Linear Models Recap

So far we have considered the following types of linear models:

Simple regression (continuous X values):

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Linear Models Recap

So far we have considered the following types of linear models:

Simple regression (continuous X values):

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

One-way (single factor) ANOVA (X values are zeros and ones)

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + e_i$$

Linear Models Recap

So far we have considered the following types of linear models:

Simple regression (continuous X values):

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

One-way (single factor) ANOVA (X values are zeros and ones)

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + e_i$$

Two-way (two factor) ANOVA without interaction (2 factors; X values are zeros and ones)

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \beta_1 X_{i3} + \beta_2 X_{i4} + e_i$$

Linear Models Recap

Today we will extend the list to include a variation on the two factor ANOVA by including what is known as an interaction term in the model.

Linear Models Recap

Today we will extend the list to include a variation on the two factor ANOVA by including what is known as an interaction term in the model.

To see why such a model may be necessary, consider the parameters for the two way ANOVA without interaction:

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + e_i$$

We'll assume as before that factor one (α) represents "city or highway" and factor 2 represents "car or truck"

Linear Models Recap

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + e_i$$

Since e_i is assumed to have a population mean of zero, the expected values for various categories are:

- car, city: $\mu + \alpha_1 + \beta_1$
- car, highway: $\mu + \alpha_2 + \beta_1$
- truck, city: $\mu + \alpha_1 + \beta_2$
- truck, highway: $\mu + \alpha_2 + \beta_2$

Linear Models Recap

- car, city: $\mu + \alpha_1 + \beta_1$
- car, highway: $\mu + \alpha_2 + \beta_1$
- truck, city: $\mu + \alpha_1 + \beta_2$
- truck, highway: $\mu + \alpha_2 + \beta_2$

With these parameters, for cars the expected difference in mileage between city and highway driving is:

$$(\mu + \alpha_1 + \beta_1) - (\mu + \alpha_2 + \beta_1) = \alpha_1 - \alpha_2$$

Linear Models Recap

- car, city: $\mu + \alpha_1 + \beta_1$
- car, highway: $\mu + \alpha_2 + \beta_1$
- truck, city: $\mu + \alpha_1 + \beta_2$
- truck, highway: $\mu + \alpha_2 + \beta_2$

With these parameters, for cars the expected difference in mileage between city and highway driving is:

$$(\mu + \alpha_1 + \beta_1) - (\mu + \alpha_2 + \beta_1) = \alpha_1 - \alpha_2$$

For cars the expected difference in mileage between city and highway driving is:

$$(\mu + \alpha_1 + \beta_2) - (\mu + \alpha_2 + \beta_2) = \alpha_1 - \alpha_2$$

Linear Models Recap

The difference between city and highway driving has to be the same for cars as it is for trucks for this model to fit the data.

The same is true of the difference between cars and trucks:

For city driving, the expected difference in mileage between cars and trucks is:

$$(\mu + \alpha_1 + \beta_1) - (\mu + \alpha_1 + \beta_2) = \beta_1 - \beta_2$$

Linear Models Recap

The difference between city and highway driving has to be the same for cars as it is for trucks for this model to fit the data.

The same is true of the difference between cars and trucks:

For city driving, the expected difference in mileage between cars and trucks is:

$$(\mu + \alpha_1 + \beta_1) - (\mu + \alpha_1 + \beta_2) = \beta_1 - \beta_2$$

For highway driving the expected difference between cars and trucks is:

$$(\mu + \alpha_2 + \beta_1) - (\mu + \alpha_2 + \beta_2) = \beta_1 - \beta_2$$

Linear Models Recap

Sometimes this restriction, which is entirely due to the structure of the two factor model without an interaction term, is not realistic.

Linear Models Recap

Sometimes this restriction, which is entirely due to the structure of the two factor model without an interaction term, is not realistic.

To eliminate this restriction, an **interaction** term, which is like a hybrid of the two main factors, can be added to the model.

Linear Models Recap

Sometimes this restriction, which is entirely due to the structure of the two factor model without an interaction term, is not realistic.

To eliminate this restriction, an **interaction** term, which is like a hybrid of the two main factors, can be added to the model.

The interaction term removes the restriction mentioned earlier, at the price of adding quite a few parameters.

Linear Models Recap

Sometimes this restriction, which is entirely due to the structure of the two factor model without an interaction term, is not realistic.

To eliminate this restriction, an **interaction** term, which is like a hybrid of the two main factors, can be added to the model.

The interaction term removes the restriction mentioned earlier, at the price of adding quite a few parameters.

This is because the interaction will be represented by one parameter for each combination of the levels of the original factors (four in this case)

Linear Models Recap

The two way model without interaction

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \beta_1 X_{i3} + \beta_2 X_{i4} + e_i$$

is expanded to include four additional parameters to become:

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \beta_1 X_{i3} + \beta_2 X_{i4} +$$

$$\gamma_{11} Xi1Xi3 + \gamma_{12} X_{i1} X_{i4} + \gamma_{21} X_{i2} X_{i3} + \gamma_{22} X_{i2} X_{i4} + e_i$$

Linear Models Recap

The two way model without interaction

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \beta_1 X_{i3} + \beta_2 X_{i4} + e_i$$

is expanded to include four additional parameters to become:

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \beta_1 X_{i3} + \beta_2 X_{i4} +$$

$$\gamma_{11} Xi1Xi3 + \gamma_{12} X_{i1} X_{i4} + \gamma_{21} X_{i2} X_{i3} + \gamma_{22} X_{i2} X_{i4} + e_i$$

There are quite a few parameters but they work pretty much the same as before.

Linear Models Recap

The expected values for the two-way models with and without interaction are:

Category	Two-way without interaction	Two-way with interaction
city, car	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
city, truck	$\mu + \alpha_1 + \beta_2$	$\mu + \alpha_1 + \beta_2 + \gamma_{12}$
highway, car	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_1 + \gamma_{21}$
highway, truck	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_2 + \gamma_{22}$

Linear Models Recap

The expected values for the two-way models with and without interaction are:

Category	Two-way without interaction	Two-way with interaction
city, car	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
city, truck	$\mu + \alpha_1 + \beta_2$	$\mu + \alpha_1 + \beta_2 + \gamma_{12}$
highway, car	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_1 + \gamma_{21}$
highway, truck	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_2 + \gamma_{22}$

The additional parameters in the model with interaction allow the expected values to exactly fit the means of the four categories.

Linear Models Recap

The expected values for the two-way models with and without interaction are:

Category	Two-way without interaction	Two-way with interaction
city, car	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
city, truck	$\mu + \alpha_1 + \beta_2$	$\mu + \alpha_1 + \beta_2 + \gamma_{12}$
highway, car	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_1 + \gamma_{21}$
highway, truck	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_2 + \gamma_{22}$

The additional parameters in the model with interaction allow the expected values to exactly fit the means of the four categories.

In the two-way model without interaction, the expected values usually do not exactly fit the category means.

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Right click on the *2009 EPA Mileage Data* link and select *copy link location*

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Right click on the *2009 EPA Mileage Data* link and select *copy link location*

This should copy the URL for the EPA .csv data file, which is:

http://www.sandgquinn.org/stonehill/MA225/notes/09tstcar.csv

Reading the EPA data into R

Go to the course web page, then the *Notes and Handouts* section.

Right click on the *2009 EPA Mileage Data* link and select *copy link location*

This should copy the URL for the EPA .csv data file, which is:

http://www.sandgquinn.org/stonehill/MA225/notes/09tstcar.csv

Carefully type the following command in R, but don't hit enter:

```
epa<-read.table("",sep=",",fill=TRUE,header=TRUE)
```

Subsetting the EPA data

Paste the URL for the EPA data between the adjacent double quotes and hit enter to load the EPA data.

Subsetting the EPA data

Paste the URL for the EPA data between the adjacent double quotes and hit enter to load the EPA data.

Since we only need three columns of the data, we'll create a subset called `twoway`.

Enter the following R command to create a new data frame called `twoway`:

```
twoway<-subset(epa,,select=c(mpg,C.H,car.truck))
```

Subsetting the EPA data

Paste the URL for the EPA data between the adjacent double quotes and hit enter to load the EPA data.

Since we only need three columns of the data, we'll create a subset called `twoway`.

Enter the following R command to create a new data frame called `twoway`:

```
twoway<-subset(epa„select=c(mpg,C.H,car.truck))
```

It should contain only the columns `mpg,C.H,car.truck`. We can verify this by entering:

```
str(twoway)
```

Subsetting the EPA data

Paste the URL for the EPA data between the adjacent double quotes and hit enter to load the EPA data.

Since we only need three columns of the data, we'll create a subset called `twoway`.

Enter the following R command to create a new data frame called `twoway`:

```
twoway<-subset(epa,,select=c(mpg,C.H,car.truck))
```

It should contain only the columns `mpg,C.H,car.truck`. We can verify this by entering:

```
str(twoway)
```

Now to simplify our code, we'll attach the new data frame. Enter:

```
attach(twoway)
```

Fitting the Models

First we'll summarize the data by computing the sample means for the four categories. Enter:

```
agd<-aggregate(twoway,
by=list(C.H,car.truck),FUN=mean)

print(agd)
```

Fitting the Models

First we'll summarize the data by computing the sample means for the four categories. Enter:

```
agd<-aggregate(twoway,
by=list(C.H,car.truck),FUN=mean)
```

```
print(agd)
```

The results indicate the sample mean of mpg for each of the four categories:

Group.1	Group.2	mpg	expected value
C	C	22.86910	$\mu + \alpha_1 + \beta_1 \ (+\gamma_{11})$
H	C	35.51319	$\mu + \alpha_2 + \beta_1 \ (+\gamma_{21})$
C	T	18.75865	$\mu + \alpha_1 + \beta_2 \ (+\gamma_{12})$
H	T	28.20533	$\mu + \alpha_2 + \beta_2 \ (+\gamma_{22})$

Fitting the Models

Now run the linear model for the two-way ANOVA **without interaction**:

```
lm0<-lm(mpg ~ C.H+truck.car)
summary(lm0)
```

Fitting the Models

Now run the linear model for the two-way ANOVA **without interaction**:

```
lm0<-lm(mpg ~ C.H+truck.car)
summary(lm0)
```

In multiple factor models, sometimes the order the factor is specified makes a difference.

A common convention is to use what are known as "Type III" sums of squares, wich essentially test each variable as if it were the last one added (i.e., with every other factor already in the model).

Fitting the Models

Now run the linear model for the two-way ANOVA **without interaction**:

```
lm0<-lm(mpg ~ C.H+truck.car)
```

```
summary(lm0)
```

In multiple factor models, sometimes the order the factor is specified makes a difference.

A common convention is to use what are known as "Type III" sums of squares, wich essentially test each variable as if it were the last one added (i.e., with every other factor already in the model).

To compute these, enter:

```
drop1(lm0,~.,test="F")
```

Fitting the Models

The results are:

	Df	Sum of Sq	RSS	F value	Pr(F)
<none>		100557			
car.truck	1	23280	123837	666.97	<2.2 e-16
C.H	1	90305	190862	2587.27	<2.2 e-16

Fitting the Models

The results are:

	Df	Sum of Sq	RSS	F value	Pr(F)
<none>		100557			
car.truck	1	23280	123837	666.97	<2.2 e-16
C.H	1	90305	190862	2587.27	<2.2 e-16

In each case the F statistic measures the significance of the model with all factors compared to a "reduced model" with all of the other factors:

Factor	F statistic	P-value	Reduced model
car.truck	666.97	<2.2e-16	$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + e$
C.H	2587.27	<2.2e-16	$Y = \mu + \beta_1 X_3 + \beta_2 X_4 + e$

Both factors are significant in this case ($P < 0.05$)

Fitting the Models

The results of `summary(lm0)` are:

Coefficients:

	Estimate	Parameters
(Intercept)	23.5898	$\mu + \alpha_1 + \beta_1$
car.truckT	-5.7063	$\beta_2 - \beta_1$
C.HH	11.1917	$\alpha_2 - \alpha_1$

Fitting the Models

The results of `summary(lm0)` are:

Coefficients:

	Estimate	Parameters
(Intercept)	23.5898	$\mu + \alpha_1 + \beta_1$
car.truckT	-5.7063	$\beta_2 - \beta_1$
C.HH	11.1917	$\alpha_2 - \alpha_1$

The interpretation in terms of expected values is:

Category	Estimate	Expected Value
car, city	23.5898	$\mu + \alpha_1 + \beta_1$
truck, city	23.5898-5.7063	$\mu + \alpha_1 + \beta_2$
car, highway	23.5898+11.1917	$\mu + \alpha_2 + \beta_1$
truck, highway	23.5898-5.7063+11.1917	$\mu + \alpha_2 + \beta_2$

Fitting the Models

Category	Estimated Expected Value	Parameters
car, city	23.5898	$\mu + \alpha_1 + \beta_1$
truck, city	23.5898-5.7063	$\mu + \alpha_1 + \beta_2$
car, highway	23.5898+11.1917	$\mu + \alpha_2 + \beta_1$
truck, highway	23.5898-5.7063+11.1917	$\mu + \alpha_2 + \beta_2$

Fitting the Models

Category	Estimated Expected Value	Parameters
car, city	23.5898	$\mu + \alpha_1 + \beta_1$
truck, city	23.5898-5.7063	$\mu + \alpha_1 + \beta_2$
car, highway	23.5898+11.1917	$\mu + \alpha_2 + \beta_1$
truck, highway	23.5898-5.7063+11.1917	$\mu + \alpha_2 + \beta_2$

Note that the estimated expected values do not exactly match the sample means:

Category	Estimated expected value	Sample mean
car, city	23.5898	22.86910
truck, city	18.5135	18.75865
car, highway	34.7815	35.51319
truck, highway	29.7052	28.20533

Fitting the Models

Next run the two-way ANOVA **with interaction**:

```
lm0<-lm(mpg ~ C.H*truck.car)
summary(lm0)
```

Fitting the Models

Next run the two-way ANOVA **with interaction**:

```
lm0<-lm(mpg ~ C.H*truck.car)
```

```
summary(lm0)
```

In multiple factor models, sometimes the order the factor is specified makes a difference.

A common convention is to use what are known as "Type III" sums of squares, wich essentially test each variable as if it were the last one added (i.e., with every other factor already in the model).

Fitting the Models

Next run the two-way ANOVA **with interaction**:

```
lm0<-lm(mpg ~ C.H*truck.car)
```

```
summary(lm0)
```

In multiple factor models, sometimes the order the factor is specified makes a difference.

A common convention is to use what are known as "Type III" sums of squares, wich essentially test each variable as if it were the last one added (i.e., with every other factor already in the model).

To compute these, enter:

```
drop1(lm0,~.,test="F")
```

Fitting the Models

The results of `summary(lm0)` are:

	Estimate	Parameters
(Intercept)	22.8691	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
car.truckT	-4.1105	$\beta_2 + \gamma_{12} - \beta_1 - \gamma_{11}$
C.HH	12.6441	$\alpha_2 + \gamma_{21} - \alpha_1 - \gamma_{11}$
car.truckT:C.HH	-3.1974	$\gamma_{22} + \gamma_{11} - \gamma_{12} - \gamma_{21}$

Fitting the Models

The results of `summary(lm0)` are:

	Estimate	Parameters
(Intercept)	22.8691	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
car.truckT	-4.1105	$\beta_2 + \gamma_{12} - \beta_1 - \gamma_{11}$
C.HH	12.6441	$\alpha_2 + \gamma_{21} - \alpha_1 - \gamma_{11}$
car.truckT:C.HH	-3.1974	$\gamma_{22} + \gamma_{11} - \gamma_{12} - \gamma_{21}$

The interpretation in terms of expected values is:

Category	Estimate	Value
car, city	(Intercept)	22.8691
truck, city	(Intercept)+car.truckT	18.7586
car, hway	(Intercept)+C.HH	35.5132
truck, hway	(Intercept)+car.truckT+C.HH	28.2053

Fitting the Models

Comparing these results to the sample means,

Category	Estimate	Sample Mean
car, city	22.8691	22.86910
truck, city	18.7586	18.75865
car, hway	35.5132	35.51319
truck, hway	28.2053	28.20533

Fitting the Models

Comparing these results to the sample means,

Category	Estimate	Sample Mean
car, city	22.8691	22.86910
truck, city	18.7586	18.75865
car, hway	35.5132	35.51319
truck, hway	28.2053	28.20533

Note that in the two factor ANOVA with interaction, the estimated expected values exactly match the sample means for each category (unlike the no interaction model)

Fitting the Models

Comparing these results to the sample means,

Category	Estimate	Sample Mean
car, city	22.8691	22.86910
truck, city	18.7586	18.75865
car, hway	35.5132	35.51319
truck, hway	28.2053	28.20533

Note that in the two factor ANOVA with interaction, the estimated expected values exactly match the sample means for each category (unlike the no interaction model)

This comes at the expense of more parameters, and a more complicated model

Fitting the Models

Comparing these results to the sample means,

Category	Estimate	Sample Mean
car, city	22.8691	22.86910
truck, city	18.7586	18.75865
car, hway	35.5132	35.51319
truck, hway	28.2053	28.20533

Note that in the two factor ANOVA with interaction, the estimated expected values exactly match the sample means for each category (unlike the no interaction model)

This comes at the expense of more parameters, and a more complicated model

Next we use the R `drop1` function to decide whether the additional complexity is justified

Fitting the Models

The results are:

	Df	Sum Sq	RSS	F value	Pr(F)
<none>		98730			
car.truck	1	6051	104780	176.499	< 2.2e-16
C.H	1	62906	161636	1835.010	<2.2 e-16
car.truck:C.H	1	1827	100557	53.303	3.682e-13

Fitting the Models

The results are:

	Df	Sum Sq	RSS	F value	Pr(F)
<none>		98730			
car.truck	1	6051	104780	176.499	< 2.2e-16
C.H	1	62906	161636	1835.010	<2.2 e-16
car.truck:C.H	1	1827	100557	53.303	3.682e-13

The F statistic measures the significance of the model with all factors compared to a "reduced model":

Factor	Reduced model
car.truck	$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + e$
C.H	$Y = \mu + \beta_1 X_3 + \beta_2 X_4 + e$
car.truck:C.H	$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 X_3 + \beta_2 X_4 + e$

All three factors are significant in this case ($P < 0.05$)

Followup Tests

Next run the Tukey HSD test. For this we need to use the `aov` function:

```
lm0<-aov(mpg ~ C.H*truck.car)
```

```
TukeyHSD(lm0)
```

Followup Tests

Next run the Tukey HSD test. For this we need to use the `aov` function:

```
lm0<-aov(mpg ~ C.H*truck.car)
```

```
TukeyHSD(lm0)
```

The results are significant for each possible comparison:

	diff	lwr	upr
T:C-C:C	-4.110452	-4.905772	-3.315132
C:H-C:C	12.644084	11.885347	13.402820
T:H-C:C	5.336223	4.542231	6.130214
C:H-T:C	16.754536	15.956461	17.552610
T:H-T:C	9.446675	8.615012	10.278338
T:H-C:H	-7.307861	-8.104611	-6.511111

Summary

The two factor ANOVA with interaction is run with either of the R commands

```
lm0<-aov(mpg ~ C.H*truck.car)
```

```
lm0<-lm(mpg ~ C.H*truck.car)
```

Summary

The two factor ANOVA with interaction is run with either of the R commands

```
lm0<-aov(mpg ~ C.H*truck.car)
```

```
lm0<-lm(mpg ~ C.H*truck.car)
```

The significance of the interaction and the two factors are tested with the command:

```
drop1(lm0,~ .,test="F")
```

Summary

The two factor ANOVA with interaction is run with either of the R commands

```
lm0<-aov(mpg ~ C.H*truck.car)
```

```
lm0<-lm(mpg ~ C.H*truck.car)
```

The significance of the interaction and the two factors are tested with the command:

```
drop1(lm0,~.,test="F")
```

Individual category differences can be tested with the Tukey HSD procedure:

```
TukeyHSD(lm0)
```

Summary

The two factor ANOVA with interaction is run with either of the R commands

```
lm0<-aov(mpg ~ C.H*truck.car)
```

```
lm0<-lm(mpg ~ C.H*truck.car)
```

The significance of the interaction and the two factors are tested with the command:

```
drop1(lm0,~.,test="F")
```

Individual category differences can be tested with the Tukey HSD procedure:

```
TukeyHSD(lm0)
```

This requires that the model be run with the `aov` command.