Lecture 9

Analysis of Variance

One-Way ANOVA

- The response variable Y is continuous
- The explanatory variable is categorical
  - We call it a factor
  - The possible values are called levels
- This is a generalization of the two-sample t-test

Data for one-way ANOVA

- Y, the response variable
- A, the factor
  - I is the number of levels
  - We sometimes refer to these as groups or treatments
- Y_{ij} is the j^{th} observation in the i^{th} group

Example

- Y is the number of cases of cereal sold
- A is the design of the cereal package
  - There are 4 levels for A because there are 4 different package designs
- i =1 to 4 levels
- j =1 to J_i stores with design i (5,5,4,5)
- Use J if it does not depend on i

Data for one-way ANOVA

```
cereal<-
read.table('ch16ta01.txt',
col.names=c("cases", "design",
           "store"));
```

The data

	cases	design	store
1	11	1	1
2	17	1	2
3	16	1	3
4	14	1	4
5	15	1	5
6	12	2	1
7	10	2	2

Notation

- For Y_{ij} we use
  - i to denote the level of the factor
  - j to denote the j^{th} observation at factor level i
- $i = 1, \ldots, I$ levels of factor A
- $j = 1, \ldots, J_i$ observations for level i of factor A

Model

- We assume that the response variable observations are
  - Normally distributed
  - With a mean that may depend on the level of the factor
  - And a variance that does not
  - Independent

Model (2)

- $Y_{ij} = \mu_i + \xi_{ij}$
  - where μ_i is the theoretical mean or expected value of all observations at level i and
  - the ξ_{ij} are iid $N(0, \sigma^2)$
  - $Y_{ij} \sim N(\mu_i, \sigma^2)$, independent
  - This is called the cell means model

Parameters

- The parameters of the model are
  - $\mu_1, \mu_2, \ldots, \mu_I$
  - σ^2

Question – Does our explanatory variable influence Y ? i.e.

Does μ_i depend on i ?

H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$

H_a: not all μ's are the same

Estimates

- Estimate μ_i by the mean of the observations at level i, $\overline{Y}_i$
- $\overline{Y}_i = (\Sigma Y_{ij})/(J_i)$
- For each level we can get an estimate of the variance
- $s_i^2 = (\Sigma(Y_{ij} - \overline{Y}_i)^2)/(J_i - 1)$
- We need to combine these to get an estimate of σ^2

Pooled estimate of σ^2

- If the J_i are all the same we would average the s_i^2
  - We would *not* average the s_i
- In general we pool the s_i^2, giving weights proportional to the df, J_i -1
- The pooled estimate is
- $s^2 = (\Sigma (J_i-1)s_i^2) / (\Sigma (J_i-1))$
- $= (\Sigma (J_i-1)s_i^2)/(n-I)$

Run proc glm

```
cereal$design=
factor(cereal$design)
obj<-aov(cases~design, cereal)
model.tables(obj, type="means")
```

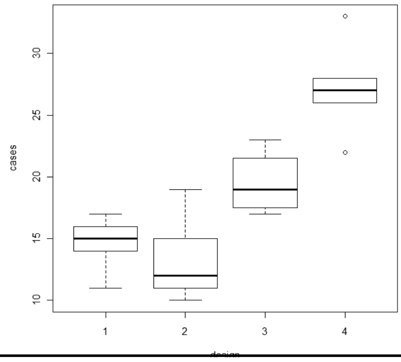
Output

```
design
        1    2    3    4
     14.6 13.4 19.5 27.2
rep   5.0  5.0  4.0  5.0
```

Plot the data

```
plot(cases~design, cereal)
```

The plot



Notation

- $Y_{i.} = (\Sigma_j Y_{ij}) / J_i$
- $Y_{..} = (\Sigma_{ij} Y_{ij}) / n$
- n is the total number of observations
- $n = \Sigma_j J_i$

ANOVA Table

Source	df	SS	MS
Model	I-1	$\Sigma_{ij}(Y_{i.}-Y_{..})^2$	SSM/dfM
Error	n-I	$\Sigma_{ij}(Y_{ij}-Y_{i.})^2$	SSE/dfE
Total	n-1	$\Sigma_{ij}(Y_{ij}-Y_{..})^2$	SST/dfT

Anova output

```
Summary(obj)
    Df  SS     MS    F      Pr(>F)
des  3 588.2 196.1 18.6 2.5e-05 *
Res 15 158.2  10.5
```

Expected Mean Squares

- $E(MSE) = \sigma^2$
- $E(MSM) = \sigma^2 + (\Sigma_i J_i (\mu_i - \mu_.)^2)/(I-1))$
  - where $\mu_. = (\Sigma_i J_i \mu_i)/n$

F test

- $F = MSM/MSE$
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
- H_1: not all of the μ_i are equal
- Under H_0, $F \sim F(I-1, n-I)$
- Reject H_0 when F is large
- Report the P-value

More output

```
obj2<-lm(cases~design, cereal)
summary(obj2)

Residual standard error: 3.248
on 15 degrees of freedom
Multiple R-squared: 0.7881,
Adjusted R-squared: 0.7457
F-statistic: 18.59 on 3 and 15
DF,  p-value: 2.585e-05
```

Factor Effects Model

- $Y_{ij} = \mu + \alpha_i + \xi_{ij}$
  - the ξ_{ij} are iid $N(0, \sigma^2)$

Parameters

- The parameters of the model are
  - $\mu, \alpha_1, \alpha_2, \ldots, \alpha_I$
  - σ^2

An example

- Suppose $I=3$; $\mu_1 = 10$, $\mu_2 = 20$, $\mu_3 = 30$
- What is an equivalent set of parameters for the factor effects model?
- We need to have $\mu + \alpha_i = \mu_i$
- $\mu = 0$, $\alpha_1 = 10$, $\alpha_2 = 20$, $\alpha_3 = 30$
- $\mu = 20$, $\alpha_1 = -10$, $\alpha_2 = 0$, $\alpha_3 = 10$
- $\mu = 5000$, $\alpha_1 = -4990$, $\alpha_2 = -4980$, $\alpha_3 = -4970$

Factor effects solution

- Put a constraint on the α_i
- $\Sigma_i\alpha_i = 0$
- This effectively reduces the number of parameters by 1

Consequences

$\mu_i = \mu + \alpha_i$

- The constraint $\Sigma_i\alpha_i = 0$ implies
- $\mu = (\Sigma_i \mu_i)/I$
- $\alpha_i = \mu_i - \mu$

Hypotheses

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
- H_1: not all of the μ_i are equal

are translated into

- H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$
- H_1: at least one α_i is not 0

Estimators of parameters

- With the constraint $\Sigma_i\alpha_i = 0$
  $\hat{\mu} = \frac{\sum Y_{i.}}{I}$
- $\hat{\alpha}_i = Y_{i.} - \hat{\mu}$

Regression Approach

- We can use multiple regression to reproduce the results based on the factor effects model
- $Y_{ij} = \mu + \alpha_i + \xi_{ij}$
- $\Sigma_i\alpha_i = 0$

Coding for Explanatory Variables

- X_{ij} = 1 if A is at level i
- = -1 if A is at level I
- = 0 if A is at any other level
- i = 1 to I-1

Means

```
model.tables(obj, type="means")
Grand mean
 18.63158
design
       1    2    3    4
    14.6 13.4 19.5 27.2
rep  5.0  5.0  4.0  5.0
```

The mean of the means

```
m<-model.tables(obj,type="means")
m<-m$tables$design
mean(m)
18.675
```

Explanatory variables

```
cereal$x1<-(cereal$design == 1)-
(cereal$design == 4);

cereal$x2<-(cereal$design == 2)-
(cereal$design == 4);

Cereal$x3<-(cereal$design == 3)-
(cereal$design == 4);
```

Output

	cases	des	x1	x2	x3
1	11	1	1	0	0
6	12	2	0	1	0
11	23	3	0	0	1
15	27	4	-1	-1	-1

Output with parameters

des	x1	x2	x3	
1	1	0	0	$\mu + \alpha_1$
2	0	1	0	$\mu + \alpha_2$
3	0	0	1	$\mu + \alpha_3$
4	-1	-1	-1	$\mu - \alpha_1 - \alpha_2 - \alpha_3$

Run the regression

```
obj3<-lm(cases~x1+x2+x3,
cereal);
summary(obj3)
```

Results

	Estimate	Std. Error	t value	Pr(>|t|)	
Int	18.6750	0.7485	24.949	1.25e-13	**
x1	-4.0750	1.2708	-3.207	0.005884	**
x2	-5.2750	1.2708	-4.151	0.000854	**
x3	0.8250	1.3706	0.602	0.556221	

- Residual standard error: 3.248 on 15 degrees of freedom
- Multiple R-squared: 0.7881, Adjusted R-squared: 0.7457
- F-statistic: 18.59 on 3 and 15 DF, p-value: 2.585e-05

Regression coefficients

Var	Est	
Int	18.675	mean of the means
x1	-4.075	$Y_1.$ - Int
x2	-5.275	$Y_2.$ - Int
x3	0.825	$Y_3.$ - Int

```
18.675-4.075 = 14.6
18.675-5.275 = 13.4
18.675+0.825 = 19.5
18.675+4.075+5.275-0.825=27.2
```

R coding for X

```
The rows are
1 1 0 0 0 for A=1 (5)
1 0 1 0 0 for A=2 (5)
1 0 0 1 0 for A=3 (4)
1 0 0 0 1 for A=4 (5)
```

- Recall, X'X does not have an inverse
- R eliminates the first variable (column) involved in the equation
- i. e. in R solution α_1=0

Some options

```
summary(obj2)

     Est  Std.  t     Pr(>|t|)
Int  14.6 1.45 10.05 4.66e-08 ***
des2 -1.2 2.05 -0.58 0.5677
des3  4.9 2.18  2.25 0.0399 *
des4 12.6 2.05  6.13 1.91e-05 ***
```

Interpretation

- If $\alpha_1 = 0$ then the corresponding estimate should be zero
- the intercept μ is estimated by the mean of the observations in group 1
- Since $\mu + \alpha_i$ is the mean of group i, the α_i are the differences between the mean of group i and the mean of group 1

Parameter estimates from means

Level of design	Mean	
		μ(hat) = 14.6
1	14.6	α_1(hat) = 14.6-14.6 = 0
2	13.4	α_2(hat) = 13.4-14.6 = -1.2
3	19.5	α_3(hat) = 19.5-14.6 = 4.9
4	27.2	α_4(hat) = 27.2-14.6 = 12.6

Confidence intervals for means

- $Y_i. \sim N(\mu_i, \sigma^2/J_i)$
- CI for μ_i is $Y_i. \pm t^* s/\sqrt{J_i}$
- t^* is computed from the t(n-I) distribution

Confidence intervals seprately in each class

```
t.test(cereal$cases
[cereal$design==1])
t.test(cereal$cases
[cereal$design==2])
t.test(cereal$cases
[cereal$design==3])
t.test(cereal$cases
[cereal$design==4])
```

Output

```
95% CI:11.74147 17.45853
mean of x :14.6
95% CI: 8.871755 17.928245
mean of x :13.4
95% CI: 15.29002 23.70998
mean of x :19.5
95% CI: 22.28013 32.11987
mean of x :27.2
```

Confidence intervals from anova

```
obj4<-lm(cases~design-1, cereal)
confint(obj4)
   fit      lwr      upr
1  14.6  11.50438 17.69562
6  13.4  10.30438 16.49562
11 19.5  16.03899 22.96101
15 27.2  24.10438 30.29562
```

Multiplicity Problem

- We have constructed 4 (in general, I) 95% confidence intervals
- The overall confidence level is less that 95%
- Many different kinds of adjustments have been proposed
- We have seen the Bonferroni (use α/I)

BONFERRONI

```
confint(obj4,  level=1-0.05/4)
```

Bonferroni CIs

```
     fit      lwr      upr
1  14.6 10.480212 18.71979
6  13.4  9.280212 17.51979
11 19.5 14.893937 24.10606
15 27.2 23.080212 31.31979
```

Hypothesis tests on individual means

- Not usually done
- Use t.test

Differences in means

- Distribution of $Y_{i.}$-$Y_{k.}$ is
- $N(\mu_i-\mu_k, (\sigma^2/J_i)+(\sigma^2/J_k))$
- CI for μ_i-μ_k is $Y_{i.}$-$Y_{k.}$ $\pm$ $t^*s(Y_{i.}-Y_{k.})$
- where $s(Y_{i.}-Y_{k.})$ =s($\sqrt{\frac{1}{J_i}+\frac{1}{J_k}}$)

t^*

- We deal with the multiplicity problem by adjusting t^*
- Many different choices are available

R uses Tukey

- Based on the studentized range distribution (max minus min divided by the standard deviation)
- $t^* = q^*/\sqrt{2}$

Example

```
TukeyHSD(obj)

    diff   lwr   upr     p adj
2-1 -1.2 -7.12  4.72 0.9352978
3-1  4.9 -1.38 11.18 0.1548895
4-1 12.6  6.68 18.52 0.0001013
3-2  6.1 -0.18 12.38 0.0582866
4-2 13.8  7.88 19.72 0.0000368
4-3  7.7  1.42 13.98 0.0142180
```

Linear Combinations of Means

- These combinations should come from research questions, not from an examination of the data
- $L = \Sigma_i w_i \mu_i$
  $\hat{L} = \Sigma_i w_i Y_{i\cdot} \sim N(L, Var(\hat{L}))$
- $Var(\hat{L}) = \Sigma_i w_i^2 Var(Y_{i\cdot})$
- Estimated by $s^2 \Sigma_i w_i^2 / J_i$

Quantitative factors

- Factor A is a quantitative variable
- Regression is a possible alternative analytical approach
- We can compare models, e.g. linear with anova; linear plus quadratic versus anova, etc.

Quantitative factors (2)

- Extra SS principle applies here
- We use the factor first as a continuous explanatory variable (regression) then as a categorical explanatory variable (anova)
- This is a test for linearity

Example

- Y is the number of acceptable units produced
- A is the number of hours of training
  - There are 4 levels for A : 6 hrs, 8 hrs, 10 hrs and 12 hrs
- i =1 to 4 levels (I=4)
- j =1 to 7 employee at each training level (J=7)

```
data2<-read.table('ch17ta04.txt',
col.names=c("product",
"trainhrs", "ind"));
data2$hrs<-2*data2$trainhrs+4;
obj<-lm(product~hrs, data2)
data2$trainhrs<-
factor(data2$trainhrs)
obj1<-lm(product~trainhrs, data2)
summary(obj1)
anova(obj,obj1)
```

Conclusion

- F-statistic: 141.5 on 3 and 24 DF,
- p-value: 2.173e-15
- Hours of training relates to product produced

Output

```
Model 1: product ~ hrs
Model 2: product ~ trainhrs
  Df RSS     Df  SS F   Pr(>F)
1 26 146.61
2 24 102.29 2 44.33 5.2 0.013 *
```

Interpretation

- **The analysis indicates that there is statistically significant lack of fit for the linear regression model (F=5.20; df=2,24; P=0.0133)**
- **Let's try a quadratic**

Quadratic Model

```
data2$hrs2<-data2$hrs^2
obj2<-lm(product~hrs+hrs2, data2)
anova(obj,obj2)
anova(obj2, obj1)
```

Output

```
Model 1: product ~ hrs
Model 2: product ~ hrs + hrs2
  Df  RSS  Df  SS     F Pr(>F)
1 26 146.61
2 25 102.86 1 43.75 10.63 0.003 *
```

Model 1: product ~ hrs + hrs2
Model 2: product ~ trainhrs

	Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	25	102.86				
2	24	102.29	1	0.57857	0.1358	0.7158

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

Example

- 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

```
rust<-read.table('ch17ta02.txt',
col.names=c("eff”,"brand","ind"));
rust$abrand=mat.or.vec(40,1)
rust$abrand[rust$brand==1]="A"
rust$abrand[rust$brand==2]="B"
rust$abrand[rust$brand==3]="C"
rust$abrand[rust$brand==4]="D"
```

Residuals to A2

```
rust$abrand=factor(rust$abrand)
obj1<-lm(eff~abrand, rust)
r<-residuals(obj1)
```

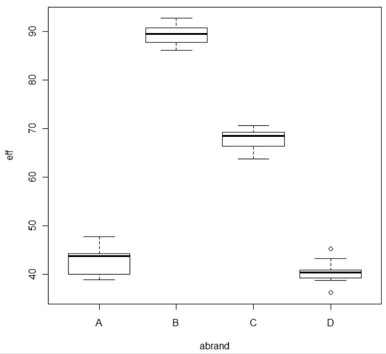
Plots

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

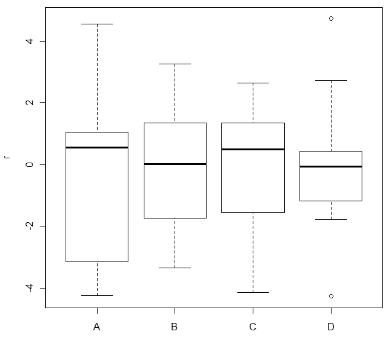
Plots vs the factor

```
plot(eff~abrand,rust)
plot(r~abrand,rust)
qqnorm(r)
```

Data vs the factor



Residuals vs the factor



The plot