DSCI561: Regression I

Lecture 4: November 27, 2017

Gabriela Cohen Freue

Department of Statistics, UBC

Review from lecture 3

- Analyze the output of lm() in relation with the mathematical formulation of the model
- Linear models with more than one categorical variable
- By default, R uses the "reference-treatment" parametrization in `lm()`
- We can test other hypotheses with "contrast"

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

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

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

$$H_0 : \tau_O = 0$$

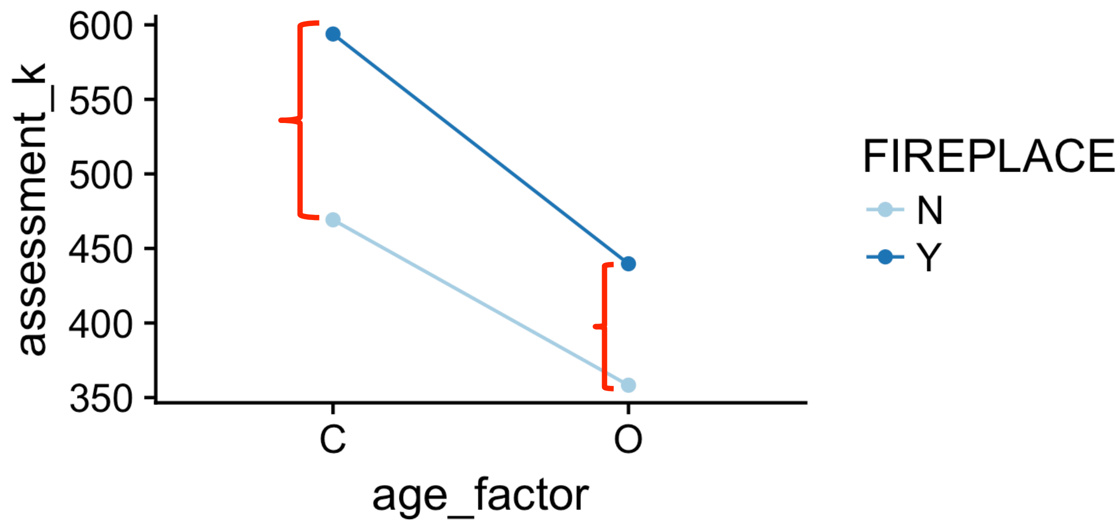
$$H_0 : \mu_{ON} = \mu_{CN}$$

Interaction 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
```

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

In today's lecture

- Linear models with a continuous independent variable
- Linear models with both continuous and categorical variables

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

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**

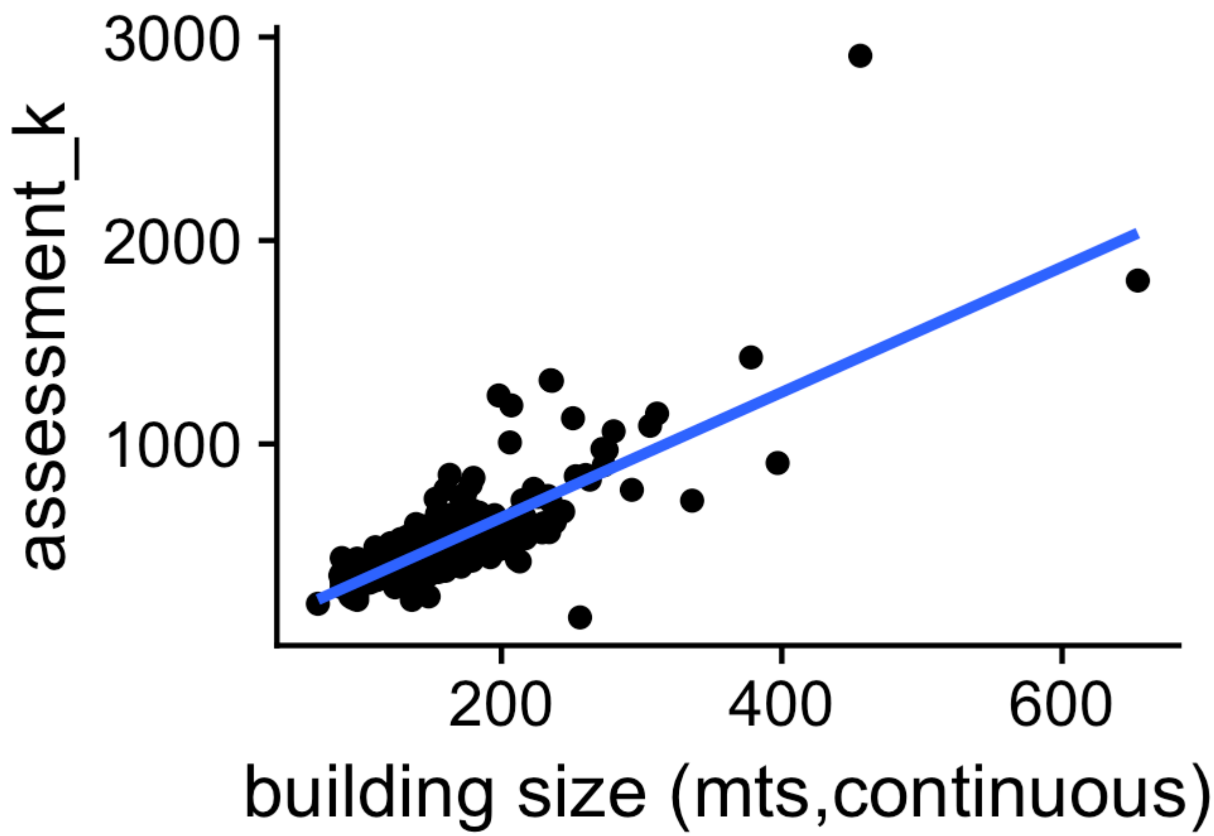
AND MANY MORE …..

Tip: ?model.matrix

Beyond categorical covariates: continuous

LINEAR REGRESSION

“BLDG_METRE” as a continuous variable



age_factor	FIREPLACE	BLDG_METRE	assessment_k
O	Y	97	354
C	Y	166	449
O	N	108	383
C	Y	217	536
C	Y	145	595
C	Y	171	449
O	Y	106	363
O	Y	160	776
O	N	99	349
O	N	104	371
O	Y	100	346
O	Y	110	358
O	Y	223	575
O	Y	168	608
C	Y	120	505
O	Y	110	329
C	Y	244	667
C	Y	226	739
O	Y	110	429
...	...	...	...

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

intercept

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{368} \end{bmatrix} = \begin{bmatrix} 1 & 97 \\ 1 & 166 \\ 1 & 108 \\ \vdots & \vdots \\ 1 & 110 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{368} \end{bmatrix}$$

slope

Simple linear regression

```r
summary(lm(assessment_k~BLDG_METRE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ BLDG_METRE, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -663.35  -62.18   -2.37   39.15 1481.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0460    19.4790   1.132    0.258
## BLDG_METRE    3.0793     0.1213  25.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.7 on 366 degrees of freedom
## Multiple R-squared:  0.638,  Adjusted R-squared:  0.637
## F-statistic:   645 on 1 and 366 DF,  p-value: < 2.2e-16
```

(usually, not of interest)

$$\boxed{H_0 : \beta_0 = 0}$$

Simple linear regression

```r
summary(lm(assessment_k~BLDG_METRE,data=dat.2))
```

```
##
## Call:
## lm(formula = assessment_k ~ BLDG_METRE, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -663.35  -62.18   -2.37   39.15 1481.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0460    19.4790   1.132    0.258
## BLDG_METRE    3.0793     0.1213  25.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.7 on 366 degrees of freedom
## Multiple R-squared:  0.638,  Adjusted R-squared:  0.637
## F-statistic:   645 on 1 and 366 DF,  p-value: < 2.2e-16
```

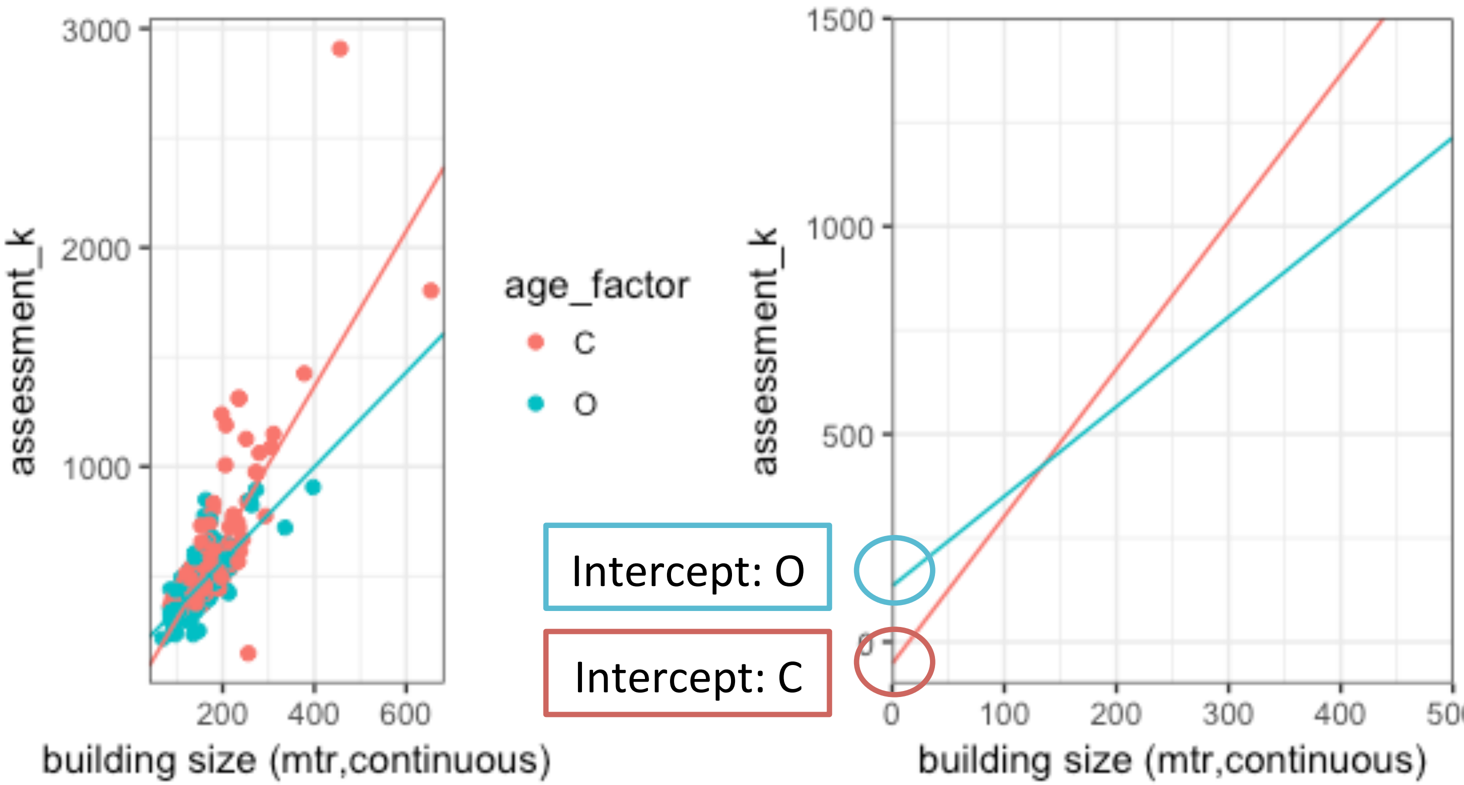
$$\boxed{H_0 : \beta_1 = 0}$$

categorical and continuous covariates

LINEAR REGRESSION

“BLDG_METRE”
continuous variable

“age_factor”
categorical variable



age_factor	BLDG_METRE	assessment_k
C	166	449
C	217	536
C	145	595
C	171	449
C	120	505
C	244	667
C	226	739
C	178	799
C	197	523
C	235	718
C	128	412
C	184	468
...	...	...
O	97	354
O	108	383
O	106	363
O	160	776
O	99	349
O	104	371
O	100	346
O	110	358
O	223	575
...	...	...

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

Intercept: C

$$\begin{bmatrix} Y_{C1} \\ Y_{C2} \\ \vdots \\ Y_{O1} \\ Y_{O2} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 166 & 0 \\ 1 & 0 & 217 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 97 & 97 \\ 1 & 1 & 108 & 108 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \beta_{C0} \\ \tau_0 \\ \beta_{C1} \\ \tau_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{C1} \\ \varepsilon_{C2} \\ \vdots \\ \varepsilon_{O1} \\ \varepsilon_{O2} \\ \vdots \end{bmatrix}$$

Slope: C

Intercept: O vs C

slope: O vs C

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

```
##
## Call:
## lm(formula = assessment_k ~ BLDG_METRE * age_factor, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -708.11  -48.09   -8.34   36.01 1343.92
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -52.3654    31.1763  -1.680   0.0939 .
## BLDG_METRE                 3.5448     0.1635  21.677  < 2e-16 ***
## age_factor0              186.8247    42.1280   4.435 1.22e-05 ***
## BLDG_METRE:age_factor0    -1.3856     0.2649  -5.230 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132.9 on 364 degrees of freedom
## Multiple R-squared:  0.6644, Adjusted R-squared:  0.6617
## F-statistic: 240.3 on 3 and 364 DF,  p-value: < 2.2e-16
```

$$\boxed{H_0 : \beta_{C1} = 0}$$

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

```
##
## Call:
## lm(formula = assessment_k ~ BLDG_METRE * age_factor, data = dat.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -708.11  -48.09   -8.34   36.01 1343.92
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -52.3654    31.1763  -1.680   0.0939 .
## BLDG_METRE               3.5448     0.1635  21.677  < 2e-16 ***
## age_factor0            186.8247    42.1280   4.435 1.22e-05 ***
## BLDG_METRE:age_factor0  -1.3856     0.2649  -5.230 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132.9 on 364 degrees of freedom
## Multiple R-squared:  0.6644, Adjusted R-squared:  0.6617
## F-statistic: 240.3 on 3 and 364 DF,  p-value: < 2.2e-16
```

$$H_0 : \tau_1 = 0$$

$$H_0 : \beta_{C1} = \beta_{O1}$$

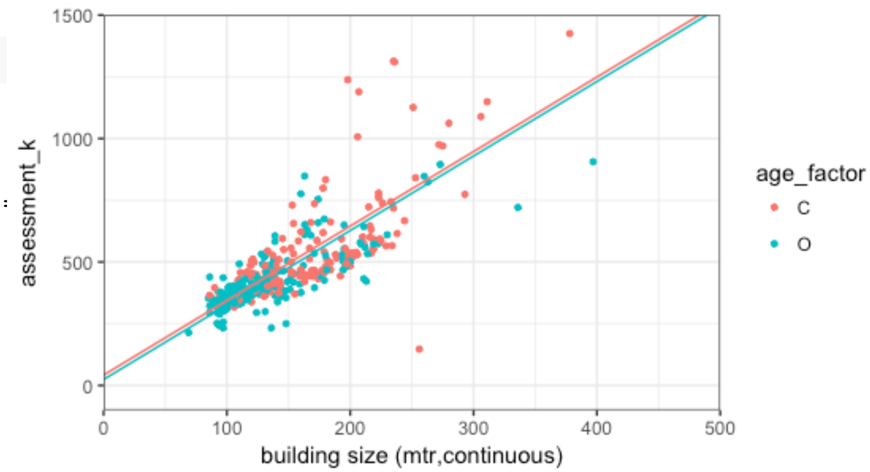
Additive models

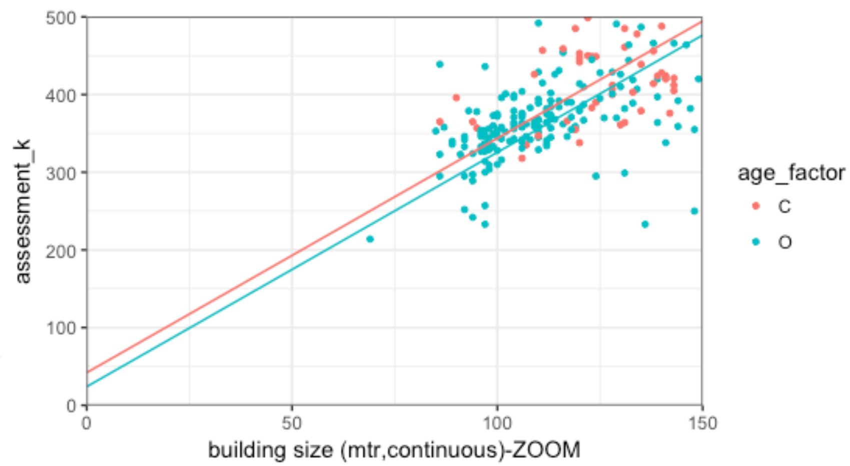
- In some applications you may want to ignore the interaction between variables

```
summary(lm(assessment_k~BLDG_METRE+age_factor,data=dat.2))
```

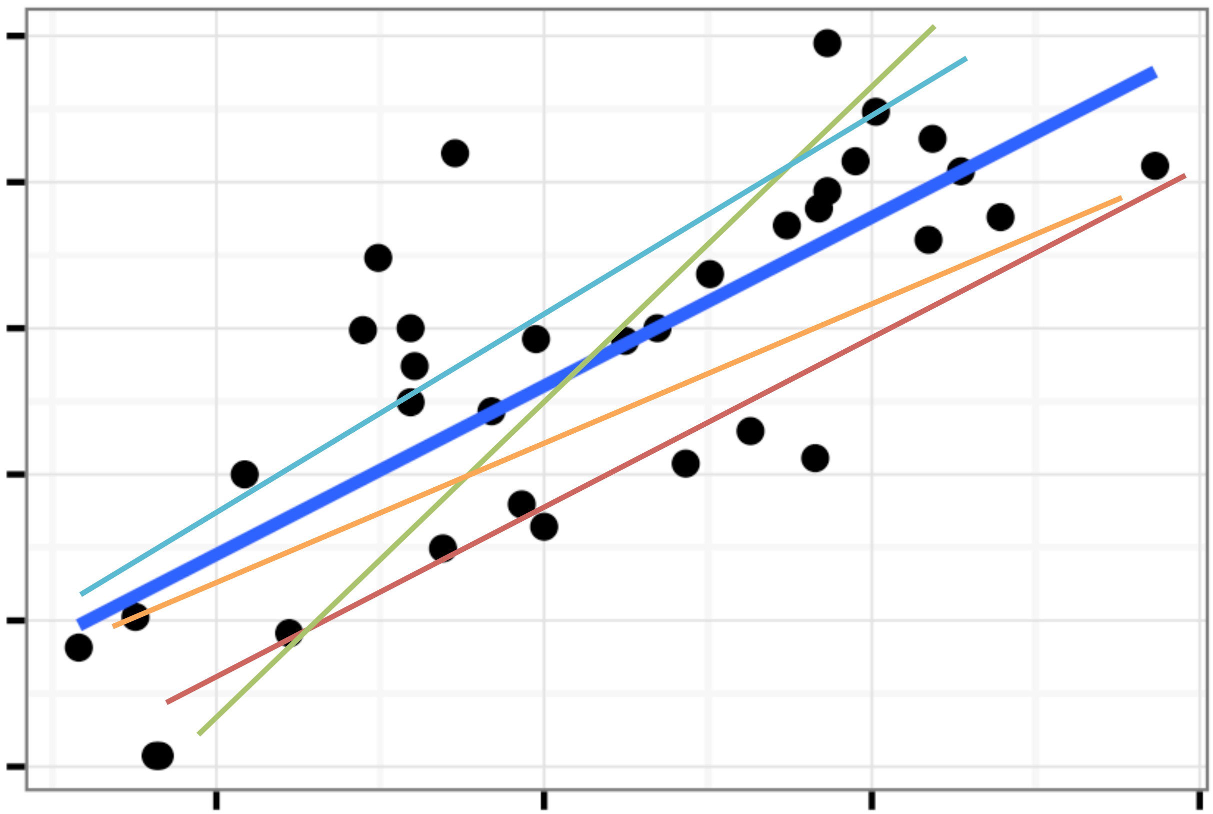
```
##
## Call:
## lm(formula = assessment_k ~ BLDG_METRE + age_factor, data =
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -661.54   -66.04    -0.50    40.24  1493.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    32.571     29.024   1.122    0.263
## BLDG_METRE      3.031      0.144  21.052   <2e-16 ***
## age_factorO   -10.890     18.189  -0.599    0.550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 141 on 295 degrees of freedom
## Multiple R-squared:  0.6505, Adjusted R-squared:  0.6481
## F-statistic: 274.5 on 2 and 295 DF,  p-value: < 2.2e-16
```

Common slope





Which one is the best line?



The error is the vertical distance between the line and the real observation

Ordinary least squares (OLS) estimates of the parameters minimize the sum of squares of the errors

Ordinary Least Square Estimator

Visual representation of the squared errors
http://setosa.io/ev/ordinary-least-squares-regression/

- The squares of the errors are represented by squared areas in the second plot:
  - select different lines by changing the intercept and the slope
  - see how the squares of the errors change
  - Which line minimizes the sum of these areas? OLS answers this question
- Move a point of the first plot along the line and away from the line. See how see how sensitive is the estimation.

Ordinary Least Square Estimator

Mathematically:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n$$

We want to find a line (i.e., an intercept and a slope) such that the sum of the squared errors is minimized

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Using results from Calculus:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

- The values β_0, β_1 that satisfies these equations are the ***OLS*** estimates of the intercept and the slope, respectively.
- Estimates are represented by a “hat” over the parameter.

- After simplification, the previous equations become

$$n(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 n \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

Thus,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

And

$$0 = \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{(n-1)s_{xy}}{(n-1)s_x^2}$$

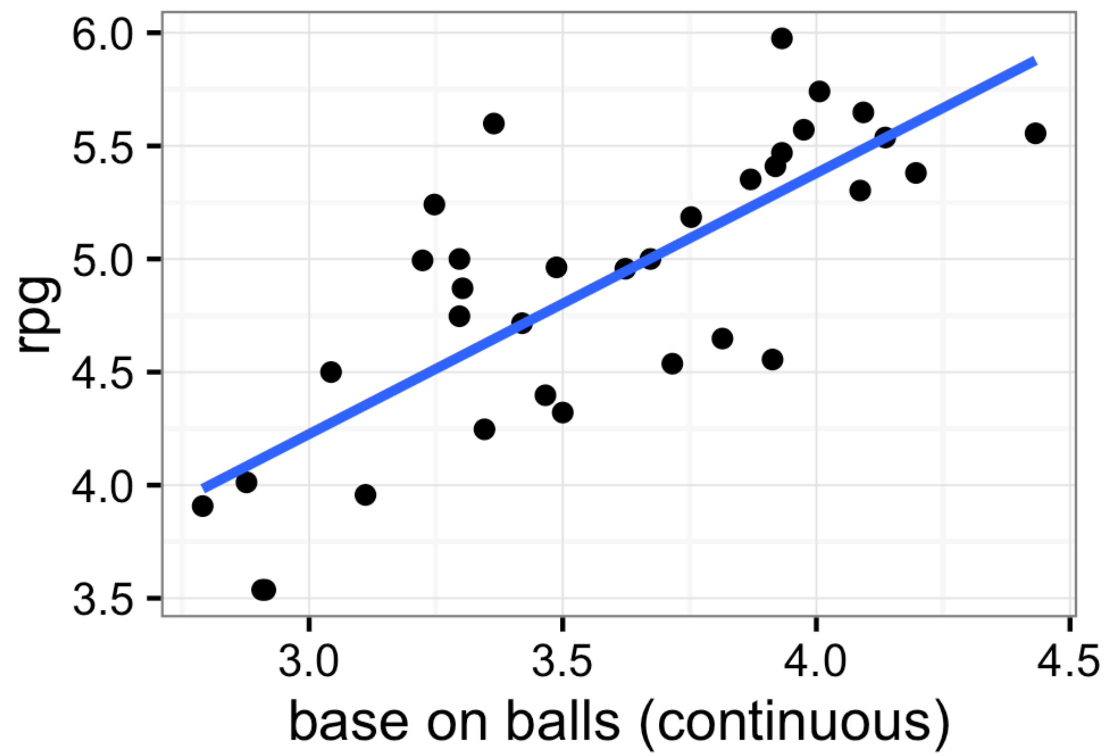
$$= \frac{r_{xy} s_x s_y}{s_x^2}$$

r_{xy} is the correlation between the response and the explanatory variable

$$= \frac{r_{xy} s_y}{s_x}$$

Sx and *Sy* are the standard deviation of the response and the explanatory variable, resp.

$$\frac{y_i - \bar{y}}{s_y} = r_{xy} \frac{x_i - \bar{x}}{s_x}$$



The linear relation between two continuous variables is characterized by their *correlation*

Simple linear regression

```r
#BB continuous
summary(lm(rpg~bbpg,data=teams.2small))
```

```
##
## Call:
## lm(formula = rpg ~ bbpg, data = teams.2small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72450 -0.35515  0.00861  0.21001  0.95257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7647     0.6003   1.274    0.212
## bbpg          1.1538     0.1666   6.926 7.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4114 on 32 degrees of freedom
## Multiple R-squared:  0.5998, Adjusted R-squared:  0.5873
## F-statistic: 47.97 on 1 and 32 DF,  p-value: 7.667e-08
```

$$\boxed{\hat{\beta}_1 = \frac{r_{xy} s_y}{s_x}}$$

Simple linear regression

```r
#BB continuous
summary(lm(rpg~bbpg,data=teams.2small))
```

```
##
## Call:
## lm(formula = rpg ~ bbpg, data = teams.2small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72450 -0.35515  0.00861  0.21001  0.95257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7647     0.6003   1.274    0.212
## bbpg          1.1538     0.1666   6.926 7.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4114 on 32 degrees of freedom
## Multiple R-squared:  0.5998, Adjusted R-squared:  0.5873
## F-statistic: 47.97 on 1 and 32 DF,  p-value: 7.667e-08
```

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$