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

Finally we considered the two factor ANOVA with interaction:

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

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

Linear Models Recap

Finally we considered the two factor ANOVA with interaction:

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

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

Today we will extend our list to include models with *both* continuous and discrete predictors.

Linear Models Recap

Finally we considered the two factor ANOVA with interaction:

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

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

Today we will extend our list to include models with *both* continuous and discrete predictors.

Historically this type of model has been called *analysis of covariance*

Linear Models Recap

Finally we considered the two factor ANOVA with interaction:

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

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

Today we will extend our list to include models with *both* continuous and discrete predictors.

Historically this type of model has been called *analysis of covariance*

If the factor has two levels and there is one continuous predictor X_{i3}, the model has the form

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

Analysis of Covariance

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

This type of model is often considered a one factor ANOVA with adjustment for the continuous variable.

Analysis of Covariance

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

This type of model is often considered a one factor ANOVA with adjustment for the continuous variable.

In our example, the factor has two levels, and the expected values of the Y_i in each case are:

- Level 1: $Y_i = \mu + \alpha_1 + \beta X_i$
- Level 2: $Y_i = \mu + \alpha_2 + \beta X_i$

Analysis of Covariance

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

This type of model is often considered a one factor ANOVA with adjustment for the continuous variable.

In our example, the factor has two levels, and the expected values of the Y_i in each case are:

- Level 1: $Y_i = \mu + \alpha_1 + \beta X_i$
- Level 2: $Y_i = \mu + \alpha_2 + \beta X_i$

It can also be thought of as fitting parallel regression lines for each level of the factor.

Analysis of Covariance

Notice that if there are no differences in the levels of the factor, ($\alpha_1 = \alpha_2 = 0$), the model

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

reduces to the simple regression model

$$Y_i = \mu + \beta X_{i3} + e_i$$

Analysis of Covariance

Notice that if there are no differences in the levels of the factor, ($\alpha_1 = \alpha_2 = 0$), the model

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

reduces to the simple regression model

$$Y_i = \mu + \beta X_{i3} + e_i$$

If the slope of the regression line is zero, ($\beta = 0$), the model

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

reduces to the one factor ANOVA,

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

Analysis of Covariance

We will perform an analysis of covariance using the EPA data in the following way:

Suppose we do a one-way ANOVA to compare mileage for cars and trucks.

Analysis of Covariance

We will perform an analysis of covariance using the EPA data in the following way:

Suppose we do a one-way ANOVA to compare mileage for cars and trucks.

However, we want to adjust for the fact that engine displacement (`cid`) has an effect on gas mileage, and trucks probably have larger engines, on average, than cars.

So we want to adjust for `cid` when we compare cars and trucks.

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 download* 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 download* link and select *copy link location*

Paste the URL between the quotes in the R command:

```
source("")
```

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 download* link and select *copy link location*

Paste the URL between the quotes in the R command:

```
source("")
```

When you hit enter, this should download the EPA data to your workspace in a data frame named epa. It also does an `attach` command for epa. Verify that you have the data by entering:

```
str(epa)
```

Subsetting the EPA data

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

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

```
cov<-subset(epa„select=c(mpg,car.truck,cid))
```

Subsetting the EPA data

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

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

```
cov<-subset(epa,,select=c(mpg,car.truck,cid))
```

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

```
str(cov)
```

Subsetting the EPA data

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

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

```
cov<-subset(epa,,select=c(mpg,car.truck,cid))
```

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

```
str(cov)
```

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

```
attach(cov)
```

Fitting the Models

First we'll summarize the data by computing the sample means for `mpg` and `cid`. Enter:

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

```
print(agd)
```

Fitting the Models

First we'll summarize the data by computing the sample means for `mpg` and `cid`. Enter:

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

```
print(agd)
```

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

Group.1	mpg	cid
C	29.14295	195.5629
T	23.49641	251.0603

Fitting the Models

First we'll summarize the data by computing the sample means for `mpg` and `cid`. Enter:

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

```
print(agd)
```

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

Group.1	mpg	cid
C	29.14295	195.5629
T	23.49641	251.0603

As expected, mileage is lower for trucks, on average, and engine displacement is higher.

Fitting the Models

Now run the linear model for the one-way ANOVA without the covariate `cid`:

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

Fitting the Models

Now run the linear model for the one-way ANOVA without the covariate `cid`:

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

The results indicate a significant difference between cars and trucks, and the model predicts the sample means:

Coefficients:

	Estimate	
(Intercept)	29.1429	
car.truckT	-5.6465	(29.1429-5.6465=23.4964)

Fitting the Models

Now run the linear model for the one-way ANOVA without the covariate `cid`:

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

The results indicate a significant difference between cars and trucks, and the model predicts the sample means:

Coefficients:

	Estimate	
(Intercept)	29.1429	
car.truckT	-5.6465	(29.1429-5.6465=23.4964)

The model without adjusting for `cid` indicates a difference of 5.65 mpg between cars and trucks.

Fitting the Models

Now we include `cid` as a continuous predictor in the model:

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

Fitting the Models

Now we include `cid` as a continuous predictor in the model:

`lm0<-lm(mpg ~ truck.car+cid)` Because we now have more than one predictor, we use the `drop1` function to test their significance:

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

Fitting the Models

Now we include `cid` as a continuous predictor in the model:

`lm0<-lm(mpg ~ truck.car+cid)` Because we now have more than one predictor, we use the `drop1` function to test their significance:

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

The results indicate both are

	Df	Sum of Sq	RSS	F value	Pr(F)
<none>		133786			
car.truck	1	3425	137211	73.759	< 2.2e-16
cid	1	57076	190862	1229.107	< 2.2e-16

Fitting the Models

Now we include `cid` as a continuous predictor in the model:

`lm0<-lm(mpg` $\sim$ `truck.car+cid)` Because we now have more than one predictor, we use the `drop1` function to test their significance:

`drop1(lm0,`$\sim$`.,test="F")`

The results indicate both are

	Df	Sum of Sq	RSS	F value	Pr(F)
<none>		133786			
car.truck	1	3425	137211	73.759	< 2.2e-16
cid	1	57076	190862	1229.107	< 2.2e-16

The F statistics are significant for both predictors ($P < 0.05$).

Fitting the Models

The results of `summary(lm0)` are:

Coefficients:

	Estimate
(Intercept)	40.814573
car.truckT	-2.334328
cid	-0.059682

Fitting the Models

The results of `summary(lm0)` are:

Coefficients:

	Estimate
(Intercept)	40.814573
car.truckT	-2.334328
cid	-0.059682

This time the estimated difference between cars and trucks is 2.33 mpg, rather than 5.65 mpg.

Fitting the Models

The results of `summary(lm0)` are:

Coefficients:

	Estimate
(Intercept)	40.814573
car.truckT	-2.334328
cid	-0.059682

This time the estimated difference between cars and trucks is 2.33 mpg, rather than 5.65 mpg.

The interpretation of these coefficients is that the difference in mileage between cars and trucks *that have the same engine displacement* is 2.33 mpg.

Fitting the Models

The results of `summary(lm0)` are:

Coefficients:

	Estimate
(Intercept)	40.814573
car.truckT	-2.334328
cid	-0.059682

This time the estimated difference between cars and trucks is 2.33 mpg, rather than 5.65 mpg.

The interpretation of these coefficients is that the difference in mileage between cars and trucks *that have the same engine displacement* is 2.33 mpg.

Note that the model corresponds to two parallel regression lines with different intercepts.

Fitting the Models

The next question that might arise is: are the regression lines really parallel? Or does `cid` have a different effect on mileage for cars and trucks?

Fitting the Models

The next question that might arise is: are the regression lines really parallel? Or does `cid` have a different effect on mileage for cars and trucks?

To answer this question, we can consider a model where we have different slopes for cars and trucks.

This is exactly equivalent to an interaction term, and our new model is:

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

Fitting the Models

The next question that might arise is: are the regression lines really parallel? Or does `cid` have a different effect on mileage for cars and trucks?

To answer this question, we can consider a model where we have different slopes for cars and trucks.

This is exactly equivalent to an interaction term, and our new model is:

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

As before, we replace the $+$ in the model with an $*$ to include all interactions:

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

Fitting the Models

The next question that might arise is: are the regression lines really parallel? Or does `cid` have a different effect on mileage for cars and trucks?

To answer this question, we can consider a model where we have different slopes for cars and trucks.

This is exactly equivalent to an interaction term, and our new model is:

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

As before, we replace the $+$ in the model with an $*$ to include all interactions:

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

Fitting the Models

Again we use `drop1` to test the significance of factors:

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

Fitting the Models

Again we use `drop1` to test the significance of factors:

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

	Df	Sum Sq	RSS	F value	Pr(F)
<none>		133736			
car.truck	1	664	134400	14.300	0.0001590
cid	1	34981	168717	753.319	< 2.2e-16
car.truck:cid	1	50	133786	1.071	0.3008058

Fitting the Models

Again we use `drop1` to test the significance of factors:

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

	Df	Sum Sq	RSS	F value	Pr(F)
<none>		133736			
car.truck	1	664	134400	14.300	0.0001590
cid	1	34981	168717	753.319	< 2.2e-16
car.truck:cid	1	50	133786	1.071	0.3008058

The result indicates that the car.truck factor and the continuous predictor cid are significant ($P < 0.05$), but the interaction term, representing a difference in the slopes of the two regression lines, is not significantly different from zero ($P = 0.30$).