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)
```

Simple Regression with EPA data

Our simple regression used the engine displacement (cid) "cubic inches displacement" as the independent variable, and gas mileage (mpg) "miles per gallon" as the dependent variable.

Simple Regression with EPA data

Our simple regression used the engine displacement (cid) "cubic inches displacement" as the independent variable, and gas mileage (mpg) "miles per gallon" as the dependent variable.

You can simplify things a bit if you "attach" the data, meaning make the column names recognizable.

```
attach(epa)
```

Simple Regression with EPA data

Our simple regression used the engine displacement (cid) "cubic inches displacement" as the independent variable, and gas mileage (mpg) "miles per gallon" as the dependent variable.

You can simplify things a bit if you "attach" the data, meaning make the column names recognizable.

```
attach(epa)
```

Now run the simple regression model, and display the results:

```
lm0<-lm(mpg ~ cid)
summary(lm0)
```

Simple Regression with EPA data

From the "Coefficients:" section, in the column labeled "Estimate", we see:

(Intercept)	40.876	This is the estimate of β_0, the interce
cid	-0.064764	This is the estimate of β_1, the slope

Simple Regression with EPA data

From the "Coefficients:" section, in the column labeled "Estimate", we see:

(Intercept)	40.876	This is the estimate of β_0, the interce
cid	-0.064764	This is the estimate of β_1, the slope

So our regression model produces a line with a slope of -0.06 and an intercept of 40.87.

Simple Regression with EPA data

From the "Coefficients:" section, in the column labeled "Estimate", we see:

(Intercept)	40.876	This is the estimate of β_0, the interce
cid	-0.064764	This is the estimate of β_1, the slope

So our regression model produces a line with a slope of -0.06 and an intercept of 40.87.

The interpretation of this model is as follows:

A car with 0 cubic inches displacement should get 40.87 mpg.

Simple Regression with EPA data

From the "Coefficients:" section, in the column labeled "Estimate", we see:

(Intercept)	40.876	This is the estimate of β_0, the interce
cid	-0.064764	This is the estimate of β_1, the slope

So our regression model produces a line with a slope of -0.06 and an intercept of 40.87.

The interpretation of this model is as follows:

A car with 0 cubic inches displacement should get 40.87 mpg.

This is not a realistic value for cid, but it does give us a kind of theoretical upper bound on mileage as you make the engine smaller.

Simple Regression with EPA data

The slope is -0.06, which says that according to the model, for every cubic inch we add to the engine, we lose $0.06mpg$ in fuel economy. We can also get predicted mpg values for various engine dispacements:

cid	predicted mpg = 40.876-0.06*cid=mpg
80	35.69488
120	33.10432
160	30.51376
200	27.9232
240	25.33264
280	22.74208
320	20.15152
360	17.56096

Simple Regression with EPA data

Another line in the summary says:

```
Multiple R-squared:  0.3578
```

The Multiple R-squared tells us the proportion of the variability in Y that the model explains. Our model explains about 35% more of the variation in Y than a model with just the mean would.

Simple Regression with EPA data

Another line in the summary says:

```
Multiple R-squared:  0.3578
```

The Multiple R-squared tells us the proportion of the variability in Y that the model explains. Our model explains about 35% more of the variation in Y than a model with just the mean would.

Another line in the summary says:

```
Residual standard error:  6.9
```

The Residual standard error is an estimate of σ_e for the model

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad \text{with} \quad e_i \sim N(0, \sigma_e)$$

Simple Regression with EPA data

A number of useful diagnostic plots can be obtained by entering:

```
plot(lm0)
```

Simple Regression with EPA data

A number of useful diagnostic plots can be obtained by entering:

```
plot(lm0)
```

Because the slope and intercept of the regression line are quite sensitive to outliers in the data, it is a good idea to inspect a plot of differences between the predicted and actual values, which are called the **residuals**:

$$r_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 * X_i$$

Simple Regression with EPA data

A number of useful diagnostic plots can be obtained by entering:

```
plot(lm0)
```

Because the slope and intercept of the regression line are quite sensitive to outliers in the data, it is a good idea to inspect a plot of differences between the predicted and actual values, which are called the **residuals**:

$$r_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 * X_i$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ represent the *estimates* of the parameters β_0 and β_1 that we obtained by fitting the model. In our case,

$$\hat{\beta}_0 = 40.876 \quad \hat{\beta}_1 = -0.064764$$

Continuous vs Discrete

So far we considered models of the form

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

where the independent or predictor variable X_i was *continuous*

Continuous vs Discrete

So far we considered models of the form

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

where the independent or predictor variable X_i was *continuous*

A model with a single continuous predictor is usually called a *simple regression* model.

Continuous vs Discrete

So far we considered models of the form

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

where the independent or predictor variable X_i was *continuous*

A model with a single continuous predictor is usually called a *simple regression* model.

A different type of model arises when we want to compare several groups.

Continuous vs Discrete

So far we considered models of the form

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

where the independent or predictor variable X_i was *continuous*

A model with a single continuous predictor is usually called a *simple regression* model.

A different type of model arises when we want to compare several groups.

In this case, there is one predictor variable for each group.

Continuous vs Discrete

So far we considered models of the form

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

where the independent or predictor variable X_i was *continuous*

A model with a single continuous predictor is usually called a *simple regression* model.

A different type of model arises when we want to compare several groups.

In this case, there is one predictor variable for each group.

The predictor variable is always one if the individual belongs to its group, zero if it does not.

Continuous vs Discrete

If we have three groups, our linear model has the form:

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

Continuous vs Discrete

If we have three groups, our linear model has the form:

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

The independent or X variables are coded as follows:

- If the i^{th} subject belongs to group 1, $X_{1i} = 1$, otherwise $X_{1i} = 0$
- If the i^{th} subject belongs to group 2, $X_{2i} = 1$, otherwise $X_{2i} = 0$
- If the i^{th} subject belongs to group 3, $X_{3i} = 1$, otherwise $X_{3i} = 0$

Continuous vs Discrete

If we have three groups, our linear model has the form:

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

The independent or X variables are coded as follows:

- If the i^{th} subject belongs to group 1, $X_{1i} = 1$, otherwise $X_{1i} = 0$
- If the i^{th} subject belongs to group 2, $X_{2i} = 1$, otherwise $X_{2i} = 0$
- If the i^{th} subject belongs to group 3, $X_{3i} = 1$, otherwise $X_{3i} = 0$

As before, e_i is assumed to have a normal distribution $N(0, \sigma_e)$

Discrete Predictors (ANOVA)

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

As before, the only random quantity on the right hand side is e_i

Discrete Predictors (ANOVA)

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

As before, the only random quantity on the right hand side is e_i

This means the expected values of the Y_i variables are:

- $E(Y_i) = \mu + \alpha_1$ If subject i is in group 1
- $E(Y_i) = \mu + \alpha_2$ If subject i is in group 2
- $E(Y_i) = \mu + \alpha_3$ If subject i is in group 3

Discrete Predictors (ANOVA)

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

As before, the only random quantity on the right hand side is e_i

This means the expected values of the Y_i variables are:

- $E(Y_i) = \mu + \alpha_1$ If subject i is in group 1
- $E(Y_i) = \mu + \alpha_2$ If subject i is in group 2
- $E(Y_i) = \mu + \alpha_3$ If subject i is in group 3

Every subject in a particular group has the same expected value for Y_i

Discrete Predictors (ANOVA)

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

As before, the only random quantity on the right hand side is e_i

This means the expected values of the Y_i variables are:

- $E(Y_i) = \mu + \alpha_1$ If subject i is in group 1
- $E(Y_i) = \mu + \alpha_2$ If subject i is in group 2
- $E(Y_i) = \mu + \alpha_3$ If subject i is in group 3

Every subject in a particular group has the same expected value for Y_i

In a sense, this model is predicting the *means* of each group

Discrete Predictors (ANOVA)

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

A model of this form, with a separate zero-one predictor for each group is usually called a *one-way analysis of variance* or *one-way ANOVA*.

Discrete Predictors (ANOVA)

$$Y_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + e_i$$

A model of this form, with a separate zero-one predictor for each group is usually called a *one-way analysis of variance* or *one-way ANOVA*.

Now we will generate artificial data that fits this model and analyze it with R.

The parameter values will be:

- $\mu = 2$
- $\alpha_1 = 3$
- $\alpha_2 = 6$
- $\alpha_3 = 9$
- $\sigma_e = 3$

Generating the ANOVA data

With discrete predictors, most statistical software generates the appropriate X values with zeros and ones in the right places automatically based on an additional variable that identifies the group.

Generating the ANOVA data

With discrete predictors, most statistical software generates the appropriate X values with zeros and ones in the right places automatically based on an additional variable that identifies the group.

First we will generate the variable of group identifiers. We will make three groups of 500 each.

Generating the ANOVA data

With discrete predictors, most statistical software generates the appropriate X values with zeros and ones in the right places automatically based on an additional variable that identifies the group.

First we will generate the variable of group identifiers. We will make three groups of 500 each.

The R code for this is (type it all on one line):

```
group<-gl(3,500,1500,
labels=c("Group1","Group2","Group3"))
```

Generating the ANOVA data

After creating `group`, entering the R command

```
table(group)
```

should list the three group labels each with a count of 500:

```
> table(group)
group
 Group1  Group2  Group3
    500     500     500
```

Generating the ANOVA data

After creating `group`, entering the R command

```
table(group)
```

should list the three group labels each with a count of 500:

```
> table(group)
group
 Group1  Group2  Group3
    500     500     500
```

Next we generate the 1500 e_i values as $N(0,3)$:

```
e<-rnorm(1500,0,3)
```

Generating the ANOVA data

Next we generate the X_1 values as: 1 for group 1, 0 otherwise:

```
x1<-c(rep(1,500),rep(0,500),rep(0,500))
```

Generating the ANOVA data

Next we generate the X_1 values as: 1 for group 1, 0 otherwise:

```
x1<-c(rep(1,500),rep(0,500),rep(0,500))
```

Now generate the X_2 values as: 1 for group 2, 0 otherwise:

```
x2<-c(rep(0,500),rep(1,500),rep(0,500))
```

Generating the ANOVA data

Next we generate the X_1 values as: 1 for group 1, 0 otherwise:

```
x1<-c(rep(1,500),rep(0,500),rep(0,500))
```

Now generate the X_2 values as: 1 for group 2, 0 otherwise:

```
x2<-c(rep(0,500),rep(1,500),rep(0,500))
```

Finally generate the X_3 values as: 1 for group 3, 0 otherwise:

```
x3<-c(rep(0,500),rep(0,500),rep(1,500))
```

Generating the ANOVA data

Now generate the parameter values:

- `mu<-2` Set $\mu = 2$
- `alpha1<-3` Set $\alpha_1 = 3$
- `alpha2<-6` Set $\alpha_2 = 6$
- `alpha3<-9` Set $\alpha_3 = 9$

Generating the ANOVA data

Now generate the parameter values:

- `mu<-2` Set $\mu = 2$
- `alpha1<-3` Set $\alpha_1 = 3$
- `alpha2<-6` Set $\alpha_2 = 6$
- `alpha3<-9` Set $\alpha_3 = 9$

Finally compute the Y values:

```
y<-mu+alpha1*x1+alpha2*x2+alpha3*x3+e
```

Generating the ANOVA data

At this point Y contains 1500 values, with the properties:

- The first 500 (Group 1) have $E(Y_i) = \mu + \alpha_1 = 2 + 3 = 5$
- The second 500 (Group 2) have $E(Y_i) = \mu + \alpha_2 = 2 + 6 = 8$
- The third 500 (Group 3) have $E(Y_i) = \mu + \alpha_3 = 2 + 9 = 11$

Generating the ANOVA data

At this point Y contains 1500 values, with the properties:

- The first 500 (Group 1) have $E(Y_i) = \mu + \alpha_1 = 2 + 3 = 5$
- The second 500 (Group 2) have $E(Y_i) = \mu + \alpha_2 = 2 + 6 = 8$
- The third 500 (Group 3) have $E(Y_i) = \mu + \alpha_3 = 2 + 9 = 11$

Finally compute the Y values:

```
y<-mu+alpha1*x1+alpha2*x2+alpha3*x3+e
```

Generating the ANOVA data

At this point Y contains 1500 values, with the properties:

- The first 500 (Group 1) have $E(Y_i) = \mu + \alpha_1 = 2 + 3 = 5$
- The second 500 (Group 2) have $E(Y_i) = \mu + \alpha_2 = 2 + 6 = 8$
- The third 500 (Group 3) have $E(Y_i) = \mu + \alpha_3 = 2 + 9 = 11$

Generating the ANOVA data

At this point Y contains 1500 values, with the properties:

- The first 500 (Group 1) have $E(Y_i) = \mu + \alpha_1 = 2 + 3 = 5$
- The second 500 (Group 2) have $E(Y_i) = \mu + \alpha_2 = 2 + 6 = 8$
- The third 500 (Group 3) have $E(Y_i) = \mu + \alpha_3 = 2 + 9 = 11$

Of course the values won't match exactly because we introduced some randomness with the e_i values, but they should be close. To check Group 1, enter:

```
mean(y[1:500])
```

Generating the ANOVA data

For Group 1: `mean(y[1:500])`

should produce something like `[1] 5.021602`

Generating the ANOVA data

For Group 1: `mean(y[1:500])`

should produce something like `[1] 5.021602`

For Group 2: `mean(y[501:1000])`

should produce something like `[1] 7.89436`

Generating the ANOVA data

For Group 1: `mean(y[1:500])`

should produce something like `[1] 5.021602`

For Group 2: `mean(y[501:1000])`

should produce something like `[1] 7.89436`

For Group 3: `mean(y[1001:1500])`

should produce something like `[1] 11.07294`

Generating the ANOVA data

Within each group, the standard deviation should be $\sigma_e = 3$ in this case:

For Group 1: `sd(y[1:500])`

should produce something like `[1] 3.013675`

Generating the ANOVA data

Within each group, the standard deviation should be $\sigma_e = 3$ in this case:

For Group 1: `sd(y[1:500])`

should produce something like `[1] 3.013675`

For Group 2: `sd(y[501:1000])`

should produce something like `[1] 3.020052`

Generating the ANOVA data

Within each group, the standard deviation should be $\sigma_e = 3$ in this case:

For Group 1: `sd(y[1:500])`

should produce something like `[1] 3.013675`

For Group 2: `sd(y[501:1000])`

should produce something like `[1] 3.020052`

For Group 3: `sd(y[1001:1500])`

should produce something like `[1] 2.934667`

Generating the ANOVA data

Within each group, the standard deviation should be $\sigma_e = 3$ in this case:

For Group 1: `sd(y[1:500])`

should produce something like `[1] 3.013675`

For Group 2: `sd(y[501:1000])`

should produce something like `[1] 3.020052`

For Group 3: `sd(y[1001:1500])`

should produce something like `[1] 2.934667`

When we fit the model, the Residual standard error, which is an estimate of σ_e based on the sample, should be close to 3.

Generating the ANOVA data

We can compute the sample *variance* σ_e^2 within each group as well:

For Group 1: `var(y[1:500])`

should produce something like `[1] 9.082236`

Generating the ANOVA data

We can compute the sample *variance* σ_e^2 within each group as well:

For Group 1: `var(y[1:500])`

should produce something like `[1] 9.082236`

For Group 2: `var(y[501:1000])`

should produce something like `[1] 9.120715`

Generating the ANOVA data

We can compute the sample *variance* σ_e^2 within each group as well:

For Group 1: `var(y[1:500])`

should produce something like `[1] 9.082236`

For Group 2: `var(y[501:1000])`

should produce something like `[1] 9.120715`

For Group 3: `var(y[1001:1500])`

should produce something like `[1] 8.612269`

Generating the ANOVA data

We can compute the sample *variance* σ_e^2 within each group as well:

For Group 1: `var(y[1:500])`

should produce something like `[1] 9.082236`

For Group 2: `var(y[501:1000])`

should produce something like `[1] 9.120715`

For Group 3: `var(y[1001:1500])`

should produce something like `[1] 8.612269`

Note that if we compute the sample variance of y without taking groups into account, we get something larger:

`var(y)`

should produce something like `15.03887`

Generating the ANOVA data

This is an important observation for the following reason:

If there are no differences between groups, we should be able to lump the three groups together into a single sample.

Generating the ANOVA data

This is an important observation for the following reason:

If there are no differences between groups, we should be able to lump the three groups together into a single sample.

This should produce a sample variance of σ_e^2 for y (assuming no differences between the groups).

Generating the ANOVA data

This is an important observation for the following reason:

If there are no differences between groups, we should be able to lump the three groups together into a single sample.

This should produce a sample variance of σ_e^2 for y (assuming no differences between the groups).

On the other hand, if there *are* differences, `var(y)` will be inflated because of the differences between the group means.

Generating the ANOVA data

This is an important observation for the following reason:

If there are no differences between groups, we should be able to lump the three groups together into a single sample.

This should produce a sample variance of σ_e^2 for y (assuming no differences between the groups).

On the other hand, if there *are* differences, `var(y)` will be inflated because of the differences between the group means.

The relative size of `var(y)` and the Residual standard error form the basis of the test for equality of the three means.

Generating the ANOVA data

This is an important observation for the following reason:

If there are no differences between groups, we should be able to lump the three groups together into a single sample.

This should produce a sample variance of σ_e^2 for y (assuming no differences between the groups).

On the other hand, if there *are* differences, `var(y)` will be inflated because of the differences between the group means.

The relative size of `var(y)` and the Residual standard error form the basis of the test for equality of the three means.

This is the reason this type of linear model has traditionally been called "analysis of variance"

Generating the ANOVA data

Now we perform the computations for the ANOVA.

We have a choice of several routines in R to accomplish this.

enter:

```
lm0<-aov(y ~ group)
summary(lm0)
```

Generating the ANOVA data

Now we perform the computations for the ANOVA.

We have a choice of several routines in R to accomplish this.

enter:

```
lm0<-aov(y ~ group)
```

```
summary(lm0)
```

The output should contain a line something like

```
Residuals 1497 13380.8 8.9
```

Generating the ANOVA data

Now we perform the computations for the ANOVA.

We have a choice of several routines in R to accomplish this.

enter:

```
lm0<-aov(y ~ group)
```

```
summary(lm0)
```

The output should contain a line something like

```
Residuals 1497 13380.8 8.9
```

The rightmost number is an estimate of σ_e^2.

Since we generated data with $\sigma_e = 3$, σ_e^2 should be close to 9, and it is.

Generating the ANOVA data

Now we perform the computations for the ANOVA.

We have a choice of several routines in R to accomplish this.

enter:

`lm0<-aov(y ~ group)`

`summary(lm0)`

The output should contain a line something like

`Residuals 1497 13380.8 8.9`

The rightmost number is an estimate of σ_e^2.

Since we generated data with $\sigma_e = 3$, σ_e^2 should be close to 9, and it is.

Generating the ANOVA data

Another line in the output looks something like this:

```
group 2 9162.5 4581.2 512.53 < 2.2e-16 ***
```

Generating the ANOVA data

Another line in the output looks something like this:

```
group 2 9162.5 4581.2 512.53 < 2.2e-16 ***
```

The three discrete variables X_1, X_2, and X_3 comprise what is known as a **factor** in this type of linear model.

Generating the ANOVA data

Another line in the output looks something like this:

```
group 2 9162.5 4581.2 512.53 < 2.2e-16 ***
```

The three discrete variables X_1, X_2, and X_3 comprise what is known as a **factor** in this type of linear model.

The results table for an ANOVA type model usually has a line for each factor (group in this case)

Generating the ANOVA data

Another line in the output looks something like this:

```
group 2 9162.5 4581.2 512.53 < 2.2e-16 ***
```

The three discrete variables X_1, X_2, and X_3 comprise what is known as a **factor** in this type of linear model.

The results table for an ANOVA type model usually has a line for each factor (group in this case)

The `F value` column lists the test statistic for the null hypthesis that all group means are zero, that is,

$$\beta_1 = \beta_2 = \beta_3 = 0$$

Generating the ANOVA data

Another line in the output looks something like this:

```
group 2 9162.5 4581.2 512.53 < 2.2e-16 ***
```

The three discrete variables X_1, X_2, and X_3 comprise what is known as a **factor** in this type of linear model.

The results table for an ANOVA type model usually has a line for each factor (group in this case)

The `F value` column lists the test statistic for the null hypthesis that all group means are zero, that is,

$$\beta_1 = \beta_2 = \beta_3 = 0$$

The `F value` in this case indicates that it is highly unlikely that this data is a sample from a population with no group differences.

Generating the ANOVA data

The results indicate group differences, but the next question is usually "which groups are different?".

Generating the ANOVA data

The results indicate group differences, but the next question is usually "which groups are different?".

One way to answer this question is with a *follow-up test* such as Tukey's test.

Generating the ANOVA data

The results indicate group differences, but the next question is usually "which groups are different?".

One way to answer this question is with a *follow-up test* such as Tukey's test.

Enter:

```
TukeyHSD(lm0)
```

Generating the ANOVA data

The results indicate group differences, but the next question is usually "which groups are different?".

One way to answer this question is with a *follow-up test* such as Tukey's test.

Enter:

```
TukeyHSD(lm0)
```

This will produce a list of upper and lower confidence bounds for the difference between each possible pair of group means.

Generating the ANOVA data

The results indicate group differences, but the next question is usually "which groups are different?".

One way to answer this question is with a *follow-up test* such as Tukey's test.

Enter:

```
TukeyHSD(lm0)
```

This will produce a list of upper and lower confidence bounds for the difference between each possible pair of group means. Intervals that **do not** include zero are significant.