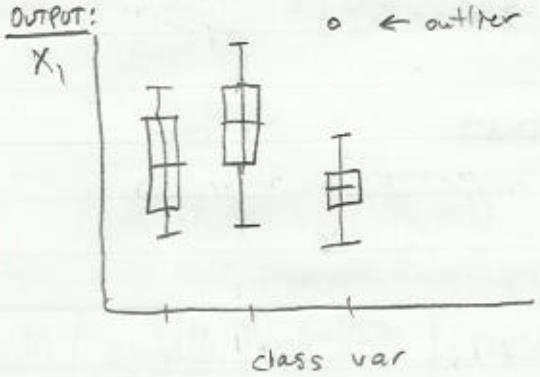
Chapter 3 Diagnostics

- Check for outlying predictor variables w/ boxplots

CODE: proc boxplot;
plot X_1* classVar / boxstyle = schematic; run;

OUTPUT:



- Residuals: • $e_i = Y - \hat{Y}_i$
  - $s^2(e_i) = \frac{\Sigma(e_i - \bar{e})}{n-2} = \frac{\Sigma e_i}{n-2} = \frac{SSE}{n-2} = MSE$
  - not independent Random Variables
- Studentized Residuals: $e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$

- We need to check for these problems
  1) The regression function isn't linear
  2) The error terms do not have constant variance
  3) The error terms are not independant
  4) The model fits all but one or a few outlier observations
  5) error terms not normally dist
  6) One or several important predictor variables have been ommited from the model

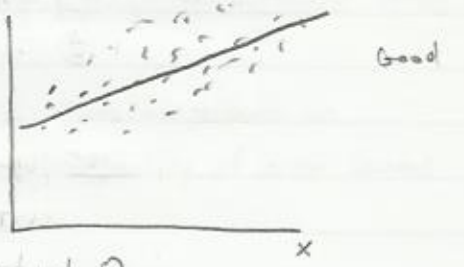
Problem One: Non-linearity of regression function

Check using:

① Scatterplot: look for a linear pattern

CODE
```
Proc sgscatter;
plot y * x / loess;
run;
```

output



② Residual Plot: look for no pattern: balanced around 0

CODE
```
Proc glm;
model y=x;
output out=glmData r=r;
run;
Proc sgscatter data=glmData;
  plot r * x;
run;
```

output

Problem Two Nonconstant variance of errors

① Residual Plot: there should be no megaphone shape

- see code above

② Absolute value of residuals Plot: There should be no pattern or triangle shape

CODE:
```
Proc Glm plots=all;
  model y=x;
  output out=glmData r=r;
  run;
data glmDataAbs;
  set glmData;
  absr = abs(r);
  run;
proc sgscatter data=glmDataAbs;
  plot absr * x;
  run;
```

output

Problem Three Presence of outliers

Check using:

① Box plot - see earlier code

② Studentized Residuals vs. predictor: look for points far from cloud for outlying observations

CODE
```
proc glm;
model y=x;
output out=glmdata rstudent=rs
run;
proc sgscatter data=glmData;
plot rs*x;
run;
```

output

Problem four Non independence of error terms: plot residuals in order of time taken, there should be no patterns

CODE
```
proc glm;
model y=x;
output out=glmData r=r;
run;
proc sgscatter;
plot r*sequence_variable;
run;
```

Problem five Non-normality of error terms

① Boxplot - see earlier code: should be semetric, no outliers

② QQ Plot - Given in glm diagnostics - should fit straight line

③ Histogram - Given in glm diagnostics - should look approx normal

If you don't like graphs there are tests! :)

Brown Forsythe Test - Tests for statistically ~~constant~~ variance across groups

CODE
```
proc glm;
class class_var;
model y = class_var;
means class_var / hovtest=BF;
run;
```

NOTE: Does not depend on normality of error terms

Output

Source	DF	Sum of Squares	Mean Square	F value	Pr > F
class_var				$\frac{MSC}{MSE}$	p value for constant var
error					

Breusch Pagan Test - Tests for constant variance

CODE;
```
proc model;
parms beta0 beta1;
y = beta0 + beta1*x;
fit y / breush=(1 x1);
run;
```

Note: Assumes normal & independent errors

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
	Breusch Pagan			test for constant variance	1, x

Normality tests

Shapiro-wilks: dependent on sample size
Kolmogorov-Smirnov: dependent on vertical distance of $F_x(x)$ and $F_{normal}(x)$
Cramer-von Mises; uses empirical distribution functions } Quadratic EDF
Anderson Darling: Based on square Difference

CODE
```
proc univariate Normal;
    var X;
run;
```

output

Test	Statistic		p value
Shapiro-Wilk	W		
Kolmogorov-Smirnov	D		
Cramer-von Mises	W-Sq		
Anderson Darling	A-Sq		

F test for Lack of Fit

- Test if model fits data well
- Assumes: independent & normal errors
  - constant variance

CODE
```
proc reg;
model y=x/lackfit;
run;
```

output

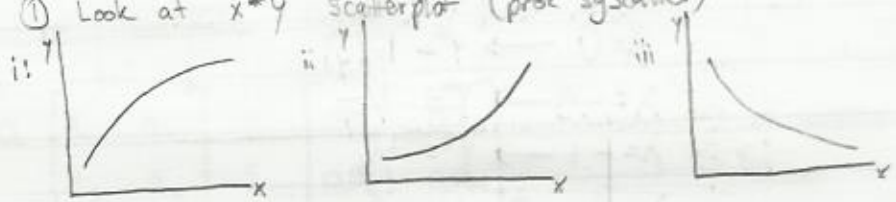
Source	DF	Sum of squares	Mean squares	F value	Pr>F
Model					All coeff=0 p value
Error					
Lack of fit					Test regression function is linear
Pure error					
Corrected total					

- How to fix Nonlinearity of regression function
  1. Add interaction terms
     i.e. $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon_i$
  2. Transform data to make it linear, or more linear
- How to fix nonconstancy of error variance
  1. Weighted least squares
  2. Transform data to make it more constant
- How to fix Nonindependence of error terms
  1. Use a model that allows this
- How to fix non normality of Error terms
  1. Transform data to make errors more normal
- How to fix outlying observations
  1. Disregard if they're not representative

Transformations

- To fix non linear relations; if error terms are approximately normal we want to transform X, not Y

1. Look at X*Y scatterplot (proc sgscatter)



2. Transform X

i: $X^* = \sqrt{X}$ or $X^* = \ln(X)$

ii: $X^* = X^2$ or $X^* = e^X$

iii: $X^* = \frac{1}{X}$ or $X^* = e^{-X}$