Multiple regression and more diagnostics

Output of ls.diag() in R [diagnostics after lsfit or lm].

An observation (row of data matrix) is influential if the value of $\hat{\boldsymbol{\beta}}$ changes a lot when this observation is deleted. influential observations only, as ideas are related to cross-validation with leave-one-out.

```
[1] "std.dev" "hat"         "std.res"  "stud.res"   "cooks"
[6] "dfits"   "correlation" "std.err"  "cov.scaled" "cov.unscaled"
```

n =sample size, $\mathbf{X}$=data matrix of explanatory variables of dimension $n \times k$ with first column of 1s for the intercept.

---

1. `std.dev`: residualSD $= \hat{\sigma} = \sqrt{\sum_i e_i^2/(n-k)}$

2. `hat`: diagonal of projection or hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, ith diagonal element denoted as P_{ii}.

3. `std.res`: vector of standardized residuals: $e_i^* = e_i/[\hat{\sigma}\sqrt{1-P_{ii}}]$; see below for the explanation of $1 - P_{ii}$.

5. `cooks`: Cook's distance = vector of inverse-covariance-matrix weighted squared distances of $\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}$ to measure influence of the observations:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})}{k\hat{\sigma}^2} = \frac{(e_i^*)^2}{k} \times \frac{P_{ii}}{1-P_{ii}},$$

   $\hat{\boldsymbol{\beta}}_{-i}$ is the vector of regression coefficients with $(\mathbf{x}_i, y_i)$ omitted. D_i is a distance measure that is invariant to scaling of explanatory variables; also other invariances.

   The covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ so the quadratic form in $(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})$ is scaled by the inverse covariance matrix.

6. `dfits`: Another measure of influence of the ith observation: $\text{dfits}_i = (\hat{y}_i - \hat{y}_{i|-i})/[\hat{\sigma}_{-i}\sqrt{P_{ii}}]$.

7. `correlation`: $V = (\mathbf{X}^T\mathbf{X})^{-1}$ converted to a correlation matrix, that is, $v_{ij} \to v_{ij}/\sqrt{v_{ii}v_{jj}}$.

8. `std.err`: vector of SEs of the $\hat{\beta}_j$ or the square roots of the diagonal of $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$.

9. `cov.scaled`: $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ = estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

10. `cov.unscaled`: $(\mathbf{X}^T\mathbf{X})^{-1}$

4. `stud.res`: vector of Studentized residuals

For your team project, if you have a continuous response variable, important diagnostics to pay attention to are Cook's distance and dfits (large absolute value indicates an influential observation for both).

Example:

```
fit=lm( ...)

diagnose=ls.diag(fit)

print(diagnose$cook); plot(diagnose$cook)

print(diagnose$dfits)
```

Check your data to understand why an observation with large Cook's distance or large absolute value of dfits is influential. Might be a typo if it is a point that is far from others.

```
# Example
set.seed(123)
x=1:10 y=1+3*x
+rnorm(10,0,0.5) x=c(x,11)
y=c(y,5)
x

# [1]  1  2  3  4  5  6  7  8  9 10 11
y
# [1]  3.719762  6.884911 10.779354 13.035254 16.064644 19.857532 22.230458
# [8] 24.367469 27.656574 30.777169  5.000000
fit=lm(y~x)
print(summary(fit))
#Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept)   6.5013     5.1369   1.266   0.2374
#x             1.6494     0.7574   2.178   0.0574 .
diagnose=ls.diag(fit)
options(digits=4)
print(diagnose$cook)
#  [1] 1.065e-01 2.729e-02 8.979e-04 5.345e-06 1.695e-03 1.043e-02 1.712e-02
#  [8] 2.889e-02 7.965e-02 1.945e-01 2.093e+00
print(diagnose$dfits)
# [1]  -0.446546  -0.222467  -0.039972  -0.003083   0.054994   0.137807
# [7]   0.177531   0.231791   0.393337   0.633971 -33.276447

print(diagnose$hat)
#  [1] 0.31818 0.23636 0.17273 0.12727 0.10000 0.09091 0.10000 0.12727 0.17273
# [10] 0.23636 0.31818
print(1-diagnose$hat)
# [1] 0.6818 0.7636 0.8273 0.8727 0.9000 0.9091 0.9000 0.8727 0.8273 0.7636
#[11] 0.6818
```

Explanation of standardized residuals

Sometimes for residual plots, standardized residuals or Studentized residuals are used instead of ordinary residuals. This is due to the random variable form of the residuals having different variances, depending on the location of $\mathbf{x}_i$.

$\epsilon_1, \ldots, \epsilon_n$ iid $N(0, \sigma^2)$ implies $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ has covariance matrix $\mathrm{Cov}\,(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$ (σ^2 on diagonal, 0 elsewhere).

Residual vector as a vector of random variables is:

$$\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{Y} - \mathbf{P}\mathbf{Y} = \mathbf{I}_n\mathbf{Y} - \mathbf{P}\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y},$$

where $\mathbf{P} \overset{\text{def}}{=} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the *projection matrix or hat matrix* (dimension $n \times n$).

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}$.

$\mathbf{P}$ projects $\mathbf{Y}$ into a vector in the column span of $\mathbf{X}$ ($\hat{\mathbf{Y}}$ is a linear combinations of the columns of $\mathbf{X}$ — to be written on whiteboard).

Why is $\mathbf{P}$ called the *hat* matrix?

$\mathbf{P} \overset{\text{def}}{=} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ satisfies $\mathbf{P}^T = \mathbf{P}$ and $\mathbf{P}^2 = \mathbf{P}$.

For the homoscedastic assumption, the covariance matrix of $\mathbf{Y}$ is $\sigma^2\mathbf{I}_n$. Hence the covariance matrix of $\mathbf{E} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$ is

$$(\mathbf{I}_n - \mathbf{P})(\sigma^2\mathbf{I}_n)(\mathbf{I}_n - \mathbf{P})^T = \sigma^2(\mathbf{I}_n - \mathbf{P})^2 = \sigma^2(\mathbf{I}_n - \mathbf{P}),$$

with ith diagonal element $\mathrm{Var}\,(E_i) = \sigma^2(1 - P_{ii})$.

If $(x_{i1}, \ldots, x_{ip})$ is closer to the edge of the x-space(?), then $1 - P_{ii}$ is smaller(?).

General result that combines probability and matrices: If $\mathbf{W}$ is an $m \times 1$ random vector with $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}}$ and $\mathbf{A}$ is a $q \times m$ matrix, then $\mathbf{V} = \mathbf{A}\mathbf{W}$ is a $q \times 1$ random vector, and its covariance matrix is $\boldsymbol{\Sigma}_{\mathbf{V}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{W}}\mathbf{A}^T$; this result was used earlier for $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.