Abbie Pearson
June Suh
Jessica Warren
STA4203
October 5, 2016

Homework 6

1. Using the uswages data, fit a model with log(wage) as the response and educ and exper as predictors.

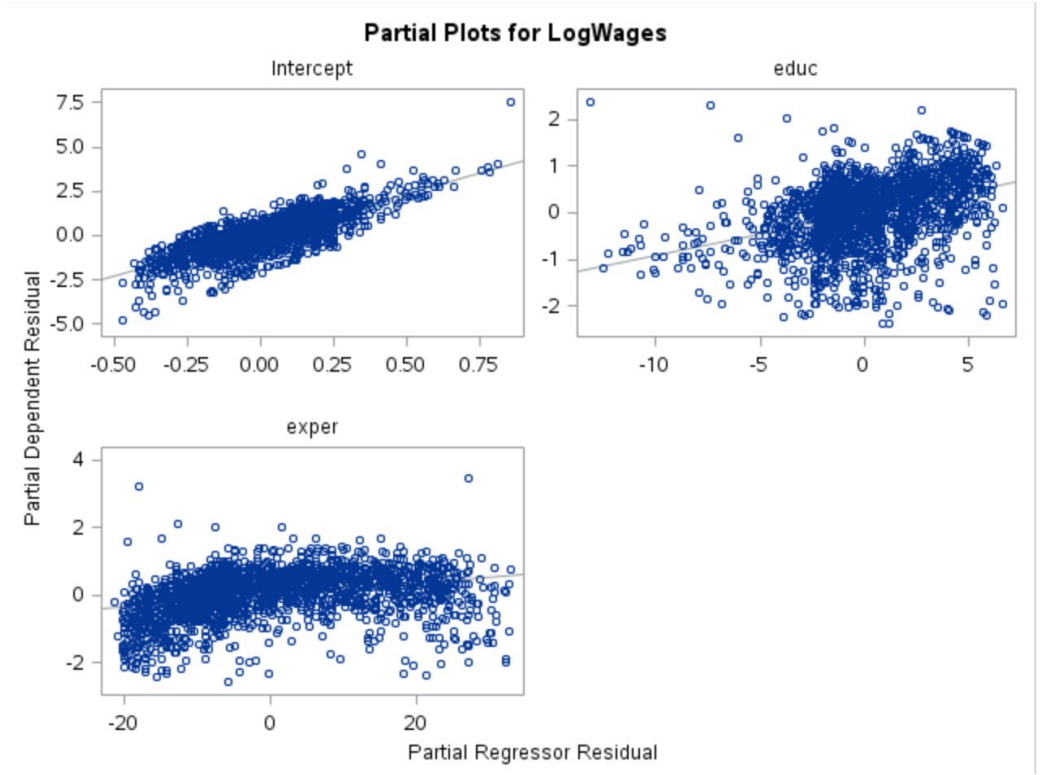
a) Draw the partial regression plots

Code:

```
proc import out=uswages
datafile="/home/jes13j0/Homework/STA 4203/uswages.csv"
dbms = csv replace;
run;

data uswages;
set uswages;
LogWages = log(wage);
run;

proc reg data=uswages;
model Logwages = educ exper / partial;
run;
```



b) Draw the partial residual plots

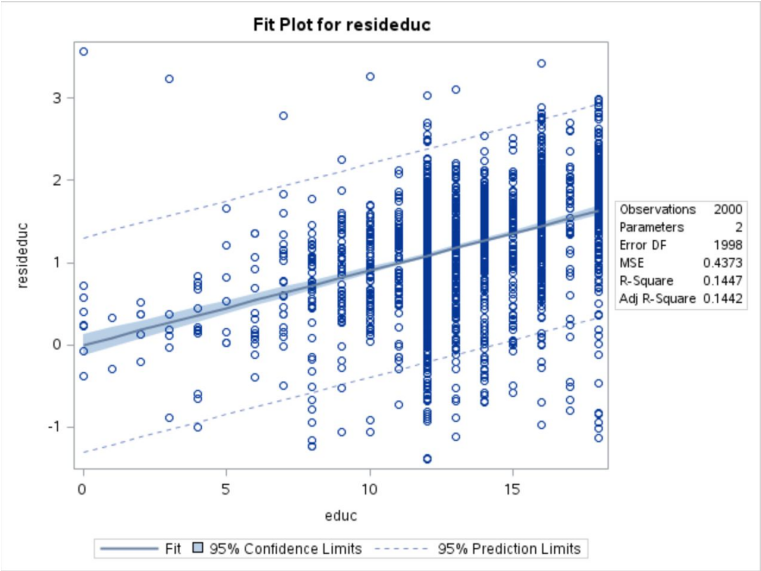
Code:

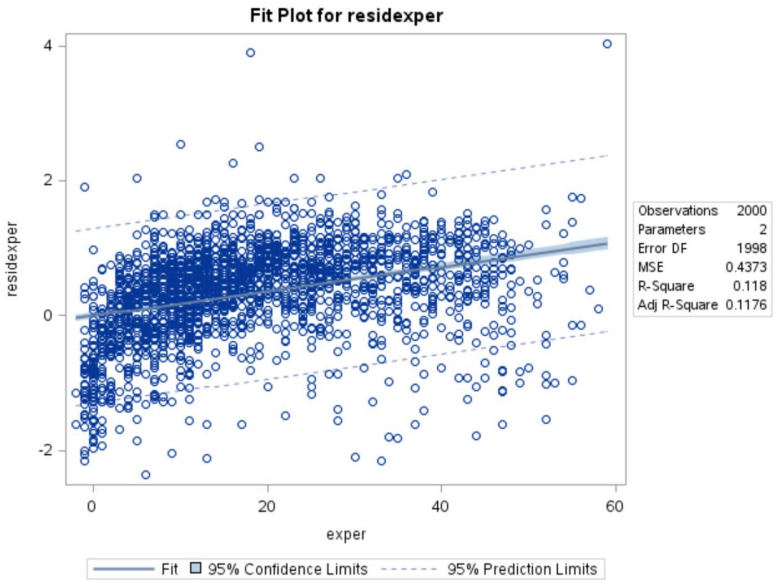
```
proc reg data=uswages;
model logwages=educ exper;
output out=new r=resid;
run;
Quit;

data new;
set new;
resideduc = resid+0.09051*educ;
residexper = resid+0.01808*exper;
run;

proc reg data=new;
model resideduc = educ;
model residexper = exper;
run;
quit;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.65032	0.07835	59.35	<.0001
educ	1	0.09051	0.00517	17.52	<.0001
exper	1	0.01808	0.00116	15.58	<.0001





c) Find the cutoff value for the outliers.

Code:

```
data quantiles;
  cuttoff=abs(tinv(0.05/(2*2000),2000-3-1));
Run;
```

cuttoff
4.2247261627

Cutoff value is 4.2247

d) Based on the cutoff, find the outliers and report how many outliers you found.

Code:

```
proc reg data=new;
model resideduc = educ;
model residexper = exper;
model logwages=educ exper;
output out=new1 rstudent=rez;
run;
quit;

data outliers;
do i=1 to 2000 by 1;
set new1 point=i;
if (abs(rez)>4.2247) then output;
end;
stop;
run;
```

Total rows: 2 Total columns: 16 Rows 1-2

esid	resideduc	residexper	rez
1367	3.5698591367	3.8952991367	5.4380861597
5482	3.234115482	4.029305482	4.5131568958

There are 2 outliers. This amount of outliers relative to the 2000 observations is not many this is probably because the data is clustered and outliers in a cluster model is hard to detect.

e) Remove the outliers, refit the model and report the R2 of the new model

Code:

```
data outliers;
set new1;
if (abs(rez)>4.2247) then delete;
run;

proc reg data=outliers;
model logwages=educ exper;
run;
```

Root MSE	0.65346	**R-Square**	0.1852
Dependent Mean	6.16720	**Adj R-Sq**	0.1844
Coeff Var	10.59570		

The new model's R2 value is 0.1852.

f) Recompute the cutoff for the model from e). Find if there are any outliers left and report how many outliers you found.

Code:

```
data quantiles1;
  cuttoff=abs(tinv(0.05/(2*1998),1998-3-1));
run;

proc reg data=outliers;
model logwages=educ exper;
output out=new2 rstudent=rez2;
run;

data outliers1;
do i=1 to 1998 by 1;
set new2 point=i;
if (abs(rez2)>4.2245) then output;
end;
stop;
Run;
```

cuttoff
4.2245087851

Total rows: 1 Total columns: 17 Rows 1-1

educ	residexper	rez	rez2
4227	-2.156614227	-4.186267829	-4.264170166

New cutoff is 4.2245. There is one outlier.

g) Remove the outliers again, refit the model and report the R2 of the new model

Code:

```
data outliers2;
set new2;
if (abs(rez2)>4.2245) then delete;
Run;

proc reg data=outliers2;
model logwages=educ exper;
Run;
```

Root MSE	0.65066	**R-Square**	0.1893
Dependent Mean	6.16822	**Adj R-Sq**	0.1885
Coeff Var	10.54861		

The R2 of the new model is 0.1893.