DSCI561: Regression I

Lecture 3: November 22, 2017

Gabriela Cohen Freue

Department of Statistics, UBC

Review from Lect 2

- Two-sample t-tests and ANOVA are special cases of linear regression: one categorical covariate (2 or more groups)
- Matrix formulation of a linear regression
- The interpretation of estimates and tests depends on the parametrization used to represent the data
- `lm()` in R uses the « reference-treatment » parametrization as a default

ANOVA-style, "cell means"

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

ANOVA-style, "ref + tx effects"

$$Y_{ij} = \theta + \tau_j + \varepsilon_{ij}, (\tau_1 = 0)$$

reference

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix} \qquad \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boxed{\theta} \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix}$$

For example:

$$Y_{11} = 1 * \theta + 0 * \tau_2 + 0 * \tau_3 + \varepsilon_{11} = \theta + \varepsilon_{11} \implies E[Y_{11}] = \theta$$

$$Y_{13} = 1 * \theta + 0 * \tau_2 + 1 * \tau_3 + \varepsilon_{13} = \theta + \tau_3 + \varepsilon_{13} \implies E[Y_{13}] = \theta + \tau_3$$

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix}$$

Reference period: C

M vs C

O vs C

reference

treatment effect

$$E[Y_{i1}] = \theta$$

$$E[Y_{i2}] = \theta + \tau_2 \implies E[Y_{i2}] - E[Y_{i1}] = \tau_2$$

$$E[Y_{i3}] = \theta + \tau_3 \implies E[Y_{i3}] - E[Y_{i1}] = \tau_3$$

differences in ***population*** means

In today's lecture

- Analyze the output of lm() in relation with the mathematical formulation of the model
- Linear models with more than one categorical variable
- Linear models with a continuous independent variable (see Lect04)
- Linear models with both continuous and categorical variables (see Lect04)

```r
#More than 2 groups

#LM with 3 age periods
summary(lm(assessment_k~age_factor,data=dat.small))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = dat.small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -250.14  -74.89  -16.97   51.36  612.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      54.10   9.631 2.87e-14 ***
## age_factorM   -92.20      62.46  -1.476    0.145
## age_factorO   -85.86      56.74  -1.513    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121 on 67 degrees of freedom
## Multiple R-squared:  0.0353, Adjusted R-squared:  0.006498
## F-statistic: 1.226 on 2 and 67 DF,  p-value: 0.3001
```

```r
#ANOVA with 3 age periods
summary(aov(assessment_k~age_factor,data=dat.small))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## age_factor   2  35867   17934   1.226    0.3
## Residuals   67 980328   14632
```

1 categorical variables

age (3 levels)

Reference period: C

difference in ***sample*** means

M *vs* C

O *vs* C

```r
#More than 2 groups

#LM with 3 age periods
summary(lm(assessment_k~age_factor,data=dat.small))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = dat.small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -250.14  -74.89  -16.97   51.36  612.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      54.10   9.631 2.87e-14 ***
## age_factorM   -92.20      62.46  -1.476    0.145
## age_factorO   -85.86      56.74  -1.513    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121 on 67 degrees of freedom
## Multiple R-squared:  0.0353, Adjusted R-squared:  0.006498
## F-statistic: 1.226 on 2 and 67 DF,  p-value: 0.3001
```

1 categorical variables

age (3 levels)

$$H_0 : \mu_M = \mu_C$$

$$H_0 : \tau_2 = 0$$

```r
#ANOVA with 3 age periods
summary(aov(assessment_k~age_factor,data=dat.small))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## age_factor   2  35867   17934   1.226    0.3
## Residuals   67 980328   14632
```

```
#More than 2 groups

#LM with 3 age periods
summary(lm(assessment_k~age_factor,data=dat.small))
```

1 categorical variables

age (3 levels)

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = dat.small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -250.14  -74.89  -16.97   51.36   612.86
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     521.00      54.10   9.631 2.87e-14 ***
## age_factorM     -92.20      62.46  -1.476    0.145
## age_factorO     -85.86      56.74  -1.513    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121 on 67 degrees of freedom
## Multiple R-squared:  0.0353, Adjusted R-squared:  0.006498
## F-statistic: 1.226 on 2 and 67 DF,  p-value: 0.3001
```

$$H_0 : \mu_C = \mu_M = \mu_O$$

$$H_0 : \tau_2 = \tau_3 = 0$$

```
#ANOVA with 3 age periods
summary(aov(assessment_k~age_factor,data=dat.small))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## age_factor   2  35867   17934   1.226    0.3
## Residuals   67 980328   14632
```

same test

Partial summary

- Two-sample t-test and ANOVA are special cases of a linear model
- We can use the `lm()` for both analyses
- By default, R uses the “reference-treatment” parametrization in `lm()`
- The t-tests in the output of `lm()` depend on the parameters of the model
- The `lm()` output includes an F-test to test a full vs an “intercept-only” model

$$Y = \boxed{X}\alpha + \varepsilon$$

This gives us a VERY FLEXIBLE framework!!

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1.22 \\ 1 & 2.02 \\ 1 & 1.42 \\ \vdots & \vdots \\ 1 & 1.89 \\ 1 & 2.01 \\ \vdots & \vdots \\ 1 & 1.56 \\ 1 & 2.17 \\ 1 & 1.51 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 1.22 & 0 \\ 1 & 0 & 2.02 & 0 \\ 1 & 0 & 1.42 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1.89 & 0 \\ 1 & 1 & 2.01 & 2.01 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1.56 & 1.56 \\ 1 & 1 & 2.17 & 2.17 \\ 1 & 1 & 1.51 & 1.51 \end{bmatrix}$$

1 categorical covariate | **2 categorical covariates** | **1 continuous covariate** | **1 continuous 1 categorical**

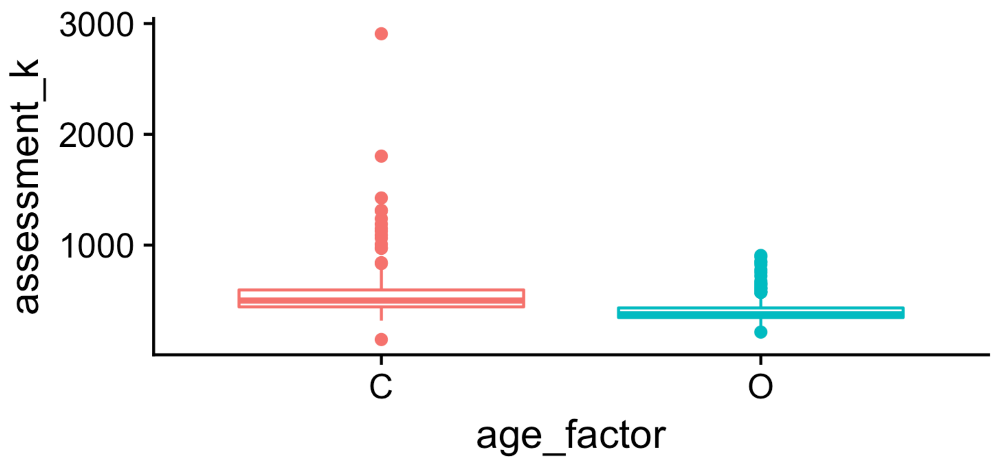
AND MANY MORE .....

Tip: ?model.matrix

More than one categorical covariate

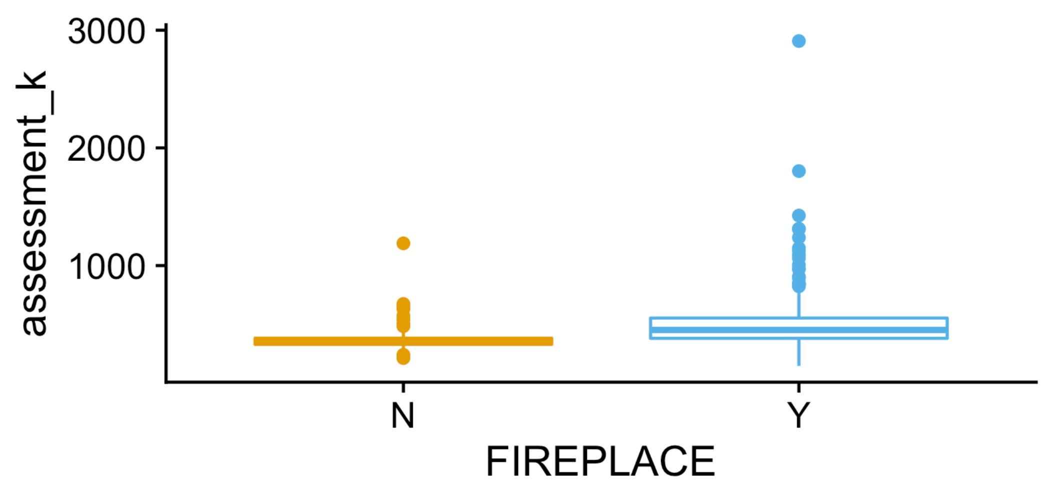
LINEAR REGRESSION

For simplicity, let's consider only 2 levels for age



Properties with a fireplace may have a higher tax value

Model with 2 categorical variables



2 categorical variables

age (2 levels) and FIREPLACE (2 levels)

age_factor	FIREPLACE	assessment_k
C	N	390
C	N	541
C	N	364
...	...	...
C	Y	449
C	Y	536
C	Y	595
C	Y	449
...	...	...
O	N	355
O	N	396
...	...	...
O	Y	354
O	Y	363
...	...	...

$$Y_{CN1}, \ldots, Y_{CN15}, \; n_{CN} = 15$$

$$Y_{CY1}, \ldots, Y_{CY136}, \; n_{CY} = 136$$

$$Y_{ON1}, \ldots, Y_{ON72}, \; n_{ON} = 72$$

$$Y_{OY1}, \ldots, Y_{OY145}, \; n_{OY} = 145$$

Two-way ANOVA: main effect

```
#Two-way ANOVA table
summary(aov(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##                       Df   Sum Sq Mean Sq F value   Pr(>F)
## age_factor             1  2536324 2536324  57.352 3.03e-13 ***
## FIREPLACE              1   509278  509278  11.516 0.000766 ***
## age_factor:FIREPLACE   1    19684   19684   0.445 0.505095
## Residuals            364 16097397   44224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Compare the 2-way and the 1-way ANOVA tables
summary(aov(assessment_k~age_factor,data=dat.2))
```

```
##             Df   Sum Sq Mean Sq F value   Pr(>F)
## age_factor   1  2536324 2536324   55.83 5.86e-13 ***
## Residuals  366 16626359   45427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

same test

but

different
MSW

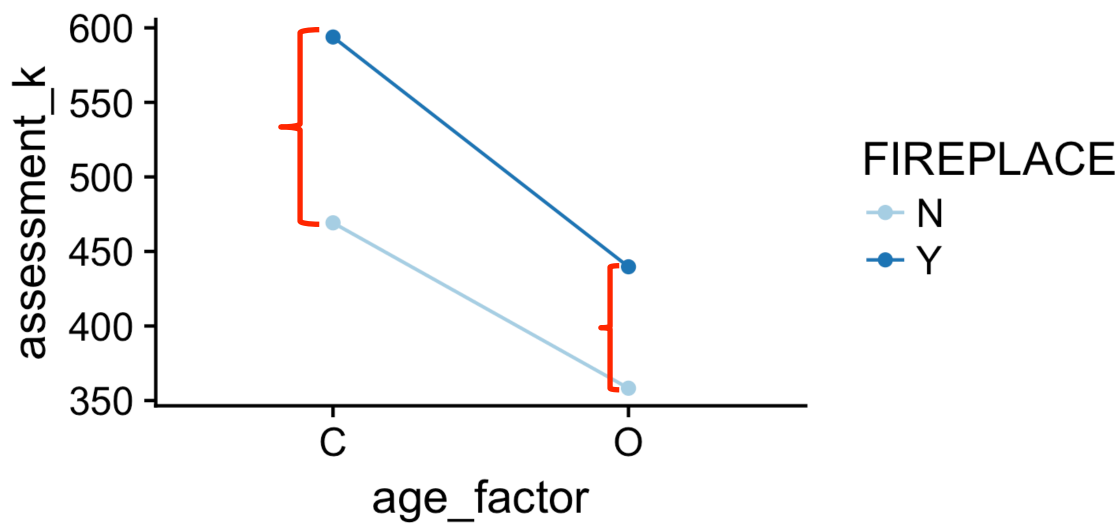
Note that both the residuals sum of squares and the degrees of freedom are different! Part of the variation is explained by “FIREPLACE” in the 2-way ANOVA

Two-way ANOVA: interaction

```
#Two-way ANOVA table
summary(aov(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##                      Df   Sum Sq Mean Sq F value   Pr(>F)
## age_factor            1  2536324 2536324  57.352 3.03e-13 ***
## FIREPLACE             1   509278  509278  11.516 0.000766 ***
## age_factor:FIREPLACE  1    19684   19684   0.445 0.505095
## Residuals           364 16097397   44224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Is the “FIREPLACE” effect the same at all age periods?



Note that the lines do not have any meaning here. These are NOT regression lines!! They just illustrate the trends

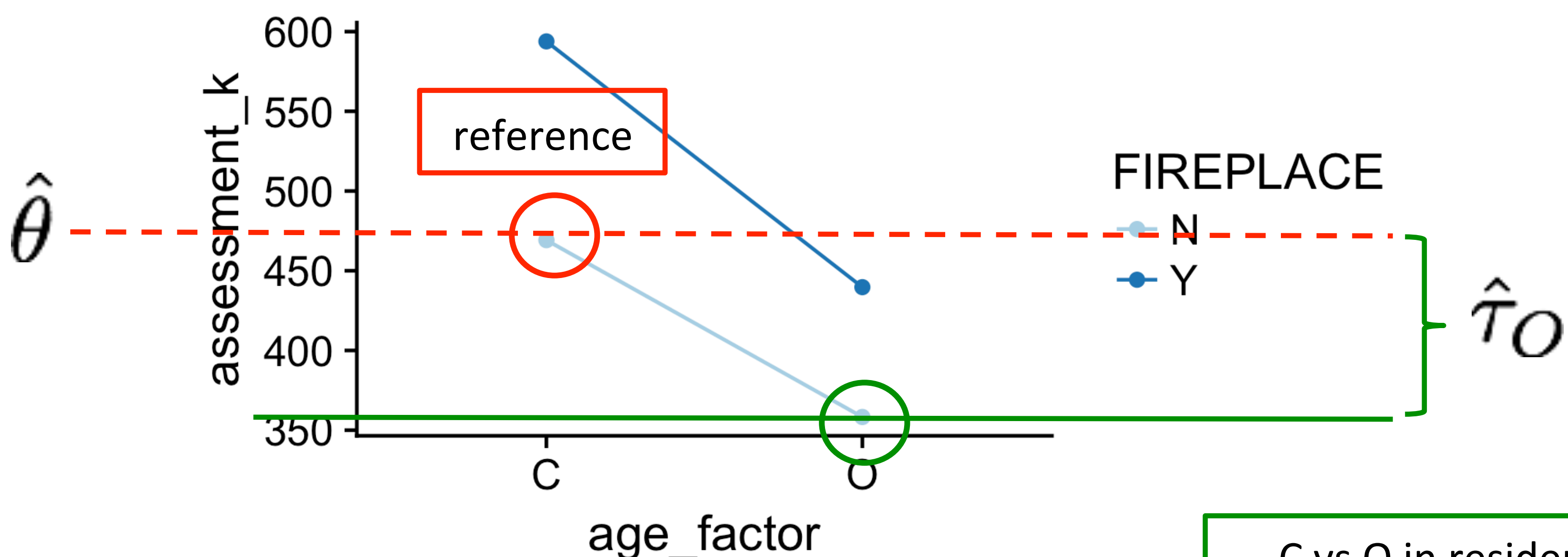
The default parametrization in lm() function

```
summary(lm(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor * FIREPLACE, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -446.84  -93.74  -44.05   21.56 2314.16
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             469.20      54.30   8.641   <2e-16 ***
## age_factor0            -110.94      59.69  -1.859   0.0639 .
## FIREPLACEY              124.64      57.21   2.178   0.0300 *
## age_factor0:FIREPLACEY  -43.20      64.75  -0.667   0.5051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 364 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:  0.153
## F-statistic:  23.1 on 3 and 364 DF,  p-value: 1.033e-13
```

Which null hypotheses are tested?

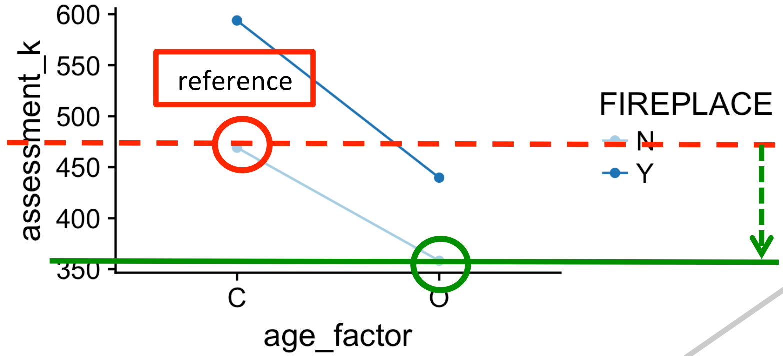
Which null hypotheses are tested by default in lm()?



C vs O in residencies *without* FIREPLACE

C *vs* O
without FIREPLACE

$\hat{\theta}$



$\hat{\tau}_O$

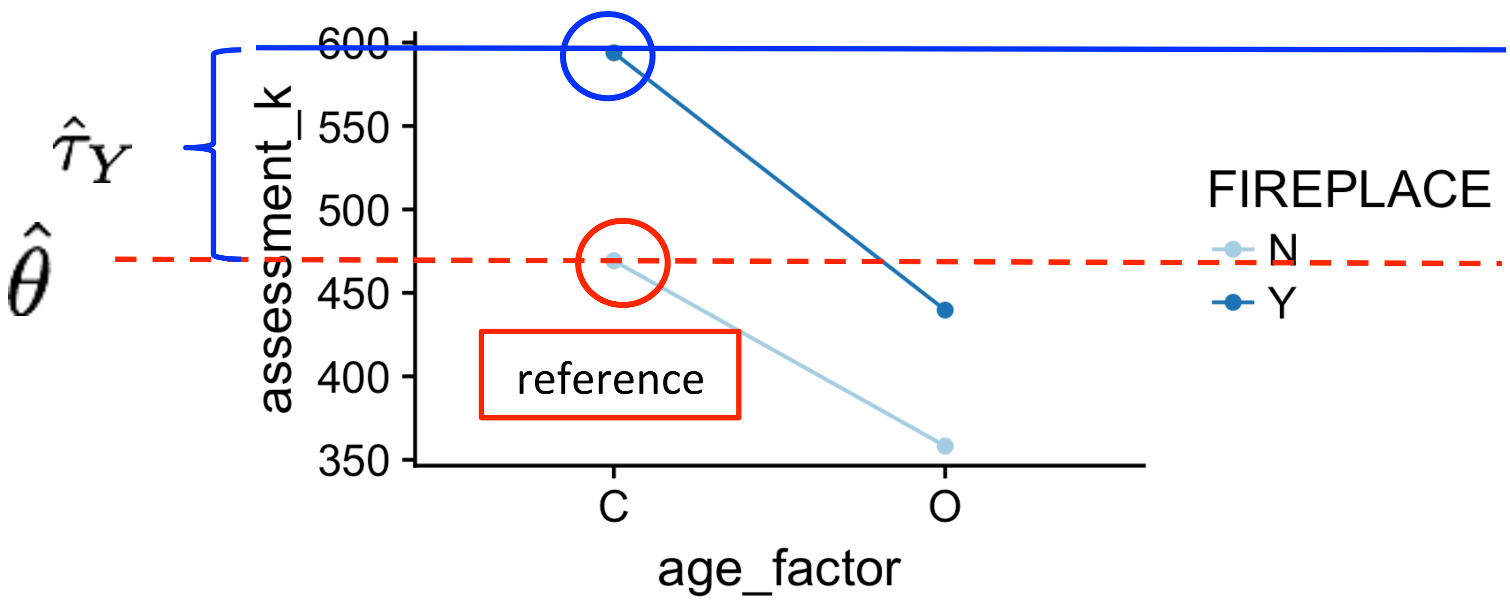
```r
summary(lm(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor * FIREPLACE, data = dat.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -446.84  -93.74  -44.05    21.56 2314.16
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              469.20      54.30   8.641   <2e-16 ***
## age_factorO             -110.94      59.69  -1.859   0.0639 .
## FIREPLACEY               124.64      57.21   2.178   0.0300 *
## age_factorO:FIREPLACEY   -43.20      64.75  -0.667   0.5051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 364 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:   0.153
## F-statistic:  23.1 on 3 and 364 DF,  p-value: 1.033e-13
```

$$H_0 : \tau_O = 0$$
$$H_0 : \mu_{ON} = \mu_{CN}$$

Note that this is not $H_0 : \mu_O = \mu_C$

FIREPLACE effect
in C residencies



```r
summary(lm(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor * FIREPLACE, data = dat.2)
##
## Residuals:
##      Min       1Q  Median      3Q     Max
## -446.84   -93.74  -44.05   21.56 2314.16
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              469.20      54.30   8.641   <2e-16 ***
## age_factorO             -110.94      59.69  -1.859   0.0639 .
## FIREPLACEY               124.64      57.21   2.178   0.0300 *
## age_factorO:FIREPLACEY   -43.20      64.75  -0.667   0.5051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 364 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:   0.153
## F-statistic:  23.1 on 3 and 364 DF,  p-value: 1.033e-13
```

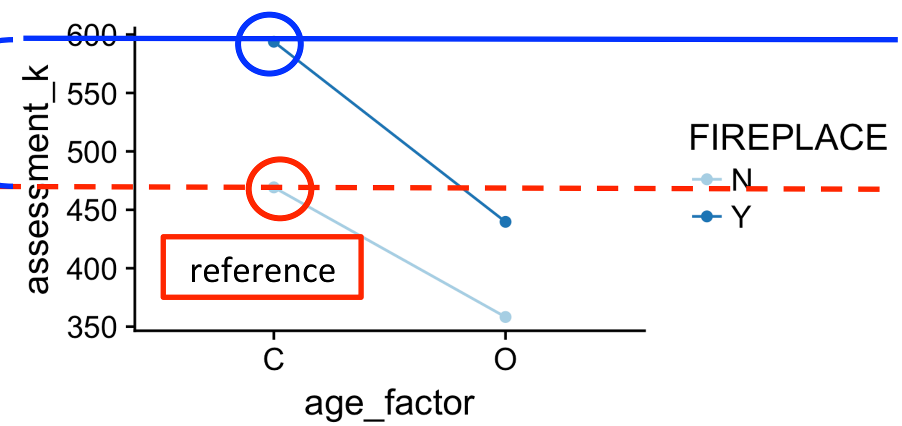
Note that this is not testing an overall « fireplace » effect

$$H_0 : \tau_Y = 0$$

$$H_0 : \mu_{CY} = \mu_{CN}$$

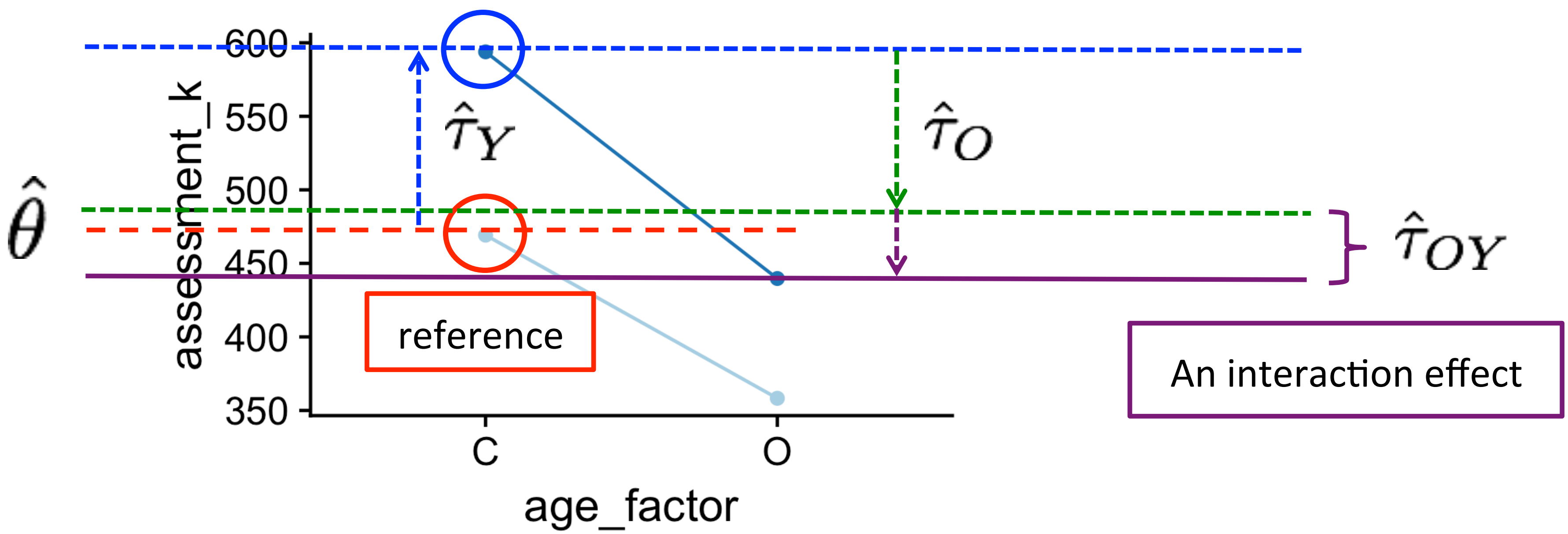
$\hat{\tau}_Y$

$\hat{\theta}$

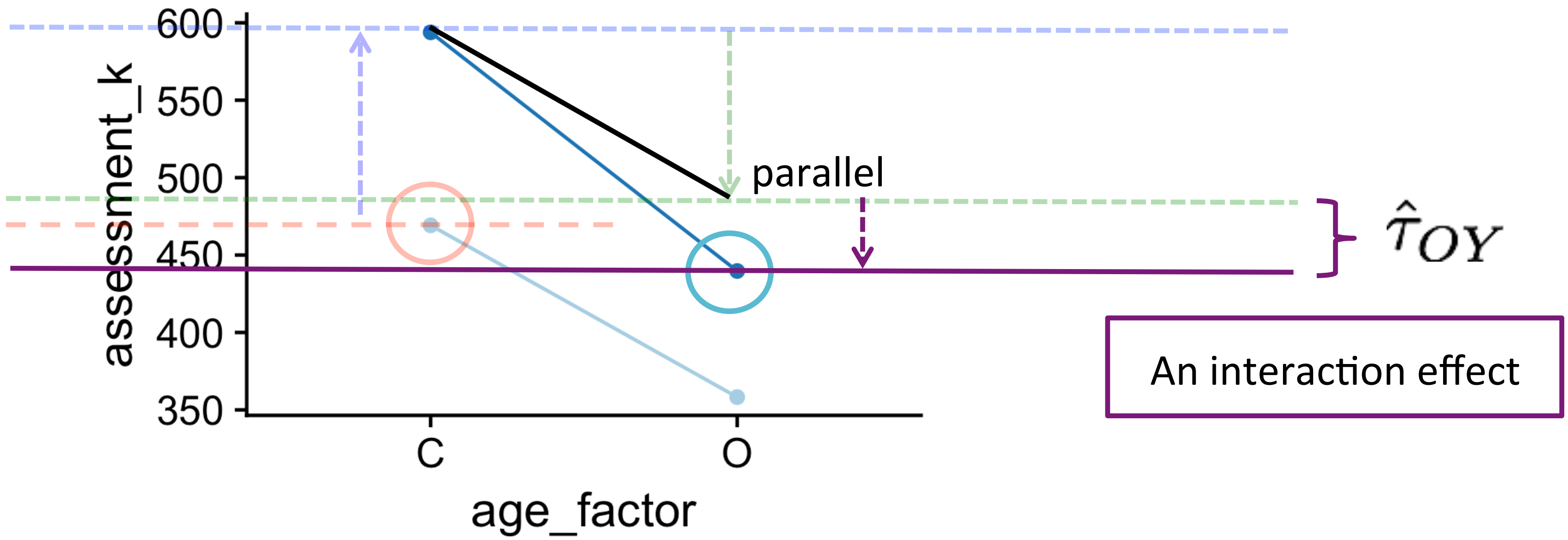


FIREPLACE effect in C residencies

Interaction: is the “FIREPLACE” effect the same at all age periods?



Interaction: can the age-effect in houses without fireplace (green arrow) be added to the fireplace-effect in contemporary houses (blue arrow) to estimate the mean tax assessment of old houses with fireplace (turquoise circle)?



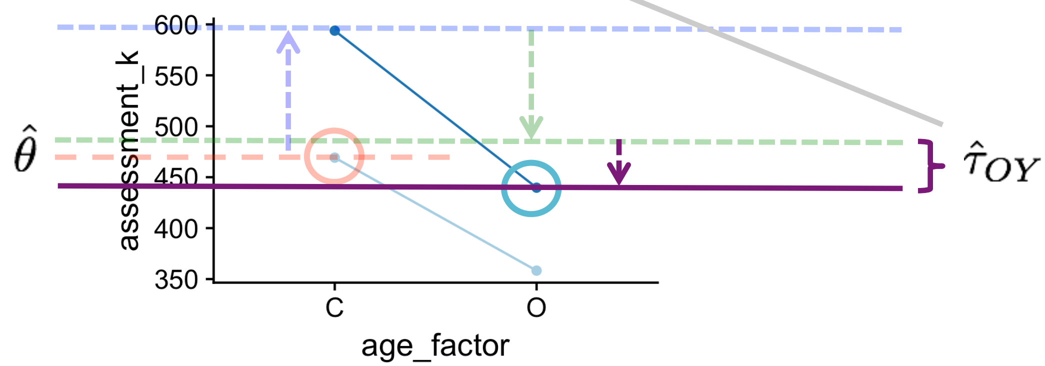
```r
summary(lm(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor * FIREPLACE, data = dat.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -446.84   -93.74   -44.05    21.56  2314.16
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              469.20      54.30   8.641   <2e-16 ***
## age_factorO             -110.94      59.69  -1.859   0.0639 .
## FIREPLACEY               124.64      57.21   2.178   0.0300 *
## age_factorO:FIREPLACEY   -43.20      64.75  -0.667   0.5051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 364 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:  0.153
## F-statistic:  23.1 on 3 and 364 DF,  p-value: 1.033e-13
```

The fireplace-effect in C is the same as that in O residencies (see pict. in slide 15). Can you state another equivalent hypothesis?

$$H_0 : \tau_{OY} = 0$$

$$H0 : \mu_{CY} - \mu_{CN} = \mu_{OY} - \mu_{ON}$$



An interaction effect

2 categorical variables

age (2 levels) and FIREPLACE (2 levels)

$$Y = X\alpha + \varepsilon$$

age_factor	FIREPLACE	assessment_k
C	N	390
C	N	541
C	N	364
...	...	...
C	Y	449
C	Y	536
C	Y	595
C	Y	449
...	...	...
O	N	355
O	N	396
...	...	...
O	Y	354
O	Y	363
...	...	...

$$\begin{bmatrix} Y_{CN1} \\ Y_{CN2} \\ \vdots \\ Y_{CY1} \\ \vdots \\ Y_{ON1} \\ \vdots \\ Y_{OY1} \\ \vdots \\ Y_{OY145} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_Y \\ \tau_O \\ \tau_{OY} \end{bmatrix} + \begin{bmatrix} \varepsilon_{CN1} \\ \varepsilon_{CN2} \\ \vdots \\ \varepsilon_{CY1} \\ \vdots \\ \varepsilon_{ON1} \\ \vdots \\ \varepsilon_{OY1} \\ \vdots \\ \varepsilon_{OY145} \end{bmatrix}$$

Reference CY

FIREPLACE in C

interaction

C vs O without FIREPLACE

Parametrizations (population)

$$Y_{CN} \sim F_1;\ E[Y_{CN}] = \mu_{CN}$$
$$Y_{CY} \sim F_2;\ E[Y_{CY}] = \mu_{CY}$$
$$Y_{ON} \sim F_3;\ E[Y_{ON}] = \mu_{ON}$$
$$Y_{OY} \sim F_4;\ E[Y_{OY}] = \mu_{OY}$$

$$E[Y_{CN}] = \theta$$
$$E[Y_{CY}] = \theta + \tau_Y$$
$$E[Y_{ON}] = \theta + \tau_O$$
$$E[Y_{OY}] = \theta + \tau_Y + \tau_O + \tau_{OY}$$

Then,

$$\theta = E[Y_{CN}]$$
$$\tau_F = E[Y_{CY}] - E[Y_{CN}]$$
$$\tau_O = E[Y_{ON}] - E[Y_{CN}]$$
$$\tau_{OF} = E[Y_{OY}] - E[Y_{CY}] - E[Y_{ON}] + E[Y_{CN}]$$

By default, lm() tests whether each of these is zero

ANOVA vs Regression: only interaction is the same

```r
#Two-way ANOVA table
summary(aov(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##                       Df   Sum Sq Mean Sq F value   Pr(>F)
## age_factor             1  2536324 2536324  57.352 3.03e-13 ***
## FIREPLACE              1   509278  509278  11.516 0.000766 ***
## age_factor:FIREPLACE   1    19684   19684   0.445 0.505095
## Residuals            364 16097397   44224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Age-effect (ignoring fireplace-effect)

$$H_0 : \mu_C = \mu_O$$

```r
summary(lm(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor * FIREPLACE, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -446.84  -93.74  -44.05   21.56 2314.16
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             469.20      54.30   8.641   <2e-16 ***
## age_factorO            -110.94      59.69  -1.859   0.0639 .
## FIREPLACEY              124.64      57.21   2.178   0.0300 *
## age_factorO:FIREPLACEY  -43.20      64.75  -0.667   0.5051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 364 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:  0.153
## F-statistic:  23.1 on 3 and 364 DF,  p-value: 1.033e-13
```

Note: aov() gives sequential type I SS. Thus, the first row ignores the fireplace-effect. The second row, tests the fireplace-effect, on average over age.

Conditional effect: C *vs* O without FIREPLACE

$$H_0 : \tau_O = 0$$

$$H_0 : \mu_{ON} = \mu_{CN}$$

ANOVA vs Regression: only interaction is the same

```r
#Two-way ANOVA table
summary(aov(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##                       Df   Sum Sq Mean Sq F value   Pr(>F)
## age_factor             1  2536324 2536324  57.352 3.03e-13 ***
## FIREPLACE              1   509278  509278  11.516 0.000766 ***
## age_factor:FIREPLACE   1    19684   19684   0.445 0.505095
## Residuals            364 16097397   44224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \tau_{OY} = 0$

```r
summary(lm(assessment_k~age_factor*FIREPLACE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor * FIREPLACE, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -446.84  -93.74  -44.05   21.56 2314.16
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 469.20      54.30   8.641   <2e-16 ***
## age_factorO                -110.94      59.69  -1.859   0.0639 .
## FIREPLACEY                  124.64      57.21   2.178   0.0300 *
## age_factorO:FIREPLACEY      -43.20      64.75  -0.667   0.5051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 364 degrees of freedom
## Multiple R-squared:   0.16,  Adjusted R-squared:  0.153
## F-statistic:  23.1 on 3 and 364 DF,  p-value: 1.033e-13
```

Equivalent tests

$H0 : \mu_{CY} - \mu_{CN} = \mu_{OY} - \mu_{ON}$

Summary

- Two-sample t-test and ANOVA are special cases of a linear model
- Two-way ANOVA is a special case of a linear model
- By default, R uses the “reference-treatment” parametrization in `lm()`
- By now, you should be able to recognize which null hypotheses are tested by default with lm()
- If you are interested in other comparisons, you can estimate and test a contrast (see companion Rmd)
- The `lm()` output includes an F-test to test a full vs an “intercept-only” model