Interactions

Irfan Kanat

August 11, 2017

So we started with a simple example of a single independent variable. First we increased the number of independent variables. Now we will see how the independent variables can work together to determine dependent variable.

What is the Big Idea?

Sometimes the variables work together (or against each other) and have an effect above and beyond their direct (sometimes also called main) effect. Sometimes the effect of a variable is moderated through another variable. This indirect effect is called an interaction (moderation).

Let us discuss the concept of interaction over an example of height of teenagers. We know the height is a function of age in teenagers. Generally speaking, the older the individual the taller he/she will be. *This is the main effect of age.*

Another factor in height of teenagers is gender. We know, generally speaking, males of our species are taller than females. *That is the main effect of gender.*

Let us say we know that males and females grow at different rates over the years. *The joint effect of gender and age is the interaction effect.*

So if you believe certain variables enhance, or dampen each other above and beyond their individual main effects, you would be interested in interaction effects.

Dataset

I will create two simulated datasets to use on this learning activity. You can safely ignore the next code block. Just run it, and don't worry too much about it. If you want to know more about it just ask me during virtual office hours.

```r
### Create a dataset for development level
set.seed(2017) # Set random number seed for replicability
# Create a simulated dataset
develop <- data.frame(gender = sample(0:1, 100, replace = T),
                      age = sample(10:18, 100, replace = T))
# Calculate the height based on other variables
develop$height <- 100 + develop$gender * -20 + (develop$age-10) * 7 +
  develop$age * develop$gender * 3 + rnorm(mean = 0, sd = 5, 100)

### Create a dataset for rent
# Set random number seed for replicability
set.seed(2017)
# Create a simulated dataset
rent <- data.frame(beds = sample(1:5, 100, replace = T),
                   baths = sample(1:2, 100, replace = T),
                   sqft = round(rnorm(100, mean = 1000, sd = 200)),
                   hood = sample(LETTERS[1:5], 100, replace = T))
```

```
# Calculate the rent based on other variables
rent$rent <- 200 + 150 * rent$beds - .50 * rent$sqft + .25 * rent$sqft * rent$beds+
  50 * rent$baths - 20 * as.numeric(rent$hood) + rnorm(100, mean=0, sd=80)
# Order the dataset by number of beds
rent <- rent[order(rent$beds, rent$sqft), ]
```

The new dataset develop has 100 observations. The dataset includes variables for age, gender and height.

```
summary(develop)
```

```
##      gender          age            height      
##  Min.   :0.00   Min.   :10.00   Min.   : 94.89  
##  1st Qu.:0.00   1st Qu.:11.00   1st Qu.:114.04  
##  Median :1.00   Median :13.00   Median :132.13  
##  Mean   :0.52   Mean   :13.45   Mean   :135.49  
##  3rd Qu.:1.00   3rd Qu.:15.25   3rd Qu.:152.96  
##  Max.   :1.00   Max.   :18.00   Max.   :196.88  
```

The new dataset rent has 100 observations. It shows the apartments for rent around Athens, OH. The dataset includes the number of bed rooms, number of bathrooms, total surface area, and rent in dollars.

```
summary(rent)
```

```
##       beds          baths            sqft        hood        rent        
##  Min.   :1.00   Min.   :1.00   Min.   : 599.0   A:22   Min.   : -83.81  
##  1st Qu.:2.00   1st Qu.:1.00   1st Qu.: 871.8   B:16   1st Qu.: 411.85  
##  Median :3.00   Median :1.00   Median :1006.5   C:18   Median : 898.77  
##  Mean   :2.98   Mean   :1.42   Mean   :1006.7   D:22   Mean   : 898.25  
##  3rd Qu.:4.00   3rd Qu.:2.00   3rd Qu.:1116.5   E:22   3rd Qu.:1376.79  
##  Max.   :5.00   Max.   :2.00   Max.   :1445.0          Max.   :1964.90  
```

How to Estimate Interaction Effects?

Development Dataset

Understanding interaction effects is a little bit easier with binary variables. So I will start with the develop dataset.

Our dependent variable is height of individuals measured in centimeters. We also have age measured in years as an independent variable.

The independent variable gender is a binary variable. It can only take one of two values (physiologically speaking). In this dataset the genders are coded as 0 for females and 1 for males. This coding is sometimes referred to as dummy coding, and these variables are sometimes called dummy variables.

We know the height is a function of age in teenagers. Generally speaking, the older the individual the taller he/she will be.

Another factor in height of teenagers is gender. We know, generally speaking, males of our species are taller than females.

Furthermore, we believe males and females grow at different rates.

Remember, we believe that the rate of growth per year is different between males and females.

1 - Estimate Main Effects Model

It is generally advisable to estimate models of increasing complexity rather than estimating the most complex model all at once. This allows you to observe how various variables act together in models. This is especially crucial in models with interactions.

Start with a main effects model, estimating role of gender and age on height.

$height = \beta_0 + \beta_1 \times gender + \beta_2 \times age + \epsilon$

```
develop_lm_0 <- lm(height ~ gender + age, data = develop)
summary(develop_lm_0)
```

```
##
## Call:
## lm(formula = height ~ gender + age, data = develop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.655  -3.653  -0.119   4.375  12.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5596     3.0278   2.497   0.0142 *
## gender       19.7794     1.1254  17.575   <2e-16 ***
## age           8.7467     0.2239  39.069   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.554 on 97 degrees of freedom
## Multiple R-squared:  0.9559, Adjusted R-squared:  0.955
## F-statistic:  1051 on 2 and 97 DF,  p-value: < 2.2e-16
```

Here we see that our model significantly explains 96 % of the variance (F(2,97)=1050.54, p<0.001). THE HIGH R^2 IS NOT SURPRISING GIVEN THIS IS A SIMULATED DATASET BUT IN REAL LIFE YOU WOULD RARELY SEE R^2 THAT HIGH.

We see that all three of our variables of interest are statistically significant. **We should just skip the interpretation of the coefficient estimates as they will doubtlessly change in the interaction model. I will still interpret the coefficients, as I want to demonstrate how binary variable is interpreted and what the interaction terms add to the model interpretation.**

We notice that males (gender=1) are about 19.78 cm taller than females. **Go back to the formula and note $\beta_1 \times gender$. Since gender = 0 for females, β_1 has no effect on females.**

Finally we see that each year, the individuals grow about 9 cm.

2 - Estimate the Interaction Model

Since both our variables are significant in the main effects model, we can go ahead and investiage the interactions. We could not investigate interactions if either one of our main effects were insignificant.

Now let's estimate the interaction model:

$height = \beta_0 + \beta_1 \times gender + \beta_2 \times age + \beta_3 \times gender \times age + \epsilon$

```
develop_lm_1 <- lm(height ~ gender + age + gender * age, data = develop)
```

Note that I used asterix (*) to create an interaction term.

```r
summary(develop_lm_1)
```

```
##
## Call:
## lm(formula = height ~ gender + age + gender * age, data = develop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.351  -3.207  -0.039   2.923  10.783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.3706     3.3056   8.280 7.24e-13 ***
## gender      -18.7426     4.6733  -4.011  0.00012 ***
## age           7.2277     0.2491  29.017  < 2e-16 ***
## gender:age    2.8723     0.3425   8.386 4.30e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.241 on 96 degrees of freedom
## Multiple R-squared:  0.9745, Adjusted R-squared:  0.9737
## F-statistic:  1224 on 3 and 96 DF,  p-value: < 2.2e-16
```

Since there are interaction effects, I can not interpret the main effect of age or gender by themselves. We will talk about how to estimate the effect size later, but for now let us focus on gender x age interaction.

Very often we would see one or both main effects turn insignificant here (you will see why in our discussion of multicollinearity). As long as interaction term is significant, you need not worry about main effects turning insignificant.

Going back to formula, you can calculate height for females as follows:

$height = \beta_0 + \beta_1 \times 0 + \beta_2 \times age + \beta_3 \times age \times 0 + \epsilon$
$height = \beta_0 + \beta_2 \times age + \epsilon$

Remember that gender=0 for females so any coefficients that have gender in their formula don't apply (multiplied by 0) to females.

So for males, the height would be calculated as:

$height = \beta_0 + \beta_1 \times 1 + \beta_2 \times age + \beta_3 \times age \times 1 + \epsilon$
$height = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \times age + \epsilon$

You can say $\beta_0 + \beta_1$ is the intercept for males and $\beta_2 + \beta_3$ is the slope of the age for males.

Basically males and females have different intercepts and slopes. Based on above calculations, we can say the females start off at $\beta_0 = 27.37$ and grow about $\beta_2 = 7.23$ cm every year. Males on the other hand start off at $\beta_0 + \beta_1 = 8.63$ and grow about $\beta_2 + \beta_3 = 10.1$ cm each year.

Going with numerical exploration:

At age 10, the height of a female would be:

```r
predict(develop_lm_1, newdata = data.frame(gender = 0, age = 10))
```

```
##        1
## 99.64719
```

Read the manual for predict function. Basically predict takes a model and a dataset and produces estimated values for dependent variables.

At age 10, the height of a male would be:

```
predict(develop_lm_1, newdata = data.frame(gender = 1, age = 10))
```

```
##        1
## 109.6274
```

At age 18, the height of a female would be:

```
predict(develop_lm_1, newdata = data.frame(gender = 0, age = 18))
```

```
##        1
## 157.4685
```

At age 18, the height of a male would be:

```
predict(develop_lm_1, newdata = data.frame(gender = 1, age = 18))
```

```
##       1
## 190.427
```

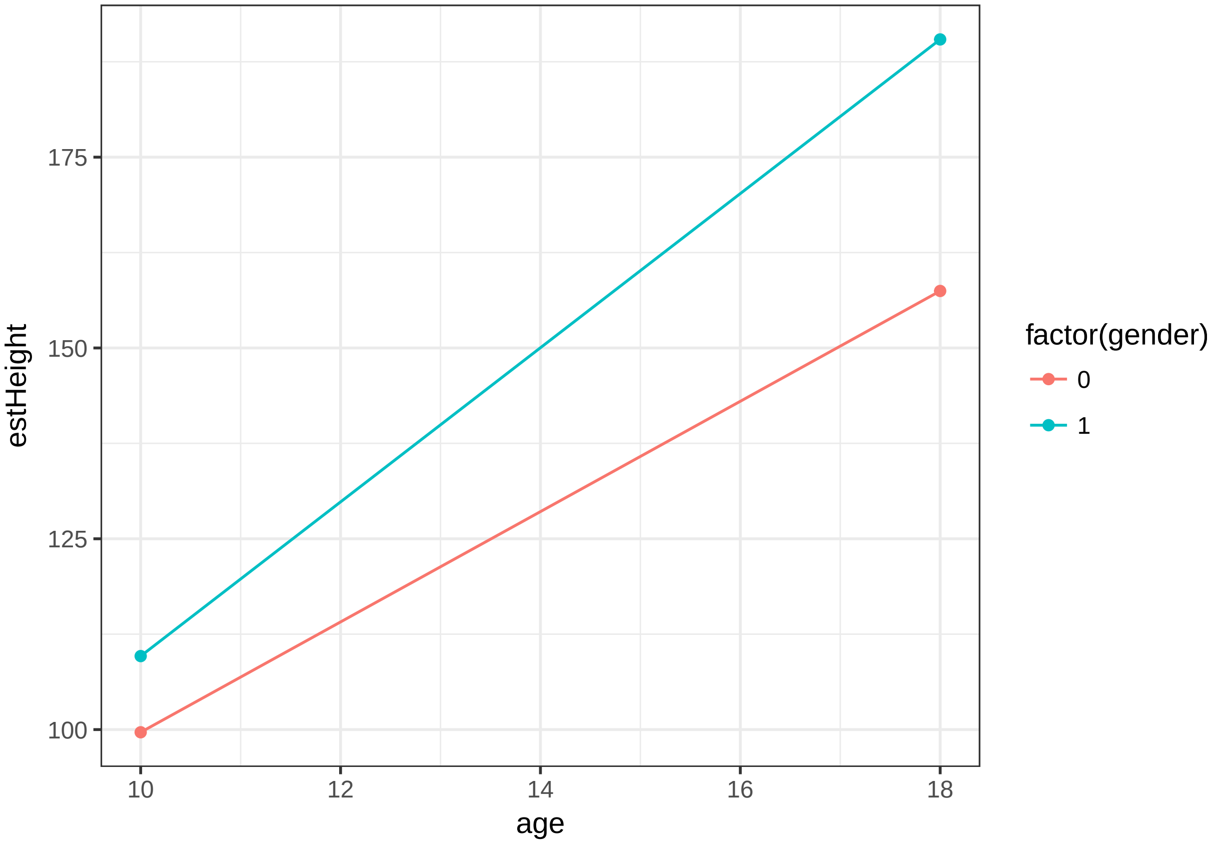
Or you can predict multiple values at once such as:

```
newDevelop <- data.frame(gender = c(0, 1, 0, 1), age = c(10, 10, 18, 18))
newDevelop$estHeight <- predict(develop_lm_1, newdata = newDevelop)
newDevelop
```

```
##   gender age estHeight
## 1      0  10  99.64719
## 2      1  10 109.62739
## 3      0  18 157.46848
## 4      1  18 190.42695
```

Let us visualize this effect:

```
qplot(x = age, y = estHeight, col = factor(gender), data = newDevelop) +
  geom_line() +
  theme_bw()
```



Easiest way to tell if there is a significant interaction is to look at the lines. If the lines are parallel, there is no significant interaction. In our case the lines are not parallel, hence the interaction is significant.

Rent Dataset

In the rent dataset we have two continuous variables we want to interact.

Talking about rent, let us assume we believe peoples preference for size of the house changes with the number of rooms. For example, let us say for a studio apartment, size beyond a certain limit actually decreases the attraction (An empty warehouse is not a very good bedroom). Whereas for houses with more rooms the bigger the area, the better. We believe there is a relation between room numbers and area that determines rent above and beyond the area or room numbers by themselves. This means we need an interaction model.

1- Estimate Main Effects Model

It is generally advisable to estimate models of increasing complexity rather than estimating the most complex model all at once. This is crucial in models with interactions. Start with a main effects model.

```r
rent_lm_0 <- lm(rent ~ beds + baths + sqft, data = rent)
summary(rent_lm_0)
```

```
##
## Call:
## lm(formula = rent ~ beds + baths + sqft, data = rent)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -289.485  -66.823   -7.703   76.096  178.084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -682.76309   67.34059 -10.139  < 2e-16 ***
## beds         406.25130    6.74765  60.206  < 2e-16 ***
## baths         50.05672   19.82308   2.525   0.0132 *
## sqft           0.29731    0.05829   5.100 1.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.18 on 96 degrees of freedom
## Multiple R-squared:  0.9747, Adjusted R-squared:  0.9739
## F-statistic:  1234 on 3 and 96 DF,  p-value: < 2.2e-16
```

Here we see that our model significantly explains 97 % of the variance (F(3,96)=1234.25, p<0.001).. THE HIGH R^2 IS NOT SURPRISING GIVEN THIS IS A SIMULATED DATASET BUT IN REAL LIFE YOU WOULD RARELY SEE R^2 THAT HIGH.

We see that all three of our variables of interest are statistically significant.

Roughly speaking, we see that each bedroom adds $ 406 to the rent, while each bathroom adds $ 50. Each square feet increases the rent by about 30 cents.

2 - Integrate the Interaction Effects

Remember we believe there is a relation between the area and the number of bedrooms. ***Since both main effects were significant in the previous model we can proceed with adding interactions.*** If the main effects were insignificant we could not proceed to second step.

```
rent_lm_1 <- lm(rent ~ beds + baths + sqft + beds * sqft, data = rent)
summary(rent_lm_1)
```

```
##
## Call:
## lm(formula = rent ~ beds + baths + sqft + beds * sqft, data = rent)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -193.452  -55.643   -2.999   55.843  168.356
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.59554  113.08206   0.377 0.707253
## beds        162.76536   33.76066   4.821 5.41e-06 ***
## baths        60.95497   16.01362   3.806 0.000250 ***
## sqft         -0.43382    0.11050  -3.926 0.000163 ***
## beds:sqft     0.24122    0.03301   7.307 8.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.16 on 95 degrees of freedom
## Multiple R-squared:  0.9838, Adjusted R-squared:  0.9831
## F-statistic:  1444 on 4 and 95 DF,  p-value: < 2.2e-16
```

This model significantly explains 98% of the variance (F(4,95)=1444.26, p<0.001).

We notice that the coefficients for all main effects and the one interaction we introduced are significant. At this stage, we can note that even if the main effects turn insignificant in the second model we can proceed. *If* the main effects were insignificant, we would have said "the main effects were fully mediated through the interaction effect." Luckily, this is a simple case where all variables are significant.

You need to note that, in an interaction model **you can't just report the main effects.** You need to take into account the interaction in any form of reporting you take on.

For now let us simply look at the sign of the interaction effect. The sign of beds x sqft interaction is positive. This means that the number of rooms and size of the house reinforce each other. The bigger the better above and beyond the main effects (note the main effect of sqft actually turned negative here). But how big is the interaction effect? Is it significant enough to counter the negative main effect of the sqft?

An easy way to understand the interaction effect is to calculate the estimates at various points to see the interaction effects. We will set the interacted variables at low and high levels and calculate rent estimates. We will fix the other variables in the model at mean levels (everything else being constant).

A - Let us hold the baths at mean and sqft at first quartile (%25).

```
sqft <- unname(quantile(rent$sqft)[2]) # check out the output of quantile function
baths <- mean(rent$baths) # fixing baths at mean values
beds <- 1:5 # Varying beds from 1 to 5
# Check out what data.frame(beds, baths, sqft) does
atQ1 <- predict(rent_lm_1, newdata = data.frame(beds, baths, sqft))
```

B - Let us hold the sqft and baths constant at mean levels (%50).

```
sqft <- mean(rent$sqft) # mean level
baths <- mean(rent$baths) # mean level
beds <- 1:5 # varying 1 to 5
# estimating rent based on new values
atMean <- predict(rent_lm_1, newdata = data.frame(beds, baths, sqft))
```

C - Let us hold baths at mean values and sqft at third quartile (%75).

```
sqft <- unname(quantile(rent$sqft)[4]) # 3rd quartile
baths <- mean(rent$baths) # mean level
beds <- 1:5 # varying
# estimating rent
atQ3 <- predict(rent_lm_1, newdata = data.frame(beds, baths, sqft))
```

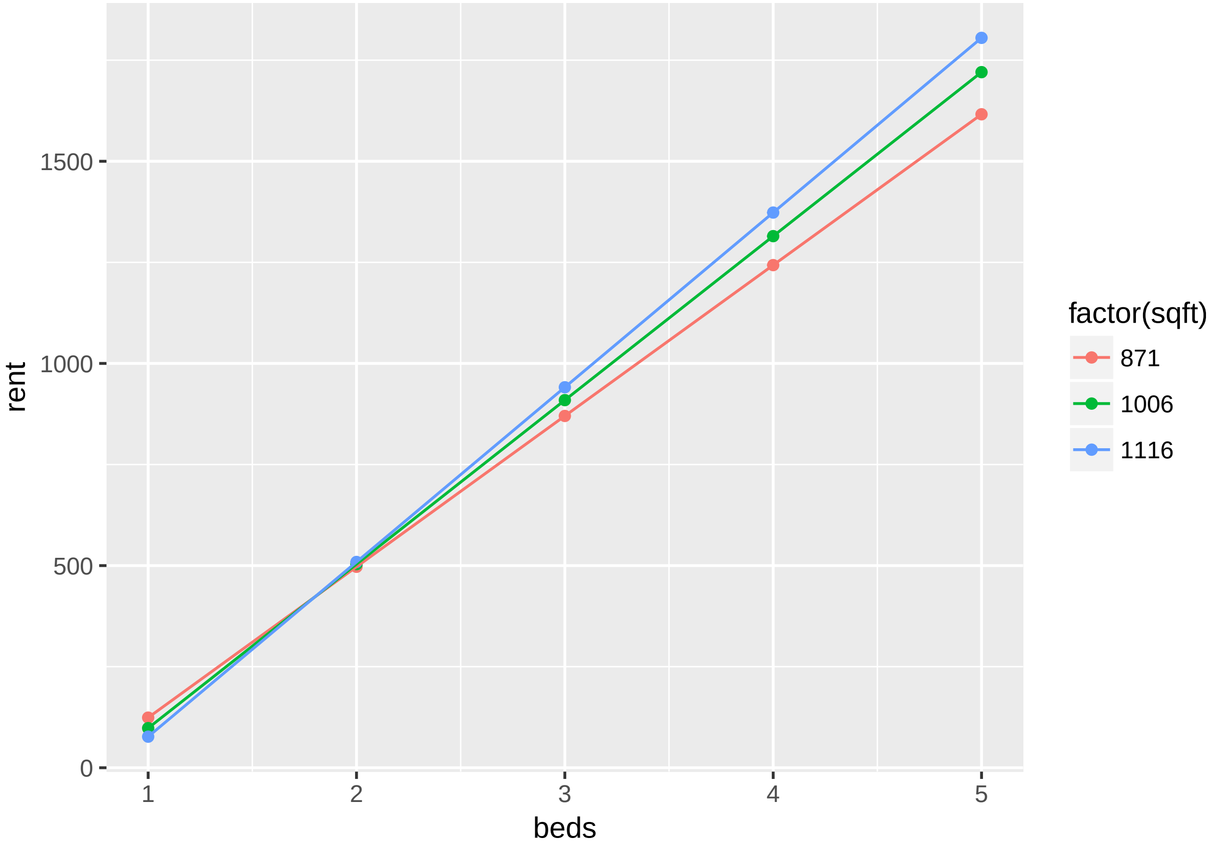
Let us compare the results of varying levels of bed rooms and sqare footage

```
## Creating a data frame to hold estimates
# This gets the rent estimates from previous estimations
estimates<-data.frame(rent = c(atQ1, atMean, atQ3),
                      # This puts the correct number of beds next to each estimate
                      beds = rep(1:5, 3),
                      # This is the sqft used in estimation (Q1, Mean, Q3)
                      sqft = rep(c(871, 1006, 1116), each=5))
# Alternatively you can create a new dataset and save predicted rent values as we did in the develop da
estimates # Let us inspect the results
```

```
##          rent beds sqft
## 1   124.01616    1  871
## 2   497.06654    2  871
## 3   870.11692    3  871
## 4  1243.16729    4  871
## 5  1616.21767    5  871
## 6    98.02452    1 1006
## 7   503.62777    2 1006
## 8   909.23101    3 1006
## 9  1314.83425    4 1006
## 10 1720.43750    5 1006
## 11   76.87682    1 1116
## 12  508.96621    2 1116
## 13  941.05559    3 1116
## 14 1373.14498    4 1116
## 15 1805.23436    5 1116
```

Looking at 15 rows of 3 columns and trying to get a meaning out of it is hard. Let us visualize.

```
qplot(data = estimates, x = beds, y = rent, col = factor(sqft)) + geom_line()
```



Easiest way to tell if there is a significant interaction is to look at the lines. If the lines are parallel, there is no significant interaction. In our case the lines are not parallel, hence the interaction is significant.

Now let's interpret the interaction. This is a bit confusing, you get better as you go.

Assuming price reflects preferences: For 1 bedroom houses, people actually prefer cozier (smaller) houses. The effect of sqft is reversed for 3 and more bedrooms, where people actually prefer spacious houses. This reversal in preferences is what we captured in interaction terms.

What is Next?

In the next learning activity we will learn about diagnostics of OLS models.

House Keeping

Forgive our dust, here I export some models so we can import them back later.

```r
save(list = ls(pattern = "develop*"), file = "data/develop_fit")
save(list = ls(pattern = "rent*"), file = "data/rent_fit")
```