Diagnostics Statistics

Durbin-Watson:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

```
DW = 2.5913, p-value = 0.9062
alternative hypothesis: true autocorrela
tion is greater than 0
```

Cook's distance measures how much the entire regression function changes when the i-th case is deleted.

Cook's distance

Y X_1 X_2

- plot(cookd(lm(prestige ~ income + education, data=Duncan)))
- Cook's distance measures how much the entire regression function changes when the i -th case is deleted.

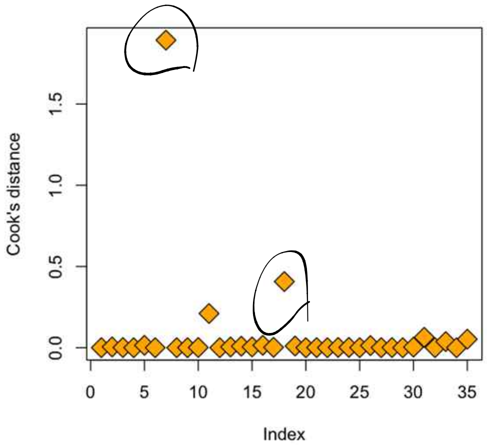
$$D_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{(p+1)\,\widehat{\sigma}^2}$$

Should be comparable to $F_{p+1,n-p-1}$: if the " p -value" of D_i is 50 percent or more, then the i -th case is likely influential: investigate further. (RABE)

Again, R has its own rules similar to the above for marking an observation as influential.

What to do after investigation? No easy answer.

plot(cooks.distance(races.lm), pch=23, bg='orange', cex=2, ylab="Cook's distance")




```
races.table[which(cooks.distance(races.lm) > 0.1),]
```

	Race	Distance	Climb	Time
7	BensofJura	16	7500	204.617
11	LairigGhru	28	2100	192.667
18	KnockHill	3	350	78.650

Differences between the betas (DFBETAS)

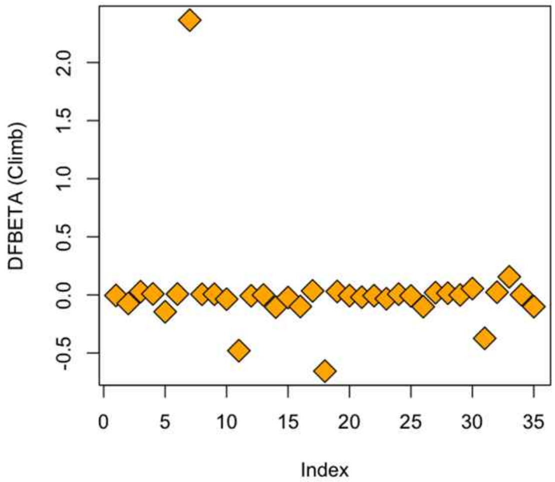
- This quantity measures how much the coefficients change when the i -th case is deleted.

$$DFBETAS_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)}(X^TX)^{-1}_{jj}}}.$$

For small/medium datasets: absolute value of 1 or greater is "suspicious". For large dataset: absolute value of $2/\sqrt{n}$.

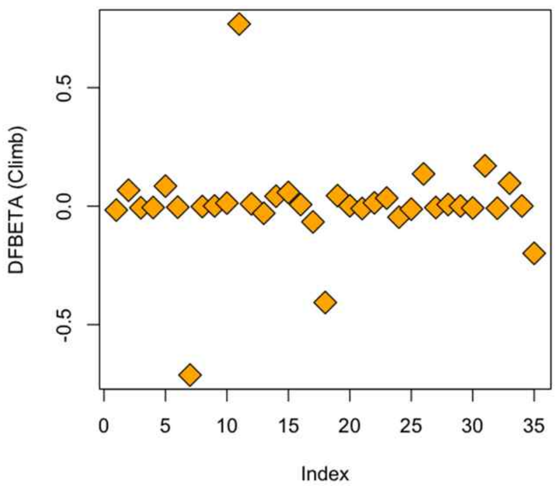
```
plot(dfbetas(races.lm)[,'Climb'], pch=23, bg='orange', cex=2, ylab="DFBETA (Climb)")
races.table[which(abs(dfbetas(races.lm)[,'Climb']) > 1),]
```

	Race	Distance	Climb	Time
7	BensofJura	16	7500	204.617




```
plot(dfbetas(races.lm)[,'Distance'], pch=23, bg='orange', cex=2, ylab="DFBETA (Climb)")
races.table[which(abs(dfbetas(races.lm)[,'Distance']) > 0.5),]
```

	Race	Distance	Climb	Time
7	BensofJura	16	7500	204.617
11	LairigGhru	28	2100	192.667



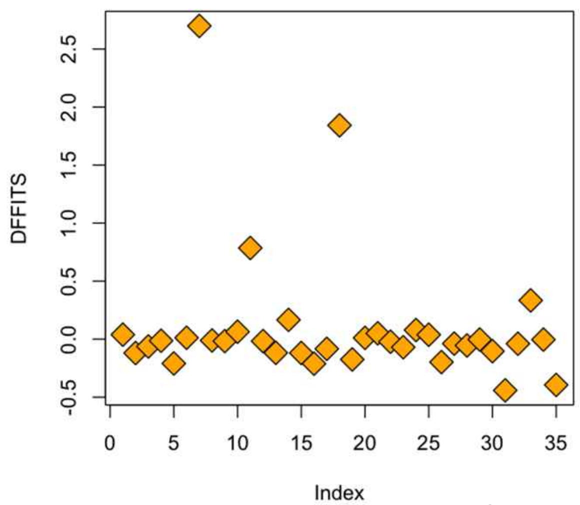
Differences between the fits (DFFITS)

$$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{H_{ii}}}$$

This quantity measures how much the regression function changes at the i -th case / observation when the i -th case / observation is deleted.

For small/medium datasets: value of 1 or greater is “suspicious” (RABE). For large dataset: value of $2\sqrt{(p+1)/n}$.

R has its own standard rules similar to the above for marking an observation as influential.



It seems that some observations had a high influence measured by DFFITS :

Not surprisingly, our longest and highest courses show up again. This at least reassures us that the leverage is capturing some of this "outlying in XX space".

Outlying X values

- One way to detect outliers in the *predictors*, besides just looking at the actual values themselves, is through their leverage values, defined by

$$\text{leverage}_i = H_{ii} = (X(X^TX)^{-1}X^T)_{ii}.$$

$\beta = (X^TX)^{-1}X^TY$

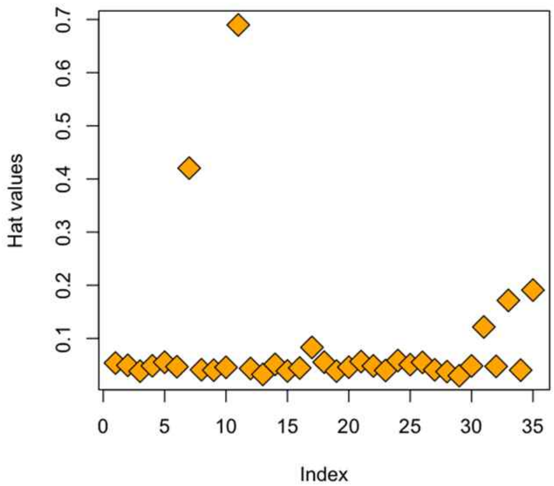
$\hat{Y} = X(X^TX)^{-1}X^TY$

$= H$

This at least reassures us that the leverage is capturing some of this "outlying in X space".

```
plot(hatvalues(races.lm), pch=23, bg='orange', cex=2, ylab='Hat values')
races.table[which(hatvalues(races.lm) > 0.3),]
```

	Race	Distance	Climb	Time
7	BensofJura	16	7500	204.617
11	LairigGhru	28	2100	192.667



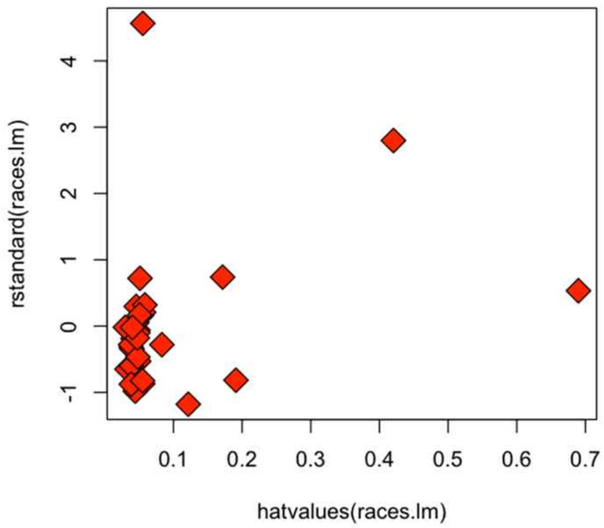
Outliers in the response

- Since rstudent are t distributed, we could just compare them to the T distribution and reject if their absolute value is too large
- Doing this for every observation results in n different hypothesis tests.
- This causes a problem: if n is large, if we “threshold” at $t(1-\alpha/2, n-p-2)$ we will get many outliers by chance even if model is correct.
- In fact, we expect to see $n \cdot \alpha$ “outliers” by this test. Every large data set would have outliers in it, even if model was entirely correct!

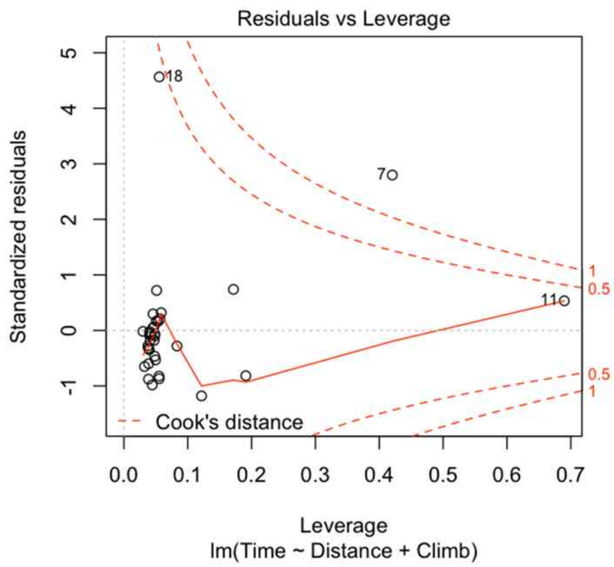
Final plot

- The last plot that R produces is a plot of residuals against leverage. Points that have high leverage and large residuals are particularly influential.

- plot(hatvalues(races.lm), rstandard(races.lm), pch=23, bg='red', cex=2)



- plot(races.lm, which=5)



Influence measures

```
influence.measures(races.lm)
```

Influence measures of
lm(formula = Time ~ Distance + Climb, data = races.table) :

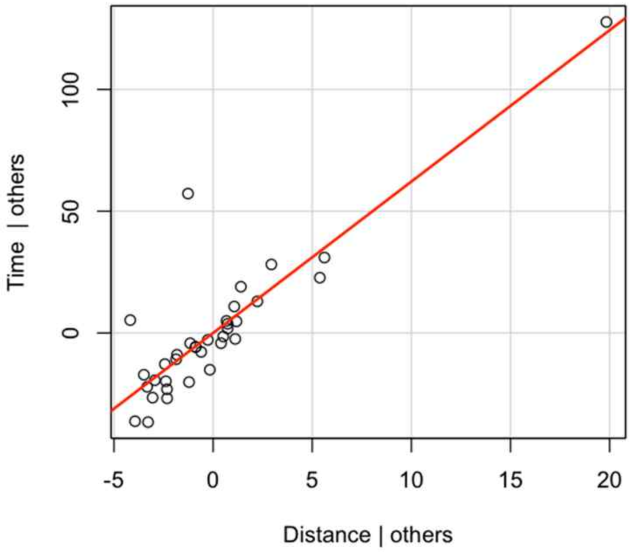
	dfb.1_	dfb.Dstn	dfb.Clmb	dffit	cov.r	cook.d	hat	inf
1	0.03781	-0.016614	-0.004744	0.03862	1.1595	5.13e-04	0.0538	
2	-0.05958	0.067215	-0.073396	-0.11956	1.1269	4.88e-03	0.0495	
3	-0.04858	-0.006707	0.028033	-0.06310	1.1329	1.37e-03	0.0384	
4	-0.00766	-0.005675	0.008764	-0.01367	1.1556	6.43e-05	0.0485	
5	-0.05046	0.084709	-0.145005	-0.20947	1.0837	1.47e-02	0.0553	
6	0.00348	-0.004316	0.007576	0.01221	1.1536	5.13e-05	0.0468	
7	-0.89065	-0.712774	2.364618	2.69909	0.8178	1.89e+00	0.4204	*
8	-0.00844	-0.001648	0.005562	-0.01115	1.1467	4.28e-05	0.0410	
9	-0.01437	0.000913	0.006161	-0.01663	1.1453	9.52e-05	0.0403	
10	0.04703	0.013057	-0.036519	0.06399	1.1431	1.41e-03	0.0457	
11	-0.30118	0.768716	-0.479849	0.78569	3.4525	2.11e-01	0.6898	*
12	-0.01149	0.009656	-0.007488	-0.01672	1.1492	9.61e-05	0.0435	
13	-0.03173	-0.029911	-0.000707	-0.11770	1.0922	4.70e-03	0.0323	
14	0.11803	0.042034	-0.104884	0.16610	1.1039	9.34e-03	0.0513	
15	-0.10038	0.057701	-0.022317	-0.11920	1.1062	4.83e-03	0.0388	
16	-0.01852	0.006789	-0.099862	-0.21135	1.0501	1.49e-02	0.0444	

Added variable plot

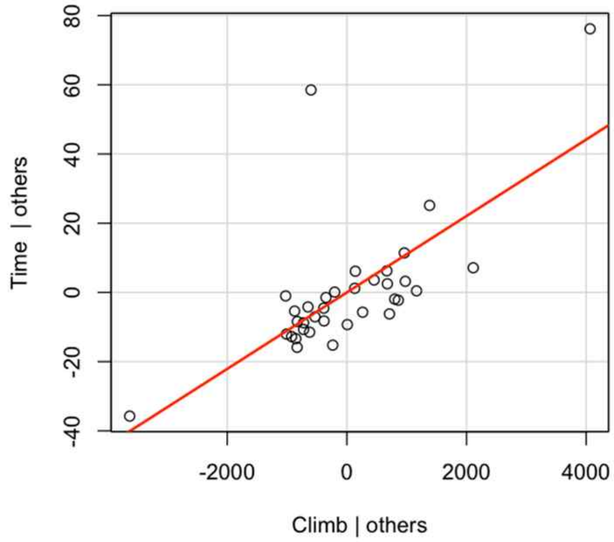
- The plots can be helpful for finding influential points, outliers. The functions can be found in the car package.

Let $\bar{e}_{X_j,i}, 1 \le i \le n$ be the residuals after regressing X_j onto all columns of X except X_j;
Let $e_{X_j,i}$ be the residuals after regressing Y onto all columns of X except X_j;
Plot $\bar{e}_{X_j}$ against e_{X_j}.
If the (partial regression) relationship is linear this plot should look linear.

avPlots(races.lm, 'Distance')



avPlots(races.lm, 'Climb')



Covariance ratio (COVRATIO) ~> de

- The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation:
- COVRATIO = $[(\det(s^2(i)\,(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}))/(\det(s^2\,(\mathbf{X}'\mathbf{X})^{-1}))]$

$$|\text{COVRATIO} - 1| \geq \frac{3p}{n}$$

where p is the number of parameters in the model and n is the number of observations used to fit the model, are worth investigation.

- P153쪽까지 적고 ... 등분산, 선형 or not, 정규성
- QQplot
- box-cox transformation
- X_i의 상관계수를 보고 다중공선성을 미리 check 할 수 있다.
  → Variance Influence factor (여기 확인(?))
- 숙제: $\boxed{5.3번, 5.5번, 5.7번}$, 금요일 실습수업 ~> 5장 전체
- 다음주 6장 변수선택부터 시작

Covariance ratio (COVRATIO)

- The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation:
- COVRATIO = [(det ($s^2(i)$ $(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}$))/(det (s^2 $(\mathbf{X}'\mathbf{X})^{-1}$))]

$$|\mathrm{COVRATIO} - 1| \geq \frac{3p}{n}$$

where p is the number of parameters in the model and n is the number of observations used to fit the model, are worth investigation.

☆ 이번주 금요일(?) ⇒ 지금까지 배운거 review하기.

☆ logistic regression ⇒ 가장 많이 쓰이는 분포

☆ weighted least square ⇒ 등분산이 아닐 때 사용 ⇒ 이런게 있다는 정도로 알아두기.

⇒ 1장 Ice cream data로 실습하면된다.

⇒ 보고서 과정: 개념+ appendix로 붙여버리기