STATS545.3 Regression diagnostics

Gunnar Stefansson

June 26, 2012

Contents

1 Verifying the assumptions of SLR

1.1 Introduction

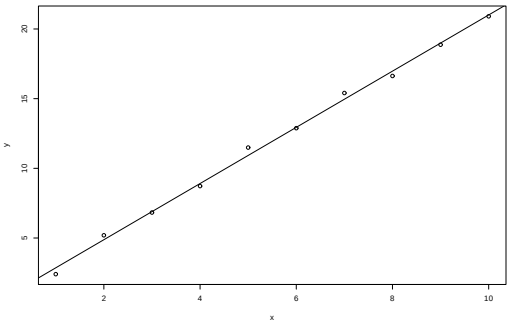


Figure 1: Simulated data

Evaluate model assumption: $Y_i \sim n(\alpha + \beta x_i, \sigma^2)$, independent.

- Linearity
- Independence
- Normality
- Constancy of variance

The simple linear regression model can be formulated succinctly as

$$Y_i \sim n(\alpha + \beta x_i, \sigma^2), \text{ independent}$$

Hence the underlying random variables are assumed to come from a Gaussian distribution, their mean is a linear function of the x-variables, their variance is constant and they are independent:

- Linearity
- Independence
- Normality
- Constancy of variance

These assumptions are all used when hypotheses are tested or confidence intervals obtained for parameters. For several other uses only some of these assumptions are required.

Naturally, each of these assumptions may be violated and some of these violations may influence the validity of any conclusions drawn.

This tutorial introduces some methods for "regression diagnostics", i.e. techniques for checking the validity of these assumptions. The first two sections/lectures contain methods appropriate for simple linear regression (SLR) whereas the subsequent sections introduce methods which are used in multiple regression.

Example: It will be useful to have a fixed example at hand to illustrate the concepts and methods. The following R commands will generate and plot data which satisfy the assumptions:

```
set.seed(19)                                # make sure we can repeat these results
alpha<-1                                    # fix the true intercept
beta<-2                                     # fix the true slope
sigma<-0.5                                  # fix the true standard deviatoin
n<-10                                       # the base sample size
x<-1:n                                      # set the base x-values
y.base<-alpha+beta*x+rnorm(10,sd=sigma) # set the base y-values

y<-y.base
```

The simulated data can be plotted along with the regression line with

```
plot(x,y)
abline(fm.base)
```

and analysed with

```
fm.base<-lm(y~x)
summary(fm.base)
```

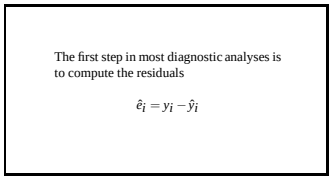
which gives

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45512 -0.16751 -0.08178  0.22318  0.56482

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.84411    0.24303   3.473   0.0084 **
x            2.01628    0.03917  51.478 2.25e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3558 on 8 degrees of freedom
Multiple R-squared: 0.997,      Adjusted R-squared: 0.9966
F-statistic:  2650 on 1 and 8 DF,  p-value: 2.247e-11
```

Note 1: Recall that if we do not specify the distributional assumptions, the simple linear regression model can be written in short-hand notation as $E[\mathbf{y}] = \beta\mathbf{X}$, where $\mathbf{X}$ is an $n \times 2$ matrix (and could in general be $n \times p$). In the following we frequently use this matrix notation.

Note 2: See influence.measures(), resid() and other functions related to lm() in R.

1.2 Residuals



When we write the linear model in terms of the data,

$$y_i = \alpha + \beta x_i + e_i,$$

the e_i are the actual residuals which generated the numbers.

The estimated residuals, based on the fitted model, are defined as

$$\hat{e}_i := y_i - \hat{y}_i$$

and these form the basis for most validation or diagnostics tests.

These are usually considered observed values of ε_i in the model

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where ε_i are usually taken to be i.i.d. $n(0, \sigma^2)$.

Note that $\bar{\hat{e}} = 0$ always holds and hence the observed residuals are not independent observations. Since the residuals in SLR correspond to estimation of two parameters, the variance of the true residuals is estimated with

$$s^2 = MSE = \frac{\sum_{i=1}^n e^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2},$$

which estimates σ^2.

Since the true residuals have variance σ^2, a natural first step in standardizing the observed residuals is with

$$\frac{e_i}{s},$$

in order to obtain observations which are close to having mean zero and variance 1, but it will be seen later that there are many alternatives to this scaling method.

Example: Continuing with the previous example, the R commands

```
ehat<-resid(fm)
plot(x,ehat)
```

will compute the residuals from the regression and plot them agains the *x*-variable.

Example: It is easy to modify the base example to simulate data which do not satisfy the linearity assumption, e.g. with

```
set.seed(19)                                # make sure we can repeat these results
alpha<-1                                    # fix the true intercept
beta<-2                                     # fix the true slope
sigma<-0.5                                  # fix the true standard deviatoin
n<-10                                       # the base sample size
x<-1:n                                      # set the base x-values
y.base<-alpha+beta*x+rnorm(10,sd=sigma) # set the base y-values
```

```
y<-y.base
#plot(x,y)
#abline(fm.base)
fm.base<-lm(y~x)
#summary(fm.base)

# residual plots with and without quadratic terms
ehat<-resid(fm.base)
#plot(x,ehat)

y.nonlin<-alpha+beta*x+0.2*x*x+rnorm(10,sd=sigma)
fm.nonlin<-lm(y.nonlin~x)
ehat.nonlin<-resid(fm.nonlin)
plot(x,ehat.nonlin)
```

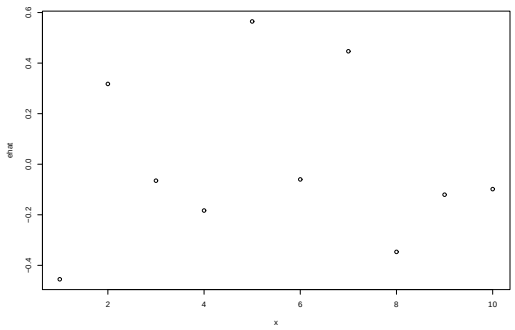
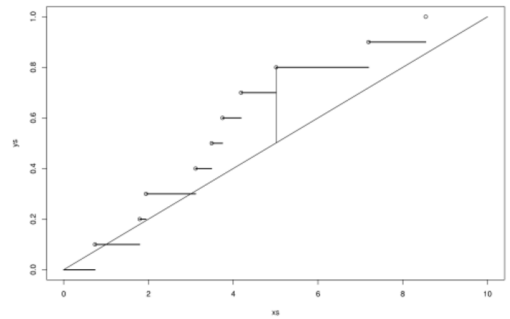


Figure 2: Residuals vs independent variable. No errors in model assumptions.

1.3 Verifying the distribution



Kolmogorov-Smirnov: Compares data to a theoretical distribution

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)}(x) \text{ for } x \in \mathbf{R}$$

$$H_0 : P\left[X_i \leq x\right] = F(x) \text{ for } x \in \mathbf{R}.$$

The statistic:

$$D := \sup_x |F_n(x) - F(x)|.$$

Given a set of measurements (possibly output such as deviations from a model) it is possible to set up an empirical distribution function. It is also possible to set up an hypothesis which provides a theoretical cumulative distribution function, such as for a Gaussian distribution. These two distribution functions can then be plotted together and compared to evaluate whether the data fit the hypothesis.

The Kolmogorov-Smirnov testing procedure computes the largest possible difference, *D*, between the empirical and cumulative distribution functions.

The probability distribution of *D* has been tabulated and the null hypothesis of e.g. normality is rejected for large enough values of *D*.

More specifically, denote by F_n the empirical distribution function (e.d.f.), so

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} I_{[x_i,\infty)}(x) \text{ for } x \in \mathbf{R}$$

where I denotes the indicator function and hence $F_n(x)$ is simply the fraction of observations which lie below the number x for $x \in \mathbf{R}$. Let F denote the proposed cumulative distribution function, so the hypothesis to be tested is

$$H_0 : P[X_i \leq x] = F(x) \text{ for } x \in \mathbf{R}.$$

The statistic to be used is

$$D := \sup_{x \in \mathbf{R}} |F_n(x) - F(x)|. \tag{1}$$

Example: Although R has a built-in function (ks.test) to evaluate the Kolmogorov-Smirnov test, it is quite useful to formally plot and evaluate such differences manually, or at least using transparent commands. The following generates data from a $U(0,10)$-distribution and then tests whether that really is the distribution.

```
x<-runif(10)*10                 # Generate some artificial ‘‘data’’
xs<-sort(x)                     # Sort the data
y<-rep(1/length(x),length(x))   # The y-axis for Fn is given
ys<-cumsum(y)                   # -- just the cumulative sum of 1/n
xgrid<-c(0,xs[1],NA,rbind(xs[1:(length(ys)-1)],xs[2:length(ys)],rep(NA,length(ys)-1)),xs[length
ygrid<-c(0,0,NA,rbind(ys[1:(length(ys)-1)] ,ys[1:(length(ys)-1)],ys[1:(length(ys)-1)]),1,1)
yhat<-xgrid/10
diff<-ygrid-yhat
adiff<-abs(diff)
D<-max(adiff,na.rm=T)
plot(xs,ys,xlim=c(0,10),ylim=c(0,1))
lines(xgrid,ygrid,lwd=2)
lines(c(0,10),c(0,1))
Dx<-xgrid[adiff==D]
Dy<-ygrid[adiff==D]
Dx<-Dx[!is.na(Dx)]
Dy<-Dy[!is.na(Dy)]
dropline<-diff[xgrid==Dx&ygrid==Dy]
dropline<-dropline[!is.na(dropline)]
lines(c(Dx,Dx),c(Dy,Dy-dropline))
```

Example: To continue with the base example where $Y_i = \alpha + \beta x_i + \varepsilon_i$ and $\varepsilon_i \sim n(0,\sigma^2)$ are independent, residuals from this model can be computed as before

```
set.seed(19)                              # make sure we can repeat these results
alpha<-1                                  # fix the true intercept
beta<-2                                   # fix the true slope
sigma<-0.5                                # fix the true standard deviatoin
n<-10                                     # the base sample size
x<-1:n                                    # set the base x-values
y.base<-alpha+beta*x+rnorm(10,sd=sigma) # set the base y-values
y<-y.base
fm.base<-lm(y~x)
ehat<-resid(fm)
```

and checked for normality by first computing residuals and finally running the ks.test function with

```
# check the normality assumption
df<-fm.base$df.residual
MSE<-sum(ehat^2)/df
s<-sqrt(MSE)
ks.test(ehat,pnorm,sd=s)
```

Note that the pnorm()-command needs to get the argument "sd=s" which is done by passing it through the ks.test-function.

Additional notes:

It should be noted that the supremum in Eq. 1 always corresponds to one of the data points. However, the e.d.f. jumps at each data point and is only continuous from the right. Since F_n is constant within each interval between measurements, but F is monotonically increasing, the individual differences in 1 (before taking absolute value) must be decreasing within each interval.

It follows that the absolute difference may either increase to a maximum from the right (at the left endpoint) or increase towards a supremum from the left.

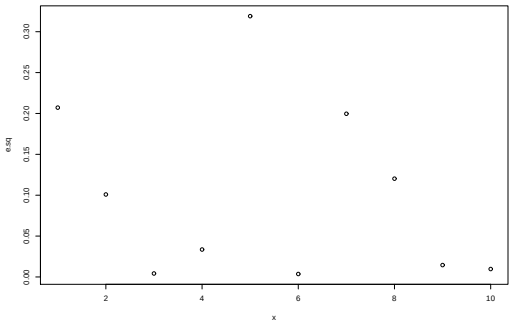
It follows that the statistic may be computed as

$$D = \max\left\{\max\left\{|F_n(x_i) - F(x_i)|, |F_n(x_i) - F(x_{i+1})|\right\}, 1 \leq i \leq n\right\},$$

i.e. simply by evaluating the differences at all the datapoints.

When there are no unknown parameters that need to be estimated in F, the statistic D has a distribution which is independent of F and the resulting test which rejects for large values of D is therefore termed a non-parametric test.

1.4 Constancy of variance



If the variance is constant, then e^2 should not show a trend in any independent variable.
Simple test: Regress e^2 on x and test in usual manner.
Slightly more advanced: Breusch-Pagan test takes properties of e^2 into account.

Figure 3: Base model with correct assumptions: e vs x.

A simple way to test whether the assumption of constant variance in regression analysis holds is to first compute the residuals from the regression and define a new variable as the squared residuals.

These squared residuals should then not show a significant trend in any way, when plotted or modelled as functions of the independent variables.

Alternative methods abound, such as splitting the data into two groups according to the levels of the x-variable and computing separately the variance in each group.

Example: Verifying the constancy of variance in the base example is done in the following

```
set.seed(19)                          # make sure we can repeat these results
```

```
alpha<-1                                    # fix the true intercept
beta<-2                                     # fix the true slope
sigma<-0.5                                  # fix the true standard deviatoin
n<-10                                       # the base sample size
x<-1:n                                      # set the base x-values
y.base<-alpha+beta*x+rnorm(10,sd=sigma) # set the base y-values

y<-y.base

fm.base<-lm(y~x)
ehat<-resid(fm)
# check whether the variance is constant
plot(x,ehat)
e.sq<-ehat^2
plot(x,e.sq)
```

The command

```
summary(lm(e.sq~x))
```

will then test whether the quadratic term is significant.

Example: It is also easy to generate an example with a variance which increases as a function of x:

```
# example with increasing variance
x.inc<-1:100
y.inc<-alpha+beta*x.inc+rnorm(100,sd=sigma*x.inc)
plot(x.inc,y.inc)
ehat.inc<-resid(lm(y.inc~x.inc))
plot(x.inc,ehat.inc)
e.sq<-(ehat.inc)^2
plot(x.inc,e.sq)
```

and a simple summary command will show a significant relationship between e^2 and x.

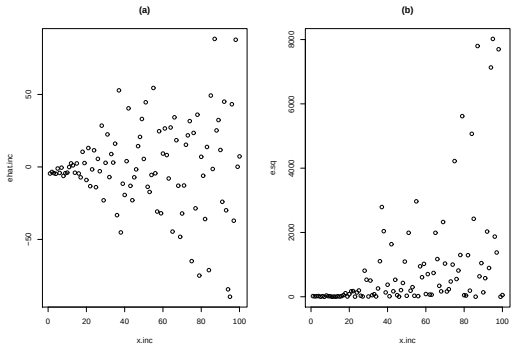


Figure 4: Example with increasing variance with x, (a) residuals, e, vs x, (b) e^2 vs x

1.5 Verifying linearity

Many tests available:

- Plot residuals against x-variable

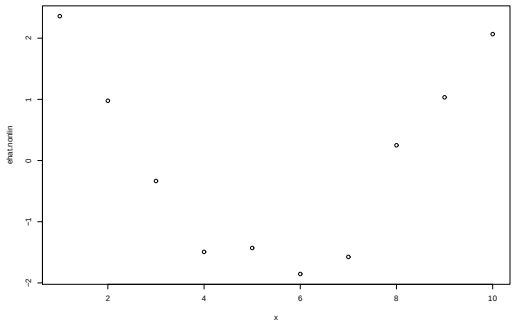


Figure 5: Residuals vs independent variable. Error in linearity assumption.

Basic:

- Plot residuals against x-variable
- Look for pattern

Later:

- Test for autocorrelation
- Multiple regression: Add a quadratic term
- Lack-of-fit tests (replace x by a factor)

- Look for pattern

  Later:
- Use factors
- Test for autocorrelation
- Multiple regression: Add a quadratic term
- Lack-of-fit test

1.6 Tests are approximate

Testing for normality etc is only approximate

Most of the tests used for diagnostics are only approximate.

The Kolmogorov-Smirnov test is derived under the assumption that the distribution is fully specified under the null hypothesis. However, the residuals in OLS are computed after fitting a model and hence they are not independent.

Similarly when plotting e^2 against x.

Note that exact tests exist, but these simple approximate tests are often adequate.

2 Further diagnostics in SLR

2.1 Outliers and influential cases

It is in particular important to search for outliers or influential cases in the x or y-measurements.
Typically use residuals and/or hat matrix:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Methods for this will be introduced.

An important part of verifying a model is to search for outliers or influential cases in the x or y-measurements.

Methods for this will be introduced.

Recall that in the regression problem with $E[\mathbf{y}] = \mathbf{X}\beta$ (and $\mathbf{X}$ of full rank), the least squares estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and the predicted values of the $\mathbf{y}$-vector are given by $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The case of simple linear regression involves an $\mathbf{X}$-matrix of dimensions $n \times 2$ but most of this tutorial also applies to more general cases (multiple regression, i.e. more than one x-variable).

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ therefore transforms the $\mathbf{y}$-vector into the predicted values $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and is therefore termed the **hat matrix**. The hat matrix is symmetrix, $\mathbf{H}' = \mathbf{H}$ and it is a projection so that $\mathbf{H}^2 = \mathbf{H}$.

The residuals are correspondingly obtained with $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ where $\mathbf{I} - \mathbf{H}$ is also symmetric and a projection matrix. It is therefore easy to see that the exact variance-covariance matrix of the observed residuals is given with $\Sigma_e = \sigma^2(\mathbf{I} - \mathbf{H})$ and in particular the variance of the i'th observed residual is

$$V(e_i) = \sigma^2(1 - h_{ii}).$$

Naturally, this is not the same as $V(\varepsilon_i) = \sigma^2$ since the e_i are functions of several $\mathbf{y}$-measurements and thus several ε's through the hat matrix.

Example: Some examples of deviations from assumptions can be set up e.g. with

```
set.seed(19)                                # make sure we can repeat these results
alpha<-1                                    # fix the true intercept
beta<-2                                     # fix the true slope
sigma<-0.5                                  # fix the true standard deviatoin
n<-10                                       # the base sample size
x.base<-1:n                                      # set the base x-values
y.base<-alpha+beta*x.base+rnorm(10,sd=sigma) # set the base y-values
par(mfrow=c(2,2))
y<-y.base
x<-x.base
plot(x,y)
fm.base<-lm(y~x)
#summary(fm.base)
abline(fm.base)
title("(a)")
# Outlier in x
y<-y.base
x<-x.base
x[n]<-2*x[n]
fm<-lm(y~x)
plot(x,y)
abline(fm,col="red")
abline(fm.base)
title("(b)")

# Outlier in y at end
y<-y.base
x<-x.base
y[n]<-2*y[n]
fm<-lm(y~x)
plot(x,y)
abline(fm,col="red")
abline(fm.base)
title("(c)")
```

```
# Outlier in y in center
y<-y.base
x<-x.base
y[floor(n/2)]<-2*y[floor(n/2)]
fm<-lm(y~x)
plot(x,y)
abline(fm,col="red")
abline(fm.base)
title("(d)")
```

The hat matrix can now be set up directly using matrix algebra in R by first setting up the X matrix and then using the usual formulas with

```
y<-y.base
x<-x.base
one<-rep(1,n)
X<-cbind(1,x)
H<-X%*%solve(t(X)%*%X)%*%t(X)
```

Upon which H contains the following diagonal elements

```
> round(diag(H),2)
 [1] 0.35 0.25 0.18 0.13 0.10 0.10 0.13 0.18 0.25 0.35
```

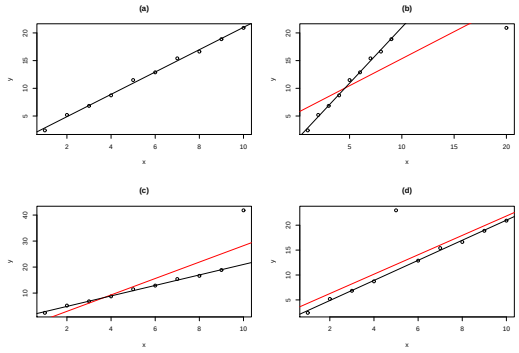


Figure 6: Effects of some outlier types on simple linear regression.

2.2 Diagnostics based on residuals

Diagnostics for residuals include tests for normality and constancy of variance.
Semistudentized residuals $(e_i/\sqrt{(MSE)})$ are commonly used but
studentized $e_i/\sqrt{(MSE)(1-h_{ii})}$
would obviously be better.

Diagnostics for residuals include tests for normality and constancy of variance.

Semistudentized residuals $(e_i/\sqrt{(MSE)})$ are commonly used. This refers to e_i being an observation of ε_i, which has variance σ^2 and MSE estimates σ^2.

Studentized $(e_i/\sqrt{(MSE)(1-h_{ii})}$ would be better since the variance of e_i is not σ^2 but $\sigma^2(1-h_{ii})$.

Example: The base SLR example has $x_i = i$ for $i = 1, \ldots, n = 10$ and we can set up the **X**-matrix in R with

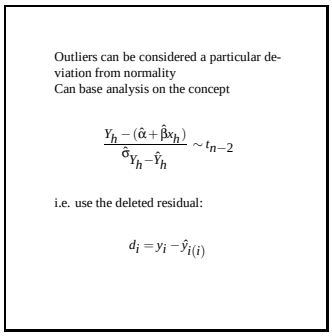
```
> one<-rep(1,n)
> X<-cbind(one,x)
```

This is then used to compute the h_{ii}-values with

```
> round(diag(X%*%solve(t(X)%*%X)%*%t(X)),2)
 [1] 0.35 0.25 0.18 0.13 0.10 0.10 0.13 0.18 0.25 0.35
```

and it is seen that there is considerable variation among these h_{ii}-values since in this example the variance to the ends is more than three times the variance in the middle.

2.3 Outliers in y - consider deleted residuals



Outliers can cause havoc with all inference in linear regression. These atypical y-values can stem from instrument failure or non-normality of the process itself. For example, if a gamma distribution is a better model than a Gaussian distribution, then outliers will be more frequent.

Recall that

$$\frac{Y_h - (\hat{\alpha} + \hat{\beta} x_h)}{\hat{\sigma}_{Y_h - \hat{Y}_h}} \sim t_{n-2}$$

for a new measurement at a location x_h, where the divisor is the "appropriate divisor".

With this in hand it is a relatively straightforward task to develop a test for outliers by simply deleting one observation at a time from the data set, refitting the regression line and evaluating the above ratio (replacing n by $n-1$ in all computations to take into account the deletion of an observation.

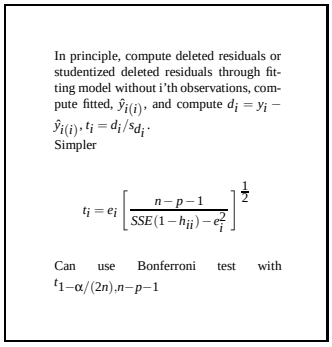
This technique is referred to as the method of deleted residuals, where the deleted residual itself is denoted

$$d_i = y_i - \hat{y}_{i(i)}$$

and the parenthesis indicates a model fit without the i'th observation.

Note that the deleted residual is a linear function of the original **y**-vector and formulae are available to compute these without refitting the regression line n times.

2.4 Computing deleted residuals



Consider deleted residuals or studentized deleted residuals through fitting model without i'th observations, compute fitted, $\hat{y}_{i(i)}$ and define $d_i = y_i - \hat{y}_{i(i)}$.

Define $t_i = d_i/s_{d_i}$ to obtain

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{\frac{1}{2}}$$

Can now use Bonferroni test with $t_{1-\alpha/(2n),n-p-1}$

Notes: The deleted residuals are based on repeatedly fitting the same model but deleting dropping a single observation each time. In principle this involves dropping observation *i* from both the **X**-matrix and the **y**-vector and fitting a model to this reduced data set. Assume the **X**-matrix is of full rank, as well as the matrices with deleted rows.

Denote the **X**-matrix without the *i*'th observation by $\mathbf{X}_{(i)}$ and the correspondingly reduced *y*-vector by $\mathbf{y}_{(i)}$. Using this reduced data set will result in a revised fitted parameter vector, $\hat{\beta}_{(i)}$. The solution to the normal equations for $\hat{\beta}_{(i)}$ will give

$$\hat{\beta}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}$$

and in order to derive the various equations of interest, expressions need to be obtained for this quantity.

To facilitate development of quantities it is useful to denote the rows of **X** by $\mathbf{x}'_i$, so that $\mathbf{X}' = (\mathbf{x}_1 \vdots \mathbf{x}_2 \vdots \ldots \vdots \mathbf{x}_n)$.

Now, some algebra can be used to see that

$$\mathbf{X}'\mathbf{X} - = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}'_i$$

and hence

$$\mathbf{X}'_{(i)}\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i$$

Further the following can be verified using simple matrix multiplication if **A** is a nonsingular matrix and **u**, **v** are column vectors, all of matching dimensions:

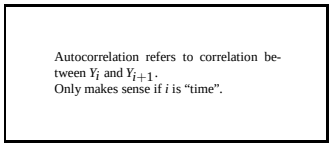
$$(\mathbf{A}+\mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{1}{1-\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}(\mathbf{A}^{-1}\mathbf{u})(\mathbf{v}'\mathbf{A}^{-1})$$

(c.f. Rao (1965), p. 33).

From this equations for $\hat{y}_{i(i)}$ and similar quantities can be derived, e.g.

$$y_i - \hat{y}_{i(i)} = \frac{e_i}{1-h_{ii}}.$$

2.5 Autocorrelation



An incorrect covariance structure can invalidate inference, since this affects the variance computations of $\hat{\beta}$, used when computing the t-statistics (or F) used for testing.

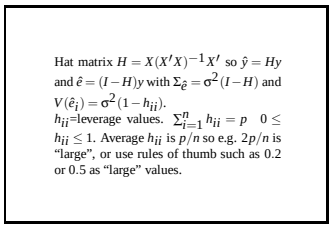
It should be noted, however, that the covariance structure was not used when deriving the result that the estimators in linear regression are unbiased.

The term autocorrelation only has meaning when some sort of order exists among the x-variables. In some cases there is an underlying time of year or day, which can be used for ordering, but in other cases length or some other variable can be used. The first question that arises in this context is whether there is a correlation between Y_i and Y_{i+1}.

Several tests for autocorrelation exist. Simple plots are useful, e.g. plotting $\hat{e}_i$ against i, followed by simple linear regression of $\hat{e}_i$ against i. Alternatively one can plot $\hat{e}_i$ vs $\hat{e}_{i-1}$ and investigate the corresponding correlations. Either approach has the problem that the observed errors are correlated in nature since they are computed based on the fitted values.

More advanced techniques are also available such as fitting a formal time series (or other) error structure to the residuals simultaneously with estimating the parameters, or implementing some two-stage method.

2.6 Leverage values



The diagonal values of the hat matrix, h_{ii} are termed **leverage values**. Each of these indicate how strongly the corresponding y_i "predicts itself".

Denote by $\mathbf{x}_i'$ the i'th row of the $\mathbf{X}$-matrix, so $X' = [\mathbf{x}_1 \vdots \mathbf{x}_p]$. Since h_{ii} is the element of the i'th row and i'th column of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, it is formed from the left through the i'th row of $\mathbf{X}$ and from the right by the i'th column of $\mathbf{X}'$, which is the transpose of the i'th row of $\mathbf{X}$. It follows that

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i.$$

Now, $\mathbf{X}'\mathbf{X}$ is always symmetric and since $\mathbf{X}$ is of rank p, $\mathbf{X}'\mathbf{X}$ is also positive definite and so is the inverse.

It follows that $h_{ii} > 0$.

Recall that $0 \leq V[\hat{e}_i] = \sigma^2(1-h_{ii})$ and hence $h_{ii} \leq 1$.

We have shown that

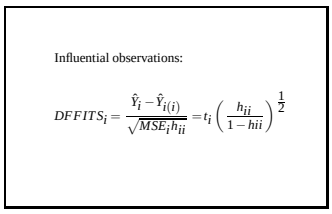
$$0 \leq h_{ii} \leq 1.$$

It is also true that

$$\sum_{i=1}^{n} h_{ii} = p.$$

The average value of the h_{ii} is therefore p/n so e.g. $2p/n$ is "large", or use rules of thumb such as 0.2 or 0.5 as "large" values.

2.7 Influential observations, DFFITS



Influential observations:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1-hii} \right)^{\frac{1}{2}}$$

where

$$t_i = \hat{e}_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - \hat{e}_i^2} \right]^{\frac{1}{2}}$$

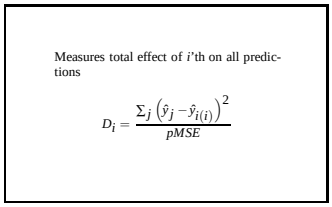
as before.

Note that $MSE_{(i)}$ refers to estimation of σ^2 while leaving out the i'th observation.

{\bf Example:} Age and live weight of lambs. Project: Complete regression analysis with detailed diagnostics.

\begin{verbatim} days weight 135 39 125 35 120 33 126 38 125 37 137 38 133 36 140 41 130 38 129 36 123 34 132 40 129 38 121 34 126 35 137 44 121 34 137 41 130 39 137 43 \end{verbatim}

2.8 Cooks distance



The DFFITS above only describe how a single observation affects the prediction of itself. Naturally one could also consider how observation i affects the prediction of the j'th observation by looking at $\hat{y}_{j(i)}$. If one were to carry this through and try to analyze how an observation affects the prediction of all other observations, this would lead to a somewhat intractable $n \times n$ matrix.

Cook's distance is a single measure describing how an individual observation affects all predictions, thus summarizing the information into an n-vector.

$$D_i = \frac{\sum_j \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{pMSE} = \frac{1}{ps^2} ||\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}||^2$$

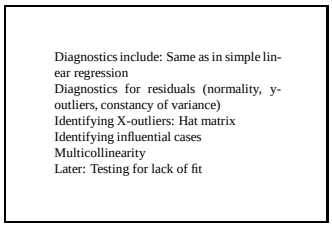
Given deletion formulas, it is not too hard to see that

$$D_i = \frac{\hat{e}_i^2}{ps^2} \frac{h_{ii}}{(1-h_{ii})^2}$$

and it is seen that this measure is large when either the residual $\hat{e}_i$ is large or the influence measure h_{ii} from the hat matrix is large.

3 Validating multiple regression models: Model diagnostics

3.1 Introduction and overview



Diagnostics include the same as in simple linear regression, but more are possible in the multiple regression setting and more things can go wrong.

Partial regression plots: Regress on x_1 and plot residual against residuals from regression of x_2 onto x_1. Provides indication of whether x_2 should be added and if so, how.

Diagnostics for residuals include tests for normality and constancy of variance. Semistudentized residuals ($\hat{e}_i/\sqrt{(MSE)}$) are commonly used but studentized ($\hat{e}_i/\sqrt{(MSE)(1-h_{ii})}$ (see below) would be better.

It is in particular important to search for outliers or influential cases in the x or y-measurements.

Note: One should always plot the residuals against $\hat{y}$ as well as against each independent variable.

Identifying y-outliers: Hat matrix $H = X(X'X)^{-1}X'$ so $\hat{y} = Hy$ and $\hat{e} = (I-H)y$ with $\Sigma_{\hat{e}} = \sigma^2(I-H)$ and $V(\hat{e}_i) = \sigma^2(1-h_{ii})$.

Further, consider deleted residuals or studentized deleted residuals through fitting model without i'th observations, compute fitted, $\hat{y}_{i(i)}$, and compute $d_i = y_i - \hat{y}_{i(i)}$, $t_i = d_i/s_{d_i}$, with

$$t_i = \hat{e}_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - \hat{e}_i^2} \right]^{\frac{1}{2}}$$

Can use Bonferroni test with $t_{1-\alpha/(2n),n-p-1}$

h_{ii}=leverage values. $\sum_{i=1}^{n} h_{ii} = p \quad 0 \leq h_{ii} \leq 1$. Average h_{ii} is p/n so e.g. $2p/n$ is "large", or use rules of thumb such as 0.2 or 0.5 as "large" values.

Influential observations:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_i h_{ii}}} = t_i \left(\frac{h_{ii}}{1-hii} \right)^{\frac{1}{2}}$$

Cook's distance...

DFBETAS - influence on coefficients

Multicollinearity...

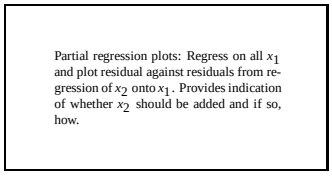
Later: Testing for lack of fit

See also section 10 in Neter et al.

See Belsley et al

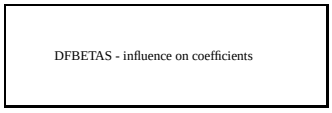
See help(dffits) in R

3.2 Partial regressions



Partial regression plots: Regress on all x_1 and plot residual against residuals from regression of x_2 onto x_1. Provides indication of whether x_2 should be added and if so, how.

3.3 DFBETAS



The variance-covariance matrix of the estimated parameter vector, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is given by $\Sigma_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Denote by c_{kk} the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ so the variance of β_k is given by

$$V[\beta_k] = c_{kk}\sigma^2.$$

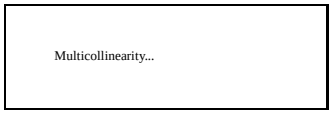
Use (i) to denote a fit without observation y_i, $0 \leq i \leq n$ and let $\hat{\beta}_k$ be one of the parameters, $1 \leq k \leq p$.

Define

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}$$

as a measure of the influence of measurement i on coefficient k.

3.4 Multicollinearity



Multicollinearity refers to the case when the columns of the **X**-matrix are "almost" linearly dependent and thus the $\mathbf{X}'\mathbf{X}$-matrix is difficult to invert. This basically means that some independent variables can be predicted based on the remaining set.

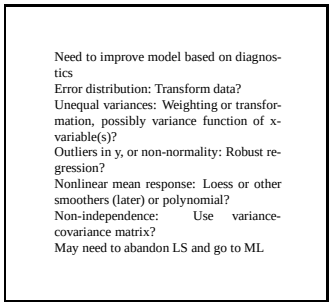
Commonly this simply reflects a poorly defined problem and sometimes it is possible to redefine the regression problem in terms of a different set of x-variables which are "less dependendent", i.e. are less correlated.

This situation arises in polynomial regression, where y_i is to be predicted based on a linear combination of x_i, x_i^2, x_i^3, etc. Naturally, if some of the x-values are larger than unity, then the raised values can become arbitrarily large and this alone will cause problems. Hence the variables are at a minimum scaled so that all raised values are limited to a reasonable magnitude. Simple scaling is not enough however, since as higher powers are used the X-matrix with columns containing 1, x_i, x_i^2 etc will become increasingly more difficult. The obvious solution here is to make the columns orthogonal by replacing

the columns with an orthonormal basis. Since the basis spans the same space as the columns of the original X-matrix, the projection (fitted values) will be the same.

The approach of replacing the original columns of the X-matrix by an orthogonal set is perfectly general and is clearly applicable whenever the interest is either in prediction or in simply testing for significance. If, however, the interest is in the parameter values themselves, then this is not sufficient.

3.5 Remedial measures



Need to improve model based on diagnostics

Error distribution: Transform data?

Unequal variances: Weighting or transformation, possibly variance function of x-variable(s)?

Outliers in y, or non-normality: Robust regression?

Nonlinear mean response: Loess or other smoothers (later) or polynomial?

Non-independence: Use variance-covariance matrix?

May need to abandon LS and go to ML

Notes: Model diagnostics will typically identify some problem areas and thus the model needs to be modified.

The investigation of the error distribution and the constancy of variances (homoscedasticity) are tightly linked. The most common problem is probably of inflated variances with increased values of an x-variable and this is commonly associated with a observed right-skewed distribution of the residuals, which again corresponds to the y-values having a right-skewed distribution, rather than Gaussian. The solution in this case may be to log-transform (at least) the y-values and perform a corresponding transformation of the model. Notably this commonly replaced an additive model with a multiplicative model and in many cases this is a reasonable approach. When looking for predictions on the original scale, however, this method is problematic and a bias is introduced.

Alternatively a weighted regression may be used for taking into account heteroscedasticity alone. A linear transformation may possibly be appropriate to take into account a correlation and variance structure (see below).

A different class of models, **generalized linear models** or **generalized additive models** may also be used to take into account different error structures without resorting to transformation of the data.

When the analysis of residuals, leverage values, deleted residuals, DFFITS or DFBETAS identifies outliers or influential observations some action is needed. It is not good practice to merely delete the corresponding values. Usually the detection of such problem values requires investigation of the initial data in order to find the source of the problem. In many real situations a data-entry problem can be found in this manner and appropriately corrected. In other cases the data point will have to be deleted since the investigation reveals that it is an impossible data point. The final resort will be to delete the

data point simply because it is in discordance with the entire rest of the data set.

Robust regression may be used to avoid the problems of outliers, influential values and non-normality.

Naturally a residual plot may indicate nonlinearity in the mean response. This may possibly be resolved using loess or other smoothers (e.g. generalized additive models) or polynomial regression?

In difficult situations a move from least squares estimation to maximum likelihood with a complex mean-variance structure may be required. A special case of this is when the residuals are found to be statistically dependent (below).

3.6 Correlated data

If $\mathbf{Y}$ are not independent, $\Sigma_{\mathbf{Y}} \neq \sigma^2\mathbf{I}$ then adjustments are needed.

The special case when the correlations are known is of particular significance. This case can be solved and is also quite common. Assume therefore that the variance-covariance matrix can be written in the form $\Sigma_{\mathbf{Y}} = \sigma^2\mathbf{B}$ with $\mathbf{B}$>0.

Use Cholesky factorisation to write $\mathbf{B} = \mathbf{T}'\mathbf{T}$ and $\mathbf{T} = \mathbf{U}^{-1}$.

Now define $\tilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$...

3.7 Further reading

Standard regression texts such as Neter et al (1996) all provide some methods for simple regression diagnostics and the theoretical foundations for these can be found or derived in theoretical texts such as Scheffe (1959).

More detailed treatment of diagnostics per se can be found in specific texts on this topics, e.g. Belsey et al (1980) which use theoretical results from linear algebra which can be found e.g. in Rao (1965).

Robust regression is handled in many textbooks, with several useful cases for robust and exploratory analysis given in Hoaglin et al (1983).

References

%T Regression diagnostics: Identifying influential data and sources of collinearity. %A Belsey, D. A., Kuh, E. and Welsch, R. E. %D 1980 %I John. Wiley and Sons, New York. %P 292pp ISBN: 0471058564

%T Understanding robust and exploratory data analysis %A Hoaglin, D. C., Mosteller, F. and Tukey, J. W. %D 1983 %I John. Wiley and Sons, New York. %P 447pp ISBN: 0471097772

%T Applied linear statistical models. %A Neter, J., A Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. %D 1996 %I McGraw-Hill %P 1408pp. ISBN: 0256117365

%T Linear statistical inference and its applications. %A Rao, C. R. %D 1965 %I John Wiley and Sons, New York. %P 625pp ISBN: 0471708232

%T The analysis of variance. %A Scheffe, H. %D 1959 %I John Wiley and Sons, Inc, New York. %P 477pp. ISBN: 0471758345