & |u| > 1.345 \end{cases}$ Bisquare: $w = \begin{cases} \left(1 - \left(\frac{u}{4.685}\right)^2\right)^2 & |u| \leq 4.685 \\ 0 & |u| > 4.685 \end{cases}$

w denotes the weight the u denotes the ***scale residual:***

$u_i = e_i/MAD,$ where $MAD,$ the median absolute deviation estimator is

$$MAD = \frac{1}{0.6745} median\{|e_i - median\{e_i\}|\}$$

Comments on the scale residual:

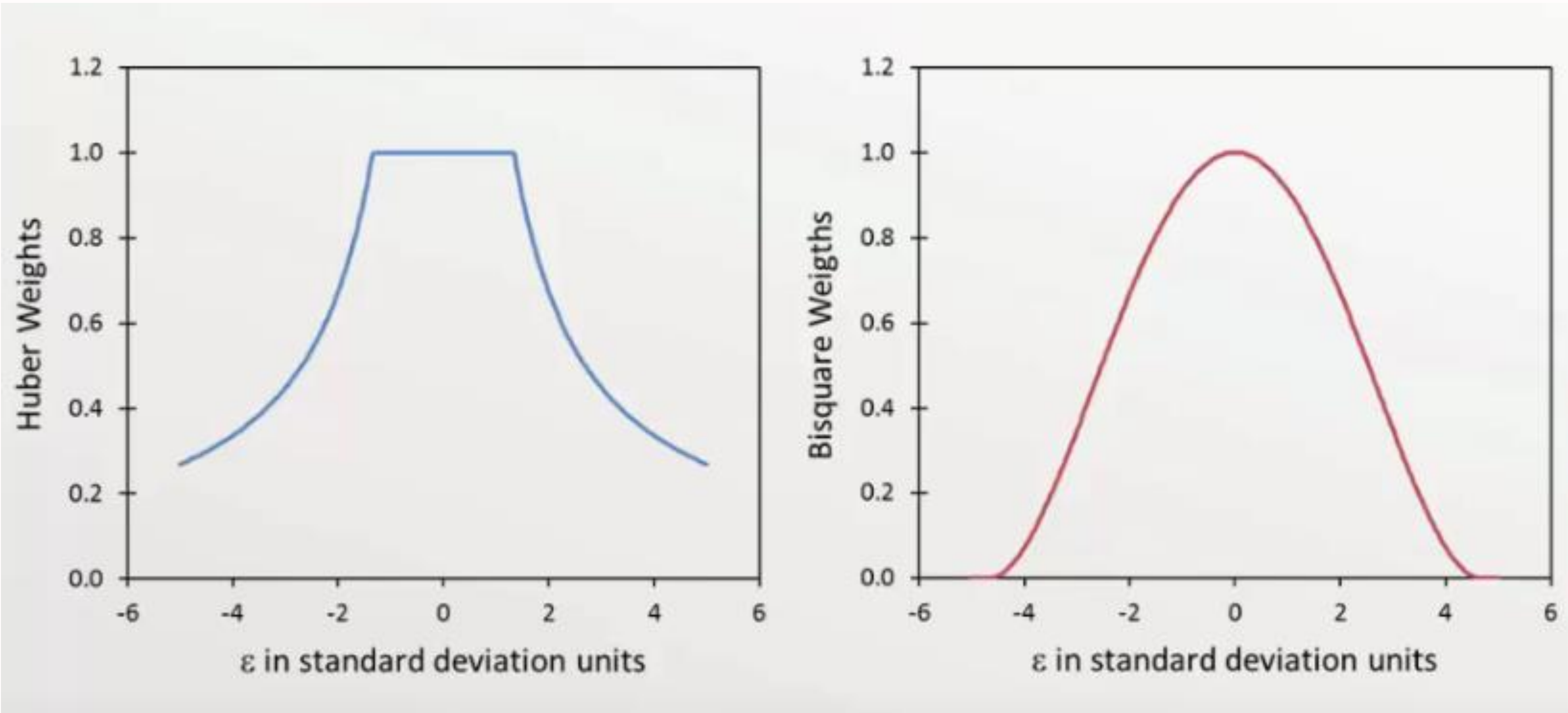
- **We want to use some measurement that is more resistant to outliers, i.e., the median.**
- **The constant 0.6745 provides an unbiased estimate of σ for independent observations from a Normal Distribution**

Weight function (Huber estimator and Bisquare estimator)

Many weight functions have been proposed for dampening the influence of outlying cases. Two widely used weight functions are the Huber and Tukey's Bisquare weight functions

Huber: $$w = \begin{cases} 1 & |u| \leq 1.345 \\ \frac{1.345}{|u|} & |u| > 1.345 \end{cases}$$

Bisquare: $$w = \begin{cases} \left(1 - \left(\frac{u}{4.685}\right)^2\right)^2 & |u| \leq 4.685 \\ 0 & |u| > 4.685 \end{cases}$$



With Bisquare we can throw in very extreme values

WLS and Robust Regression

- Both methods use weight function to adjust the influence of the observation on the Estimations. Both can handle unequal variance of in the error terms.

- WLS applies re-weighting on each observation in the sample, assuming the errors have a known variance structure, $|e_i| \sim U_i' X + \phi_i$. If the primary issue is heteroscedasticity, WLS should be considered.

- On the other hand, robust regression uses the weight function to trim the influence of outliers or influential observations based on their residuals. but not for other observations in the sample. If the main issue is the presence of outliers or influential observations, robust regression should be considered.

- In some cases, both robust regression and WLS may be used together.

Case study (Math proficiency)

The educational testing service study *America's smallest school: the family* investigated the relation of educational achievement of students to their home environment. Data on average mathematics proficiency (Mathprof, Y) and five home environment variables were obtained. The sample size **n=40**

Parents (X1): percentage of eighth-grade students with both parents living at home
Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home
Reading (X3): percentage of eighth-grade students who read more than 10 pages a day
Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day
Absences (X5): percentage of eighth-grade students absent three days or more last month

state	math proficiency	parents	home library	reading	TV watch	absence
Alabama	252	75	78	34	18	18
Arizona	259	75	73	41	12	26
Arkansas	256	77	77	28	20	23
California	256	78	68	42	11	28
Colorado	267	78	85	38	9	25
Connecticut	270	79	86	43	12	22

The initial model

Parents (X1): percentage of eighth-grade students with both parents living at home
Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home
Reading (X3): percentage of eighth-grade students who read more than 10 pages a day
Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day
Absences (X5): percentage of eighth-grade students absent three days or more last month

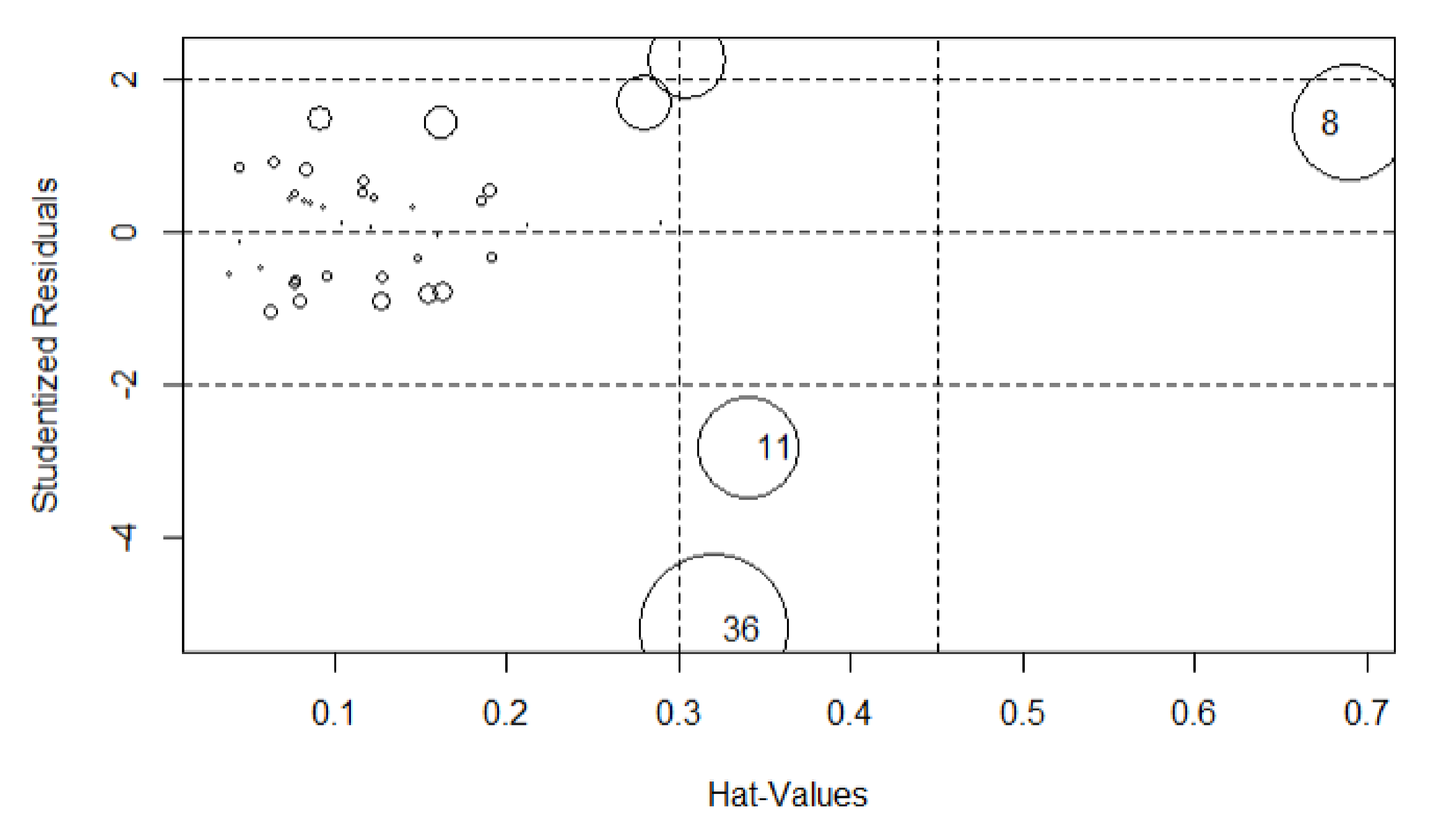
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 155.0304    36.2383   4.278 0.000145 ***
x1            0.3911     0.2571   1.521 0.137399    
x2            0.8639     0.1797   4.807 3.05e-05 ***
x3            0.3616     0.2690   1.345 0.187679    
x4           -0.8467     0.3525  -2.402 0.021927 *  
x5            0.1923     0.2636   0.729 0.470718    
```

```
Residual standard error: 5.268 on 34 degrees of freedom
Multiple R-squared:  0.861,	Adjusted R-squared:  0.8406 
F-statistic: 42.13 on 5 and 34 DF,  p-value: 1.276e-13
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)    
x1         1 3732.4  3732.4 134.4896 2.303e-13 ***
x2         1 1647.0  1647.0  59.3468 5.863e-09 ***
x3         1  290.5   290.5  10.4693  0.002705 ** 
x4         1  161.6   161.6   5.8245  0.021341 *  
x5         1   14.8    14.8   0.5321  0.470718    
Residuals 34  943.6    27.8                       
```

Diagnostics for the outlying and influential case



	State	studRes	Hat	CookD
8	D.C.	1.41	0.69	0.72
11	Guam	-2.83	0.34	0.57
36	Virgin_islands	-5.21	0.32	1.21

A case i is considered influence point if D_i>

Major influence if $> F(0.5; 6, 34)$ =0.91
Moderate influence if less than 0.91 but greater than
$F(0.2; 6,34) = 0.51$

Any influence case?

D.C. Guam and Virgin-Islands

Best model selection

Parents (X1): percentage of eighth-grade students with both parents living at home
Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home
Reading (X3): percentage of eighth-grade students who read more than 10 pages a day
Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day
Absences (X5): percentage of eighth-grade students absent three days or more last month

	p	1	2	3	4	5	SSEp	r2	r2.adj	Cp	AICp	SBCp	PRESSp
1	2	0	0	0	1	0	1609.4257	0.7629677	0.7567300	21.992880	151.7901	155.1679	1883.644
2	3	0	1	0	1	0	1071.3398	0.8422157	0.8336868	4.603884	137.5114	142.5781	1392.568
3	4	0	1	1	1	0	1008.8965	0.8514122	0.8390299	4.353845	137.1093	143.8648	1412.810
4	5	1	1	1	1	0	958.3394	0.8588581	0.8427276	4.532109	137.0529	145.4973	1629.298
5	6	1	1	1	1	1	943.5723	0.8610330	0.8405966	6.000000	138.4317	148.5650	1832.519

We consider the model: $\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Robust regression

Parents (X1): percentage of eighth-grade students with both parents living at home
Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home
Reading (X3): percentage of eighth-grade students who read more than 10 pages a day
Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day
Absences (X5): percentage of eighth-grade students absent three days or more last month

We consider the model: $\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

The Robust Model Summary

```
library(MASS)
r<-rlm(y~x2+x3+x4, data=mathpro, psi=psi.bisquare)
```

```
Coefficients:
               Value    Std. Error t value
(Intercept) 207.6806  17.6965    11.7357
x2            0.7972   0.1399     5.6982
x3            0.1609   0.2209     0.7282
x4           -1.1692   0.2231    -5.2412

Residual standard error: 4.342 on 36 degrees of
freedom
```

OLS Model Summary

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 199.6107    21.5289   9.272 4.50e-11 ***
x2            0.7804     0.1702   4.585 5.29e-05 ***
x3            0.4012     0.2688   1.493  0.14423
x4           -1.1565     0.2714  -4.261  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.294 on 36 degrees of freedom
Multiple R-squared:  0.8514,	Adjusted R-squared:  0.839
F-statistic: 68.76 on 3 and 36 DF,  p-value: 5.646e-15
```

- **The residual standard error for the OLS is 5.294, the robust model is better with a smaller s of 4.342.**
- **Access the robust model based on the residuals in a similar way as the OLS.**

Remedial Measures for evaluating precision in Nonstandard situations: Bootstrapping

Bootstrap method introduction

Conceptually simple but extremely powerful, nonparametric method for estimating precision when the standard approaches are unavailable.

Bootstrap methods allow approximate estimation of

- Confidence and prediction intervals in weighted regression, robust regression, or ridge regression
- Correct intervals when the errors are strongly non-normal

Robust regression

Parents (X1): percentage of eighth-grade students with both parents living at home
Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home
Reading (X3): percentage of eighth-grade students who read more than 10 pages a day
Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day
Absences (X5): percentage of eighth-grade students absent three days or more last month

We consider the model: $\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

```
library(MASS)
r<-rlm(y~x2+x3+x4, data=mathpro, psi=psi.bisquare)
```

```
Coefficients:
            Value    Std. Error t value
(Intercept) 207.6806 17.6965    11.7357
x2            0.7972  0.1399     5.6982
x3            0.1609  0.2209     0.7282
x4           -1.1692  0.2231    -5.2412

Residual standard error: 4.342 on 36 degrees of
freedom
```

The confidence interval for the coefficients after the Robust regression

```
Coefficients:
             Value    Std. Error t value
(Intercept) 207.6806  17.6965    11.7357
x2            0.7972   0.1399     5.6982
x3            0.1609   0.2209     0.7282
x4           -1.1692   0.2231    -5.2412

Residual standard error: 4.342 on 36 degrees of
freedom
```

CI for the linear impact for X2, e.g., β_1:

$$b_1 \pm t(0.975, 36)S(b_1) = 0.7972 \pm 2.028(0.1399)$$

$$= 0.7972 \pm 0.2837 = (0.5135, \quad 1.0809)$$

- (IMPORTANT) The T method here is still based on the assumption

We now evaluate the precision of the estimate b1=0.7972 by the bootstrap method.

Basic bootstrap algorithm to evaluate the precision of the estimated coefficients

1. Randomly resample the available data ***with replacement*** to generate a new bootstrap sample with n equal to the original sample

2. Run regression on the bootstrap sample and save $\boldsymbol{b}$ or $\hat{\boldsymbol{Y}}$

3. Repeat Steps 1-2 B times to obtain an empirical sampling distribution for the parameters or fitted values (the bootstrap samples) $b^*_{1\ 1}\ b^*_{1\ 2}\ .....\ b^*_{1\ B}$

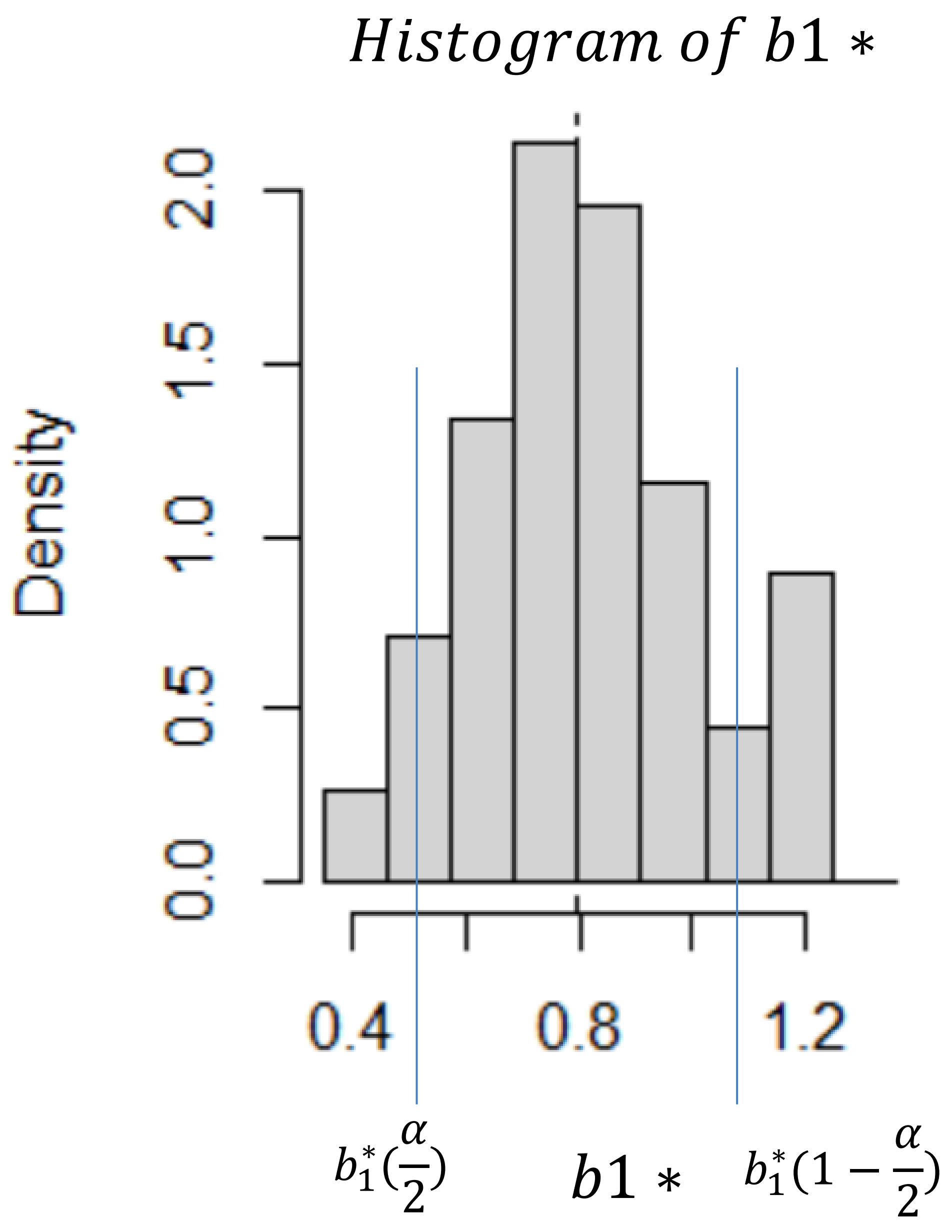
4. The standard deviation of the bootstrapped samples estimates standard error $s^*\{b^*_1\}$, and the quantiles of the bootstrapped values give approximate confidence intervals

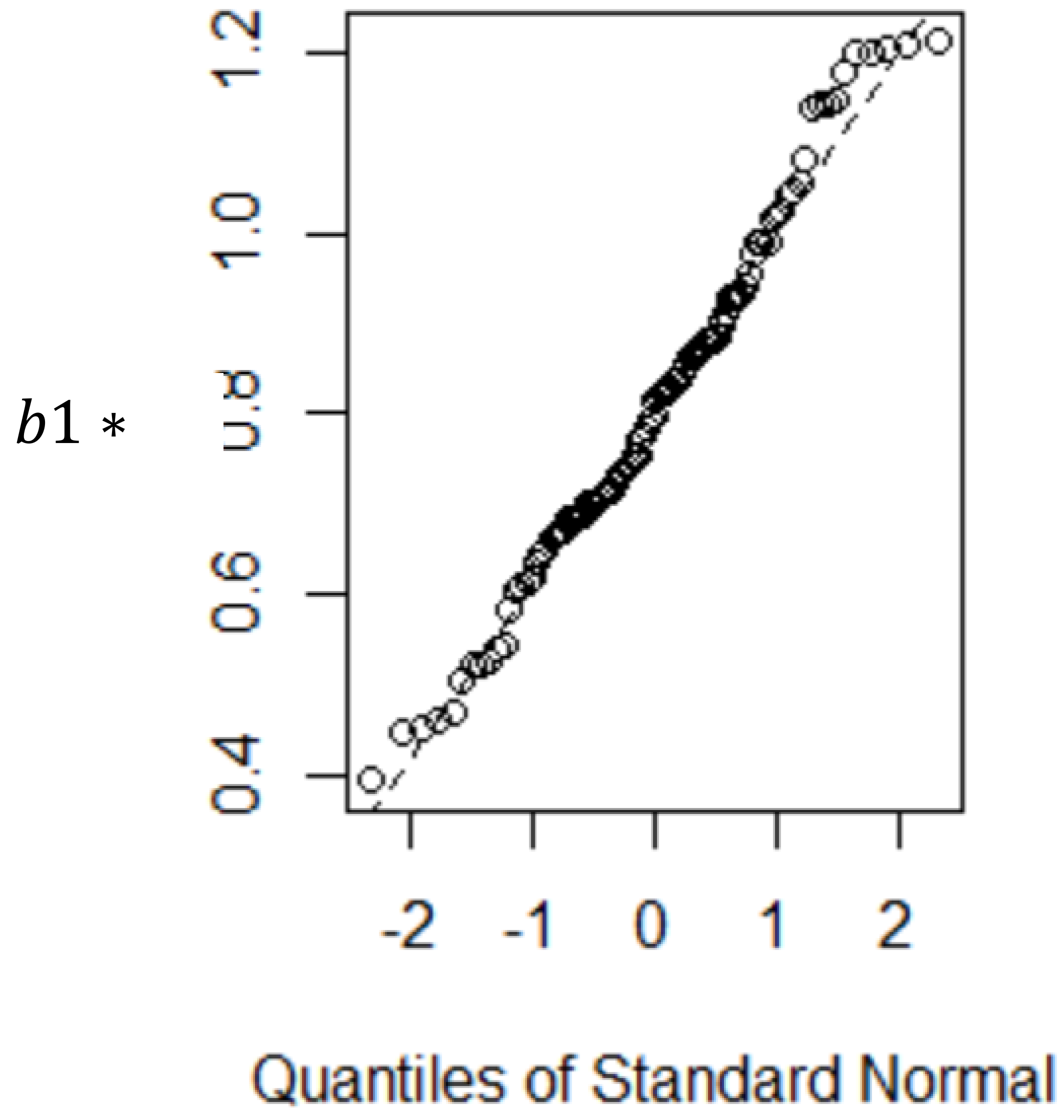
For example, the 90% confidence interval is given by
(the 5th percentile, the 95% percentile) denoted by $(b^*_1(0.05), b^*_1(0.95))$

In comparison, the 90% confidence interval, under Normal distribution, is given by a symmetric interval:
$(\ b_1 - tSE(b_1),\ \ b_1 + tSE(b_1)\ )$

The bootstrap resampling distribution of b1

```
plot(mathpro.boot, index=2)
```





```r
library(boot)
boot.huber <- function(data, indices, maxit=100){
  data <- data[indices,] # select obs. in bootstrap sample
  mod <- rlm(y ~ x2+x3+x4, data=data, maxit=maxit)
  coefficients(mod) # return coefficient vector
}

mathpro.boot<-boot(data=mathpro,statistic=boot.huber, R=100,maxit=100)
```

```
Bootstrap Statistics :
        original        bias    std. error
t1* 207.6806290 -3.10601076  22.6057370
t2*   0.7971940  0.01232840   0.1951210
t3*   0.1608632  0.04356760   0.2686475
t4*  -1.1692169  0.02979697   0.2174511
```

- “original” is the value of the estimates computed from Robust model.
- “bias” is the difference between the average of the bootstrap samples and the original.
- “std. error” is the standard deviation of the bootstrap samples.
- Check out the course website for R markdown file for examples on how to apply Bootstrapping on WLS and Ridge model.

The bootstrap confidence interval for β_1

```
boot.ci(mathpro.boot, index=2, type="perc")
```

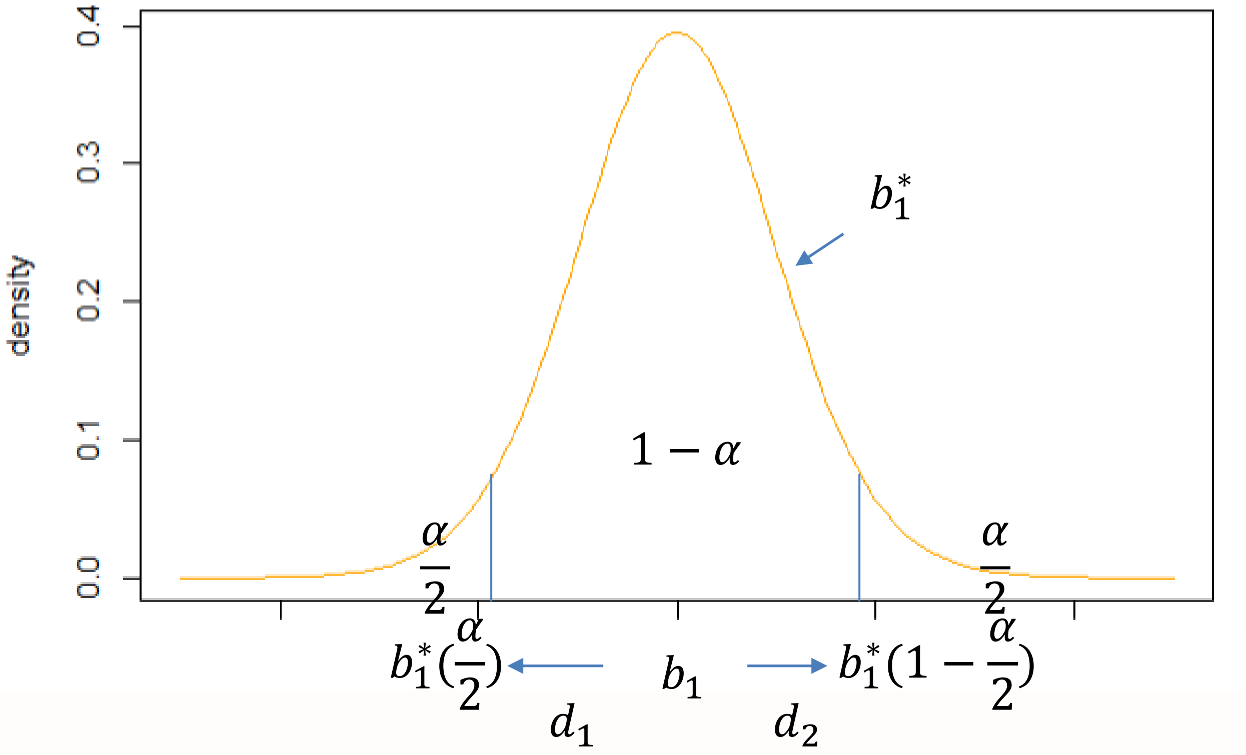
```
Intervals :
Level      Percentile
95%   ( 0.4493,  1.2068 )
Calculations and Intervals on Original Scale
Some percentile intervals may be unstable
```

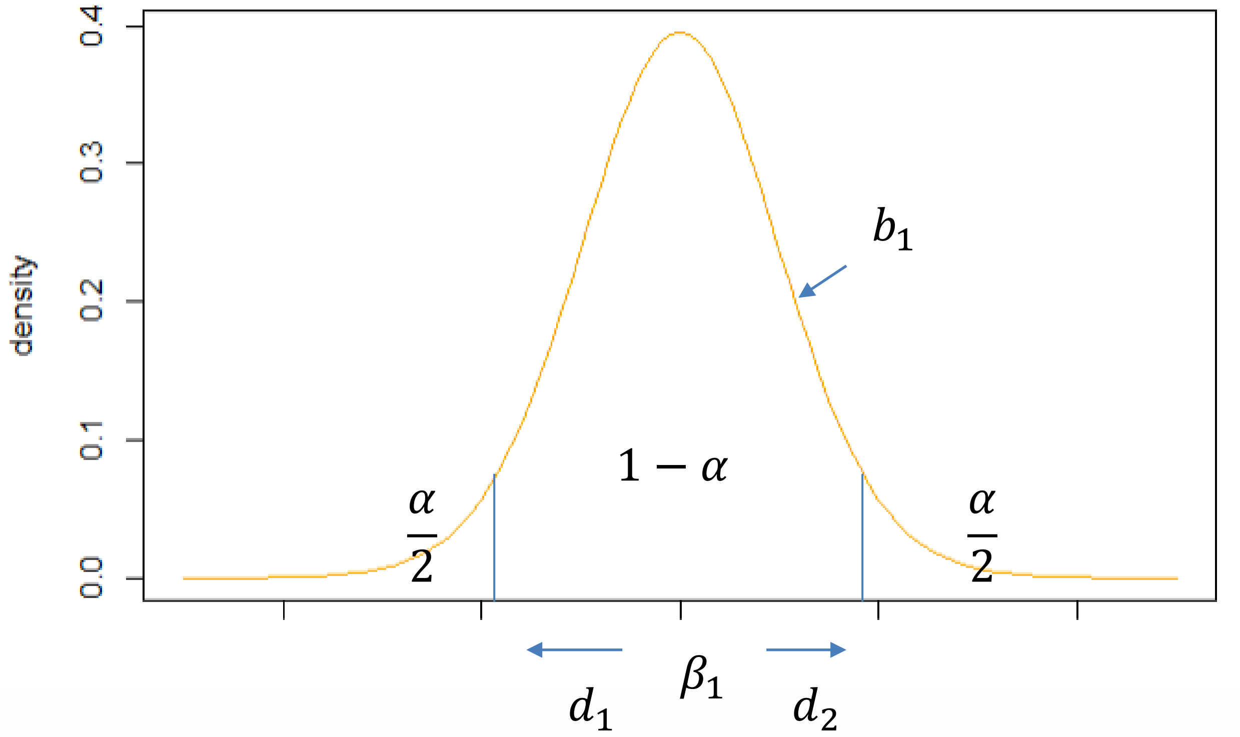
← Computing by hand is not required.

Comparing to the Robust CI for β_1: $b_1 \pm t(0.975, 36)S(b_1) = 0.7972 \pm 2.028(0.1399)$

$$= 0.7972 \pm 0.2837 = (0.5135, \quad 1.0809)$$

(Optional) Use the reflection method to estimate the empirical confidence interval for $\boldsymbol{\beta_1}$





$$d_1 = b_1 - b_1^*(\frac{\alpha}{2}) \qquad d_2 = b_1^*\left(1-\frac{\alpha}{2}\right) - b_1$$

$$\beta_1 - d_1 \le b_1 \le \beta_1 + d_2$$

Hence the $1-\alpha$ confidence interval for β_1 is $b_1 - d_2 \le \beta_1 \le b_1 + d_1$