Regression model in practice

Week 3 assignment

Multiple regression model

I chose **allhealth** as my data set. The relationship between general health (H1GH1), weight (H1GH60) and height (H1GH59) were analyzed. General health is a categorical variable. It has five levels. From 1 to 5 represent excellent, very good, good, fair and poor. Therefore it can be used as the response variable. Both the weight and height are quantitative variables.

Theoretically, the larger weight, the less health a person is.

So a **regression model between general health and weight** was processed.

The GLM Procedure

Number of Observations Read	6504
Number of Observations Used	6347

The GLM Procedure
Dependent Variable: H1GH1 general health

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	95.159300	95.159300	119.93	<.0001
Error	6345	5034.436729	0.793449		
Corrected Total	6346	5129.596030			

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1.587452435	0.04753610	33.39	<.0001	1.494265622	1.680639247
H1GH60	0.003586634	0.00032751	10.95	<.0001	0.002944609	0.004228659

General health = 0.0036 * weight + 1.59

The result corresponds with the hypothesis. **P-value is less than 0.001**, therefore at 95% confidence level, the linear regression model is statistically significant.

To show if height is confounding to the correlation between general health and weight, a **multiple regression model** was processed.

The GLM Procedure

Number of Observations Read	6504
Number of Observations Used	6290

The GLM Procedure
Dependent Variable: H1GH1 general health

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	279.654916	139.827458	183.52	<.0001
Error	6287	4790.123145	0.761909		
Corrected Total	6289	5069.778060			

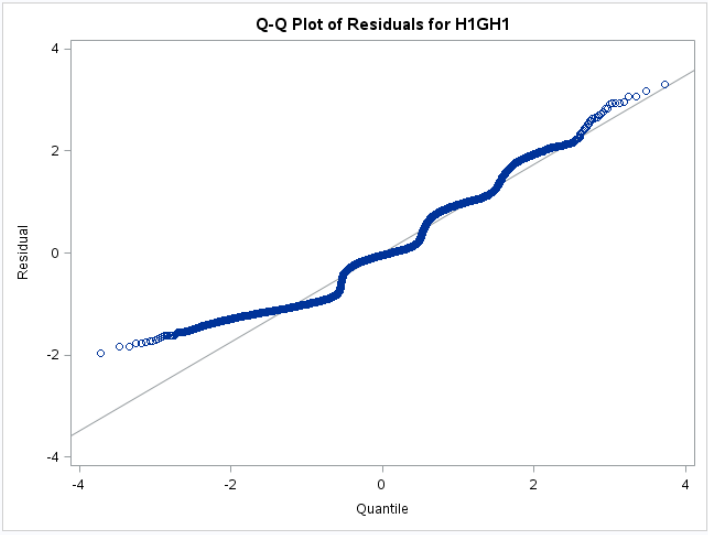
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	4.454094782	0.19069482	23.36	<.0001	4.080267837	4.827921728
H1GH60	0.007207489	0.00039665	18.17	<.0001	0.006429913	0.007985066
H1GH59	-0.050988189	0.00328260	-15.53	<.0001	-0.057423210	-0.044553168

General health = 0.0072 * weight – 0.05 * height + 4.45

After adding height to the regression model, the intercept and estimated parameter for weight changed, which **provides the evidence of confounding**. **P-value is less than 0.0001**, therefore at 95% confidence level, the correlation is statistically significant.

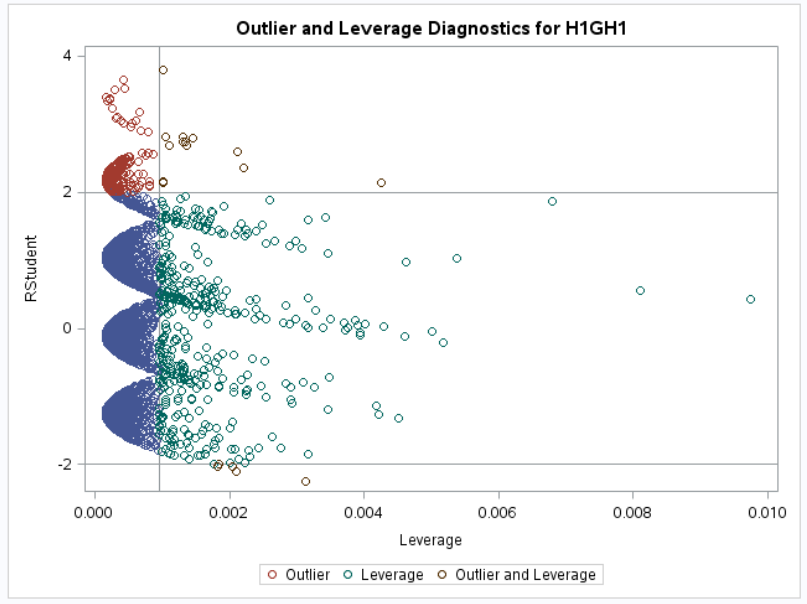
To evaluate the regression model, several diagnostic plots were obtained.

1. Q-Q plot



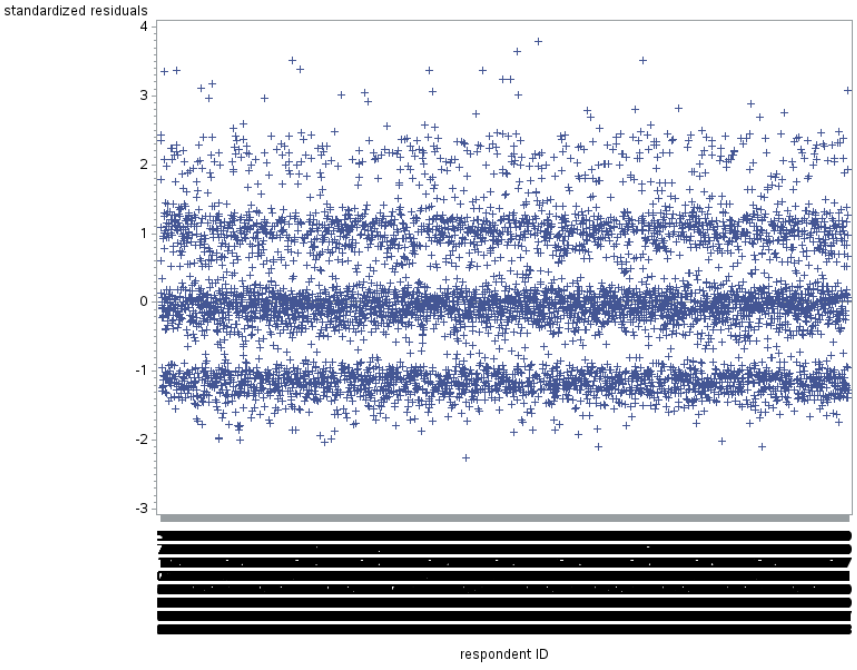
The residuals do not follow the straight line very well. So the residuals do not follow the normal distribution perfectly, showing the **poor fit** of this regression model.

2. Leverage plot



There are many outliers, but they have small leverage values, which means although they are outlying observations, they do not have significant effect on the regression model. However, there are many observations that are **both high leverage and outliers**, which shows the **poor fit** of this regression model.

3. Plot of standardized residuals for all observations



Most standardized residuals are centered between -1 and 1, but several of them are **larger than 2**, which show **poor fit** of this regression model.

Code:

```
*load data;
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
data new; set mydata.addhealth_pds;
*set aside missing values;
if H1GH1=6 then H1GH1=.; if H1GH1=8 then H1GH1=.;
if H1GH59A=96 then H1GH59A=.; if H1GH59A=98 then H1GH59A=.;
if H1GH59A=99 then H1GH59A=.;
if H1GH59B=96 then H1GH59B=.; if H1GH59B=98 then H1GH59B=.;
if H1GH59B=99 then H1GH59B=.;
if H1GH60=996 then H1GH60=.; if H1GH60=998 then H1GH60=.;
if H1GH60=999 then H1GH60=.;
*calculate the height;
H1GH59=H1GH59A * 12 + H1GH59B;
*add labels;
label AID="respondent ID"
      H1GH1="general health"
*sort by AID;
proc sort; by AID;

*relationship between general health and weight;

ods graphics on;
proc glm plots(maxpoints=none);
model H1GH1=H1GH60/solution clparm;
run;
ods graphics off;

*multiple regression;
ods graphics on;
proc glm plots(maxpoints=none);
model H1GH1=H1GH60 H1GH59/solution clparm;
run;
ods graphics off;

*diagnostic plot: Q-Q plot and leverage plot;
ods graphics on;
proc glm plots(unpack maxpoints=none)=all;
model H1GH1 = H1GH60 H1GH59/solution clparm;
output residual=res student=stdres out=results;
run;
ods graphics off;

*diagnostic plot: standardized residuals for all observations;
proc gplot;
label stdres="standardized residuals";
plot stdres*AID/vref=0;
run;
```