ANOVA

This document is to try to clarify some things about ANOVA.

Firstly, what does the ANOVA give us? It provides information about how SST, the total sum of squares is related to SSE, the sum of squared errors, and SSR, the sum of squares explained by the regression in our particular model(s) of interest. $SST \equiv SSR + SSE$. That is, the total variability in the differences between the Y_i and the mean of Y can be divided into variability that can be explained by our model and leftover variability represented by the residuals.

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

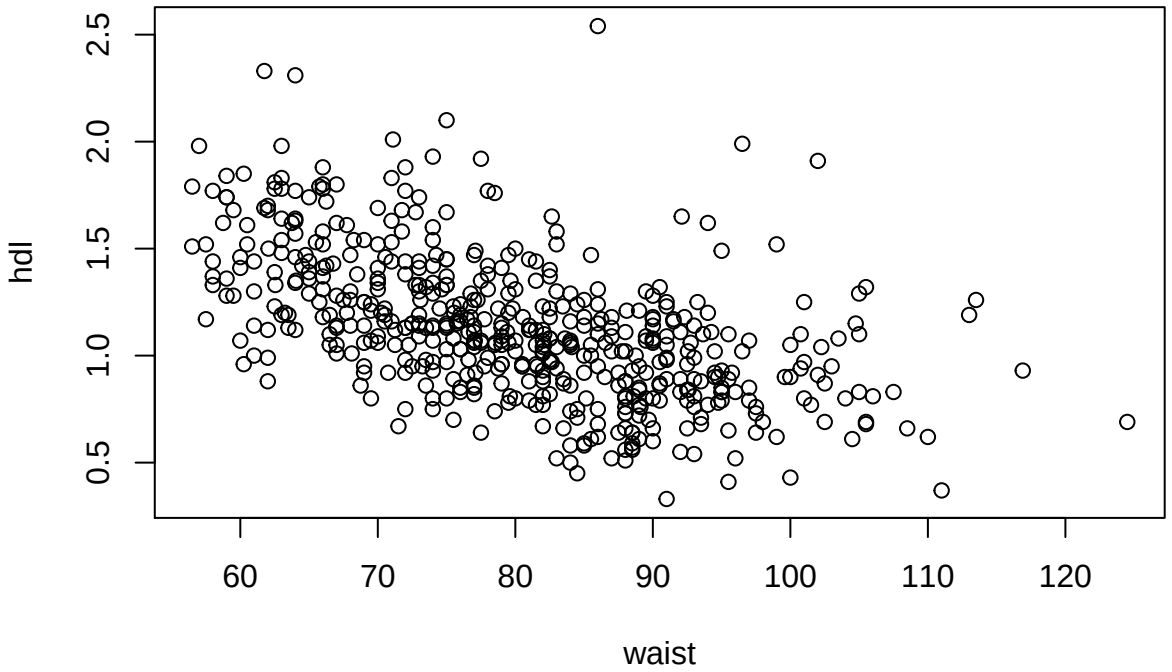
Let's do a demonstration on the SCCS2_v12 dataset.

Simple Linear Regression

Let's say that we are interested in the relationships between HDL cholesterol and waist measurement. HDL is also known as 'good' cholesterol, and higher values are generally better.

First, let's see if this even looks linear:

```
plot(hdl ~ waist, data = SCCS2)
```



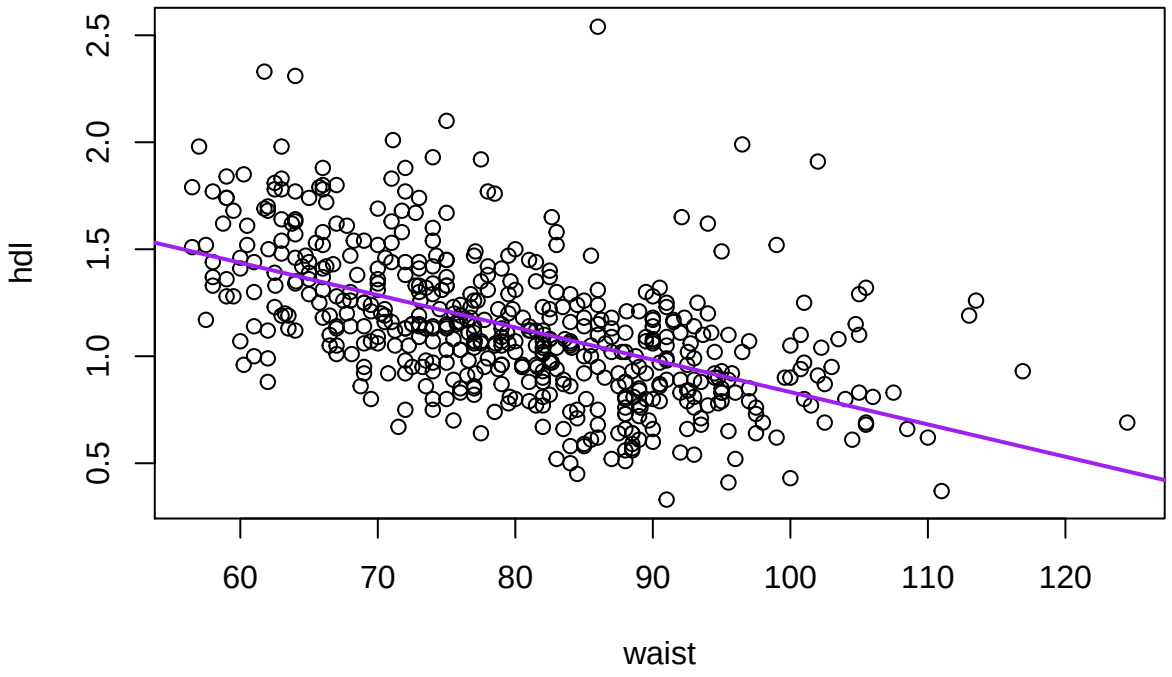
Not bad! There may be some outliers, but this is something we can work with.

Let's do the simple linear regression and plot that line onto our original scatterplot:

```
mod.waist = lm(hdl ~ waist, data = SCCS2)
summary(mod.waist)
```

```
##
## Call:
## lm(formula = hdl ~ waist, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63841 -0.17918 -0.03267  0.16685  1.49604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.343384   0.082239   28.50   <2e-16 ***
## waist       -0.015110   0.001016  -14.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2848 on 525 degrees of freedom
## Multiple R-squared:  0.2964, Adjusted R-squared:  0.2951
## F-statistic: 221.2 on 1 and 525 DF,  p-value: < 2.2e-16
```

```
plot(hdl ~ waist, data = SCCS2)
abline(mod.waist, lwd=2, col = "purple")
```



This is saying that for every 1 cm increase in waist measurement, mean HDL cholesterol decreases by 0.015 units.

Adding a second covariate

Let's add a bit to this model. Gender is a confounder for the relationship between HDL and waist measurement. Why? Because, biologically, women generally have smaller waists than men. Also, women generally have higher HDL levels than men, even after accounting for waist measurement. This means that it meets the criteria of a confounder and should be included in our model. How does it affect things?

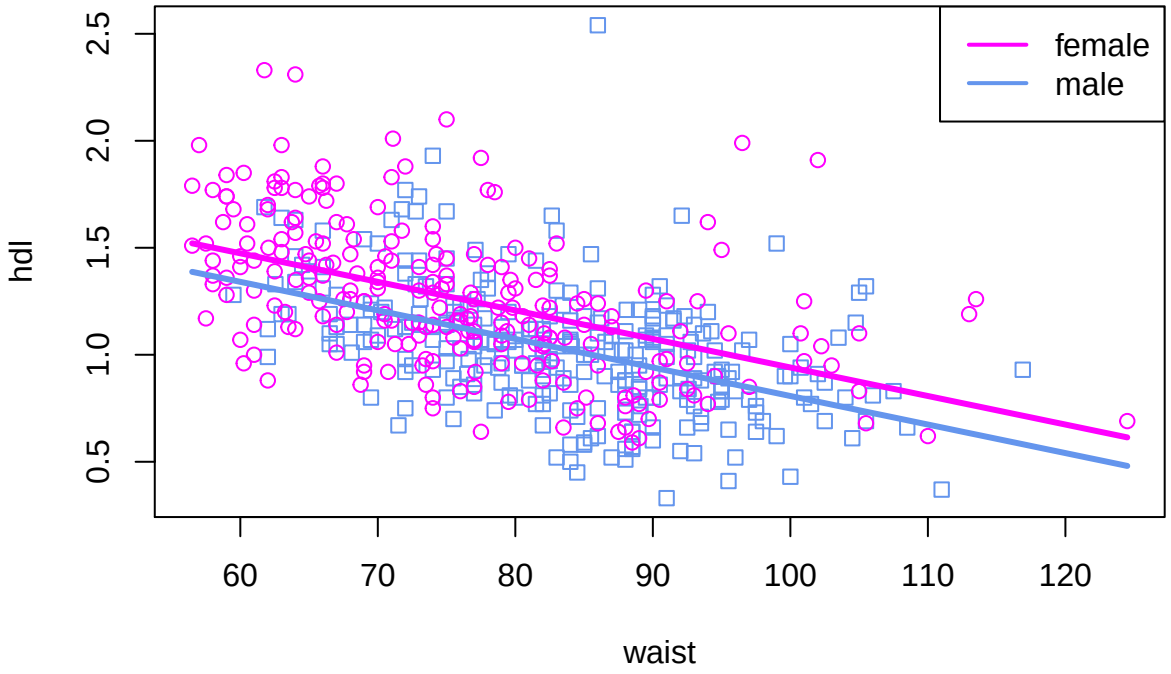
```
mod.waist.gender = lm(hdl ~ waist + gender, data = SCCS2)
summary(mod.waist.gender)
```

```
##
## Call:
## lm(formula = hdl ~ waist + gender, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60078 -0.17252 -0.02402  0.14918  1.54593
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.141204   0.089263  23.988  < 2e-16 ***
## waist       -0.013339   0.001049 -12.716  < 2e-16 ***
## gender       0.133335   0.025729   5.182 3.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.278 on 524 degrees of freedom
## Multiple R-squared:  0.3307, Adjusted R-squared:  0.3282
## F-statistic: 129.5 on 2 and 524 DF,  p-value: < 2.2e-16
```

Note that the estimate for waist changes from -0.015 to -0.013, which is more than a 10% change. This shows that there was bias from not including gender in the model. In the first simple linear regression, every 1 cm increase in waist measurement is associated with a 0.015 unit decrease in mean HDL cholesterol. In the second one, after accounting for gender, every 1 cm increase in waist measurement is associated with a 0.013 unit decrease in mean HDL. Also, female gender is associated with 0.133 unit higher mean HDL cholesterol, after accounting for waist size.

Let's look at this in plots:

```
plot(hdl ~ waist, pch = ifelse(gender == 1, 1, 0),
     col = ifelse(gender == 1,'magenta','cornflowerblue'), data = SCCS2)
curve(coef(mod.waist.gender)[1] + coef(mod.waist.gender)[2]*x +
        coef(mod.waist.gender)[3], col="magenta", lwd=3, add=T)
curve(coef(mod.waist.gender)[1] + coef(mod.waist.gender)[2]*x,
      col = "cornflowerblue", lwd=3, add=T)
legend('topright', lty = 1, lwd = 2, col = c("magenta", "cornflowerblue"),
       legend = c('female', 'male'))
```



Adding a categorical covariate with 3 levels

Now let's include ethnicity as another confounder. Importantly, I first turned it into a categorical variable! R will automatically make this 3-level categorical variable into the necessary indicator/dummy variables, but let's also show that. We will end up with the linear model:

$$\text{HDL}_i = \beta_0 + \beta_1 \text{waist}_i + \beta_2 \text{gender}_i + \beta_3 \text{I(ethnic} = 2)_i + \beta_4 \text{I(ethnic=3)}_i + \epsilon_i$$

```
#first we'll do this in R after turning ethnic into a categorical variable
SCCS2$ethcat = as.factor(SCCS2$ethnic) #makes a categorical variable of ethnicity
summary(SCCS2$ethcat)
```

```
##   1   2   3
## 277 120 130
```

So there are 277 people with ethnicity 1, 120 people with ethnicity 2, and 130 people with ethnicity 3.

```
m3 = lm(hdl ~ waist + gender + ethcat, data = SCCS2)
summary(m3)
```

```
##
## Call:
## lm(formula = hdl ~ waist + gender + ethcat, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60060 -0.17403 -0.02212  0.15258  1.49496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.083516   0.089342  23.321  < 2e-16 ***
## waist       -0.012075   0.001075 -11.238  < 2e-16 ***
## gender       0.145755   0.025484   5.720 1.80e-08 ***
## ethcat2     -0.078532   0.030057  -2.613  0.00924 **
## ethcat3     -0.126371   0.030278  -4.174 3.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2736 on 522 degrees of freedom
## Multiple R-squared:  0.3545, Adjusted R-squared:  0.3495
## F-statistic: 71.65 on 4 and 522 DF,  p-value: < 2.2e-16
```

Looks like ethnicity is a significant predictor, too!

Now I will make two separate indicator variables: one will be true when ethnic = 2, the other when ethnic = 3.

```
# make a variable that is true when ethnic = 2
SCCS2$eth2 = as.factor(SCCS2$ethnic == 2)
summary(SCCS2$eth2)
```

```
## FALSE  TRUE
##   407   120
```

```
#make a new variable that is true when ethnic = 3
SCCS2$eth3 = as.factor(SCCS2$ethnic == 3)
summary(SCCS2$eth3)
```

```
## FALSE  TRUE
##   397   130
```

```
#now include both of these variables into the model
m3.manual = lm(hdl ~ waist + gender + eth2 + eth3, data = SCCS2)
summary(m3.manual)
```

```
##
## Call:
## lm(formula = hdl ~ waist + gender + eth2 + eth3, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60060 -0.17403 -0.02212  0.15258  1.49496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.083516   0.089342  23.321  < 2e-16 ***
## waist       -0.012075   0.001075 -11.238  < 2e-16 ***
## gender       0.145755   0.025484   5.720 1.80e-08 ***
## eth2TRUE    -0.078532   0.030057  -2.613  0.00924 **
## eth3TRUE    -0.126371   0.030278  -4.174 3.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2736 on 522 degrees of freedom
## Multiple R-squared:  0.3545, Adjusted R-squared:  0.3495
## F-statistic: 71.65 on 4 and 522 DF,  p-value: < 2.2e-16
```

```
#additionally, we could have turned ethnic into a factor variable in the lm command:
m3.inline = lm(hdl ~ waist + gender + as.factor(ethnic), data = SCCS2)
```

Wow - that's exactly the same as before!

But what would have happened if we hadn't turned it into a categorical variable?

```
m3.wrong = lm(hdl ~ waist + gender + ethnic, data = SCCS2)
summary(m3.wrong)
```

```
##
## Call:
## lm(formula = hdl ~ waist + gender + ethnic, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59788 -0.17603 -0.02469  0.15540  1.49732
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.143941   0.087776   24.43  < 2e-16 ***
## waist       -0.012053   0.001073  -11.23  < 2e-16 ***
## gender       0.145915   0.025464    5.73 1.70e-08 ***
## ethnic      -0.064672   0.014867   -4.35 1.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2734 on 523 degrees of freedom
## Multiple R-squared:  0.3541, Adjusted R-squared:  0.3504
## F-statistic: 95.58 on 3 and 523 DF,  p-value: < 2.2e-16
```

This makes no sense. This model is saying that, after adjusting for gender, for every 1 unit increase in 'ethnic', mean HDL decreases by 0.065. Make sure that you don't do this!

ANOVA

Now, after all of that, let's finally see what ANOVA can do for us. For comparison purposes, let's run an intercept-only model (which is just the mean).

```
int.only = lm(hdl ~ 1, data = SCCS2)
summary(int.only)
```

```
##
## Call:
## lm(formula = hdl ~ 1, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80425 -0.23425 -0.03425  0.19575  1.40575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.13425    0.01477   76.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3392 on 526 degrees of freedom
```

```
mean(SCCS2$hdl)
```

```
## [1] 1.13425
```

Now, let's get the total sum of squares using R.

```
sst = sum((SCCS2$hdl - mean(SCCS2$hdl))^2)
sst
```

```
## [1] 60.51068
```

And let's run our first ANOVAs! In R, we do this by putting whatever we have saved our model as in the anova command.

```
#now let's look at the model predicting HDL just from waist measurement
anova(mod.waist)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## waist       1 17.938 17.9376   221.2 < 2.2e-16 ***
## Residuals 525 42.573  0.0811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#and compare this to the intercept-only model:
anova(int.only)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 526 60.511 0.11504
```

Note that $SST = SSE + SSR : 60.511 = 17.938 + 42.573$ and that SSE from the intercept-only model is the SST. Why would that be? Because $\hat{Y}_i = \bar{Y}$ in the intercept model! From before, $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. Substituting in $\bar{Y}$ for $\hat{Y}_i$ clearly gives us the same thing. The SSR here is how much of the original variability is explained by the waist variable.

From ANOVA to R squared

What if we now compare this to a 'bigger' model? Let's look at the second model (HDL ~ waist + gender) and compare that to the model with just waist as a covariate:

```
anova(mod.waist.gender)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## waist       1 17.938 17.9376 232.096 < 2.2e-16 ***
## gender      1  2.076  2.0756  26.857 3.132e-07 ***
## Residuals 524 40.497  0.0773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note here that SSR from waist is the same in both: 17.938. But now more of our data is explained by including gender, and SSE has decreased from 42.573 to 40.497.

How is this increase in information reflected in the R^2? We can get the R^2 from the lm summary:

```
summary(mod.waist)$r.squared
```

```
## [1] 0.2964369
```

```
summary(mod.waist.gender)$r.squared
```

```
## [1] 0.3307389
```

So the R-squared is 0.2964 for the waist-only model and 0.3307 for the model with waist and gender.

And for comparison using the SSR and SST from ANOVA:

```
#waist only model
17.938 / 60.511
```

```
## [1] 0.296442
```

```
#waist and gender
(17.938 + 2.076) / 60.511
```

```
## [1] 0.3307498
```

Wow, they're the same! That's because $R^2 = \frac{SSR}{SST}$

Getting SSR and SSE

Now, how do we manually get those individual SSR and SSE numbers from R outputs? SSE is easy: we just add up the squared residuals for our model. For example, in the model including both waist and gender, the sum of the squared residuals is the SSE which can be seen under 'Sum Sq' at the 'Residuals' line and equals 40.497.

```
sum(mod.waist.gender$residuals^2)
```

```
## [1] 40.49744
```

```
anova(mod.waist.gender)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## waist       1 17.938 17.9376 232.096 < 2.2e-16 ***
## gender      1  2.076  2.0756  26.857 3.132e-07 ***
## Residuals 524 40.497  0.0773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR is a little trickier in a multiple regression. The ANOVA for a model containing waist and gender (in that order!) has two SSRs given: one for waist alone, and then the second to represent the advantage we gained by adding gender to the model with waist.

To get these numbers, let's think first about what this means:

$$\begin{aligned} SST &= SSR + SSE \\ &= SSR_{\text{waist}} + SSE_{\text{waist}} \\ &= SSR_{\text{waist}} + SSR_{\text{gender|waist}} + SSE_{\text{waist and gender}} \end{aligned}$$

$$\begin{aligned} \implies SSR_{\text{waist}} &= SST - SSE_{\text{waist}} \\ SSR_{\text{gender|waist}} &= SST - SSR_{\text{waist}} - SSE_{\text{waist and gender}} \end{aligned}$$

That looks like this in R:

```
#remember that mod.waist is a simple linear regression for hdl ~ waist
ssr.waist = sst - sum(mod.waist$residuals^2) #SSR = SST - SSE from SLR
#and that mod.waist.gender is hdl ~ wait + gender
ssr.gender = sst - ssr.waist - sum(mod.waist.gender$residuals^2)

ssr.waist
```

```
## [1] 17.9376
```

```
ssr.gender
```

```
## [1] 2.07564
```

```
anova(mod.waist.gender)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## waist       1 17.938 17.9376 232.096 < 2.2e-16 ***
## gender      1  2.076  2.0756  26.857 3.132e-07 ***
## Residuals 524 40.497  0.0773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we cannot get it by just calculating the SSR from a model for gender alone:

```
mod.gender = lm(hdl ~ gender, data = SCCS2)
ssr.gender.WRONG = sst - sum(mod.gender$residuals^2)

ssr.gender.WRONG
```

```
## [1] 7.517193
```

Why is this wrong? Because when waist is already in the model, gender adds less information than when it is the only predictor. This is because waist and gender are associated with each other and thus have overlapping information.

Also, if we had started with a model of just gender and then looked at the effect of adding waist, we get something different in the ANOVA:

```
anova(mod.gender)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## gender      1  7.517  7.5172  74.472 < 2.2e-16 ***
## Residuals 525 52.993  0.1009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod.gender.waist = lm(hdl ~ gender + waist, data = SCCS2)
anova(mod.gender.waist)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## gender      1  7.517  7.5172  97.266 < 2.2e-16 ***
## waist       1 12.496 12.4960 161.687 < 2.2e-16 ***
## Residuals 524 40.497  0.0773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that mod.waist.gender and mod.gender.waist both have the same total SSR and the same SSE, because they have the same covariates. The difference between them is which model we are comparing to for the SSR attributable to each covariate: for anova(mod.waist.gender), we added gender to a model containing waist, so the waist SSR comes from the model with waist alone and the gender SSR is how much additional information is explained by adding gender. In other words, we compare the model with both waist and gender to the model with waist alone.

For anova(mod.gender.waist), we have added waist to a model containing gender. Here, the gender SSR comes from the simple linear regression with gender alone and the waist SSR is the additional information gained by adding waist to that SLR with only gender.

Multi-level Categorical Variables in ANOVA

Lastly, let's go back to m3, which is the mode where we added ethnicity as a categorical variable.

```
summary(m3)
```

```
##
## Call:
## lm(formula = hdl ~ waist + gender + ethcat, data = SCCS2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60060 -0.17403 -0.02212  0.15258  1.49496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.083516   0.089342  23.321  < 2e-16 ***
## waist       -0.012075   0.001075 -11.238  < 2e-16 ***
## gender       0.145755   0.025484   5.720 1.80e-08 ***
## ethcat2     -0.078532   0.030057  -2.613  0.00924 **
## ethcat3     -0.126371   0.030278  -4.174 3.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2736 on 522 degrees of freedom
## Multiple R-squared:  0.3545, Adjusted R-squared:  0.3495
## F-statistic: 71.65 on 4 and 522 DF,  p-value: < 2.2e-16
```

```
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## waist       1 17.938 17.9376 239.7048 < 2.2e-16 ***
## gender      1  2.076  2.0756  27.7373 2.035e-07 ***
## ethcat      2  1.435  0.7176   9.5889 8.133e-05 ***
## Residuals 522 39.062  0.0748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we get the total information provided by the ethnicity variable. Remember that variables with more than two categories require $n-1$ indicator variables, where n is the number of possible values that variable can take. Here, ethnic can take on three possible values, so it requires two indicator variables to convey all of the information. This means that we have actually added two variables to our model, not just one, by adding ethnic. summary(m3) gives us information about how each ethnicity compares to the baseline ethnicity, but doesn't give us information about the total information provided by ethnicity. ANOVA can do that for us!

Hope that helps!