

Checking Regression Assumptions

1 Testing for Autocorrelation and Normality

This section we will learn some techniques to test for two of our regression assumptions. Recall that the four main simple linear regression assumptions are:

1. The residuals are centered around zero for all predicted values (Linearity). Can be checked using loess lines.
2. $Var(\epsilon_i) = \sigma^2$ ("Constant Error Variance"). Also can be check using loess lines.
3. ϵ_i are independent (or uncorrelated), $i = 1, 2, \dots n$.
4. ϵ_i are Normally Distributed.

Today we will go over ways to test and check the last two assumption's above. The forth assumption can be checked using **PROC UNIVARIATE**, with methods previously learned. While testing for autocorrelation can be tested for using **PROC ARIMA**.

Here is an example, which test's for all four:

```
proc glm data=tmp1.school2;
model gpahs = iq;
output out=dat r=ry p=py;
```

```
proc gplot;
plot ry*(py iq)/vref=0;
```

```
proc univariate normal;
var ry;
qqplot ry;
histogram ry/normal;
```

```
proc arima;
identify var=ry;
run;
quit;
```

Let's look at two examples which violate the certain assumptions:

```
data ali;
do i = 1 to 1000 by 1;
x1 = normal(0)*20 + 100;
x2 = normal(0)*20 + 100;
y = 14 + 4*x1 + 8*x2 + normal(0)*x1**2;
output;
end;
run;
```

```
proc glm data=ali;
model y = x1 x2;
output out=dat r=ry p=py;
```

```
proc gplot;
plot ry*(py x1)/vref=0;
```

```
proc univariate normal;
var ry;
qqplot ry;
histogram ry/normal;
```

```
proc arima;
identify var=ry;
run;
quit;
```

```
data ali;
do i = 1 to 1000 by 1;
x1 = normal(0)*20 + 100;
x2 = normal(0)*20 + 100;
y = 14 + 4*x1 + i*x2 + normal(0);
output;
end;
run;
```

```
proc glm data=ali;
model y = x1 x2;
output out=dat r=ry p=py;
```

```
proc gplot;
plot ry*(py x1)/vref=0;
```

```
proc univariate \textcolor[rgb]{1.00,0.00,0.00}{normal};
var ry;
histogram ry/normal;
qqplot ry;
```

```
proc arima;
identify var=ry;
run;
quit;
```

2 Calculating ICC with PROC MIXED

PROC MIXED is a procedure that we will use in detail in the upcoming semester. For today we'll only use it to find the ICC for the UIS data set on blackboard.

Recall that:

$$ICC = \frac{\tau}{\tau + \sigma^2} \quad (1)$$

Here's how we will calculate it:

```
data uis;
infile "Z:/uis.dat";
input ID AGE BECK IVHX NDRUGTX RACE TREAT SITE DRFREE;
run;

proc sort;
by site;
run;

PROC MIXED;
Class site;
Model beck = IVHX NDRUGTX RACE;
Random int /subject = site;
run;
quit;
```