Experimental Design and Data Analysis, Lecture 8

Eduard Belitser

VU Amsterdam

Lecture overview

1. strategies to choose the variables (lasso method in the next lecture)
   - step up
   - step down
2. diagnostics in linear regression
3. problems in linear regression
   - outliers and influence points
   - collinearity

strategies to choose the variables

Strategies to choose the variables

An important issue in multiple linear regression is how to find a suitable model. That is, how to select explanatory variables $X_1, \ldots, X_p$ such that

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{n2} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n,$$

is a good model for the given data. → we want relevant var.s included

A good model should be as precise and as concise as possible. It should

- contain all explanatory variables X_j that are essential in explaining Y
- not contain any variable X_j that does not contribute significantly.

no point in including too many var.s = each exp. var. should carry info about response

Common strategies to build a model are:

- step-up
- step-down
- lasso (next lecture)

The coefficient of determination $R^2 \in [0, 1]$ yields a global check on the linear regression model. The higher R^2 the more variation the model explains.

rough check whether your lin. regr. model fits data well → if this num. close to 1, then you have good explan. of data

this num. grows to 1 when you add more var.

$R^2 = 1$ ← perfect fit but many feat.s = needs to be balanced

Two strategies for finding a good model

In practice we need a strategy for building a model. We consider two strategies.

The step up method:

1. start with the background model $Y = \beta_0 + e$; → doesn't have any var.s, slope
2. take the variable (that is not in the model) that yields the maximum increase in R^2; → we look at all poss. var.s and R^2s = we include the var.s and look at the R^2 one by one → we get the var. that gives biggest increase in R^2 and include in model if var. is signif.
3. if this variable is significant (t-test) add it to the model and go to step 2, otherwise stop. → we stop when either we exhaust all var.s or we can't include any more signif. var.

The step down method:

1. start with the full model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e$; → incl. all var.s
2. test all variables by using the t-test;
3. if the largest p-value is larger than 0.05, remove the corresponding variable and go back to step 2. → = least relevant var.; → we continue until there is no insig. var.

Step up (1)

We apply the step up strategy to the bodyfat data:

```
> summary(lm(Fat~Triceps))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
Triceps       0.8572     0.1288   6.656 3.02e-06 ***
```

Multiple R-squared: 0.7111

```
> summary(lm(Fat~Thigh))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
Thigh         0.8565     0.1100   7.786  3.6e-07 ***
```

β_0 β_1

significant

Multiple R-squared: 0.771

biggest increase in R^2 → include in model

```
> summary(lm(Fat~Midarm))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.6868     9.0959   1.615    0.124
Midarm        0.1994     0.3266   0.611    0.549
```

Multiple R-squared: 0.02029

Thus, the first variable to add is Thigh.

Step up (2)

The second step:

```
> summary(lm(Fat~Thigh+Triceps))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
Thigh         0.6594     0.2912   2.265   0.0369 *
Triceps       0.2224     0.3034   0.733   0.4737
```

not significant → we cannot include in model

```
Multiple R-squared:  0.7781
```

biggest increase in R^2

```
> summary(lm(Fat~Thigh+Midarm))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.99695    6.99732  -3.715  0.00172 **
Thigh         0.85088    0.11245   7.567 7.72e-07 ***
Midarm        0.09603    0.16139   0.595  0.55968
```

not significant → we cannot include in model

```
Multiple R-squared:  0.7757
```

model only consists of this var.

Resulting model: Fat $=-23.6345 + 0.8565$*Thigh $+$ error, with $R^2 = 0.771$.

intercept $= \beta_0$; $\beta_1 =$ slope

Step down (1)

We now apply the step down strategy to the bodyfat data:

```
> summary(lm(Fat~Triceps+Thigh+Midarm))
            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

Multiple R-squared: 0.8014
```

we start w/ biggest model = all 3 var.s are incl.

biggest p-val. → acc. to step down we remove it

We see that none of the variables is significant. The first variable to remove is Thigh, which has the highest p-value.

Step down (2)

The second step:

```
> summary(lm(Fat~Triceps+Midarm))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7916   β0       4.4883    1.513   0.1486
Triceps       1.0006   β1       0.1282    7.803 5.12e-07 ***
Midarm       -0.4314   β2       0.1766   -2.443   0.0258 *
```

all var.s are signif.

```
Multiple R-squared: 0.7862
```

All remaining variables are significant. → no insig. var.s to remove = we stop

Resulting model: Fat = 6.7916 + 1.0006*Triceps -0.4314*Midarm + error with $R^2 = 0.7862$.

β_0 β_1 β_2

Step up or step down?

★ thigh and triceps are dependent = one needs to be removed b/c both carry more or less same info abt. response → collinear.

Now we are left with two different models.

Model 1 with $R^2 = 0.771$ and $\hat{\sigma} = 2.51$:

```
Fat = -23.6345 + 0.8565*Thigh + error
```

Model 2 with $R^2 = 0.7862$ and $\hat{\sigma} = 2.496$:

```
Fat = 6.7916 + 1.0006*Triceps -0.4314*Midarm + error
```

R^2 doesn't suffer too much

Question: which one do we prefer, and why?

Answer: Model 1 is preferred, because it has less variables and a comparable value of R^2. Also the term -0.4314*Midarm in the second model is not well interpretable.

slightly smaller than 1

negative coeff = could be badly interpret. e.g. bigger the mid arm, less fat you have = not true

Remember that one needs to check the model assumptions for the resulting model.

prediction

The predicted value

For the x-values in the data set, the fitted (predicted) values are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_p X_{pi}, \quad i = 1, \ldots n.$$

there might be other X's which we would like to predict our response but not available yet.

est. of response

predictor based on est. β's

The $\hat{Y}_i$'s can be computed by the R-command `fitted(model)`. Notice that these are in general different from the observed Y_i's.

b/c they are fitted

```
> fitted(bodyfatlm)
       1        2        3        4        5
14.85499 20.21884 20.98668 23.12732 11.75761
...
      16       17       18       19       20
23.72747 22.97360 26.78590 18.52628 20.48791
```

this formula might be useful when predicting response for new X's but not for obs. X's

Once all $\hat{\beta}_i$'s are computed, one can predict the y-value for a new measurement of the p explanatory variables: $x_{new} = (x_{new,1}, \ldots, x_{new,p})$ as

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new,1} + \ldots \hat{\beta}_p x_{new,p}.$$

predict. val. of Y

This is done by the command `predict(model,newxdata,...)`.

dataframe created for new X = vect. of new X's

fitted model

Confidence and prediction intervals

We can also construct two types of intervals for $\hat{Y}_{new}$ for given x_{new}-values:

- confidence interval for Y_{new}: an interval for the mean Y_{new}-value for given x_{new}-values. (This is interval $x_{new}^T\hat{\beta} \pm t_{\alpha/2,n-(p+1)}\sqrt{s^2(x_{new}^T(X^TX)^{-1}x_{new})}$.)
- prediction interval for Y_{new}: an interval for an individual Y_{new}-observation for given x_{new}-values. This interval is larger as the error is also taken into account. (This is interval $x_{new}^T\hat{\beta} \pm t_{\alpha/2,n-(p+1)}\sqrt{s^2(1+x_{new}^T(X^TX)^{-1}x_{new})}$.)

To summarize, confidence is for the population mean, prediction is for an individual observation.

In R: `predict(lm(y~x1+...+xk),newxdata,interval=...,level=...)`

true resp. = you don't see it

est. response

est. of std. dev. of mean

you give CI and on top of that you take uncertainty = error

est. of our mean Y_{new}

interval w/ higher CL = bigger interval

we increase our margin for err. → done by adding another s^2 to formula

Example: bodyfat data

Prediction intervals for the body fat data for new data can be found by

- designing a data.frame with the new x-values
- applying predict to this data.frame and specify the type of interval.

```
> newxdata=data.frame(Triceps=24.5,Thigh=51.3,Midarm=28.7)
> predict(bodyfatlm,newxdata)
 13.97372
> predict(bodyfatlm,newxdata,interval="prediction")
        fit      lwr      upr
 13.97372 3.053481 24.89396
> predict(bodyfatlm,newxdata,interval="prediction",level=0.95)
        fit      lwr      upr
 13.97372 3.053481 24.89396
> predict(bodyfatlm,newxdata,interval="confidence")
        fit      lwr      upr
 13.97372 4.402296 23.54515
```

we didn't have these val.s e.g. suppose it's a new patient and we want to use our model to predict body fat of this person

P.I. for new resp. w/ new xdata

est. for resp.

CI (95% by def.)

def. sfg. lvl. of interval

same pred.

diff. interval = smaller

The prediction interval is indeed larger!

valid b/c CI is interval for mean val. only while PI is interval for mean val. and err. is taken into acc.

Discussion

Finding different models by different strategies is exemplary for linear regression: there is no golden strategy to resolve this.

In such a case one should compare

- R^2 values of both models (higher is better),
- plots of fitted values versus residuals of both plots (should be no specific structure),
- ★ the number of explanatory variables in both models (fewer is better),
- the character of the explanatory variables in both models (easy to measure?),
- interpretation of both models,
- ...

and choose the one that is most appropriate.

but keep in mind that you can increase R^2 in cost of adding more var.s even irrelevant ones

you want your models to be interpret.

some var. might be imp. to keep for clear interpret.

need to have as little var. as poss. for compact.

remove irrelevant, uninformative var.s

diagnostics in linear regression

Example

Checking the fit in the linear regression by looking at the (adjusted) R^2 is not sufficient, we need to check the model assumptions: the linearity of the relation and the normality of the errors. We consider both graphical and numerical tools.

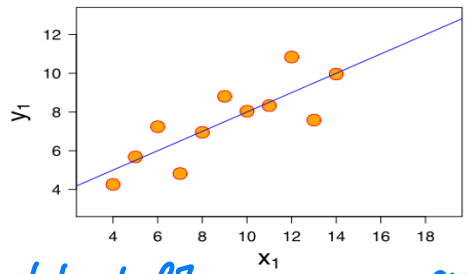
In the following 4 examples of artificial data, the fitted model is `y = 3.0 + 0.5*x + error`, $\hat{\sigma}^2 = 1.5$ and $R^2 = 0.67$.

we take diff. choice of X-Y pair, for each pair we get $\hat{\sigma}^2 = 1.5$ and $R^2 = 0.67$

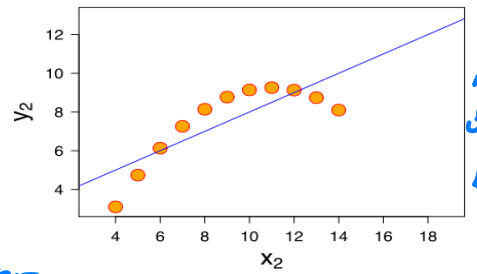
The differences between the 4 situations illustrate the need for a diagnostic tool, apart from $R^2, \hat{\sigma}$.

1. The first looks ok.
2. No lin. relation between X, Y.
3. Outlying point in Y.
4. Only one X is different.

same $\hat{\sigma}^2$ and R^2

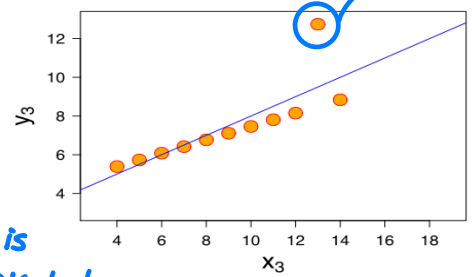


reasonable dataset, fit

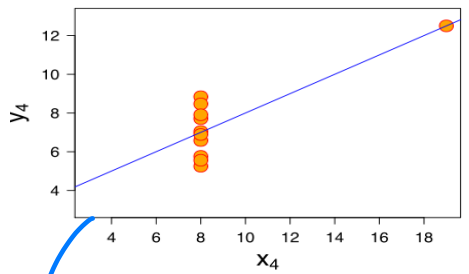


not good or bad

outlier



data is not OK but $\hat{\sigma}^2$ and R^2 val.s are same.



all X's are same except 1

Diagnostic plots

To check the model quality look at

1. scatter plot: plot Y against each X_k separately (this yields overall picture, and shows outlying values)
2. scatter plot: plot residuals against each X_k in the model separately (look at pattern (curved?) and spread)
3. added variable plot (partial regression plot, see Velleman and Welsch (1981)): plot residuals of X_j against residuals of Y with omitted X_j (to show the effect of adding X_j to the model.) (Or, to show the relationship between Y and X_j, once all other predictors have been accounted for.)
4. scatter plot: plot residuals against each X_k not in the model separately (look at pattern — linear? then include!)
5. scatter plot: plot residuals against Y and $\hat{Y}$ (look at spread)
6. normal QQ-plot of the residuals (check normality assumption)

response var.

some X var.

It should be all over place

how X_j influences the response w/all others taken into acc.

Indicates that X_k affects outcome

residuals should behave like normal

Example: bodyfat data (1)

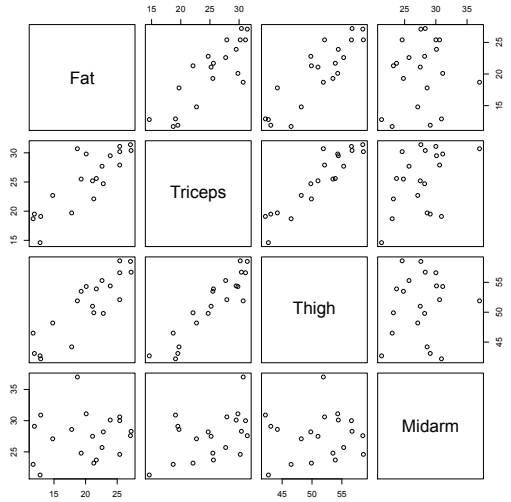
Read in the data.

```
>bodyfat=read.table("bodyfat.txt",header=T)
>attach(bodyfat)
```

1. Scatter plot of Y against each X_k separately.

```
> pairs(bodyfat)
```

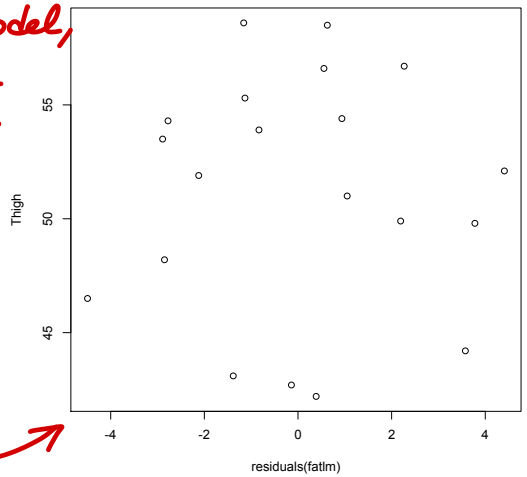
plots each col. against each col.



2. Scatter plot of residuals against each X_k in the model separately.

```
> bodyfatlm=lm(Fat~Thigh)
> plot(residuals(bodyfatlm),Thigh)
```

create model, then we take this residuals and plot against explan. var.

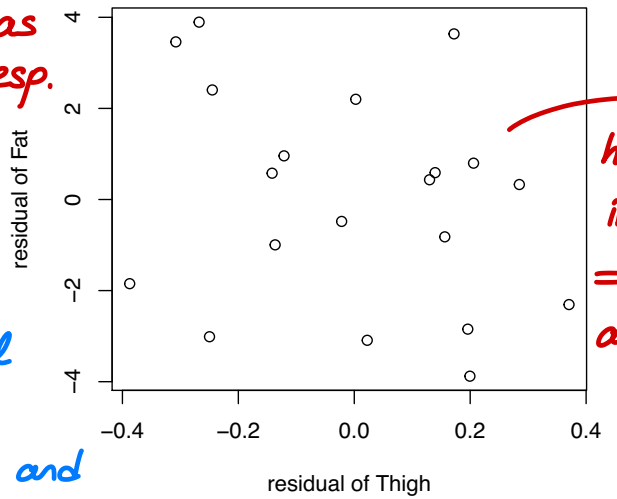


If a curved pattern is visible, include, e.g., X_j^2 or transform X_j (e.g., $\log(X_j)$, $\sqrt{X_j}$).

Example: bodyfat data (2)

3. Added variable plot of residuals of X_j against residuals of Y with omitted X_j.

```
> x=residuals(lm(Thigh~Midarm+Triceps))
> y=residuals(lm(Fat~Midarm+Triceps))
> plot(x,y,main="Added variable plot for
  + Thigh", xlab="residual of Thigh",
  + ylab="residual of Fat")
```



The slope in this plot is the regression coefficients β_j from the original multiple regression model, and the residuals in this plot are precisely the residuals from the original multiple regression. Outliers and heteroskedasticity (caused by X_j) can be identified by looking at the plot of this simple rather than multiple regression model.

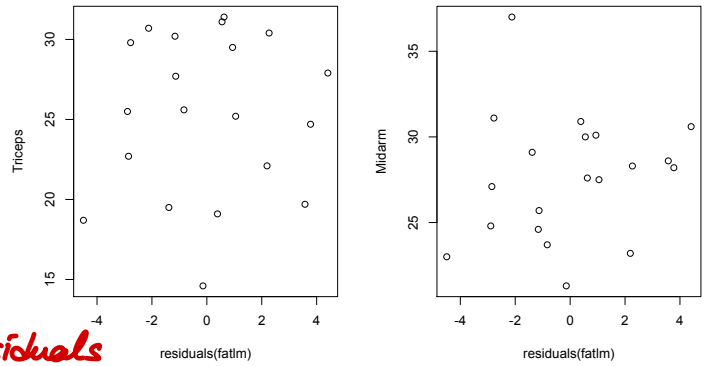
All the added variable plots can be obtained from the full model `mod` by the command `avPlots(mod)` from the package `car`.

Example: bodyfat data (3)

4. Scatter plot of residuals against each X_k not in the model separately.

```
> plot(residuals(bodyfatlm),Triceps)
> plot(residuals(bodyfatlm),Midarm)
```

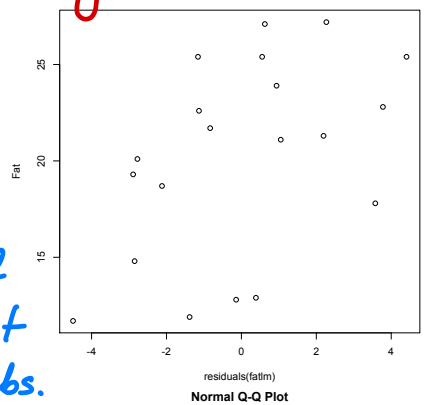
we take residuals of model and plot against X_k that is not in model



5. Scatter plot of residuals against Y (and $\hat{Y}$).

```
> plot(residuals(bodyfatlm),Fat)
> plot(res(bodyfatlm),fitted(bodyfatlm))
```

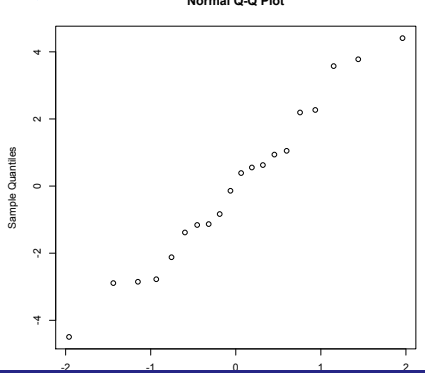
take residuals of model and plot against obs. val. = Y or fitted val. = $\hat{Y}$



6. Normal QQ-plot of the residuals.

```
> qqnorm(residuals(bodyfatlm))
```

check for norm. assump.



Also: `shapiro.test(residuals(bodyfatlm))`. If residuals are not normally distributed, go back to scatter plots and start with different model, possibly apply transforms.

outliers and influence points

Outliers, leverage points, influence points

very big diff. in responses, can be seen in boxplot

- An outlier is an observation with an extremely high or low response value, compared to what is expected under the model.
- A leverage (or potential) point is an observation with an outlying value in the explanatory variable.

when val. of exp. var. is too diff. than other var.s

- To study the effect of a leverage point one can fit the model with and without that data point. If the estimated parameters change drastically by deleting the leverage point, the observation is called an influence point.
- The Cook's distance D_i quantifies the influence of observation i on the predictions:

# of exp. var.s

we remove point i from responses and apply our model again → then we prod. fit. val.s

$$D_i = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{j=1}^{n} (\hat{Y}_{(i),j} - \hat{Y}_j)^2,$$

jth fit. resp. in full model

jth fit. val. w/ ith point removed

could influence a lot b/c some val.s of X are too big or small → you multiply your coeff.s w/them and get bigger/smaller resp.

with $\hat{Y}_{(i),j}$ the predicted j-th response based on the model without the i-th data point.

- Rule of thumb: if the Cook's distance for some data point is larger than 1, it is considered to be an influence point.

→ this point influences dataset too much, makes too much diff. in responses.

Outlier: Forbes' data

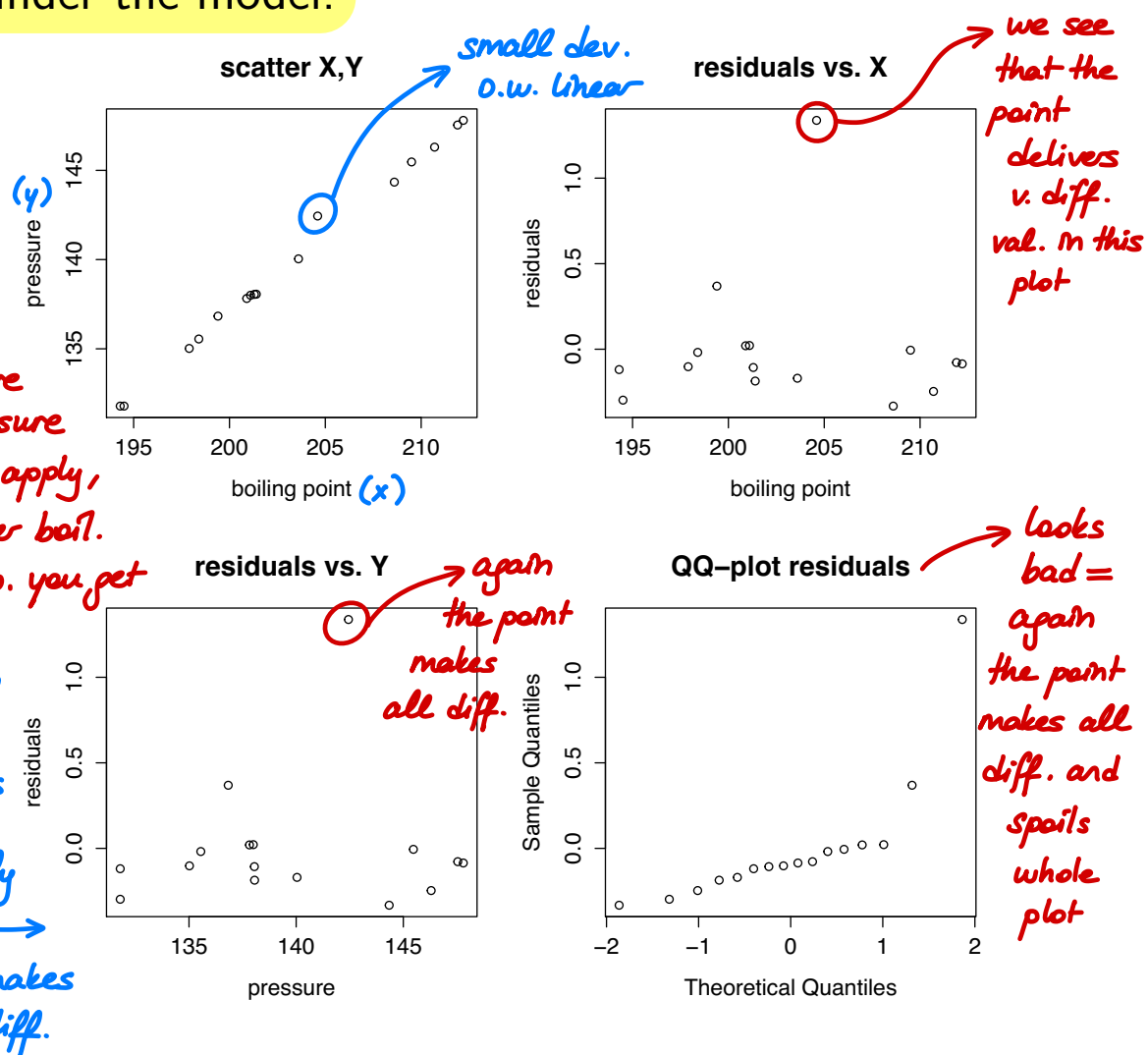
An outlier is an observation with an extremely high or low response value, compared to what is expected under the model.

Consider Forbes' data (in file `forbes.txt`) which describe the relation between boiling point of water and pressure.

```
> x=forbes[,2];y=forbes[,3]
> forbeslm=lm(y~x)
```

One outlier point "spoils" all the plots, its value deviates too much from what it is expected under the model.

Residuals are for the simple linear regression model.



Outlier: Forbes' data

```
> order(abs(residuals(forbeslm)))
 [1] 12  4  6  7 15 16  3  9  2 10  8 14  1 13  5 11
```

outlier

The 11-th data point seems to be an outlier. The command `order(abs(residuals(model)))` gives the indices of the ordered absolute values of residuals from smallest to largest.

The mean shift outlier model can be applied to test whether the k-th point significantly deviates from the other points in a linear regression setting.

```
> u11=rep(0,16); u11[11]=1; u11
 [1] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
> forbeslm11=lm(y~x+u11); summary(forbeslm11)
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.787278   1.530216 -26.655 9.87e-13 ***
x             0.888534   0.007533 117.950  < 2e-16 ***
u11           1.433143   0.177565   8.071 2.03e-06 ***
...
```

we introduce artif. col./dummy var.s

all 0's except 1 is 1 at index of dt. pt. that makes all diff.

we include dummies in model

reflects imp. of point = it makes diff. b/c coeff. is sig. diff. from 0.

Since the coefficient for explanatory variable u11 is significantly different from 0, the outlier is significant.

Leverage/influence points: Huber's data

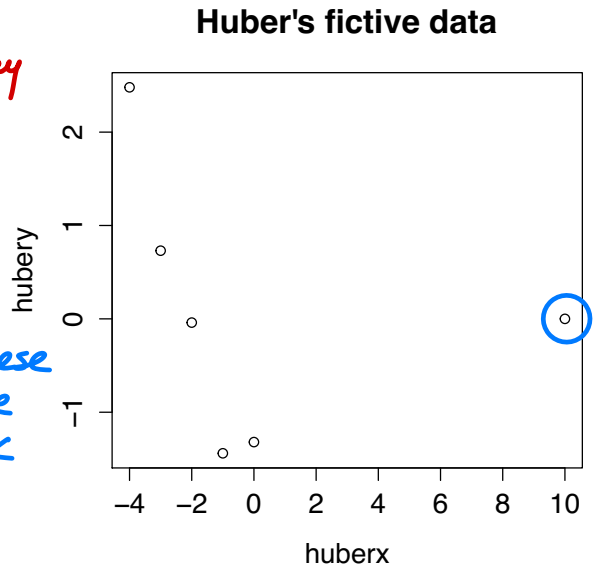
Consider Huber's fictive data.

Question: what is the influence of the observation with value x=10 of the explanatory variable?

far away from rest

```
> xh = c(-4:0,10)
> yh = c(2.48,.73,-.04,-1.44,-1.32,0)
> huberlm = lm(yh ~ xh)
```

these are OK

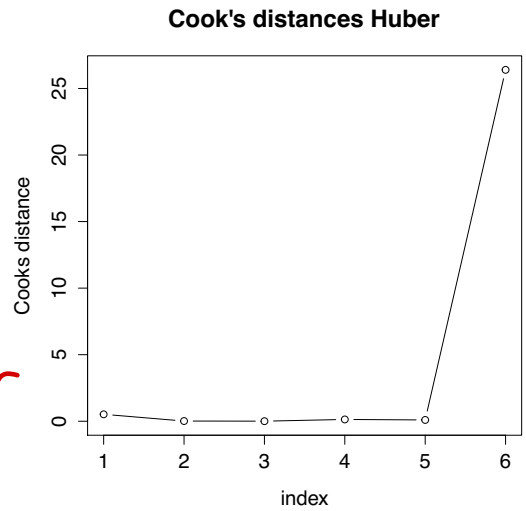


Compute and plot the Cook's distances.

```
> round(cooks.distance(huberlm),2)
    1    2    3    4    5    6
 0.52 0.01 0.00 0.13 0.10 26.40
> plot(1:6,cooks.distance(huberlm),type="b")
```

x = 10

huge when all others are small



Here we see an influence point: the Cook's distance is 26.40 for the leverage point.

bigger than 1

collinearity

Collinearity

Collinearity is the problem of linear relations between explanatory variables. A straight line in a scatter plot of two variables means they explain the same.

Example. Suppose we have a response variable Y and one explanatory variable X_1. Now we add a second explanatory variable $X_2 = 2X_1$. Can we do a meaningful analysis using the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$? No, in this model we cannot uniquely estimate β_1 and β_2, because

$X_2 = 2X_1$ → all info we have in X_2 is same as in X_1 = including X_2 has no point b/c X_2 doesn't bring anything extra

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e = \beta_0 + (\beta_1 + 2\beta_2) X_1 + e$$

and only the sum $\beta_1' = \beta_1 + 2\beta_2$ is estimable. There are many choices β_1 and β_2 giving the same $\beta_1' = \beta_1 + 2\beta_2$ (e.g., $1 = \beta_1' = 0 + 2 \cdot 0.5 = 1 + 2 \cdot 0$).

β_1 and β_2 are not identif. → you can only estim. their comb. but not them indiv.ly

If X_1 and X_2 are close to collinear then β_1 and β_2 are difficult to estimate. This is reflected in large variances and large confidence intervals of $\hat{\beta}_1$ and $\hat{\beta}_2$.

If the confidence interval of $\hat{\beta}_j$ is large, the estimate is not reliable.

We can have collinearity amongst a set of more than two explanatory variables (multicollinearity).

you shouldn't include exp. var.s which are linear comb. of each other → makes model overload w/param which cannot be recovered = collin.

Ways to investigate and remove collinearity

you will only see that one is lin. comb. of the other

Graphical ways to investigate collinearity:

- scatter plot of X_i against X_j for all i, j (only pairwise collinearities visible).

Numerical way to investigate collinearity:

- pairwise linear correlation of X_i and X_j for all combinations i, j.
- variance inflation factor of β_j for all j (check whether these are high). ← easier

There are more advanced numerical ways to investigate collinearity (special packages in R like `car`), e.g.: condition indices, variance decomposition.

When there is collinearity amongst the explan. variables $X_1, \ldots, X_p$ one should

- avoid having two collinear explanatory variables in the model
- choose a model with a small number of explanatory variables
- choose a model that intuitively/practically makes sense

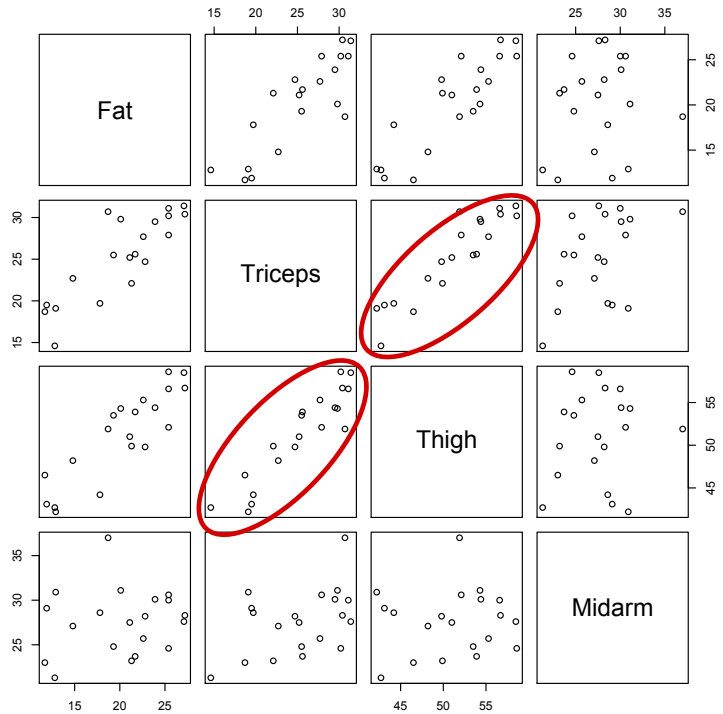
Recognizing multicollinearity among a set of explanatory variables is not necessarily easy. For pairwise collinearity, we can simply examine the scatterplots or the correlations between the variables, but we may miss more subtle forms of multicollinearity.

Example: bodyfat data

Apply these checks to the bodyfat data:

```
> round(cor(bodyfat),2)
        Fat Triceps Thigh Midarm
Fat    1.00    0.84  0.88   0.14
Triceps 0.84   1.00  0.92   0.46
Thigh  0.88    0.92  1.00   0.08
Midarm 0.14    0.46  0.08   1.00
```

```
> pairs(bodyfat)
```



we should only incl. 1 of them b/c there is a pairwise r.s. between them

Clearly Triceps and Thigh are collinear, both from the plot and from the correlation value of 0.92.

Variance inflation factor

A more useful approach is to examine the variance inflation factors (VIF) of the explanatory variables. The VIF for the j-th independent variable is given by

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, \ldots, k,$$

if there is lin. comb., then R_j^2 will be close to 1 → then VIF_j will be big.

where R_j^2 the determination coefficient R^2 from the regression of the j-th explanatory X_j (as response) variable on the remaining explanatory variables.

The VIF of an explanatory variable indicates the strength of the linear relationship between the variable X_j and the remaining explanatory variables.

Rule of thumb: VIF_j's larger than 5 (equivalent to $R_j^2 > 0.8$) give some cause for concern.

X_j is a lin. comb. of others

Remark: these values do not give information about which variables are in the same collinear group of variables.

Example: bodyfat data

We compute the *VIF*-values for the bodyfat data.

```
> bodyfatlm=lm(Fat~Thigh+Triceps+Midarm, data=bodyfat)
> library(car); vif(bodyfatlm)
   Thigh  Triceps   Midarm
564.3434 708.8429 104.6060
> bodyfatlm2=lm(Fat~Triceps+Midarm, data=bodyfat)
> vif(bodyfatlm2)
 Triceps   Midarm
1.265118 1.265118
> bodyfatlm3=lm(Fat~Thigh, data=bodyfat)
> vif(bodyfatlm23)
Error in vif.default(bodyfatlm3) : model contains fewer than 2 terms
```

all 3 var.s are incl.

huge val.s = one of 3 is lin. comb. of the others b/c all are bigger than 5

remove thigh

all smaller than 5

err. for 1 var. = there is only 1 col. so it cannot be lin. comb. of anything else

If we fit the full model all 3 VIF's are large, so there is a collinearity problem (as we saw in the scatter plots). The other 2 models are ok with respect to collinearity problems.

To finish

Today we dicussed:

- strategies to choose the variables (step up, step down)
- diagnostics in linear regression
- problems in linear regression (outliers and influence points, collinearity)

Next time: Lasso, ANCOVA, multiple testing, FDR control.