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

Two factor ANOVA with interaction:

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

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

Linear Models Recap

Two factor ANOVA with interaction:

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

$$\gamma_{11} X i 1 X i 3 + \gamma_{12} X_{i1} X_{i4} + \gamma_{21} X_{i2} X_{i3} + \gamma_{22} X_{i2} X_{i4} + e_i$$

Models with both continuous and discrete predictors (analysis of covariance)

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

Linear Models Recap

Two factor ANOVA with interaction:

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

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

Models with both continuous and discrete predictors (analysis of covariance)

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

Now we consider models with multiple continuous predictors (multiple regression)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i$$

Multiple Regression

We will consider a model with the following 5 continuous predictors:

- vehicle weight `etw`
- engine displacement `cid`
- horsepower `rhp`
- compression ratio `cmp`
- coast down time `cstdwn`

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 *code to download 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 *code to download the 2009 EPA mileage data* link and select *copy link location*

Paste the link between the quotes in a `source("")` command

Reading the EPA data into R

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

Right click on the *code to download the 2009 EPA mileage data* link and select *copy link location*

Paste the link between the quotes in a `source("")` command Verify the download by typing `str(epa)`

Subsetting the EPA data

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

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

```
mreg<-subset(
epa„select=c(mpg,etw,cid,rhp,cmp,cstdwn))
```

Subsetting the EPA data

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

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

```
mreg<-subset(
epa,,select=c(mpg,etw,cid,rhp,cmp,cstdwn))
```

Subsetting the EPA data

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

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

```
mreg<-subset(
epa„select=c(mpg,etw,cid,rhp,cmp,cstdwn))
```

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

```
attach(mreg)
```

Fitting the Models

We use `lm` to run the model:

```
lm0<-lm(mpg   etw+cid+rhp+cmp+cstdwn)
```

Fitting the Models

We use `lm` to run the model:

`lm0<-lm(mpg   etw+cid+rhp+cmp+cstdwn)` Because we have several predictors, we use the `drop1` function to test their significance:

`drop1(lm0,∼.,test="F")`

Fitting the Models

The result of the command

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

is

	Df	Sum of Sq	RSS	F value	Pr(F)
<none>		71987			
etw	1	3753.5	75741	90.1531	< 2.2e-16
cid	1	1200.5	73188	28.8338	8.955e-08
rhp	1	297.2	72284	7.1379	0.007618
cmp	1	1339.5	73327	32.1714	1.651e-08
cstdwn	1	22.4	72009	0.5378	0.463460

Fitting the Models

The result of the command

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

is

	Df	Sum of Sq	RSS	F value	Pr(F)
<none>			71987		
etw	1	3753.5	75741	90.1531	< 2.2e-16
cid	1	1200.5	73188	28.8338	8.955e-08
rhp	1	297.2	72284	7.1379	0.007618
cmp	1	1339.5	73327	32.1714	1.651e-08
cstdwn	1	22.4	72009	0.5378	0.463460

The F statistics are significant ($P < 0.05$) for all except `cstdwn`.

Interpreting the Coefficients

The results of `summary(lm0)` are:

Coefficients:

	Estimate
(Intercept)	37.9108831
etw	-0.0030264
cid	-0.0254872
rhp	-0.0078682
cmp	0.9761919
cstdwn	0.0152935

Interpreting the Coefficients

The results of `summary(lm0)` are:

Coefficients:

	Estimate
(Intercept)	37.9108831
etw	-0.0030264
cid	-0.0254872
rhp	-0.0078682
cmp	0.9761919
cstdwn	0.0152935

Evidently `etw`, `cid`, and `rhp` have a negative effect on mileage, while `cmp` and `cstdwn` have a positive effect, although `cstdwn` is not significantly different from zero.

Interpreting the Coefficients

The linear model for the expected mpg is:

$$\text{mpg} = 37.91 - 0.003\text{etw} - 0.025\text{cid} - 0.0079\text{rhp} + 0.976\text{cmp} + 0.015\text{cstdwn}$$