Lecture 13

Diagnostics

Diagnostics overview

- We will take the diagnostics and remedial measures that we learned for regression and adapt them to the ANOVA setting
- Many things are essentially the same
- Some things require modification

Residuals

- Predicted values are cell means, $\hat{Y}_{ij} = Y_{i\cdot}$.
- Residuals are the differences between the observed values and the cell means $Y_{ij}-Y_{i\cdot}$.

Basic plots

- Plot the data vs the factor levels (the values of the explanatory variables)
- Plot the residuals vs the factor levels
- Construct a normal quantile plot of the residuals

KNNL Example

- KNNL p 734
- Compare 4 brands of rust inhibitor (A has I=4 levels)
- Response variable is a measure of the effectiveness of the inhibitor
- There are 10 units per brand (J=10)

Data

```
data a1;
infile '../data/ch17ta02.txt';
   input  eff brand;
run;
```

Recode the factor

```
data a1; set a1;
   if brand eq 1 then abrand='A';
   if brand eq 2 then abrand='B';
   if brand eq 3 then abrand='C';
   if brand eq 4 then abrand='D';
 run;
```

Residuals to A2

```
proc glm data=a1;
   class abrand;
   model eff=abrand;
   output out=a2 r=resid;
run;
```

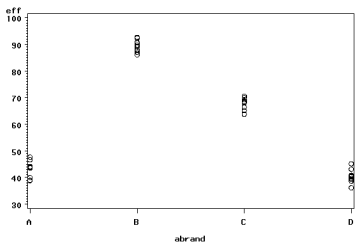
Plots

- Data versus the factor
- Residuals versus the factor
- Normal quantile plot fo the residuals

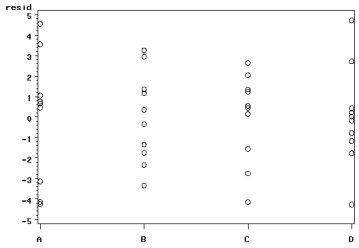
Plots vs the factor

```
symbol1 v=circle i=none;
proc gplot data=a2;
plot (eff resid)*abrand;
run;
```

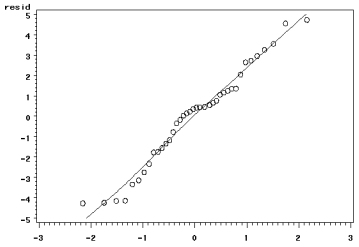
Data vs the factor



Residuals vs the factor



The plot



Homogeneity tests (1)

- Homogeneity of variance (homoscedasticity)
- H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2$
- H_1: not all σ_i^2 are equal
- Several significance tests are available

Homogeneity tests (2)

- SAS has several including Bartlett's (essentially the likelihood ratio test) and several versions of Levene

Homogeneity tests (3)

- There is a problem with assumptions
  - Anova is robust with respect to moderate deviations from normality
  - Anova results can be sensitive to the homogeneity of variance assumption
- Some homogeneity tests are sensitive to the normality assumption

Levene's Test

- Do anova on the squared residuals
- Modified Levene's test uses absolute values of the residuals
- Modified Levene is recommended

KNNL Example

- KNNL p 783
- Compare the strengths of 5 types of solder flux (A has I=5 levels)
- Response variable is the pull strength, force in pounds required to break the joint
- There are 8 solder joints per flux (J=8)

Levene's Test

```
proc glm data=a1;
   class type;
   model strength=type;
   means type/
    hovtest=levene(type=abs);
run;
```

Output

```
Levene's Test
ANOVA of Absolute Deviations

Source DF  F Value    Pr > F
type    4     3.07    0.0288
Error  35
```

Means and SDs

Level type	N	strength Mean	Std Dev
1	8	15.42	1.23
2	8	18.52	1.25
3	8	15.00	2.48
4	8	9.74	0.81
5	8	12.34	0.76

Remedies

- **Delete outliers**
- **Use weights**
- **Transformations**
- **Nonparametric procedures**

Weighted least squares

- **Here we can compute the variance for each level**
- **Use these as weights in PROC GLM**
- **We will illustrate with the soldering example from KNNL (p 783)**

Obtain the variances and weights

```
proc means data=a1;
   var strength;
   by type;
   output out=a2 var=s2;
data a2; set a2; wt=1/s2;
```

NOTE. Data set a2 has 5 cases

Merge and then use the weights in PROC GLM

```
data a3; merge a1 a2;
   by type;
proc glm data=a3;
   class type;
   model strength=type;
   weight wt;
run;
```

Output

```
Source     DF   F Value   Pr > F
Model       4     81.05   <.0001
Error      35
Total      39
```

Transformation Guides

- When σ_i^2 is proportional to μ_i, use $\sqrt{Y}$
- When σ_i is proportional to μ_i, use log(y)
- When σ_i is proportional to μ_i^2, use 1/y
- For proportions, use 2arcsin $\sqrt{Y}$
- (arsin(y) in a SAS data step)

Nonparametric approach

- Based on ranks
- SAS procedure NPAR1WAY