Lab 04

Jaime Montana

24/9/2021

Reminder of last session

- Opening CSV files with `data.table`'s `fread()` function.
- Intro to `ggplot2` package. (Syntax)
- Visual inspection of two (three) variables
- Beyond visual inspection: correlation
- Beyond visual inspection: linear regression
- Output from linear regression

Libraries to use in todays lab.

Exercise: Use the `library()` command to load the following libraries in your session. - `data.table` - `stargazer` - `ggplot2` - `doBy`

Tip: Recall to install the package so the library is available to be used by Rstudio.

Libraries to use in todays lab.

Exercise: Use the `library()` command to load the following libraries in your session. - `data.table` - `stargazer` - `ggplot2`

Tip: Recall to install the package so the library is available to be used by Rstudio.

```
#install.packages("data.table")
#install.packages("stargazer")
#install.packages("ggplot2")
#install.packages("doBy")
library(data.table)
library(stargazer)
library(ggplot2)
library(doBy)
```

Kitchen recipe

0. Define a question (feasible, relevant, interesting)
1. What are the available variables? Which of them are useful for your analysis?
   - read the codebook
   - inspect visually the head and bottom of your data.
2. Define a sample for your analysis
3. What is the model that allows me to evaluate my hypothesis using the available data?
4. Visual inspection (Plots)
5. Summary statistics (your sample, and relevant sub-samples)
6. Correlations
7. Testing your model: linear regression
8. Interpreting all the *relevant* outputs

Case Study:

The DirectMarketing data set shows data from a direct marketer. The direct marketer sells her products (e.g. clothing, books, or sports gear) via direct mail exclusively; she sends catalogs with product descriptions to her customers, and the customers order directly from the catalogs.

She is interested in mining her customers' data in order to better customize the marketing process. In particular, she is interested in understanding **what factors drive some customers to spend more money than others**.

The dataset:

Customer records:

- `Age`: young, middle, and old
- `Gender`: female/male
- `OwnHome`: own home or rented home
- `Married`: single or married
- `Location`: whether the customer is close or far from the nearest brick-and-mortar store selling similar products
- `Salary`: yearly salary (in US dollars)
- `Children`: how many children the customer has (between 0 and 3)
- `History`: past purchasing history (low, medium, or high, or NA if the customer has not purchased anything in the past)
- `Catalogs`: the number of catalogs she has sent to that customer
- `amountspent`: the amount of money the customer has spent (in US dollars)

Some intuition

What could drive a customer spending?

- Earnings
- Considering were the customer lives.
- ...

Exercise

Recall: in order to evaluate the hypothesis, we need to establish a model (the relationship) that we ant to estimate.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- In the left hand side y_i is the dependent variable. - The right hand side of the model contains the intercept , the independent variables (x_i), and the error term (ϵ_i).

The values β_k are k values to be estimated.

Propose a model (using the available variables) that allow you to evaluate the question of the case study.

Proposed model

- Dependent variable we use `amountspent`
- Independent variables: `salary` or `location`, or ...

Then the first model would be:

$$amountspent_i = \beta_0 + \beta_1 salary_i + \epsilon_i$$

And the second model:

$$amountspent_i = \beta_0 + \beta_1 location_i + \epsilon_i$$

...

Import only the desired variables load()

```
load(file = "direct_marketing.RData") # data.frame
dt.marketing <- data.table(dt.mktg)   # transform in DT
rm(dt.mktg)                           # Remove object
setnames(dt.marketing, tolower(names(dt.marketing))) #lowe
```

Inspect visually the data

Three commands:

- `head()` and `tail()`
- `summary()`
- `stargazer()`

```
head(dt.marketing)
```

```
##       age gender ownhome married location salary children history catalogs
## 1:    Old Female     Own  Single      Far  47500        0    High        6
## 2: Middle   Male    Rent  Single    Close  63600        0    High        6
## 3:  Young Female    Rent  Single    Close  13500        0     Low       18
## 4: Middle   Male     Own Married    Close  85600        1    High       18
## 5: Middle Female     Own  Single    Close  68400        0    High       12
## 6:  Young   Male     Own Married    Close  30400        0     Low        6
##    amountspent
## 1:         755
## 2:        1318
## 3:         296
## 4:        2436
## 5:        1304
## 6:         495
```

Inspect visually the data

```
summary(dt.marketing)
```

```
##      age         gender     ownhome       married     location
##  Middle:508   Female:506   Own :516   Married:502   Close:710
##  Old   :205   Male  :494   Rent:484   Single :498   Far  :290
##  Young :287
##
##
##
##      salary          children        history        catalogs      amountspent
##  Min.   : 10100   Min.   :0.000   High  :255   Min.   : 6.00   Min.   :  38.0
##  1st Qu.: 29975   1st Qu.:0.000   Low   :230   1st Qu.: 6.00   1st Qu.: 488.2
##  Median : 53700   Median :1.000   Medium:212   Median :12.00   Median : 962.0
##  Mean   : 56104   Mean   :0.934   NA's  :303   Mean   :14.68   Mean   :1216.8
##  3rd Qu.: 77025   3rd Qu.:2.000                3rd Qu.:18.00   3rd Qu.:1688.5
##  Max.   :168800   Max.   :3.000                Max.   :24.00   Max.   :6217.0
```

Inspect visually the data

```r
stargazer(dt.marketing, type = "text")
```

```
##
## ========================================================================
## Statistic     N      Mean     St. Dev.   Min   Pctl(25) Pctl(75)   Max
## ------------------------------------------------------------------------
## salary      1,000 56,103.900 30,616.310 10,100  29,975   77,025  168,800
## children    1,000   0.934      1.051      0       0        2        3
## catalogs    1,000   14.682     6.623      6       6        18      24
## amountspent 1,000 1,216.770   961.069     38    488.2   1,688.5   6,217
## ------------------------------------------------------------------------
```

Why are the two results different?

Inspect visually the data

```r
stargazer(dt.marketing,
          type = "text",
          nobs = FALSE,
          mean.sd = TRUE,
          median = TRUE,
          iqr = TRUE,
          no.space = TRUE)
```

```
##
## ============================================================================
## Statistic      Mean     St. Dev.   Min   Pctl(25) Median Pctl(75)   Max
## ----------------------------------------------------------------------------
## salary      56,103.900 30,616.310 10,100  29,975  53,700  77,025  168,800
## children      0.934      1.051      0       0       1       2        3
## catalogs      14.682     6.623      6       6       12      18      24
## amountspent 1,216.770   961.069     38    488.2    962   1,688.5   6,217
## ----------------------------------------------------------------------------
```

Cross tabulations (means by group)

```r
#mean salary by age group
summaryBy(salary ~ age,    # V. to describe ~ variable to groupBy
          data=dt.marketing,    # name of data object
          FUN = c(mean,min,max),# The function for summary statistics
          na.rm=TRUE)           # Remove missing observations
```

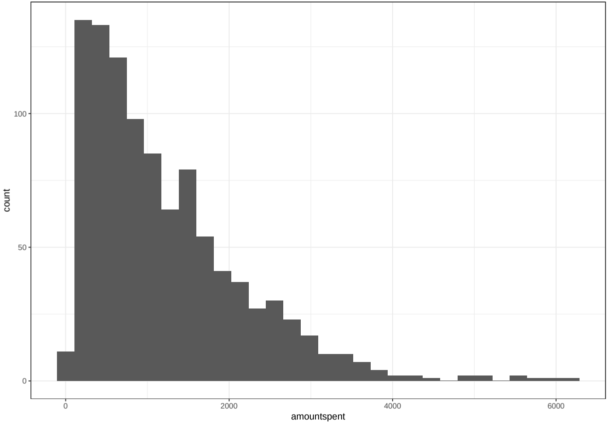
```
##       age salary.mean salary.min salary.max
## 1: Middle    72036.42      25300     140700
## 2:    Old    56365.85      10100     168800
## 3:  Young    27715.68      10200      80700
```

Exercise: calculate the mean salary grouping by `history`. Use the NA.

Visual inspection (plots)

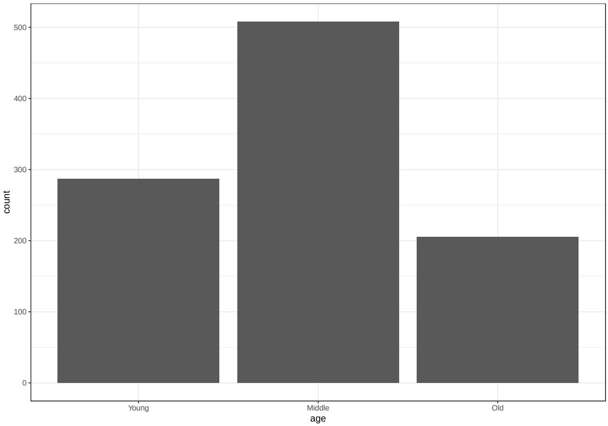
Last class we use an **histogram** to visualize the distribution of one variable. Is always good to see the distribution of hour dependent variable.

```
ggplot(data = dt.marketing,            # plot layer with data
       aes(x = amountspent)) +         # mapping if common to all layers
  geom_histogram() +                   # Type of graph
  theme_bw()                           # Theme of the plot
```



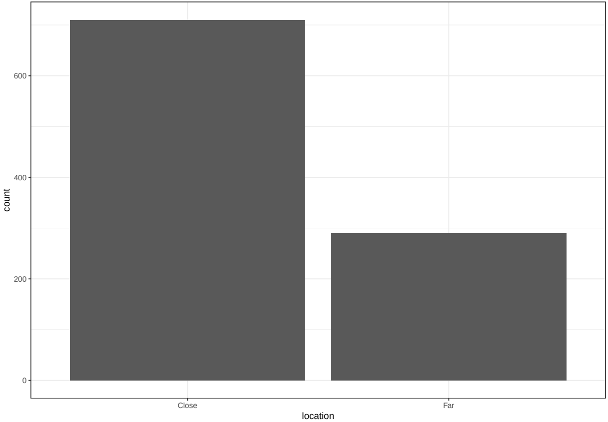
Bar plot

```
ggplot(data = dt.marketing,          # plot layer with data
       aes(x = age)) +          # mapping
  geom_bar() +                  # Type of graph
  xlim("Young","Middle","Old") +# Order of categories
  theme_bw()                    # Theme of the plot
```



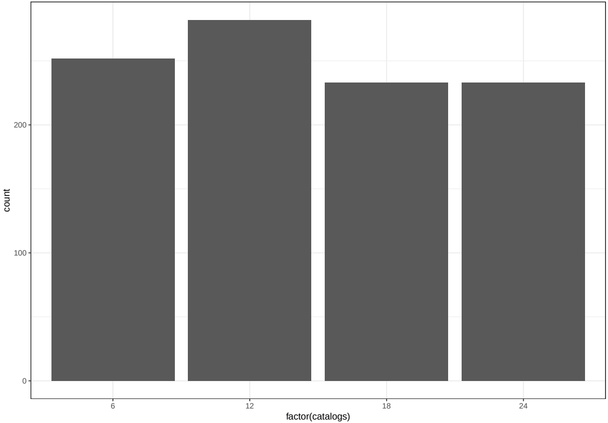
Bar plot

```
ggplot(data = dt.marketing,            # plot layer with data
       aes(x = location)) +         # mapping
  geom_bar() +                    # Type of graph
  theme_bw()                      # Theme of the plot
```



Bar plot

```
ggplot(data = dt.marketing,           # plot layer with data
       aes(x = factor(catalogs))) +        # mapping
  geom_bar() +                  # Type of graph
  theme_bw()                    # Theme of the plot
```



Boxplot

source

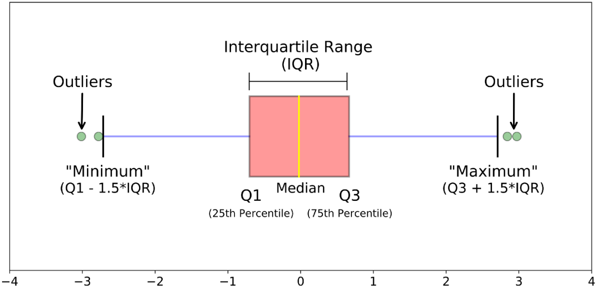
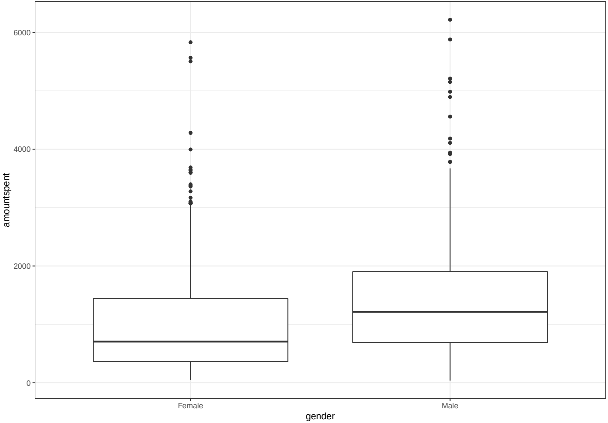


Figure 1: How to read a boxplot

Using boxplots on different allow us to explore different **customer segments**.

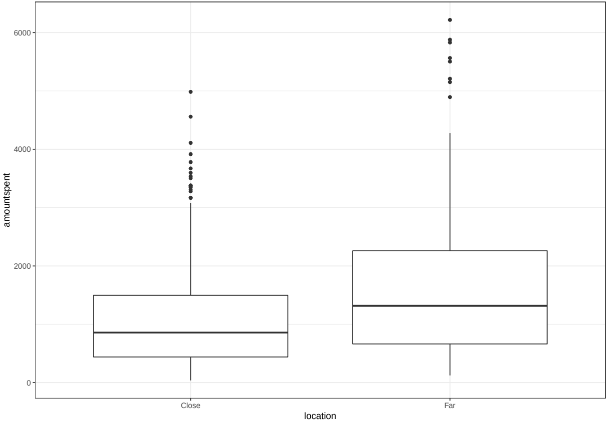
Visual inspection - Boxplot - Spending (Age segment)

```
ggplot(data = dt.marketing,           # plot layer with data
       aes(x = gender, y = amountspent)) +        # mapping
  geom_boxplot() +                 # Type of graph
  theme_bw()                   # Theme of the plot
```



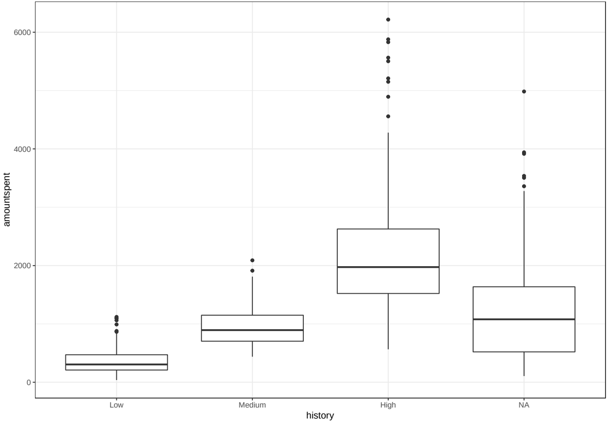
Visual inspection - Boxplot - Spending (Location segment)

```
ggplot(data = dt.marketing,            # plot layer with data
       aes(x = location, y = amountspent)) +        # mapping
  geom_boxplot() +                  # Type of graph
  theme_bw()                  # Theme of the plot
```



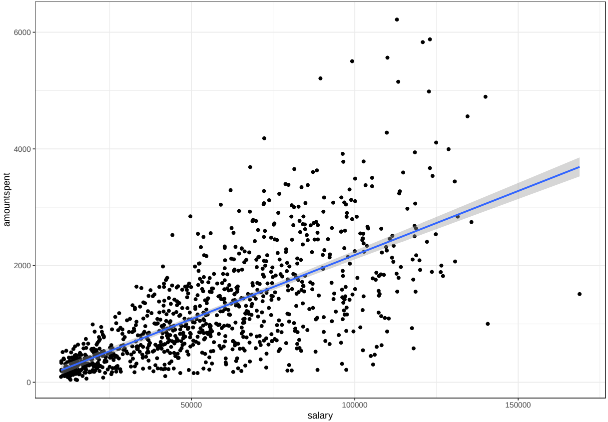
Visual inspection - Boxplot - Spending (History segment)

```
ggplot(data = dt.marketing,           # plot layer with data
       aes(x = history, y = amountspent)) +        # mapping
  geom_boxplot() +                # Type of graph
  xlim("Low", "Medium", "High", NA) +# Levels on `x`
  theme_bw()                    # Theme of the plot
```



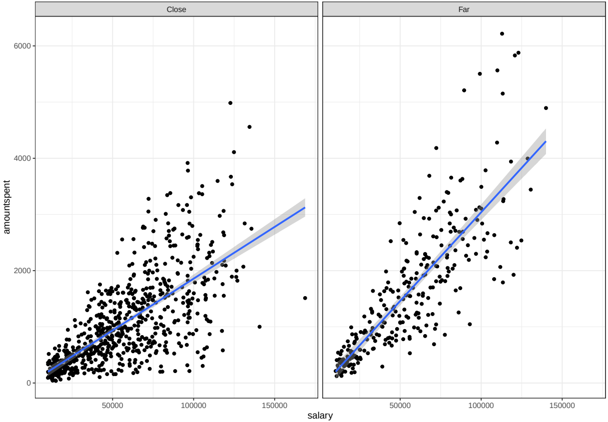
Visual inspection - scatterplot (relation between variables)

```
ggplot(data = dt.marketing,           # plot layer with data
       aes(x = salary, y = amountspent)) +        # mapping
  geom_point() +                  # Type of graph
  theme_bw()    +                # Theme of the plot
  geom_smooth(method = "lm")    # git a linear model and draw regression line
```



Visual inspection - scatterplot (relation between variables)

```
ggplot(data = dt.marketing,           # plot layer with data
       aes(x = salary, y = amountspent)) +        # mapping
  geom_point() +                  # Type of graph
  theme_bw()    +                # Theme of the plot
  geom_smooth(method = "lm")  +  # git a linear model and draw regression line
  facet_grid( ~ location)
```



Beyond visual inspections - Linear regression (OLS)

CASE 1: - y_i, the dependent variable is **continuous** - x_i, the dependent variable is **continuous**

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{salary}_i + \epsilon$$

```
formula_case1 <- as.formula(amountspent ~ salary)   # Define the formula of the model
lm.case1   <- lm(formula = formula_case1,
                data = dt.marketing)
stargazer(lm.case1, type = "text", no.space = TRUE)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             amountspent
## -----------------------------------------------
## salary                       0.022***
##                               (0.001)
## Constant                     -15.318
##                              (45.374)
## -----------------------------------------------
## Observations                  1,000
## R2                            0.489
## Adjusted R2                   0.489
## Residual Std. Error     687.065 (df = 998)
## F Statistic          956.694*** (df = 1; 998)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Beyond visual inspections - Linear regression (OLS)

CASE 2: - y_i, the dependent variable is **continuous** - x_i, the dependent variable is **categorical**

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{location}_i + \epsilon$$

```r
formula_case2 <- as.formula(amountspent ~ location)   # Define the formula of the model
lm.case2   <- lm(formula = formula_case2,
                data = dt.marketing)
stargazer(lm.case2, type = "text", no.space = TRUE)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             amountspent
## -----------------------------------------------
## locationFar                 534.773***
##                              (64.837)
## Constant                   1,061.686***
##                              (34.916)
## -----------------------------------------------
## Observations                  1,000
## R2                            0.064
## Adjusted R2                   0.063
## Residual Std. Error     930.364 (df = 998)
## F Statistic           68.028*** (df = 1; 998)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Beyond visual inspections - Linear regression (OLS)

$\beta_0 = 1,061.686$ is the average amount spent by customers who are "close" (where "close"" is the omitted category of the variable `location`). You can confirm this by computing it directly from the sample.

```
dt.marketing[location == "Close", mean(amountspent)]
```

```
## [1] 1061.686
```

$\beta_1 = 534.7736$. By adding $\beta_0 + \beta_1$ we get the average amount spent by customers who are "far". You can confirm this by calculating the mean.

```
dt.marketing[location == "Far", mean(amountspent)]
```

```
## [1] 1596.459
```

Beyond visual inspections - Linear regression (OLS)

CASE 2A: - y_i, the dependent variable is **continuous** - x_i, the dependent variable is **categorical** (*more than 2 categories*)

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{history}_i + \epsilon$$

```
formula_case2A <- as.formula(amountspent ~ history)   # Define the formula of the model
lm.case2A   <- lm(formula = formula_case2A,
                data = dt.marketing)
stargazer(lm.case2A, type = "text", no.space = TRUE)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             amountspent
## -----------------------------------------------
## historyLow                 -1,829.050***
##                              (56.917)
## historyMedium              -1,235.736***
##                              (58.174)
## Constant                   2,186.137***
##                              (39.196)
## -----------------------------------------------
## Observations                    697
## R2                             0.610
## Adjusted R2                    0.608
## Residual Std. Error     625.902 (df = 694)
## F Statistic         541.884*** (df = 2; 694)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Multiple regression

I can also define a model that have multiple independent variables.

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{location}_i + \beta_2 \times \text{salary}_i + \beta_3 \times \text{children}_i + \beta_4 \times \text{catalogs}_i + \epsilon$$

```r
formula_m1 <- as.formula(amountspent ~ location + salary + children + catalogs)   # Define the formula of
lm.spend1   <- lm(formula = formula_m1,
                 data = dt.marketing)
stargazer(lm.spend1, type = "text", no.space = TRUE)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             amountspent
## -----------------------------------------------
## locationFar                 508.076***
##                              (36.217)
## salary                       0.021***
##                               (0.001)
## children                   -203.479***
##                              (15.625)
## catalogs                     42.719***
##                               (2.544)
## Constant                   -539.806***
##                              (49.592)
## -----------------------------------------------
## Observations                   1,000
## R2                             0.715
## Adjusted R2                    0.714
## Residual Std. Error     514.246 (df = 995)
## F Statistic         623.563*** (df = 4; 995)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Interpretation:

The interpretation of the coefficients:

- Continuous variables: the coefficient gives you the unit change in the expected value of your dependent variable that results from a unit change in your independent variable, *ceteris paribus*.
- Categorical variables: The coefficients of dummy variables tell you how people in that category behave differently from people in the corresponding omitted category, *ceteris paribus*.

Multiple regression

Model 2

I can also define a model that have multiple independent variables.

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{location}_i + \beta_2 \times \text{salary}_i + \beta_3 \times \text{children}_i + \beta_4 \times \text{catalogs}_i + \beta_5 \times \text{history}_i + \epsilon$$

```
formula_m2 <- as.formula(amountspent ~ location + salary + children + catalogs + history)   # Define the f
lm.spend2   <- lm(formula = formula_m2,
                data = dt.marketing)
stargazer(lm.spend1, lm.spend2, type = "text", no.space = TRUE)
```

Model 2

```
##
## =====================================================================
##                                  Dependent variable:
##                      ------------------------------------------------
##                                      amountspent
##                                (1)                      (2)
## ---------------------------------------------------------------------
## locationFar                 508.076***               615.362***
##                              (36.217)                 (43.776)
## salary                       0.021***                 0.018***
##                              (0.001)                  (0.001)
## children                   -203.479***              -274.246***
##                              (15.625)                 (22.743)
## catalogs                    42.719***                40.126***
##                              (2.544)                  (2.860)
## historyLow                                          -240.503***
##                                                       (86.755)
## historyMedium                                       -346.815***
##                                                       (59.781)
## Constant                   -539.806***               -199.410*
##                              (49.592)                (107.546)
## ---------------------------------------------------------------------
## Observations                  1,000                     697
## R2                            0.715                    0.787
## Adjusted R2                   0.714                    0.785
## Residual Std. Error    514.246 (df = 995)        463.306 (df = 690)
## F Statistic         623.563*** (df = 4; 995) 425.757*** (df = 6; 690)
## =====================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

Missing values

- The variable history have missing values.

In order not to lose observations, we can create a new variable (let's call it `newH`) that is equal to our `history` variable, but instead of having missing data has an extra category called "NewCust" — we presume that clients for which there is no past purchasing behavior are new customers. This is a good example of how to use the `ifelse` function.

```
dt.marketing[, newH := ifelse(is.na(history), "NewCust", pa
```

Model 3

Using the new variable we define the model:

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{location}_i + \beta_2 \times \text{salary}_i + \beta_3 \times \text{children}_i + \beta_4 \times \text{catalogs}_i + \beta_5 \times \text{newH}_i + \epsilon$$

```
formula_m3 <- as.formula(amountspent ~ location + salary + children + catalogs + newH)   # Define the form
lm.spend3   <- lm(formula = formula_m3,
                 data = dt.marketing)
stargazer(lm.spend1, lm.spend2, lm.spend3, type = "text", no.space = TRUE)
```

Model 3

Table 1: Regression Results

	Dependent variable:		
	amountspent		
	(1)	(2)	(3)
locationFar	508.076***	436.304***	436.304***
	(36.217)	(35.893)	(35.893)
salary	0.021***	0.019***	0.019***
	(0.001)	(0.001)	(0.001)
children	−203.479***	−169.448***	−169.448***
	(15.625)	(16.647)	(16.647)
catalogs	42.719***	41.652***	41.652***
	(2.544)	(2.453)	(2.453)
newHLow		−350.929***	−350.929***
		(65.442)	(65.442)
newHMedium		−409.901***	−409.901***
		(52.413)	(52.413)
newHNewCust		−1.875	−1.875
		(51.100)	(51.100)
Constant	−539.806***	−244.589***	−244.589***
	(49.592)	(79.393)	(79.393)
Observations	1,000	1,000	1,000
R^2	0.715	0.746	0.746
Adjusted R^2	0.714	0.744	0.744
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01		

Model 4

We add also gender to the model.

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{location}_i + \beta_2 \times \text{salary}_i + \beta_3 \times \text{children}_i + \beta_4 \times \text{catalogs}_i +$$

$$+\beta_5 \times \text{newH}_i + \beta_6 \times \text{gender}_i + \epsilon$$

```
formula_m4 <- as.formula(amountspent ~ location + salary + children + catalogs + newH + gender)   # Define
lm.spend4   <- lm(formula = formula_m4,
                  data = dt.marketing)
stargazer(lm.spend1, lm.spend2, lm.spend3, lm.spend4, type = "text", no.space = TRUE)
```

Model 4

Table 2: Regression Results

	Dependent variable:			
	amountspent			
	(1)	(2)	(3)	(4)
locationFar	508.076***	436.304***	436.304***	436.046***
	(36.217)	(35.893)	(35.893)	(35.860)
salary	0.021***	0.019***	0.019***	0.019***
	(0.001)	(0.001)	(0.001)	(0.001)
children	−203.479***	−169.448***	−169.448***	−171.982***
	(15.625)	(16.647)	(16.647)	(16.699)
catalogs	42.719***	41.652***	41.652***	41.746***
	(2.544)	(2.453)	(2.453)	(2.452)
newHLow		−350.929***	−350.929***	−355.056***
		(65.442)	(65.442)	(65.427)
newHMedium		−409.901***	−409.901***	−408.813***
		(52.413)	(52.413)	(52.368)
newHNewCust		−1.875	−1.875	−0.035
		(51.100)	(51.100)	(51.064)
genderMale				−54.284*
				(32.171)
Constant	−539.806***	−244.589***	−244.589***	−228.384***
	(49.592)	(79.393)	(79.393)	(79.898)
Observations	1,000	1,000	1,000	1,000
R^2	0.715	0.746	0.746	0.747
Adjusted R^2	0.714	0.744	0.744	0.745
Note:				*p<0.1; **p<0.05; ***p<0.01

Dependent variable:

Model 5

We remove `salary` to the model.

$$\text{amountspend}_i = \beta_0 + \beta_1 \times \text{location}_i + \beta_3 \times \text{children}_i + \beta_4 \times \text{catalogs}_i +$$

$$+\beta_5 \times \text{newH}_i + \beta_6 \times \text{gender}_i + \epsilon$$

```
formula_m5 <- as.formula(amountspent ~ location  + children + catalogs + newH + gender)   # Define the for
lm.spend5   <- lm(formula = formula_m5,
                data = dt.marketing)
stargazer(lm.spend1, lm.spend2, lm.spend3, lm.spend4,lm.spend5, type = "text", no.space = TRUE)
```

Model 5

Table 3: Regression Results

	Dependent variable:				
	amountspent				
	(1)	(2)	(3)	(4)	(5)
locationFar	508.076***	436.304***	436.304***	436.046***	208.594***
	(36.217)	(35.893)	(35.893)	(35.860)	(46.247)
salary	0.021***	0.019***	0.019***	0.019***	
	(0.001)	(0.001)	(0.001)	(0.001)	
children	−203.479***	−169.448***	−169.448***	−171.982***	−7.448
	(15.625)	(16.647)	(16.647)	(16.699)	(20.658)
catalogs	42.719***	41.652***	41.652***	41.746***	42.774***
	(2.544)	(2.453)	(2.453)	(2.452)	(3.249)
newHLow		−350.929***	−350.929***	−355.056***	−1,490.437***
		(65.442)	(65.442)	(65.427)	(67.159)
newHMedium		−409.901***	−409.901***	−408.813***	−1,041.889***
		(52.413)	(52.413)	(52.368)	(62.311)
newHNewCust		−1.875	−1.875	−0.035	−712.983***
		(51.100)	(51.100)	(51.064)	(58.262)
genderMale				−54.284*	105.681**
				(32.171)	(41.940)
Constant	−539.806***	−244.589***	−244.589***	−228.384***	1,262.729***
	(49.592)	(79.393)	(79.393)	(79.898)	(77.608)
Observations	1,000	1,000	1,000	1,000	1,000
R^2	0.715	0.746	0.746	0.747	0.555
Adjusted R^2	0.714	0.744	0.744	0.745	0.552
Note:					*p<0.1; **p<0.05; ***p<0.01

Predict amount spent by new customer

Now let's predict the amount spend for a new customer using our initial model:

```
new.client <- data.table(gender = "Male",
                         location = "Close",
                         salary = 53700,
                         children = 1,
                         catalogs = 12)
```

```
my.pred <- predict(lm.spend1, newdata = new.client, level = .95, interval = "confidence")
my.pred
```

```
##        fit      lwr      upr
## 1 891.2053 851.8992 930.5114
```

We can also get the estimated residuals $(y - \hat{y})$ by using the function `residual`.

```
my.res <- residuals(lm.spend1)
head(my.res)
```

```
##          1          2          3          4          5          6
## -461.92002  272.80626 -215.16979  622.04862  -97.78629  143.39655
```