Lecture 4: Multiple linear regression with categorical variables and interaction terms

Contents

Last class ...

- Estimation and inference of a SLR
- (Set and tested hypotheses of interest about the regression coefficients)
- Predictions using the estimated regression

Lecture 4 Learning Objectives

- Write a computer script to estimate and test coefficients of additive and multiplicative LR models
- Interpret each estimated coefficient in the presence of other variables in the model, including interaction terms
- Define and calculate residuals and predictions from the estimated multiple linear regression (MLR)
- Identify hypotheses tested by the lm function in R
- Reflect about the methodology used by lm to test hypotheses
- Write a computer script to test hypotheses about the regression coefficients
- Interpret the results of inference (CI and tests) correctly given the data and method used

In many applications, we observe *many* explanatory variables

For example, both the size (continuous) and the age (discrete) of a property may affect its value

$$Y_i = \beta_0 + \beta_1 \text{size}_i + \beta_2 \text{age}_i + \varepsilon_i$$

1.1 Conceptual notes: population assumptions

- The conditional expectation is the best predictor of Y given a set of explanatory variables $X_1, X_2, \ldots, X_p$
- The linear regression models the conditional expectation as a *linear combination* of the predictors
  - we'll focus later on the meaning of *linear*

$$E[Y|X_1, \ldots, X_p] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

- If the data is jointly Normal, the conditional expectation is truly linear! Otherwise, this is just an assumption.

Question: Does linearity imply normality?

So, are we ready to predict?

As before, we need to estimate the (population) regreesion coefficients

$$\beta_0, \beta_1, \text{ and } \beta_2$$

... Back to sample, data and estimation!

```
library(tidyverse)
library(repr)
options(repr.plot.width=7, repr.plot.height=4)
library(ggplot2)
library(broom)
library(gridExtra)
library(latex2exp)
library(GGally)
library(moderndive)

dat <- read.csv("data/Assessment_2015.csv")
dat <- dat %>% filter(ASSESSCLAS=="Residential")  %>%
        mutate(assess_val = ASSESSMENT / 1000, age=2020-YEAR_BUILT)


# Our sample
set.seed(561)
dat_s <- sample_n(dat, 1000, replace = FALSE)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 -
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
```

```
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() -
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all con
```

```
Attaching package: ‘gridExtra’
```

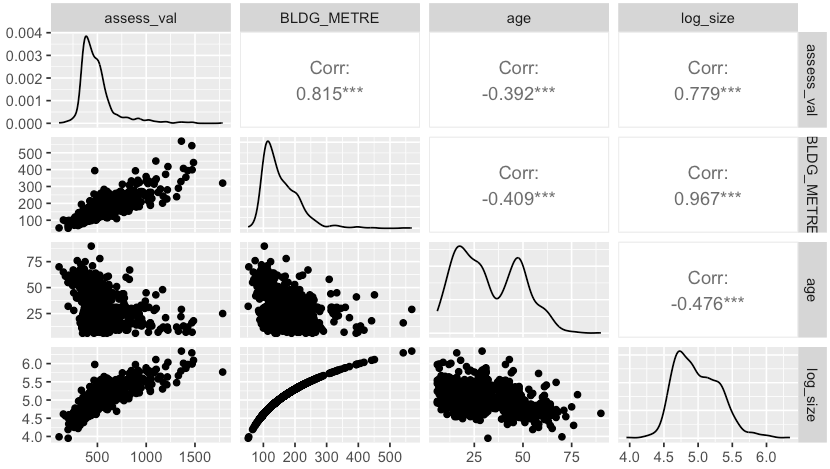
```
The following object is masked from ‘package:dplyr’:

    combine
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

1.2 Visualizing the relation between the variables in the sample

We can start by visualizing the relation between these 3 variables (all quantitative) using pair-wise scatter plots. Take some time to reflect on the relationship between the variables.

```
dat_s %>% select(assess_val, BLDG_METRE,age) %>% mutate(log_size = log(BLDG_ME
    ggpairs()
```



We observe a linear association between size and value and a negative association between age and value. However, there's a lot of variation in value for newer houses, perhaps because they vary a lot in size

The distribution of `age` seems bimodal and a normality assumption doesn't seem reasonable here! We can transform variables to get closer to the assumed model (more about this in Regression II).

2. SLR vs. MLR

SLR: one variable at a time

```
#Size
lm_size <- lm(assess_val~BLDG_METRE,data=dat_s)
tidy(lm_size)
```

A tibble: 2 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	90.768749	9.79347555	9.268288	1.132776e-19
BLDG_METRE	2.617547	0.05880332	44.513599	2.897913e-239

```
#Age
lm_age <- lm(assess_val~age,data=dat_s)
tidy(lm_age)
```

A tibble: 2 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	647.413924	12.3523539	52.41219	7.601564e-289
age	-4.576485	0.3403857	-13.44500	5.376037e-38

MLR: more than one predictor

Should we look at `BLDG_METRE` and `age` together? —> YES

```
lm_size_age <- lm(assess_val~BLDG_METRE + age, data=dat_s)
tidy(lm_size_age)
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	131.508644	15.1982597	8.652875	1.986205e-17
BLDG_METRE	2.526029	0.0640819	39.418755	1.342602e-205
age	-0.814666	0.2333191	-3.491639	5.011087e-04

```
tidy(lm_size_age)  %>% select(term,estimate)  %>% rename(MLR=estimate) %>%
    merge(tidy(lm_size)  %>% select(term,estimate),by="term",all=T)  %>% renam
    merge(tidy(lm_age)  %>% select(term,estimate),by="term",all=T)  %>% rename
```

A data.frame: 3 × 4

term	MLR	SLR_size	SLR_age
<chr>	<dbl>	<dbl>	<dbl>
(Intercept)	131.508644	90.768749	647.413924
age	-0.814666	NA	-4.576485
BLDG_METRE	2.526029	2.617547	NA

The slope for `BLDG_METRE` remains almost the same but that of `age` is very different

> usually a sign of multicollinearity or confounding

3. MLR with continuous predictors

$$Y_i = \beta_0 + \beta_1 \text{size}_i + \beta_2 \text{age}_i + \varepsilon_i$$

MLR simultaneously models the association of multiple predictors with the response.

An important model assumption:

If we hold the size of the property fixed, say at 155mts:

- the relationship between the value and the age of the property is *linear*

$$Y_i = \beta_0 + \beta_1 155 + \beta_2 \text{age}_i + \varepsilon_i$$

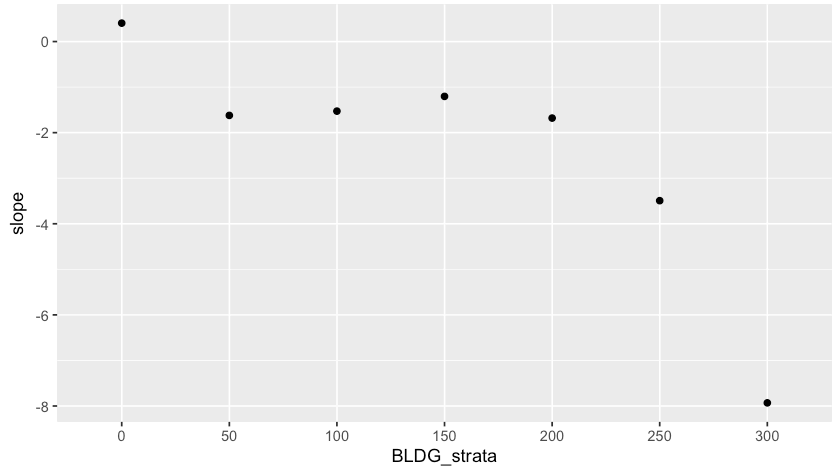
- the slope β_2 *does not depend* on the value held constant

(and same for holding `age` fixed)

These type of models are called ***additive models***.

A peek at our population...

```
dat  %>% filter(BLDG_METRE <= 300)  %>%
    mutate(BLDG_strata = factor(50*round(BLDG_METRE/50))) %>%
    group_by(BLDG_strata)  %>%
    summarize(slope=cor(assess_val,age)*sd(assess_val)/sd(age))  %>%
    ggplot(aes(BLDG_strata,y=slope))+ geom_point()
```



In general, the slope of `age` does not depend on the value of size we hold fixed

> Note: I've created groups of houses of similar sizes to have enough points to look at the slope. At the tail, this assumption doesn't seem to hold.

3.1 Estimating the MLR with `lm`

`R` Code:

```
lm_s <- lm(assess_val~BLDG_METRE + age,data=dat_s)
```

> variables are added to the model using `+`

`lm` uses a method called least squares to estimate the coefficients from a random sample

- Least squares estimators minimizes the mean squared error (MSE) loss function

The tidy table

```
tidy(lm_size_age) %>% mutate_if(is.numeric, round, 3)
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	131.509	15.198	8.653	0.000
BLDG_METRE	2.526	0.064	39.419	0.000
age	-0.815	0.233	-3.492	0.001

As before, we get estimates, SE of the estimators, and results of classical hypothesis tests

How do we interpret these values? Which hypotheses are we testing?

Interpretation of the estimated coefficients:

Two estimated slopes: $\hat{\beta}_1 = 2.526$ and $\hat{\beta}_2 = -0.815$

- Holding age constant, for every additional square meter in the size of a house, we predict an increase of 2526 dollars in the assessed value.
- Holding size constant, for every additional year a house has, we predict a decrease of 815 dollars in the assessed value.

> Is it reasonable to assume the that slope does not depend on the value being held constant?

3.2 Confidence Intervals

- CI can be constructed using bootstrapping,either to estimate the SE or the sampling distribution of any coefficient
- The `lm` function uses results from asymptotic theory to construct confidence intervals

```
tidy(lm_size_age, conf.int = TRUE)
```

A tibble: 3 × 7

term	estimate	std.error	statistic	p.value	conf.low	cor
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
(Intercept)	131.508644	15.1982597	8.652875	1.986205e-17	101.684396	161.33
BLDG_METRE	2.526029	0.0640819	39.418755	1.342602e-205	2.400278	2.65
age	-0.814666	0.2333191	-3.491639	5.011087e-04	-1.272519	-0.35

3.3 Prediction using the regression line and residuals

As before, the *predicted* assessed value is a point on the estimated hyper-plane; the *residual* is the distance to the observed value

$$\hat{y}_i = 131.51 + 2.526 \times x_{i1} - 0.815 \times x_{i2};\ r_i = y_i - \hat{y}_i$$

They are random variables since they are functions of the estimated coefficients

> Therefore, we can estimate their standard error and CI! (Note: $\varepsilon_i \neq r_i$)

```
augment(lm_size_age,dat_s)  %>% head(3)
```

the_geom	TAX_YEAR	ROLL_NUM	ADDRESS	YEAR_BUILT	ASSES
<chr>	<int>	<dbl>	<chr>	<int>	
POINT (-113.26828568594681 53.53364376673626)	2015	-2147483648	107 GLENCOE BLVD	1990	Res
POINT (-113.31770257850987 53.516152401699564)	2015	-2147483648	116 WILLOW ST	1959	Res
POINT (-113.28777490571461 53.52360710210846)	2015	-2147483648	45 RAVEN DR	1966	Res

NOTE: the `augment` function is very useful to add in-sample predicted values to the original dataset!

3.4 Hypothesis tests on the regression coefficients

Same as before, just more coefficients!

As a default, `lm` tests if the intercept and the slopes are 0:

- $H_0 : \beta_0 = 0$ (null) *vs* $H_1 : \beta_0 \neq 0$ (default alternative)
- $H_0 : \beta_1 = 0$ (null) *vs* $H_1 : \beta_1 \neq 0$ (default alternative)
- $H_0 : \beta_2 = 0$ (null) *vs* $H_1 : \beta_2 \neq 0$ (default alternative)

Under the null hypothesis:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

> Recall from DSCI552: when the estimator is the sample mean, the $SE(\bar{X}) = \hat{\sigma}/\sqrt{n}$, other estimators have different standard errors.

Mathematical result: the t-statistic follows a t-distribution with $n - q$ degrees of freedom (df), where n is the sample size and q the number of parameters (in this case $q = 3 = p + 1 = 2 + 1$).

```
tidy(lm_size_age)%>%mutate_if(is.numeric, round, 3)

summary(lm_size_age)
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	131.509	15.198	8.653	0.000
BLDG_METRE	2.526	0.064	39.419	0.000
age	-0.815	0.233	-3.492	0.001

```
Call:
lm(formula = assess_val ~ BLDG_METRE + age, data = dat_s)

Residuals:
    Min      1Q  Median      3Q     Max 
-620.36  -67.04   -3.69   35.24  864.53 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 131.50864   15.19826   8.653  < 2e-16 ***
BLDG_METRE    2.52603    0.06408  39.419  < 2e-16 ***
age          -0.81467    0.23332  -3.492 0.000501 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 108.7 on 997 degrees of freedom
Multiple R-squared:  0.6691,    Adjusted R-squared:  0.6684 
F-statistic:  1008 on 2 and 997 DF,  p-value: < 2.2e-16
```

Question

True or False: The `lm` function in R shows results of a hypothesis test for the slope of the LR where the null hypothesis states that the estimated coefficient equals 0.

A. TRUE

B. FALSE

Question

Confidence intervals (CI) created by `lm` are centered at the true population coefficients. True or False?

A. TRUE

B. FALSE

4. MLR: categorical variables with many levels

In lecture 2 we showed that we can add categorical variables in the LR using (numerical) dummy variables

Recall: using a dummy variable X_2

$$X_2 = \begin{cases} 1 & \text{if FIREPLACE} = \text{Y}; \\ 0 & \text{if FIREPLACE} = \text{N} \end{cases}$$

Then,

$\beta_0 = E[Y|X_2 = 0]$, and

$\beta_2 = E[Y|X_2 = 1] - E[Y|X_2 = 0]$

Remember that R creates this dummy variable for us if `FIREPLACE` is a factor!

- R calls the dummy variable `FIREPLACEY` (name of the variable followed by the level corresponding to 1
- Note that we only need *one* dummy variable for 2 levels (the reference level = 0, the other level = 1)
- The reference level (dummy variable = 0, level "left out") is chosen alphabetically. You can change that if you want!

But how can we include a categorical variable with *more than 2* levels?

We just need more dummy variables!

$$E[Y|X_2, X_3] = \beta_0 + \beta_2 X_2 + \beta_3 X_3$$

> Note: I'm using the name X_2 to keep it consistent with lecture 02 but you can use any other name for the variables and the coefficients.

For example, you may prefer:

$$E[Y|X_2, X_3] = \beta_0 + \beta_M M + \beta_N N$$

where M and N are dummy variables.

The dataset we are working with does not have a categorical variable with more levels so let's create one:

`generation`: a discrete variable with 3 levels: "old" for houses built before 1980, "modern" for houses built between 1981-2000, "new" for houses built after 2000

```
dat <- dat  %>% mutate(generation = cut(YEAR_BUILT,
                    breaks=c(-Inf,1980,2000,+Inf),
                    labels=c("Old","Modern","New")))

set.seed(561)
dat_s <- sample_n(dat, 1000, replace = FALSE)

head(dat_s)
```

	the_geom	TAX_YEAR	ROLL_NUM	ADDRESS	YEAR_BUILT	A
	<chr>	<int>	<dbl>	<chr>	<int>	
1	POINT (-113.26828568594681 53.53364376673626)	2015	-2147483648	107 GLENCOE BLVD	1990	
2	POINT (-113.31770257850987 53.516152401699564)	2015	-2147483648	116 WILLOW ST	1959	
3	POINT (-113.28777490571461 53.52360710210846)	2015	-2147483648	45 RAVEN DR	1966	
4	POINT (-113.27617960277775 53.55193363362361)	2015	-2147483648	8 CRANBERRY WY	2004	
5	POINT (-113.21561364238943 53.41664801825034)	2015	1220717001	157 51309 RNG RD 225	1982	
6	POINT (-113.27839485086774 53.54226897623362)	2015	-2147483648	28 CLAIRE CL	2006	

```
cbind(generation = c("Old","Modern","New"),dat %>% summarize(pop_count = summa
dat_s  %>% summarize(sample_count = summary(generation)))
```

```
Warning message:
“Returning more (or less) than 1 row per `summarise()` group was deprecated in
dplyr 1.1.0.
ℹ Please use `reframe()` instead.
ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.”
```

```
Warning message:
“Returning more (or less) than 1 row per `summarise()` group was deprecated in
dplyr 1.1.0.
ℹ Please use `reframe()` instead.
ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.”
```

A data.frame: 3 × 3

generation	pop_count	sample_count
<chr>	<int>	<int>
Old	10748	384
Modern	8799	331
New	8152	285

$$\text{M} = X_2 = \begin{cases} 1 & \text{if generation} = ``\text{Modern}"; \\ 0 & \text{if otherwise} \end{cases}$$

$$\text{N} = X_3 = \begin{cases} 1 & \text{if generation} = ``\text{New}"; \\ 0 & \text{if otherwise} \end{cases}$$

Note that we need *two* dummy variable for 3 levels (2 levels compared to one reference level). The reference level is the same for both dummy variables. Both dummy variables equal to 0 when the observation corresponds to the reference level.

- if generation = "old": $E[Y|X_2 = 0, X_3 = 0] = \beta_0$
- if generation = "modern": $E[Y|X_2 = 1, X_3 = 0] = \beta_0 + \beta_2$
- if generation = "new": $E[Y|X_2 = 0, X_3 = 1] = \beta_0 + \beta_3$

Then,

$$\beta_2 = E[Y|\text{mod}] - E[Y|\text{old}] \text{ and } H_0 : \beta_2 = 0$$

$$\beta_3 = E[Y|\text{new}] - E[Y|\text{old}] \text{ and } H_0 : \beta_3 = 0$$

Code

If `generation` is a factor, R creates dummy variables for them!

- R calls these dummy variables `generationModern` and `generationNew` (name of the variable followed by the levels corresponding to 1)

- The reference level (dummy variable = 0, level "left out") is "Old"

```
dat  %>% select(YEAR_BUILT,age,generation)  %>% str()
```

```
'data.frame':   27699 obs. of  3 variables:
 $ YEAR_BUILT: int  1950 1983 1981 1967 1968 1962 1967 1983 1965 1969 ...
 $ age       : num  70 37 39 53 52 58 53 37 55 51 ...
 $ generation: Factor w/ 3 levels "Old","Modern",..: 1 2 2 1 1 1 1 2 1 1 ...
```

The tidy table:

```
tidy(lm(assess_val~generation,dat_s))
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	409.5833	8.928084	45.87584	6.039893e-248
generationModern	131.7883	13.121917	10.04337	1.126496e-22
generationNew	159.5219	13.678828	11.66196	1.497539e-29

Interpretation of the estimated coefficients

- The estimated **intercept**, $\hat{\beta}_0 = 409.58$, is the average assessed value for old properties!

> It is the sample version of the conditional expectation (mean of the reference group)

- The estimated **"slopes"** (we have 2, one per dummy variable) are the differences of sample means:
  - $\hat{\beta}_2 = 131.79$: difference between sample mean of Modern houses and sample mean of Old houses. In other words, on average, modern houses are valued 131788 dollars higher than old houses.
  - $\hat{\beta}_3 = 159.5219$: difference between sample mean of New houses and sample mean of Old houses. In other words, on average, new houses are valued 159521 dollars

higher than old houses.

> It is the sample version of the difference of the conditional expectations (or group means)

In-class Exercise

What if my objective is to compare average costs of “Modern” and “New” generation homes? Recode your model and interpret the slopes (Hint: use `fct_relevel`).

```
dat_s$generation <- fct_relevel(dat_s$generation, "Old", after = 3 )
levels(dat_s$generation)

tidy(lm(assess_val~generation,dat_s))
```

'Modern' · 'New' · 'Old'

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	541.37160	9.616342	56.297044	7.288920e-312
generationNew	27.73366	14.137669	1.961686	5.007703e-02
generationOld	-131.78827	13.121917	-10.043370	1.126496e-22

5. MLR: Interaction effects

Interaction effects occur when the effect of one explanatory variable on a response *depends* on another explanatory variable.

Example: the effect of drug dosage on reducing blood pressure may depend on weight.

We can incoorporate interaction effects into our model as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \times X_{i2} + \varepsilon_i.$$

In `lm` we can just add variable in the model using `*` if you want a model with an interaction term.

> An important difference to note between the additive model and the model with interaction is that we are not assuming that groups share a slope!

- Let's reconsider a linear regression model with `BLDG_METRE` and `FIREPLACE` as explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

```
fit1 <- lm(assess_val ~ BLDG_METRE + FIREPLACE, data = dat_s)
tidy(fit1)%>%mutate_if(is.numeric, round, 3)
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	78.482	10.812	7.259	0.000
BLDG_METRE	2.567	0.062	41.660	0.000
FIREPLACEY	24.685	9.327	2.647	0.008

We can compute estimated regression lines for homes without a fireplace ($x_2 = 0$) and for homes with a fireplace ($x_2 = 1$):

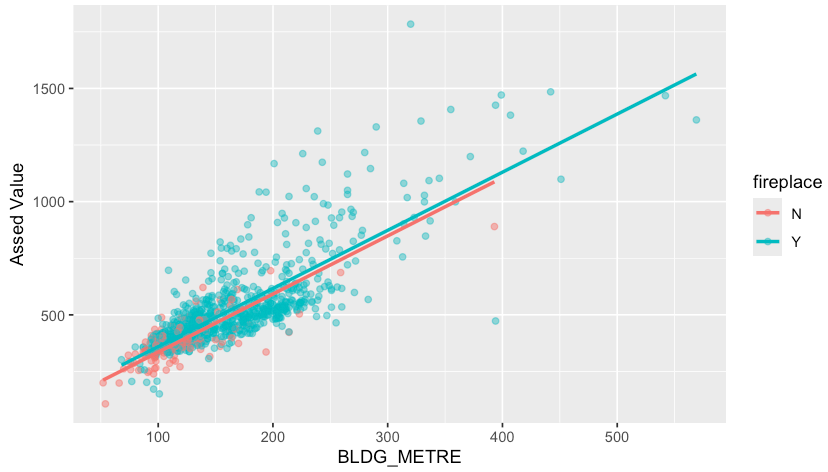
- if fireplace = "N":

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \times 0 = 78.482 + 2.567 x_1$$

- if fireplace = "Y":

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \times 1 \\ &= 78.482 + 2.567 x_1 + 24.685 \\ &= 103.167 + 2.567 x_1 \end{aligned}$$

```
ggplot(dat_s, aes(x = BLDG_METRE, y = assess_val, color = FIREPLACE)) +
  geom_point(alpha=0.5) +
  labs(x = "BLDG_METRE", y = "Assed Value", color = "fireplace") +
  geom_parallel_slopes(se = FALSE)
```



Class Discussion: Think, Pair, Share

- What do you notice about the slopes and intercepts?
- Is this model additive?

Let's look at predicting the home's value by considering the size of the home and whether it not it has a fireplace, and consider a potential interaction effect between these variables.

```
# Alternative:
# fit <- lm(assess_val ~ BLDG_METRE + fireplace + BLDG_METRE* fireplace, data=

fit <- lm(assess_val ~ BLDG_METRE*FIREPLACE, data= dat_s)

tidy(fit)%>%mutate_if(is.numeric, round, 3)
```

A tibble: 4 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	155.064	28.075	5.523	0.000
BLDG_METRE	1.917	0.228	8.393	0.000
FIREPLACEY	-60.217	30.208	-1.993	0.046
BLDG_METRE:FIREPLACEY	0.701	0.237	2.954	0.003

Now,

- if fireplace = "N":

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \times 0 + \hat{\beta}_3 x_1 \times 0 = 155.064 + 1.917 x_1$$

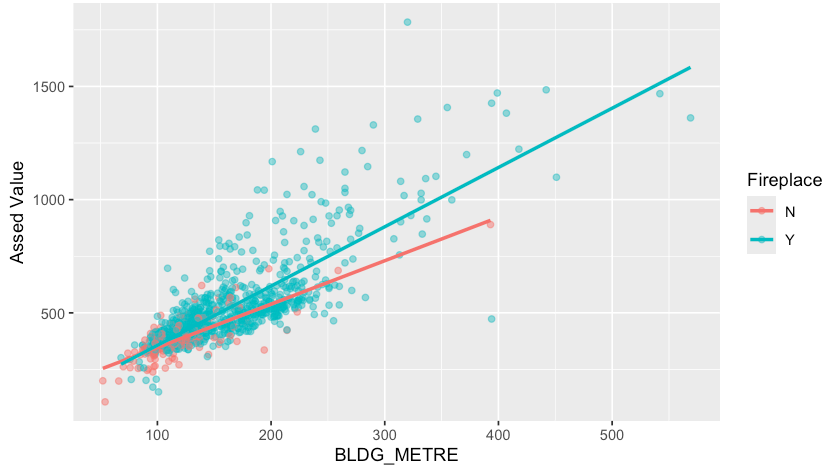
- if fireplace = "Y":

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \times 1 + \hat{\beta}_3 x_1 \times 1 \\ &= \hat{\beta}_0 + \hat{\beta}_2 + (\hat{\beta}_1 + \hat{\beta}_3) x_1 \\ &= (155.064 - 60.217) + (1.917 + 0.701) x_1 \\ &= 94.847 + 2.618 x_1\end{aligned}$$

What do you notice about these estimated regression lines?

```
ggplot(dat_s, aes(x = BLDG_METRE, y = assess_val, color = FIREPLACE)) +
  geom_point(alpha=0.5) +
  labs(x = "BLDG_METRE", y = "Assed Value", color = "Fireplace") +
  geom_smooth(method="lm",se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Class Discussion: Think, Pair, Share

- What do you notice about the slopes and intercepts?
- Is this model additive?
- If we can accomodate different slopes with an interaction term, why would you ever force the lines to be parallel? (*Hint: Occam's Razor*)

Visually, the two slopes still appear to be roughly parallel, and this is because the interaction term here is not significant. Next class, we will see an example where the interaction term has a clear effect.

Meaning of Coefficients:

- `Intercept`: The average value of a home (in $1000s) of zero square meters without a fireplace (155.064; not a practical interpretation).
- `BLDG_METRE`: the slope for homes without a fireplace (1.917). That is, an increase in `BLDG_METRE` by one unit is associated with an expected increase in assessed value by $1917 for homes without a fireplace.
- `fireplaceY`: holding the size of the home constant, the difference in average value of a home between homes with an without a fireplace (-60.217).
- `BLDG_METRE:fireplaceY`: This can be interpreted as the difference in the slope of homes with a fireplace compared to the slope of homes without a fireplace (i.e., the offset in slope for homes with a fireplace is 0.701). That is, an increase of `BLDG_METRE` by one unit is associated with an expected increase in assessed home value of $1.917 + 0.701 = 2.618$ for homes with a fireplace.

What might be an explanation as to why the sign flipped from (+) to (-) for the fireplace coefficient?

If the presence of a fireplace tends to be associated with larger homes, the main effect might show a negative coefficient due to the interaction capturing the true positive relationship in combination with `BLDG_METRE`. Adding interaction effects into the model can complicate the interpretation of the main effects.

Summary: what have we learned today?

- MLR including more than one predictor in our model, including interaction terms
- Interpreting coefficients for categorical data with more than 2 levels
- Interpreting coefficients for MLR (i.e., need to hold all other variables constant)
- How to relevel a factor
- Using ggpairs() to visualize relationships between variables
- Difference between additive and multiplicative models

> This lecture references material from Chapter 6 of ModernDive.