Statistical Data Analysis, Lecture 12

dr. Dennis Dobler

Vrije Universiteit Amsterdam

13 May 2020

Topics in this course

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Chapter 8: Linear regression analysis

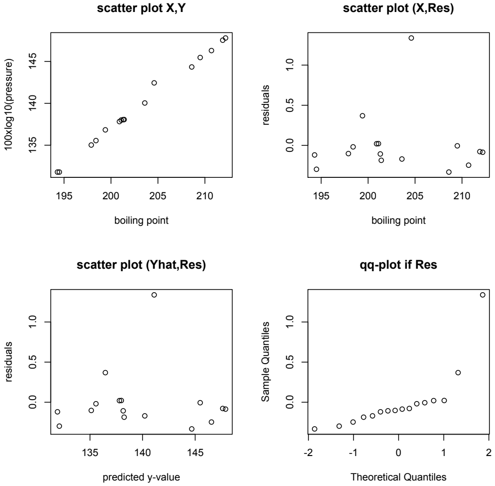
Contents of Chapter 8:

1. The multiple linear regression model
   - parameter estimation
   - selection of explanatory variables
2. Diagnostics
   - plots
   - outliers
   - leverage points
   - influence points
3. Collinearity

outliers

Definition outlier

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



Forbes' data on the relation between boiling point of water and pressure.

Residuals are for the univariate linear regression model.

Test for outliers (1)

In order to jugde whether the k^{th} point significantly deviates from the other points, a mean shift outlier model can be applied:

$$Y_i = \begin{cases} x_i^T\beta + e_i, & i \neq k \\ x_i^T\beta + \delta + e_i, & i = k, \end{cases}$$

In other words: for the k^{th} observation the mean is shifted by δ. In matrix notation:

$$Y = X\beta + u\delta + e = (X, u)\begin{pmatrix} \beta \\ \delta \end{pmatrix} + e,$$

with $u_i = 0$ for $i \neq k$ and $u_k = 1$.

Test for outliers (2)

The significance of the added parameter δ can be tested in $H_0 : \delta = 0$ using the common t-test:

$$T_{p+1} = \frac{\hat{\delta}}{\sqrt{\widehat{\mathrm{Cov}}(\hat{\beta}, \hat{\delta})_{p+1,p+1}}}$$

which has under H_0 the t_{n-p-2}-distribution.

It is common to apply this test one-sided — we know whether the Y-value is very small ($\delta < 0$) or very big ($\delta > 0$).

Example

We apply this test to Forbes' data.

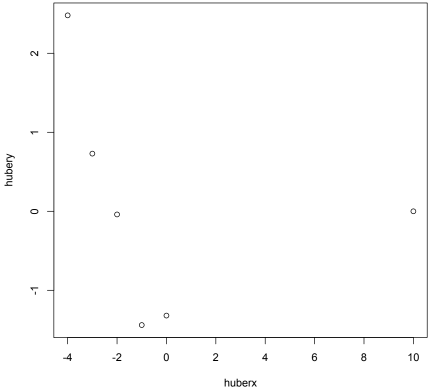
```
> round(residuals(lm(y~x)),2)
    1     2     3     4     5     6     7     8
-0.30 -0.12 -0.10 -0.02  0.37  0.02  0.02 -0.19
    9    10    11    12    13    14    15    16
-0.11 -0.17  1.34 -0.01 -0.33 -0.25 -0.08 -0.09
> u=c(rep(0,10),1,rep(0,5))
> msolm=lm(y~x+u)
> summary(msolm)
...
Coefficients:
              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 ***
u             1.433143   0.177565   8.071 2.03e-06 ***
...
```

Obviously we have encountered an outlier: $H_0 : \delta \leq 0$ is rejected with $2p = 0.00000203$ (in the output there is the two-sided p-value!)

leverage points

Definition leverage point

A leverage point or potential point is an observation with an outlying value in the explanatory variable; also see next slide.



Huber's fictive data.

Question What is the influence of the observation with x=10?

Investigate leverage points

Consider the predicted response $\hat{Y}$:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

with

$$H = X(X^T X)^{-1} X^T$$

the so called hat matrix. Its diagonal elements h_{ii} are in [0,1]. It turns out (see syllabus) that the variance of the i^{th} residual is equal to

$$\mathrm{var}(R_i) = \sigma^2(1 - h_{ii})$$

This means that if h_{ii} is close to 1 the i^{th} residual will be small. In other words, the fit will be pulled towards a perfect fit for the i^{th} observation, regardless of the value of Y_i! It depends on the value Y_i whether this has a large influence on the fitted parameters.

Typical rule: $h_{ii} > 2 \cdot (p+1)/n \;\Rightarrow i$th observation is a leverage point.

Example

We compute h_{ii}-values for Huber's data.

```
> xh = c(-4:0, 10)
> yh = c(2.48, .73, -.04, -1.44, -1.32, 0)
> huberlm = lm(yh ~ xh)
> round(hatvalues(huberlm),3)
    1     2     3     4     5     6
0.290 0.236 0.197 0.174 0.167 0.936
```

Value h_{66} is very close to 1. This was expected from the plot.
Also, $0.936 > 2 \cdot (p+1)/n = 2 \cdot 2/6 = 0.666$
So, the 6^{th} observation is a leverage point.

influence points

Definition influence point

To study the effect of a leverage point (or other points) one can fit the model with and without that data point. If the estimated parameters change drastically by deleting the single point, the observation is called an influence point. For a leverage point, this is not necessarily the case, it depends on the Y-value of the point.

The Cook's distance for the i^{th} data point is

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T(\hat{Y}_{(i)} - \hat{Y})}{(p+1)\hat{\sigma}^2}$$

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

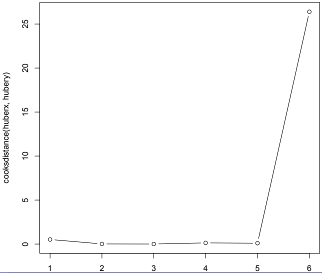
Rule of thumb If the Cook's distance for some data point is close to or larger than 1, it is considered an influence point.

Example

We compute the Cook's distances for Huber's data set:

```
> 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(humberlm))
```

Here we clearly have encountered an influence point: the Cook's distance is 26.40 for the data point (which we previously found to be a leverage point). A plot is usually insightful.



collinearity

Definition collinearity

Definition Explanatory variables X_1 and X_2 are called collinear if there is a linear relationship between X_1 and X_2.

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

Example Suppose we have a response variable Y and one explanatory variable X_1 (m). Now we add a second explanatory variable, $X_2 = 100X_1$ (cm).

Question Can we do a meaningful analysis using the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$?

Definition collinearity

Example Suppose we have a response variable Y and one explanatory variable X_1 (m). Now we add a second explanatory variable, $X_2 = 100X_1$ (cm).

Question Can we do a meaningful analysis using the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$?

In this model we cannot uniquely estimate β_1 and β_2. Only the sum $\beta_1 + 100\beta_2$ is identifiable, because X_1 and X_2 are perfectly collinear.

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

Ways to investigate collinearity

Graphical ways to investigate collinearity:

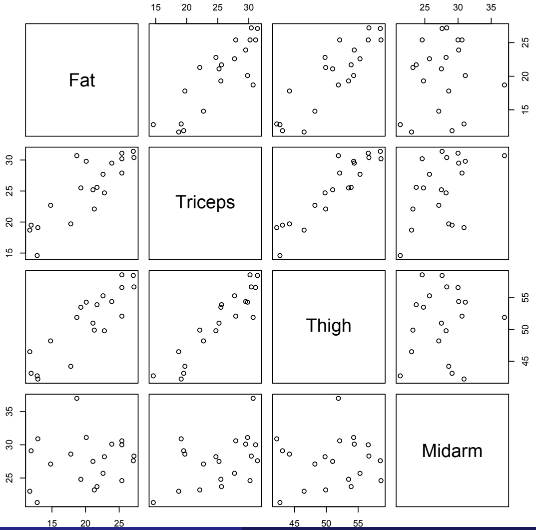
- scatter plot of X_j against X_k for all combinations j, k (check pairwise collinearity)

Numerical ways to investigate collinearity:

- pairwise linear correlation of X_j and X_k for all combinations j, k (check whether these are far from 0)
- variance inflation factor of all β_j (check which $\hat{\beta}_j$ are unreliable)
- condition number of design matrix X (check whether there is collinearity)
- condition indices of design matrix X (check the number of collinearities)
- variance decomposition of $\hat{\beta}_j$ (check which X_k are involved in collinearities)

Scatter plots

We look at the pairwise scatter plots of the `bodyfat` data:



Y= Fat
X_1= Triceps
X_2= Thigh
X_3= Midarm

X_1 and X_2 look very collinear.

Pairwise correlations

We compute the pairwise correlations of the bodyfat data.

```
> round(cor(bodyfat[,2:4]),2)
        Triceps Thigh Midarm
Triceps    1.00  0.92   0.46
Thigh      0.92  1.00   0.08
Midarm     0.46  0.08   1.00
```

We see that the correlation between Triceps and Thigh is indeed very high (0.92). This is in agreement with the scatter plots (of course!).

Variance inflation factor

To see which variables (columns in X) are involved in collinearities we can look at $R_{X_j}(X_{-j})$ — the residuals of X_j regressed on the other explanatory variables (cf. added variable plot). If these residuals are very small, then the j^{th} column of X is (nearly) a linear combination of other columns.

This is quantified in the variance inflation factor

$$VIF_j = \frac{1}{1 - \mathcal{R}_j^2}, \qquad j = 1, \ldots, p,$$

with $\mathcal{R}_j^2$ the determination coefficient of the mentioned regression.

If VIF_j is (much) larger than 1 (its minimum) $\hat{\beta}_j$ is unreliable. However, these values do not give information about which variables are in the same collinear group of variables.

Example

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

```
> head(bodyfat)
   Fat Triceps Thigh Midarm
1 11.9    19.5  43.1   29.1
2 22.8    24.7  49.8   28.2
3 18.7    30.7  51.9   37.0
4 20.1    29.8  54.3   31.1
5 12.9    19.1  42.2   30.9
6 21.7    25.6  53.9   23.7
> varianceinflation(bodyfat[,2:4])
[1] 708.8429 564.3434 104.6060
```

All 3 *VIF*'s are large! So there is a collinearity problem here (as we saw in the scatter plots). We need to determine which variables are involved, in which groups (the intercept might also be part of the problem...).

Condition number of X

Perfect collinear columns in X lead to $rank(X) < p + 1$. This implies that $rank(X^TX) < p + 1$ as well. This matrix will have at least one eigenvalue λ_j equal to 0.

Since $\hat{\beta} = (X^TX)^{-1}X^TY$ involves the inverse of X^TX, $\hat{\beta}$ cannot be determined if X^TX has not full rank.

If the smallest eigenvalue of X^TX is close to zero, we are already in trouble. The variance of one or more $\hat{\beta}_j$'s will be large (remember $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$).

This problem is quantified in the condition number of X:

$$\kappa(X) = \sqrt{\frac{\max_j \lambda_j}{\min_j \lambda_j}}$$

If $\kappa(X) > 30$ you should investigate collinearity problems further.

Condition indices of X

We can look at all eigenvalues of $X^T X$ and their ratios to the largest. This yields the condition indices of X:

$$\kappa_k(X) = \sqrt{\frac{\max_j \lambda_j}{\lambda_k}}$$

Again the threshold of 30 is recommended.

Note that the eigenvalues depend on the scaling of the columns. It is wise to scale all columns in X to the same Euclidean length (e.g. 1).
Be careful: it often does not make sense to apply the rescaling to the linear model! This might only be useful for the diagnostics!

Example

We compute the condition indices of the bodyfat data (with unscaled columns):

```
> conditionindices(bf[,2:4])
[1]     1.00000    16.62115    26.04727 11482.12116
```

The condition number is equal to the biggest condition index. It is much larger than 30!

Scaling the columns does not change much, since the scales of the columns are not that different.

Singular value decomposition

The (rectangular) design matrix X can be decomposed in a singular value decomposition (SVD)

$$X = UDV^T$$

with U a full $n \times (p+1)$, D a diagonal $(p+1) \times (p+1)$ matrix and V a full $(p+1) \times (p+1)$ matrix. This SVD is comparable to an eigenvalue decomposition for square matrices.

The singular values of X are the entries of D
and their squares are the eigenvalues $\lambda_1, \ldots, \lambda_{p+1}$ of $X^T X$.

If the k^{th} eigenvalue λ_k of $X^T X$ is small, then $\sum_{j=0}^{p} v_{jk} X_j \approx 0$. In other words, the linear combination of explanatory variables that corresponds to the (multi)collinearity is given by the coefficients v_{jk} of the matrix V.

Variance decomposition

One can show that

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \sum_{k=1}^{p+1} \frac{v_{jk}^2}{\lambda_k}.$$

In each of the terms in the sum one λ_k is involved, corresponding to one condition index $\kappa_k(X)$.

Consider the relative contribution of each term in the sum. These contributions are called the variance decomposition proportions of $\hat{\beta}_j$.

Example (1)

We compute the variance decomposition for the bodyfat data.

```
> round(vardecomposition(bodyfat[,2:4]),3)
     conditionindices 0     1     2     3
[1,]            1.000 0 0.000 0.000 0.000
[2,]           16.621 0 0.000 0.000 0.007
[3,]           26.047 0 0.005 0.001 0.000
[4,]        11482.121 1 0.995 0.998 0.993
```

The first column contains the $\kappa_k(X)$. The other columns contain the relative contributions to the variance of $\hat{\beta}_j$ for $j = 0, 1, 2, 3$. Each column sums to 1 (apart from rounding).

Last row: variance of all four $\hat{\beta}_j$'s is dominated by last condition index. Interpretation: all three explanatory variables and intercept are involved in one and the same multicollinearity.

Example (2)

This problem is also illustrated in the estimated standard errors for the $\hat{\beta}_j$'s in the full model for these data:

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

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,Adjusted R-squared: 0.7641
```

All standard deviations are large, none of the parameters is significant. Nevertheless, the determination coefficient is 80%.

This is an indication of trouble! Note This explains why we got two different models last week, explaining the same `Fat`-values.

Remedies against collinearity

There is no standard fix against collinearity. Try:

- plots, plots, plots!
- scaling the columns (if that makes sense)
- deleting explanatory variables
- something else ...

Read the last examples in the syllabus, which illustrate these problems nicely.

to finish

To wrap up

Today we discussed

1. The multiple linear regression model
2. Diagnostics
   - plots
   - outliers
   - leverage points
   - influence points
3. Collinearity

Overview

In this course we covered:

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Exam on Wednesday 27 May (GOOD LUCK)