

## Overview

- In Topic 5, we discussed diagnostics that can be used to assess whether there are any gross violations of assumptions
- The assumptions are
  - 1 Model is correct (not missing important predictors)
  - 2 Independent observations (often difficult to test)
  - 3 Linearity
  - 4 Errors normally distributed
  - 5 Constant variance

$$\begin{aligned} Y_i &= \hat{\mu}_{Y|X_i} + (Y_i - \hat{\mu}_{Y|X_i}) \\ Y_i &= \hat{Y}_i + \hat{E}_i \\ \text{observed} &= \text{predicted} + \text{residual} \end{aligned}$$

- Assessment of constant variance, normality, and linearity are all based on analyzing the residuals
- Will now discuss other types of residuals that are used in this assessment process.

## Regression Diagnostics II

Applied Regression and Other Multivariable Methods  
Sections 12-1 - 12-4

14

14-1

## Types of Residuals

- Define the *residual* to be  $e_i = Y_i - \hat{Y}_i$  (drop  $\hat{E}_i$  notation)
- Properties of the residuals
  - $\bar{e} = 0$ . Recall that this is due to the regression line going through  $(\bar{X}, \bar{Y})$
  - $\sum e_i^2 = \text{SSE}$ .
  - The  $\{e_i\}$  are not independent (i.e.,  $\bar{e} = 0$ ). When  $n$  is large relative to  $k$ , this property, however, can be ignored.
- Define the *standardized residual* to be

$$z_i = \frac{e_i - 0}{\sqrt{\text{MSE}}}$$

This residual also has a mean of zero but a variance of one so it is easier to use empirical rules (i.e., 95% of standardized residuals should be between  $\pm 2$ ).

14-2

## Types of Residuals

- Define the *studentized residual* to be

$$r_i = \frac{e_i}{\sqrt{1 - h_i} \sqrt{\text{MSE}}} = \frac{z_i}{\sqrt{1 - h_i}}$$

This residual also has a mean of zero but follows a  $t$  distribution with df  $n - k - 1$  if the underlying assumptions are true.

- The quantity  $h_i$  is the **leverage** of observation  $i$ 
  - Measures the importance of obs  $i$  in determining model fit
  - Is the geometric distance of point  $\{X_1, X_2, \dots, X_k\}$  from center  $\{\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k\}$
  - Varies between zero and one
- Define the *jackknife residual* to be

$$r_{(-i)} = r_i \sqrt{\frac{\text{MSE}}{\text{MSE}_{(-i)}}} = r_i \sqrt{\frac{n - k - 2}{n - k - 1 - r_i^2}}$$

The quantity  $\text{MSE}_{(-i)}$  is the residual variance with the  $i$ th obs deleted. If assumptions met, each  $r_{(-i)}$  follows a  $t$  with  $n - k - 1 - 1$  df.

14-3

## Studentized and Jackknife Residuals

- All residuals look similar when assumptions met
- Only  $e_i$  residuals sensitive to change in units
- When problems develop (especially outliers), suspect values look more obvious with studentized/jackknife residuals
  - If  $e_i$  large,  $MSE_{(-i)}$  will be smaller
  - Observations with large  $h_i$  will stand out more
- Can do usual graphical and numerical summaries with these residuals
- Table A-8A and A-8B provide critical values for jackknife and studentized residuals respectively (tells you obs is unusual, not that you should remove)
- Table A-9 provides crit values for the leverages,  $h_i$ 's
- **Cook's Distance** measures the extent to which the regression coefs change when the  $i$ th obs is deleted. This is a function of  $h_i$  and  $r_i$ . Table A-10 presents critical values for this summary.
- **PRESS** (predicted residual SS) is the sum of the squared jackknife residuals. Can compare this to the observed SSE.

14-4

## Diagnostics

- Normality
  - Histogram/Boxplot of residuals
  - Normal probability plot / QQ plot
  - Shapiro-Wilks/Kolmogorov-Smirnov Test
- Variance
  - Plot residuals vs  $\hat{Y}_i$  (residual plot)
- Independence
  - Plot residuals vs run order or distance
  - Runs test/Durbin-Watson Test
- Outliers
  - Is it influential? With and without analysis
  - Formal tests (e.g. studentized residuals)
  - Formal tests (e.g. Cook's distance, leverage)

14-5

## Example

Consider Problem 19 of Chapter 12. We are looking at asthmatic individuals and trying to describe FEV in terms of AGE, HGT, and WGT.

The following SAS commands were used to create the following output.

```
options nocenter ls=75;

data table8;
  infile 'I:\www\datasets502\EX1219.DAT' firstobs=2 dlm='09'x;
  input sub age sex $ hgt wgt fev;

proc reg corr;
  model fev = hgt wgt age / r influence;
  plot rstudent.*p. / nostat nomodel;
  plot rstudent.*nqq. / nostat nomodel;
  plot h.*p. / nostat nomodel;
  plot cooks.d.*obs. / nostat nomodel;
  output out=new1 p=pred rstudent=res;

proc univariate;
  var res;
run;
quit;
```

14-6

Correlation				
Variable	hgt	wgt	age	fev
hgt	1.0000	0.1291	0.2530	0.2257
wgt	0.1291	1.0000	-0.1996	0.1807
age	0.2530	-0.1996	1.0000	-0.0552
fev	0.2257	0.1807	-0.0552	1.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.11738	0.37246	0.44	0.7275
Error	15	12.68789	0.84586		
Corrected Total	18	13.80526			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-1.93683	5.72267	-0.34	0.7397
hgt	1	0.03015	0.03410	0.88	0.3906
wgt	1	0.01118	0.02151	0.52	0.6110
age	1	-0.01264	0.03840	-0.33	0.7465

The parameter t-tests are similar to variables-added-last  $F$  tests. In this case, none of the predictors appear to help describe FEV. Height has the strongest correlation and this is only 0.2257.

14-7

Obs	Dep Var	Predicted Value	Output Statistics		Std Error Residual	Student Residual
			Mean	Predict		
1	4.7000	3.9074	0.2917	0.7926	0.872	0.909
2	4.3000	3.5490	0.3615	0.7510	0.846	0.888
3	3.5000	3.9598	0.5874	-0.4598	0.708	-0.650
4	4.0000	3.4837	0.3952	0.5163	0.830	0.622
5	3.2000	3.4655	0.3896	-0.2655	0.833	-0.319
6	4.7000	3.8232	0.4339	0.8768	0.811	1.081
7	4.3000	4.2548	0.4611	0.0452	0.796	0.0568
8	4.7000	3.7299	0.2546	0.9701	0.884	1.098
9	5.2000	4.0684	0.4424	1.1316	0.806	1.403
10	4.2000	4.2325	0.4613	-0.0325	0.796	-0.0408
11	3.5000	3.9491	0.3451	-0.4491	0.853	-0.527
12	3.2000	3.4489	0.4542	-0.2489	0.800	-0.311
13	2.6000	3.8326	0.3136	-1.2326	0.865	-1.426
14	2.0000	3.8244	0.3418	-1.8244	0.854	-2.137
15	4.0000	3.8974	0.2900	0.1026	0.873	0.118
16	3.9000	3.5540	0.4485	0.3460	0.803	0.431
17	3.0000	3.8060	0.5010	-0.8060	0.771	-1.045
18	4.5000	3.5968	0.6427	0.9032	0.658	1.373
19	2.4000	3.5166	0.3881	-1.1166	0.834	-1.339

The standard error of a residual  $e_i$  is  $\sqrt{MSE(1-h_i)}$ . Thus the studentized residuals (last column) are the residuals divided by this quantity (second to last column). Only one studentized residual is outside  $\pm 2$  which is not unusual ( $\approx 5\%$  of the time outside  $\pm 2$ ). Using Table A-8B, the critical value is 2.73, so there doesn't appear to be any unusual observations.

Moments			
N	19	Sum Weights	19
Mean	-0.0072	Sum Observations	-0.1368001
Std Deviation	1.07663163	Variance	1.15913566
Skewness	-0.5915404	Kurtosis	-0.1125272
Uncorrected SS	20.8654268	Corrected SS	20.8644419
Coeff Variation	-14953.206	Std Error Mean	0.24699623

Basic Statistical Measures			
Location		Variability	
Mean	-0.00720	Std Deviation	1.07663
Median	0.05486	Variance	1.15914
Mode	.	Range	3.92958
		Interquartile Range	1.53981

Tests for Location: Mu0=0			
Test	-Statistic-	-----p Value-----	
Student's t	t -0.02915	Pr >  t	0.9771
Sign	M 0.5	Pr >=  M	1.0000
Signed Rank	S 5	Pr >=  S	0.8596

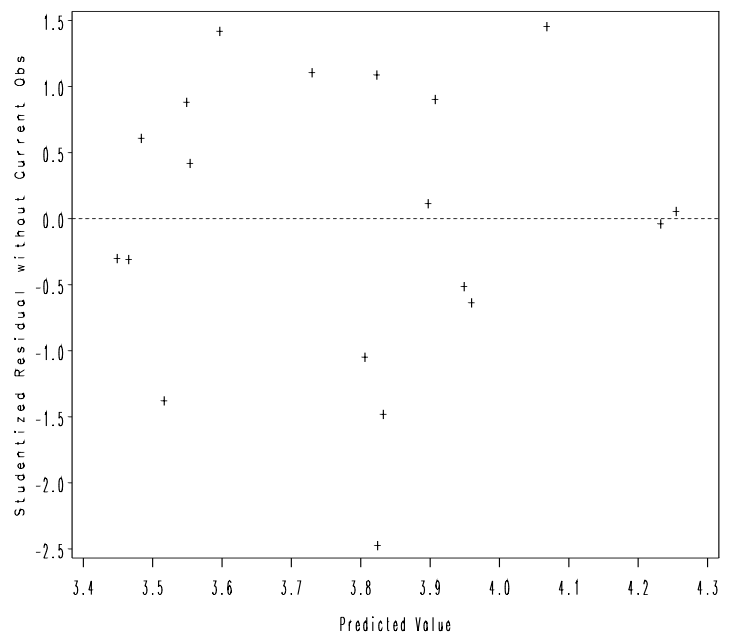
Nothing here jumps out in terms of the distribution of residuals. The following plots do not indicate any violations either. Observation 18 is pretty obvious. The last page of output has the analysis when this observation is removed.

Obs	Output Statistics					Hat	Diag	Cov	Ratio	DFFITS
	-2	-1	0	1	2					
1			*			0.023	0.9031	0.1006	1.1682	0.3020
2			*			0.036	0.8814	0.1545	1.2558	0.3768
3			*			0.073	-0.6367	0.4079	1.9855	-0.5285
4			*			0.022	0.6084	0.1846	1.4560	0.2895
5						0.006	-0.3089	0.1795	1.5630	-0.1445
6			**			0.084	1.0879	0.2226	1.2253	0.5822
7						0.000	0.0549	0.2514	1.7588	0.0318
8			**			0.025	1.1058	0.0766	1.0210	0.3186
9			**			0.148	1.4547	0.2314	0.9763	0.7981
10						0.000	-0.0394	0.2516	1.7601	-0.0229
11			*			0.011	-0.5137	0.1408	1.4234	-0.2079
12						0.008	-0.3017	0.2439	1.6982	-0.1713
13			**			0.067	-1.4813	0.1163	0.8329	-0.5373
14			****			0.183	-2.4749	0.1381	0.3581	-0.9906
15						0.000	0.1137	0.0994	1.4579	0.0378
16						0.014	0.4189	0.2378	1.6449	0.2339
17			**			0.115	-1.0485	0.2967	1.3849	-0.6811
18			**			0.450	1.4184	0.4883	1.5051	1.3856
19			**			0.097	-1.3788	0.1781	0.9635	-0.6418

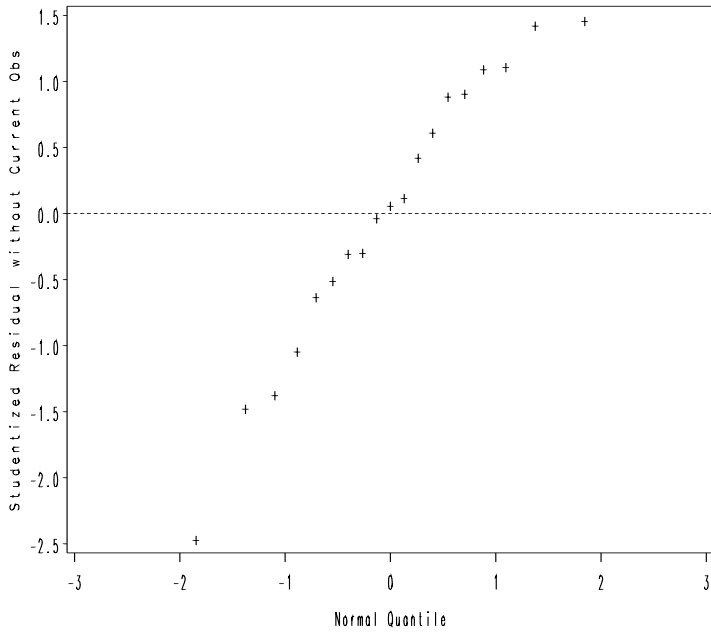
Rstudent is the jackknife residual. Using Table A-8A, the critical value is 3.62. The largest jackknife residual is -2.47 which is well below this critical level.

For the leverages, the critical value is 0.602 (Table A-9). The H column's largest value is .4883 which is below. For the Cook's D column, the critical value is  $15.49/(19-3)=1.03$  (Table A-10). Interestingly, it is obs 18 that has the highest leverage and Cook's D. This subject is the oldest by 10 years. Its leverage, however, is not large enough to expect a different relationship if it is removed.

Residual plot using Jackknife Residuals

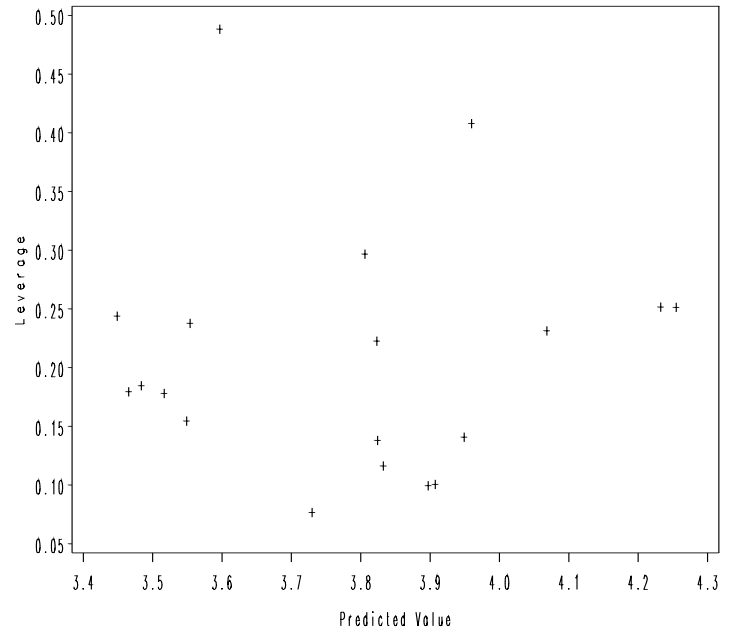


Normal QQplot of jackknife residuals



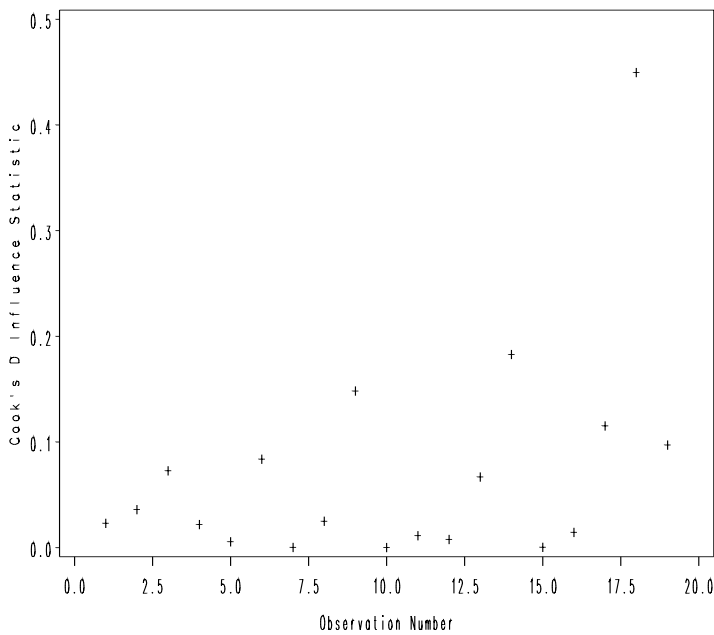
14-12

Plot of Leverages



14-13

Plot of Cook's D's



14-14

Correlation				
Variable	hgt	wgt	age	fev
hgt	1.0000	0.1506	0.2349	0.2075
wgt	0.1506	1.0000	-0.1298	0.2184
age	0.2349	-0.1298	1.0000	-0.2617
fev	0.2075	0.2184	-0.2617	1.0000

Analysis of Variance					
Sum of Mean					
Source	DF	Squares	Square	F Value	Pr > F
Model	3	2.17080	0.72360	0.91	0.4597
Error	14	11.09364	0.79240		
Corrected Total	17	13.26444			

Root MSE	0.89017	R-Square	0.1637
Dependent Mean	3.74444	Adj R-Sq	-0.0156
Coeff Var	23.77309		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-1.26437	5.55914	-0.23	0.8234
hgt	1	0.03331	0.03308	1.01	0.3310
wgt	1	0.01162	0.02083	0.56	0.5855
age	1	-0.05889	0.04944	-1.19	0.2534

The MSE is reduced and the parameter estimate for AGE becomes more negative. Age now has the highest correlation but it is still not statistically helpful in describing FEV

14-15