07-Régression

PRO1036 – Analyse des données scientifiques en R

Maribel Diaz

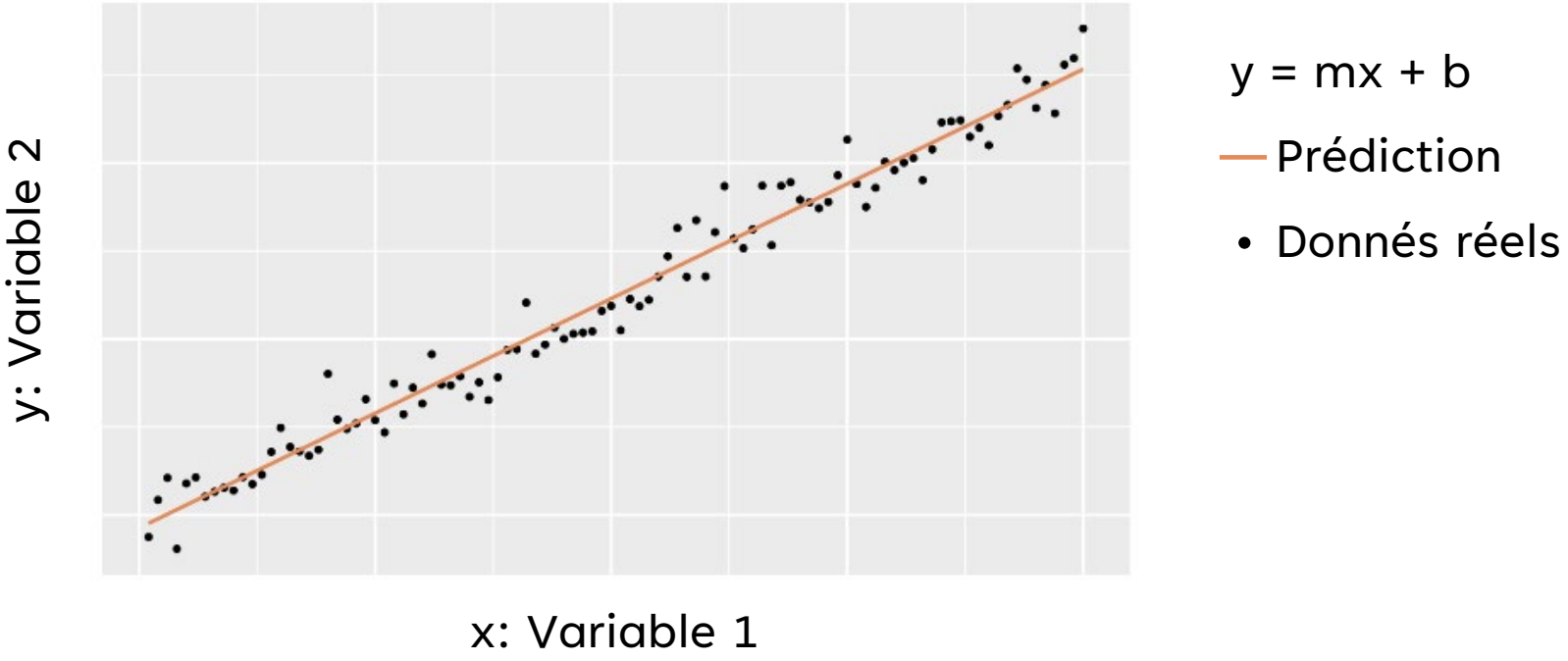
Coordinateur: Tim Bollé

11 novembre 2024

Les modèles

La modélisation

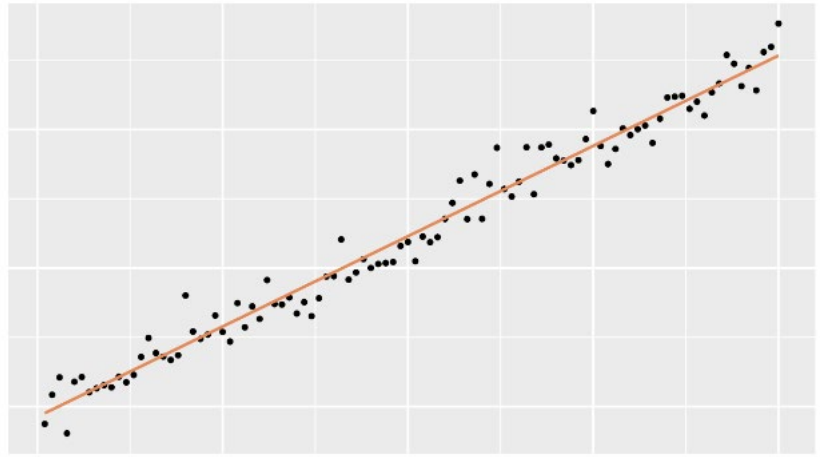
- On utilise des modèles pour expliquer la relation entre les variables et faire des prédictions



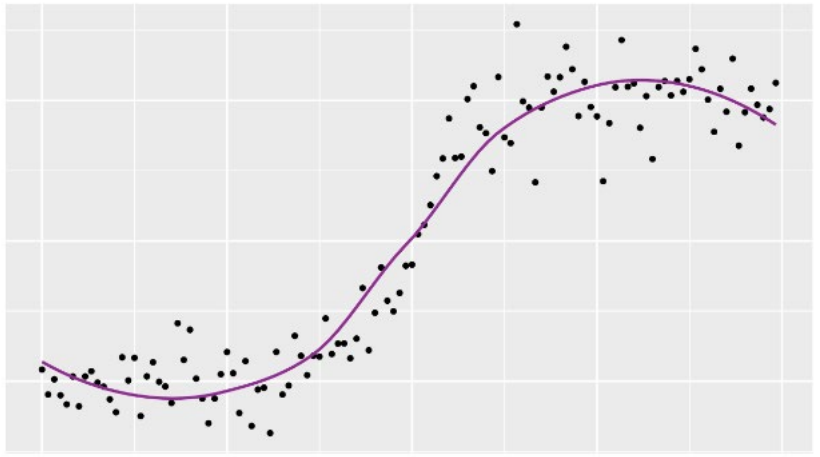
La modélisation

Nous nous concentrerons sur les modèles **linéaires** (mais n'oubliez pas qu'il existe de *nombreux* autres types de modèles !)

Linéaire



Non linéaire



Donnés à utiliser: « Paris paintings »

« Paris paintings »

- Source : Catalogues imprimés de 28 ventes aux enchères à Paris, 1764 - 1780

- Les conservateurs de données Sandra van Ginhoven et Hilary Coe Cronheim ont traduit et compilé les catalogues.

- 3393 peintures, leurs prix et les détails descriptifs des catalogues de vente sur 60 variables.

```
pp <- read_csv("data/paris-paintings.csv", na = c("n/a", "", "NA"))
```

```
pp %>%
filter(name == "R1777-89a") %>%
glimpse()
```

```
#  Rows: 1 Columns: 61
#  $ name <chr> "R1777-89a"
#  $ sale <chr> "R1777"
#  $ lot <chr> "89"
#  $ position <dbl> 0.3755274
#  $ dealer <chr> "R"
#  $ year <dbl> 1777
#  $ origin_author <chr> "D/FL"
#  $ origin_cat <chr> "D/FL"
#  $ school_pntg <chr> "D/FL"
#  $ diff_origin <dbl> 0
#  $ logprice <dbl> 8.575462
#  $ price <dbl> 5300
#  $ count <dbl> 1
#  $ subject <chr> "D\x8epart pour la chasse"
#  $ authorstandard <chr> "Wouwerman, Philips"
#  $ artistliving <dbl> 0
#  $ authorstyle <chr> NA
#  $ author <chr> "Philippe Wouwermans"
#  $ winningbidder <chr> "Langlier, Jacques for Poullain, Antoine"
#  $ winningbiddertype <chr> "DC"
#  $ endbuyer <chr> "C"
#  $ Interm <dbl> 1
#  $ type_intermed <chr> "D"
#  $ Height_in <dbl> 17.25
#  $ Width_in <dbl> 23
#  $ Surface_Rect <dbl> 396.75
#  $ Diam_in <dbl> NA
#  $ Surface_Rnd <dbl> NA
#  $ Shape <chr> "squ_rect"
#  $ Surface <dbl> 396.75
#  $ material <chr> "bois"
#  $ mat <chr> "b"
#  $ materialCat <chr> "wood"
#  $ quantity <dbl> 1
#  $ nfigures <dbl> 0
#  $ engraved <dbl> 0
#  $ original <dbl> 0
#  $ prevcoll <dbl> 1
#  $ othartist <dbl> 0
#  $ paired <dbl> 1
#  $ figures <dbl> 0
#  $ finished <dbl> 0
#  $ lrgfont <dbl> 0
#  $ relig <dbl> 0
#  $ landsALL <dbl> 1
#  $ lands_sc <dbl> 0
#  $ lands_elem <dbl> 1
#  $ lands_figs <dbl> 1
#  $ lands_ment <dbl> 0
#  $ arch <dbl> 1
#  $ mytho <dbl> 0
#  $ peasant <dbl> 0
#  $ othgenre <dbl> 0
#  $ singlefig <dbl> 0
#  $ portrait <dbl> 0
#  $ still_life <dbl> 0
#  $ discauth <dbl> 0
#  $ history <dbl> 0
#  $ allegory <dbl> 0
#  $ pastorale <dbl> 0
#  $ other <dbl> 0
```

Fonctions (mathématique)

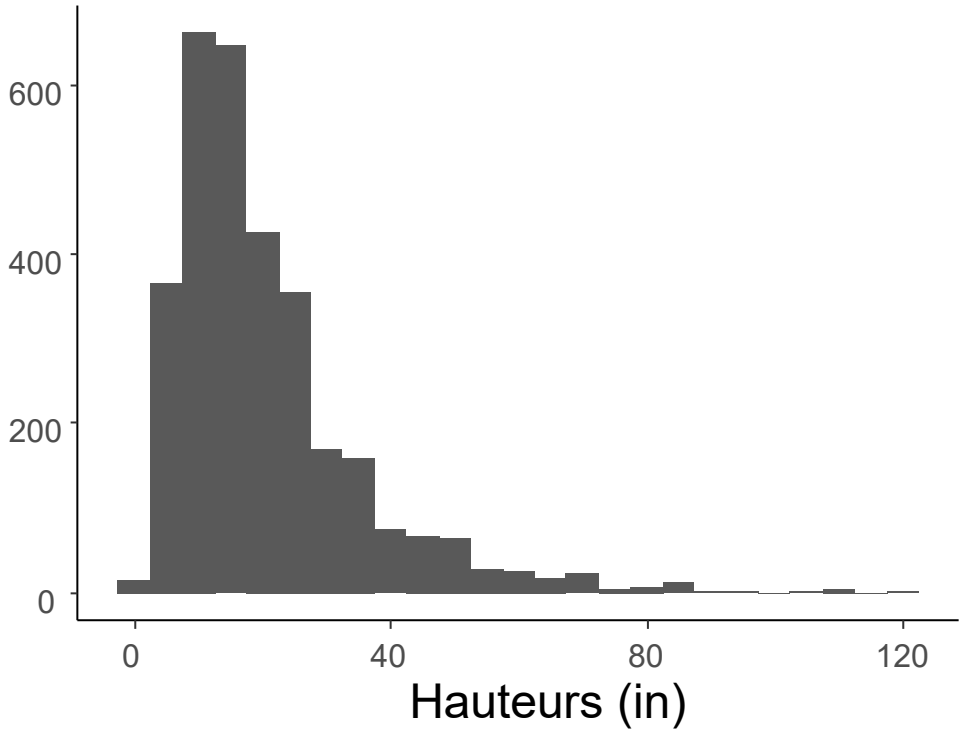
- Nous pouvons représenter les relations entre les variables à l'aide de **fonctions**

- Une fonction est un concept mathématique : la relation entre une sortie et une ou plusieurs entrées.

- Exemple : La formule y = 3x + 7 est une fonction avec une entrée x et une sortie y.
  - Si x est 5, y = ?

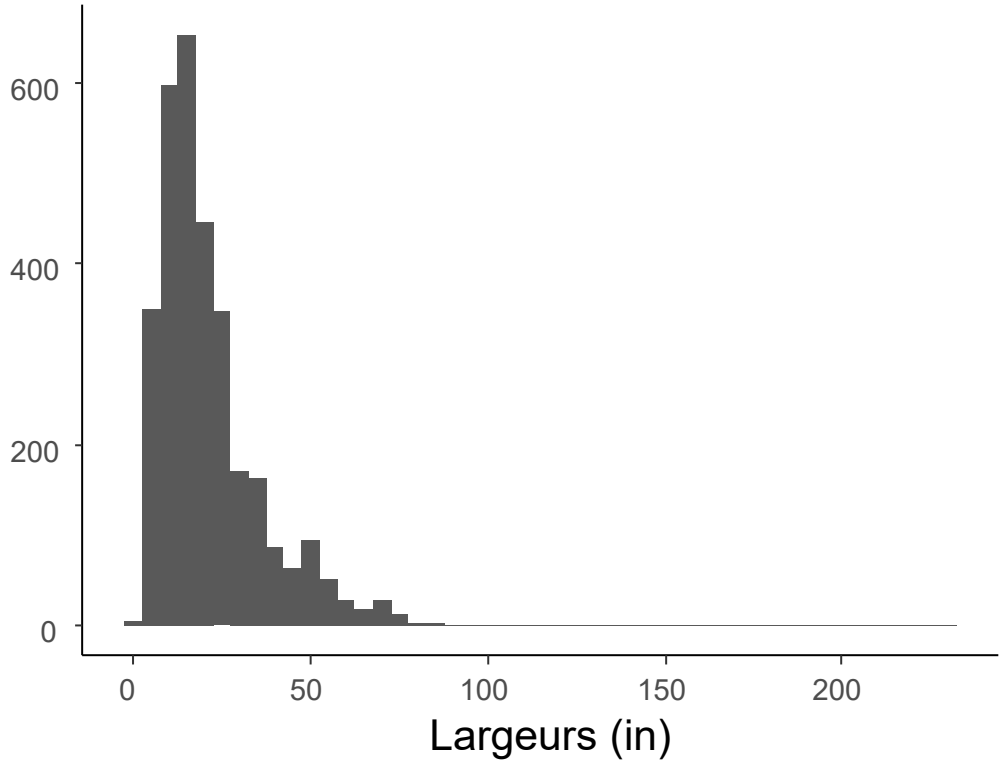
Hauteurs

```
ggplot(data = pp, aes(x = Height_in)) +
geom_histogram(binwidth = 5) +
labs(x = "Hauteurs (in)", y = NULL) +
theme_classic()
```

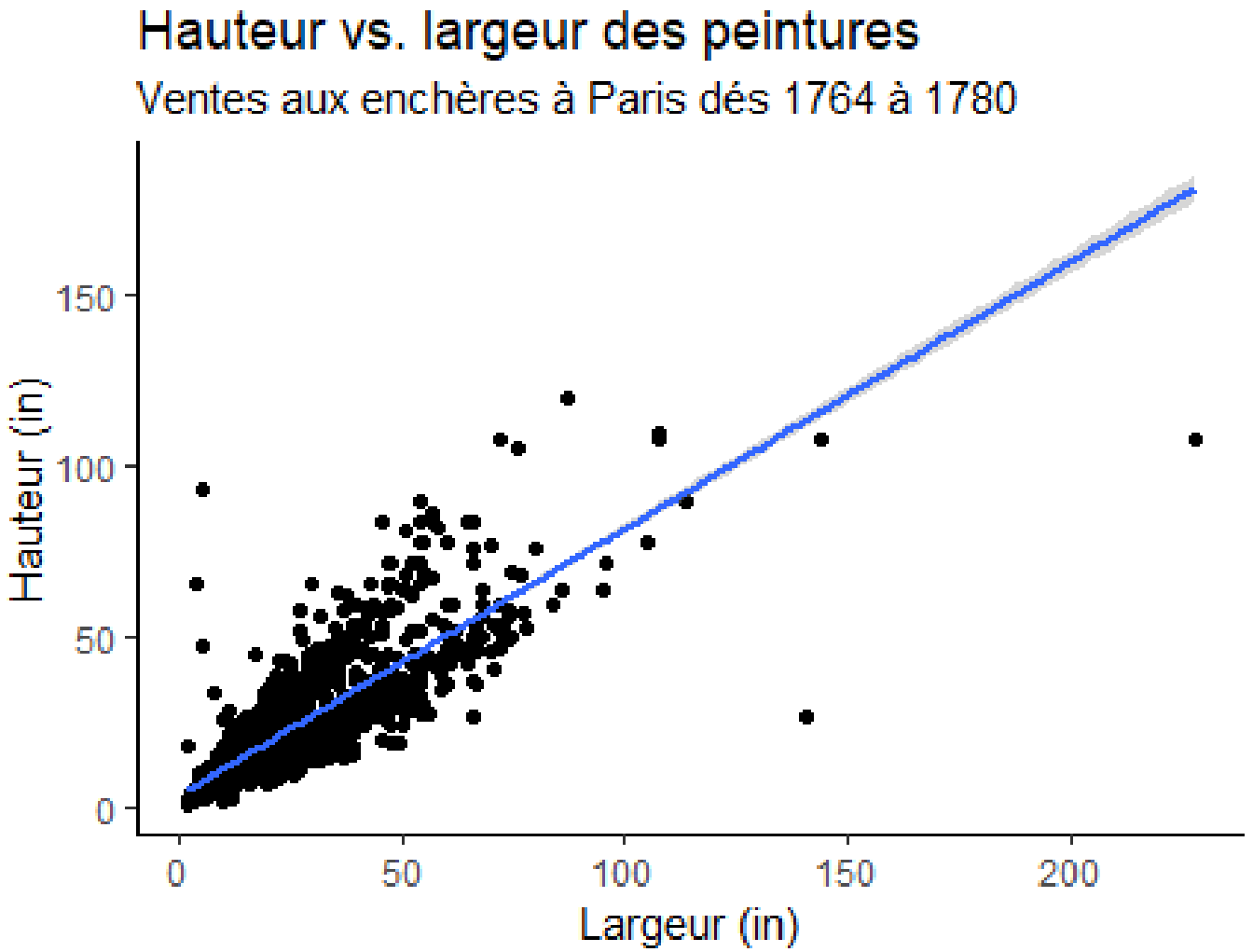


Largeurs

```
ggplot(data = pp, aes(x = Width_in)) +
geom_histogram(binwidth = 5) +
labs(x = “Largeurs (in)", y = NULL)+
theme_classic()
```



Hauteur en fonction de la largeur

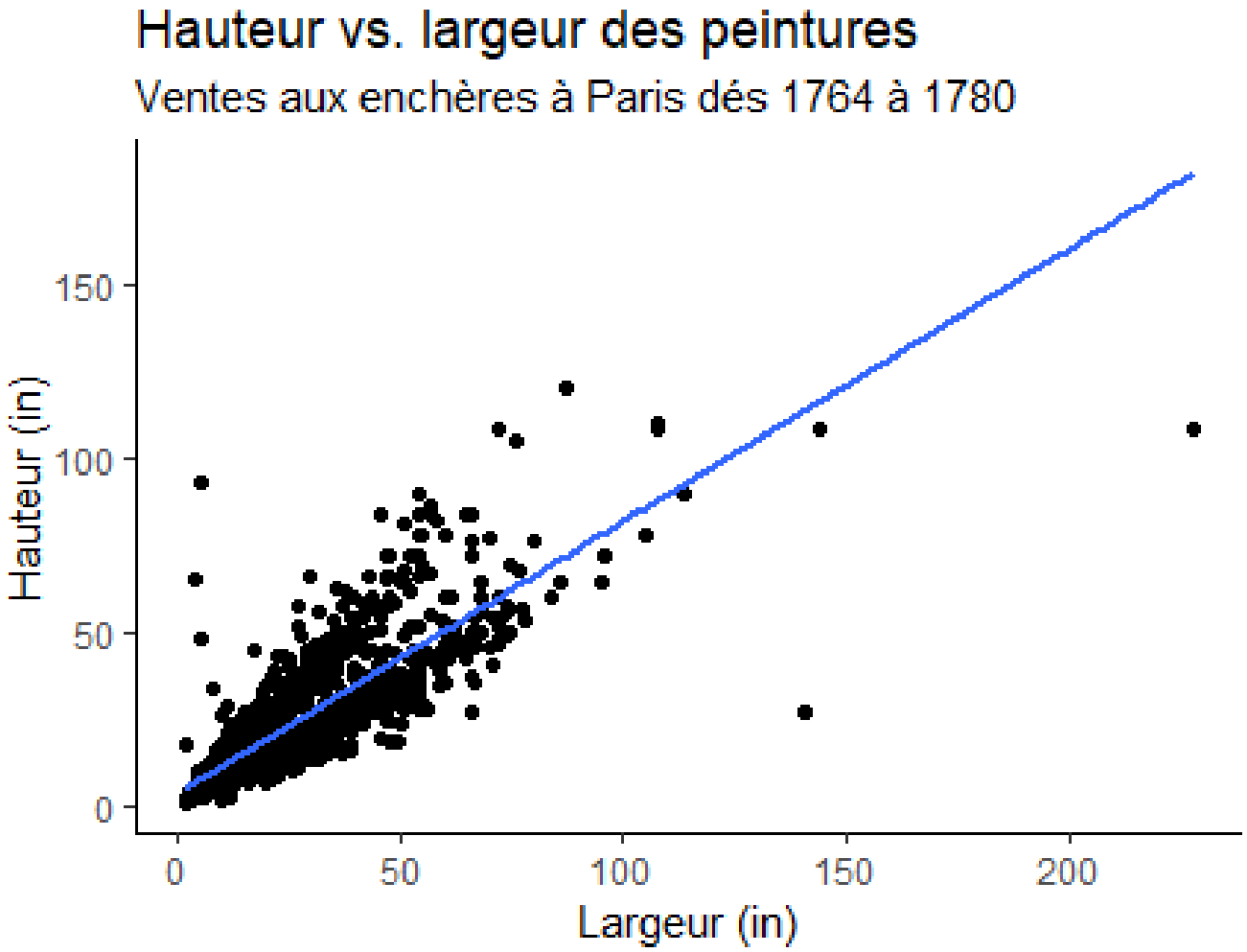


Hauteur en fonction de la largeur

Code:

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
geom_point() +
geom_smooth(method = "lm") +
labs(
title = "Hauteur vs. largeur des peintures",
subtitle = "Ventes aux enchères à Paris dés 1764 à 1780",
x = "Largeur (in)",
y = "Hauteur (in)"
)+
theme_classic()
```

...sans la mesure d'incertitude

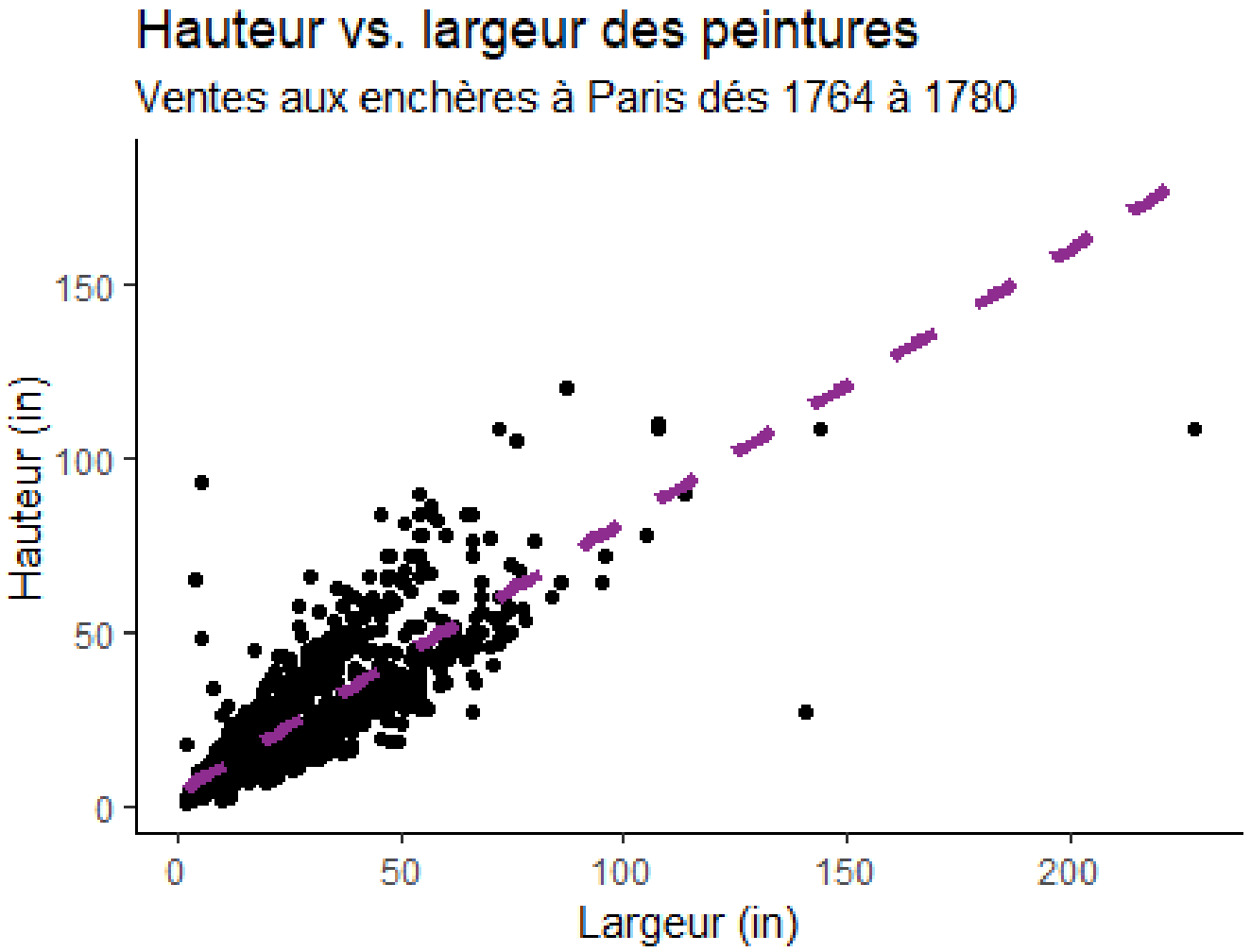


...sans la mesure d'incertitude

Code:

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
labs(
title = "Hauteur vs. largeur des peintures",
subtitle = "Ventes aux enchères à Paris dés 1764 à 1780",    x =
"Largeur (in)",
y = "Hauteur (in)"
)+
theme_classic()
```

...avec différents choix esthétiques



...avec différents choix esthétiques

Code:

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
      geom_point() +
      geom_smooth(method = "lm", se = FALSE,
                  color = "#8E2C90", linetype = "dashed",linewidth = 1.5) +
      labs(
            title = "Hauteur vs. largeur des peintures",
            subtitle = "Ventes aux enchères à Paris dés 1764 à 1780",
            x = "Largeur (in)",
            y = "Hauteur (in)"
      )+
      theme_classic()
```

Concepts

- **Variable de réponse :** Variable dont vous essayez de comprendre le comportement ou la variation, sur l'axe des y.

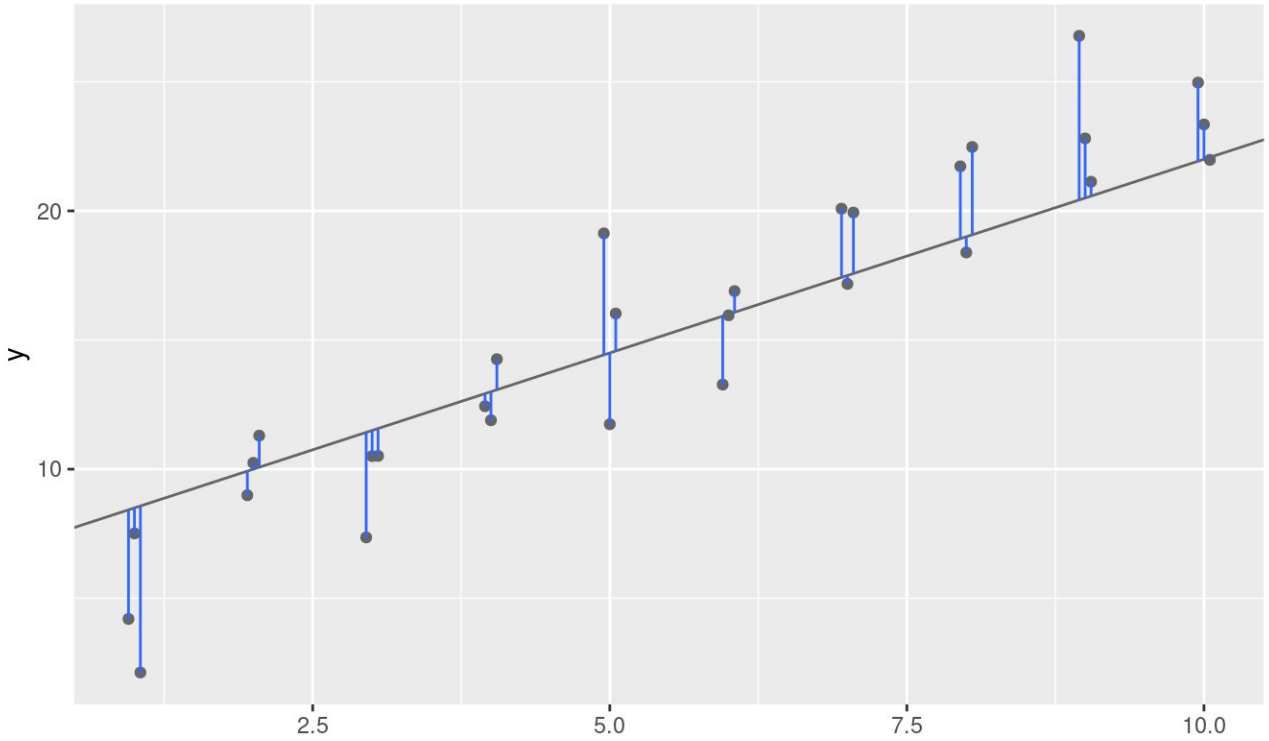
- **Variables explicatives :** Autres variables que vous souhaitez utiliser pour expliquer la variation de la réponse, sur l'axe des x.

- **Valeur prédite :** Sortie de la **fonction du modèle**

    - La fonction de modèle donne la valeur typique (attendue) de la variable de réponse *en* fonction des variables explicatives.

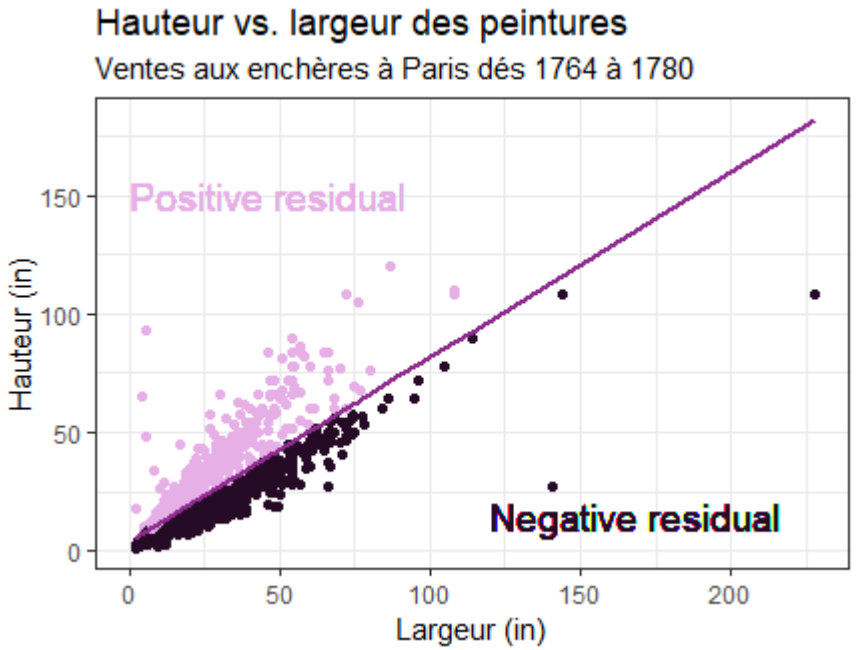
Résidus

- Mesure de la distance entre chaque cas et la valeur prédite (sur la base d'un modèle particulier)
- Résidu = Valeur observée - Valeur prédite
- Indique dans quelle mesure chaque cas est supérieur ou inférieur à la valeur attendue.



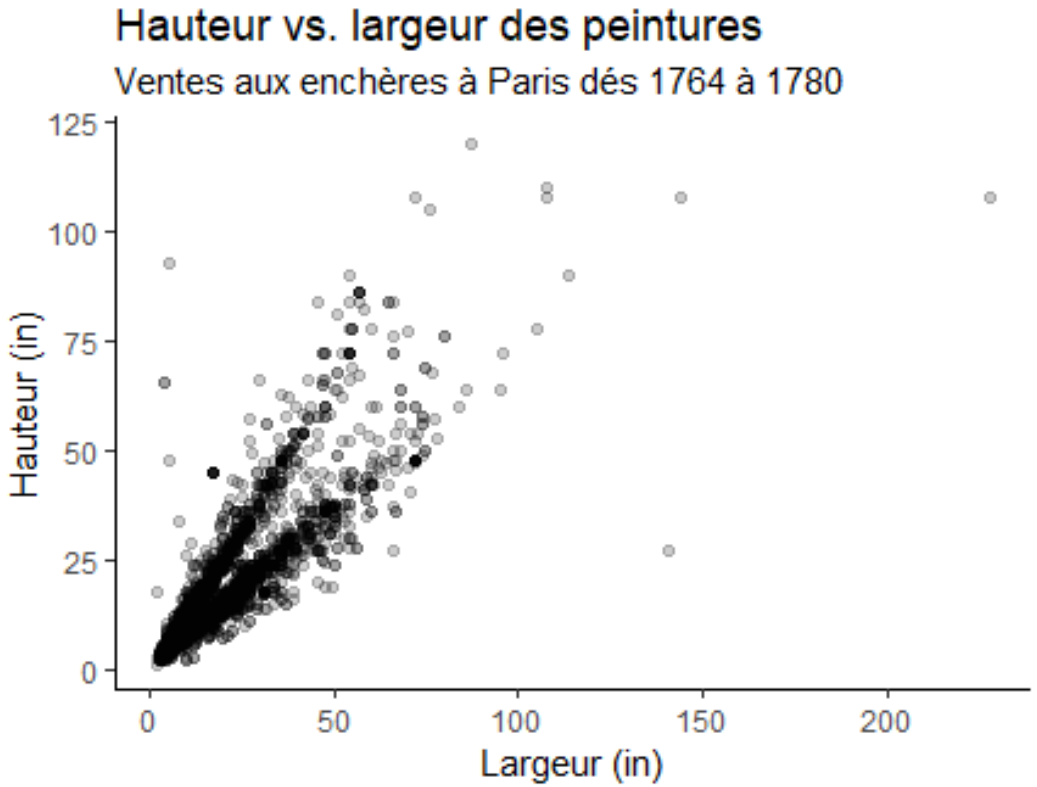
Résidus

- Mesure de la distance entre chaque cas et la valeur prédite (sur la base d'un modèle particulier)
- Résidu = Valeur observée - Valeur prédite
- Indique dans quelle mesure chaque cas est supérieur ou inférieur à la valeur attendue.



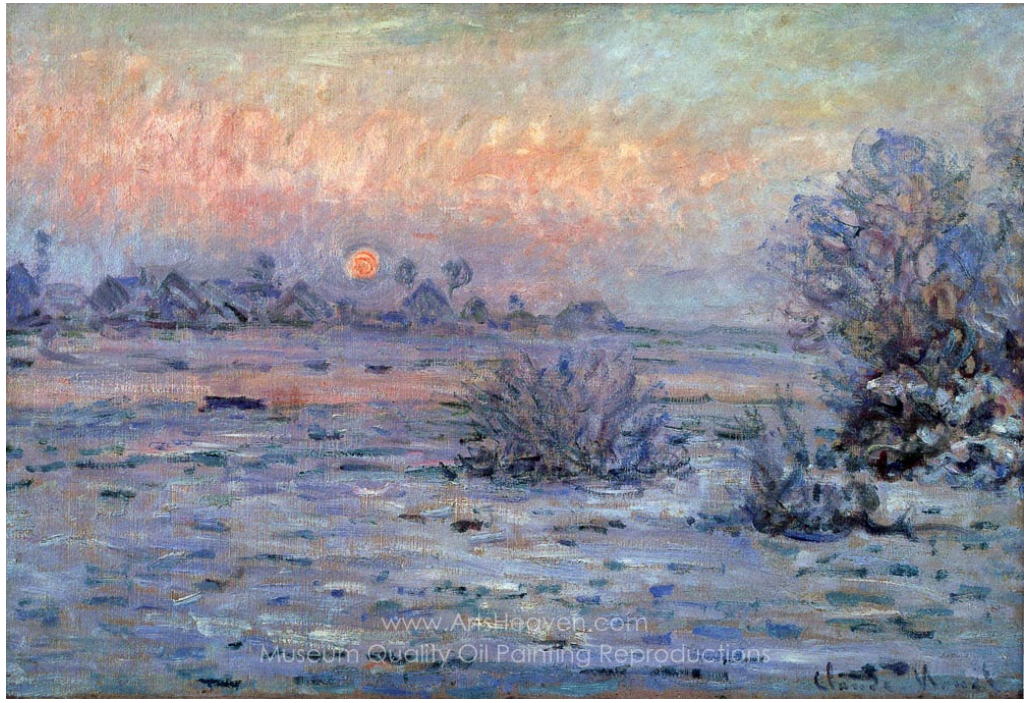
Modification de la valeur d'alpha (0.2) pour rendre les points (valeurs observées) transparent.

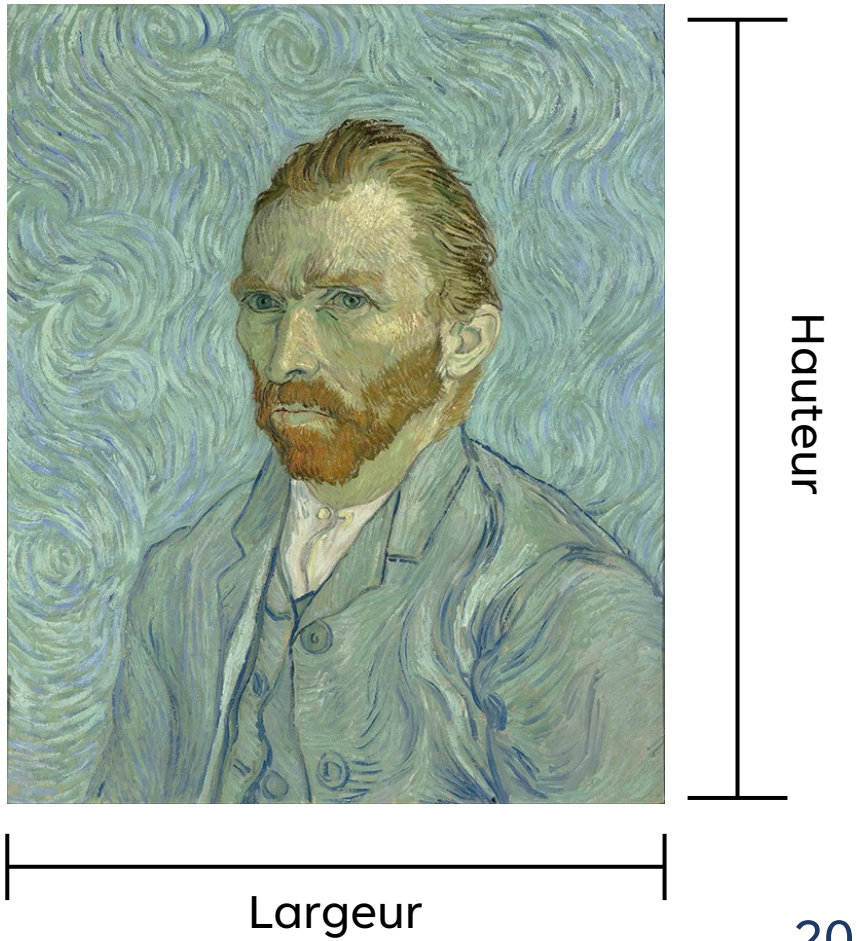
- Quelle est la caractéristique apparente de ce graphique qui n'était pas aussi apparente dans les graphiques précédents ?
- Quelle pourrait être la raison de cette particularité ?



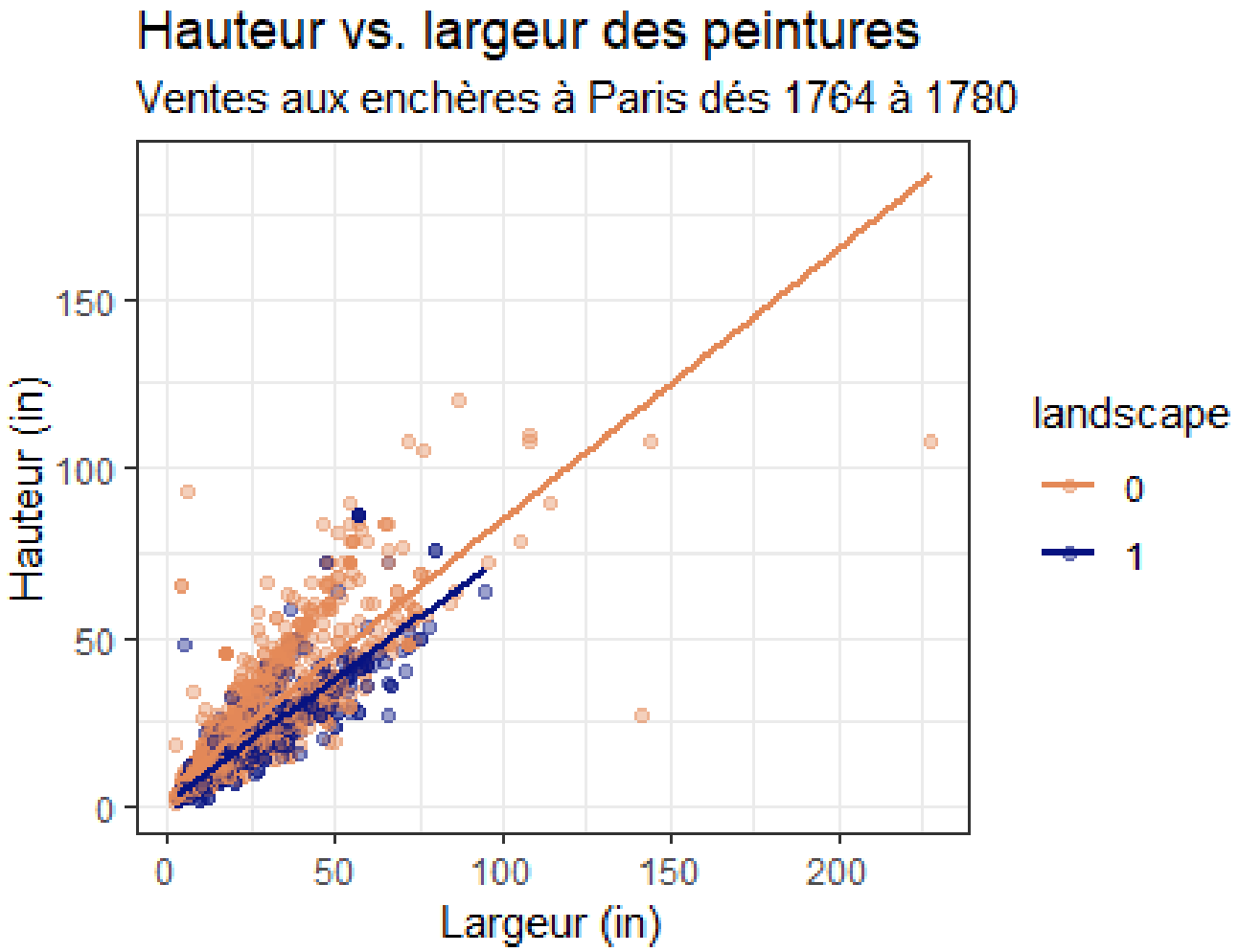
Peinture de paysage vs portrait

Représentation dans l'art de paysages naturels tels que les montagnes, les vallées, les arbres, les rivières et les forêts.





Multiples variables explicatives

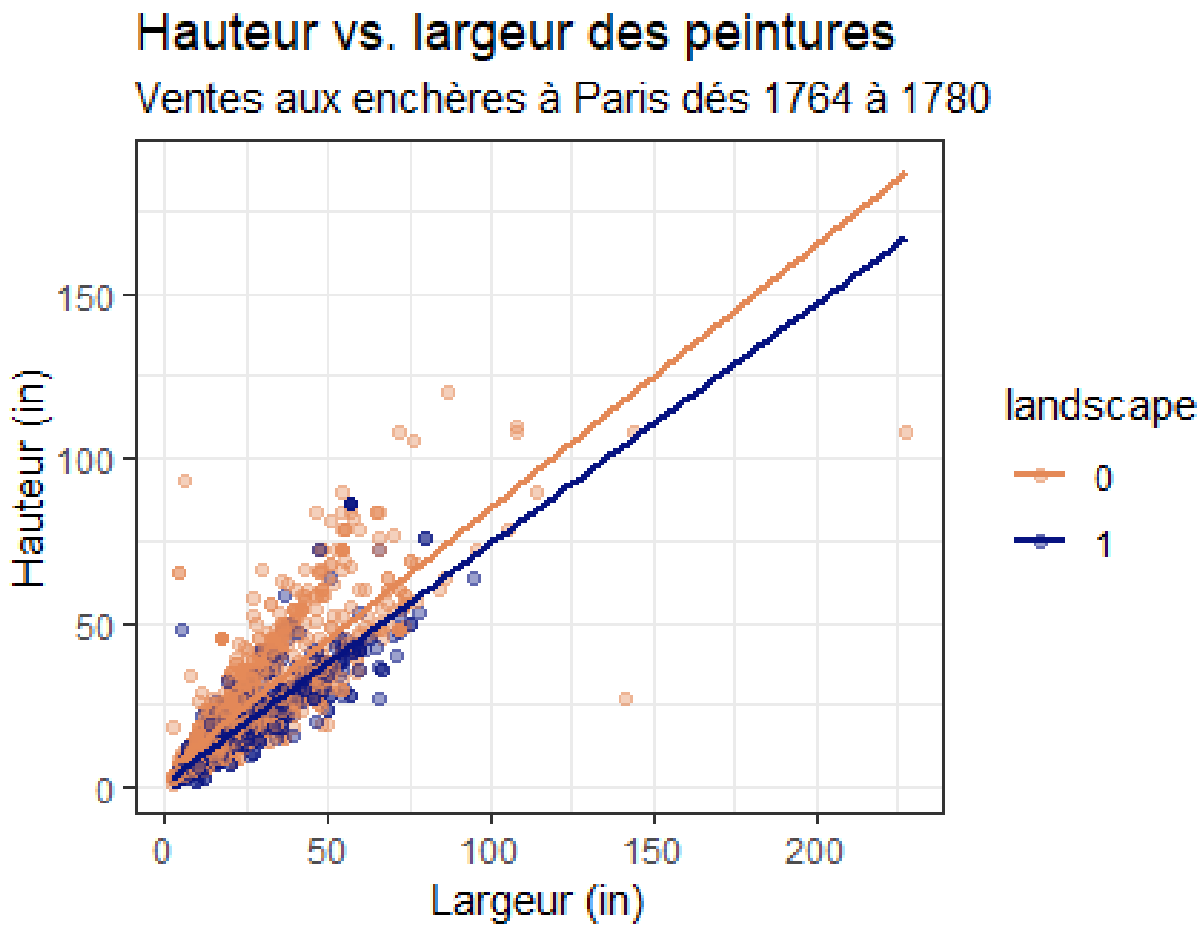


Multiples variables explicatives

Code:

```
ggplot(data = pp, aes(x = Width_in, y = Height_in, color = factor(landsALL)))+
      geom_point(alpha = 0.4) +
      geom_smooth(method = "lm", se = FALSE) +
      labs(
            title = "Hauteur vs. largeur des peintures",
            subtitle = "Ventes aux enchères à Paris dés 1764 à 1780",
            x = "Largeur (in)",
            y = "Hauteur (in)"
      )+
      scale_color_manual(values = c("#E48957", "#071381"))+
      theme_bw()
```

Extension des courbes de régression



Extension des courbes de régression

Code:

```
ggplot(data = pp, aes(x = Width_in, y = Height_in, color = factor(landsALL)))+
      geom_point(alpha = 0.4) +
      geom_smooth(method = "lm", se = FALSE, fullrange = TRUE) +
      labs(
            title = "Hauteur vs. largeur des peintures",
            subtitle = "Ventes aux enchères à Paris dés 1764 à 1780",
            x = "Largeur (in)",
            y = "Hauteur (in)"
      )+
      scale_color_manual(values = c("#E48957", "#071381"))+
      theme_bw()
```

Les modèles – avantages et inconvénients

- Les modèles peuvent parfois révéler des patrons qui ne sont pas évidents dans un graphique de données.
- Il existe toutefois un risque qu'un modèle impose une structure qui n'existe pas réellement dans la dispersion des données.

La variation autour du modèle...

est tout aussi importante que le modèle, si ce n'est plus !

La statistique est l'explication de la variation dans le contexte de ce qui reste inexpliqué.

- La dispersion suggère que d'autres facteurs pourraient expliquer une grande partie de la variabilité d'une peinture à l'autre, ou peut-être simplement que le hasard joue un rôle important.

- L'ajout de variables explicatives à un modèle peut parfois réduire utilement la taille de la dispersion autour du modèle.

Questions ?

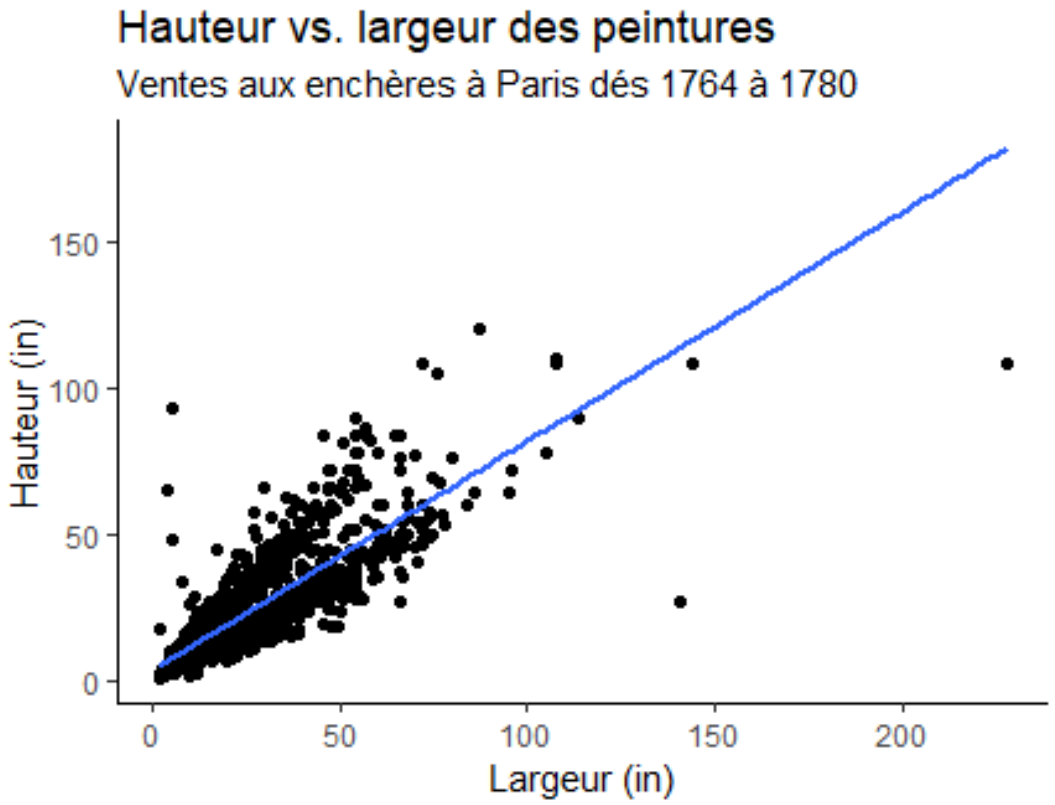
Exercices – Demo

Compléter les exercices de 1-4.

Ajustement et interprétation des modèles

Objectif : prédire la hauteur à partir de la largeur

$$\widehat{hauteur} = \beta_0 + \beta_1 \times largeur_i$$



Étape 1 : Spécifier le modèle

```
linear_reg()
## Linear Regression Model Specification (regression)
## Computational engine: lm
```

Étape 2 : Définir le moteur d'ajustement du modèle

```
linear_reg() %>%
  set_engine("lm") # lm: linear model
## Linear Regression Model Specification (regression)
## Computational engine: lm
```

Étape 3 : Ajuster le modèle et estimer les paramètres

... en utilisant la syntaxe de la formule

```
linear_reg() %>%
  set_engine("lm") %>%
  fit(Height_in ~ Width_in, data = pp)
## parsnip model
##
## object Call:
## stats::lm(formula = Height_in ~ Width_in, data = data)
##
## Coefficients:
## (Intercept) Width_in
##  3.6214      0.7808
```

Un examen plus approfondi des résultats du modèle

$$\widehat{hauteur} = 3.6214 + 0.7808 \times largeur_i$$

```
## parsnip model
##
## object Call:
## stats::lm(formula = Height_in ~ Width_in, data = data)
##
## Coefficients:
## (Intercept) Width_in
##  3.6214      0.7808
```

En ajoutant « tidy »

$$\widehat{hauteur} = 3.6214 + 0.7808 \times largeur_i$$

```
linear_reg() %>%
  set_engine("lm") %>%
  fit(Height_in ~ Width_in, data = pp) %>%
  tidy()
# A tibble: 2 × 5
# term        estimate std.error statistic p.value
# <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)   3.62      0.254       14.3   8.82e-45
2  Width_in     0.781   0.00950       82.1      0
```

La pente et l'ordonnée à l'origine

$$\widehat{hauteur} = 3.6214 + 0.7808 \times largeur_i$$

- **La pente :** Pour chaque pouce supplémentaire de largeur du tableau, la hauteur devrait être plus élevée, en moyenne, de 0,781 pouce.
- **L’ordonnée à l'origine :** Les tableaux de 0 pouce de large devraient avoir une hauteur moyenne de 3,62 pouces. (Cela a-t-il un sens ?)

LA CORRÉLATION N'IMPLIQUE PAS LA CAUSALITÉ

Estimation des paramètres

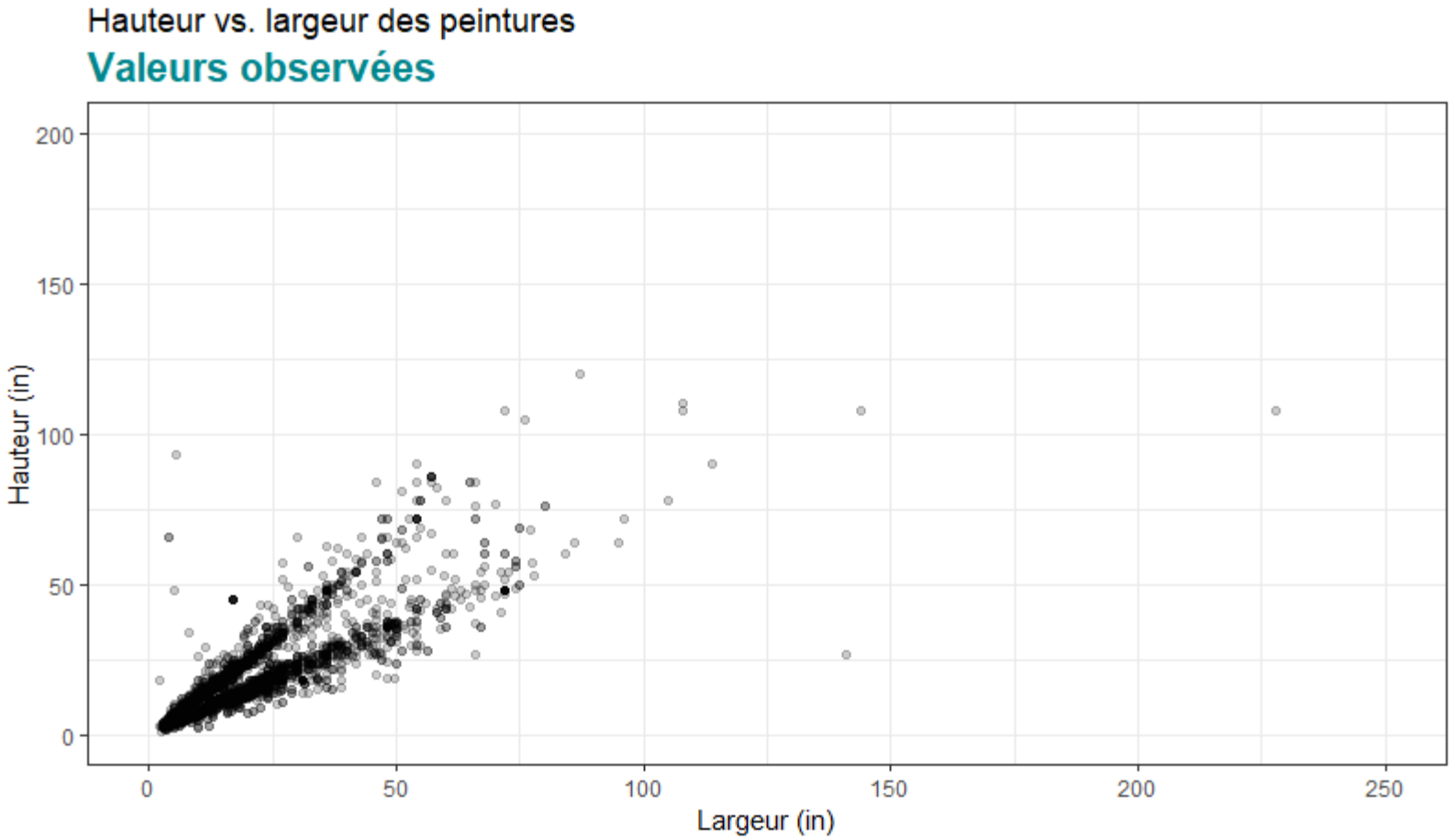
Modèle linéaire avec un prédicteur unique: Régression par la méthode des moindres carrés

- Nous nous intéressons à β_0 (paramètre de population pour l'ordonnée à l'origine) et β_1 (paramètre de population pour la pente) dans le modèle suivant :

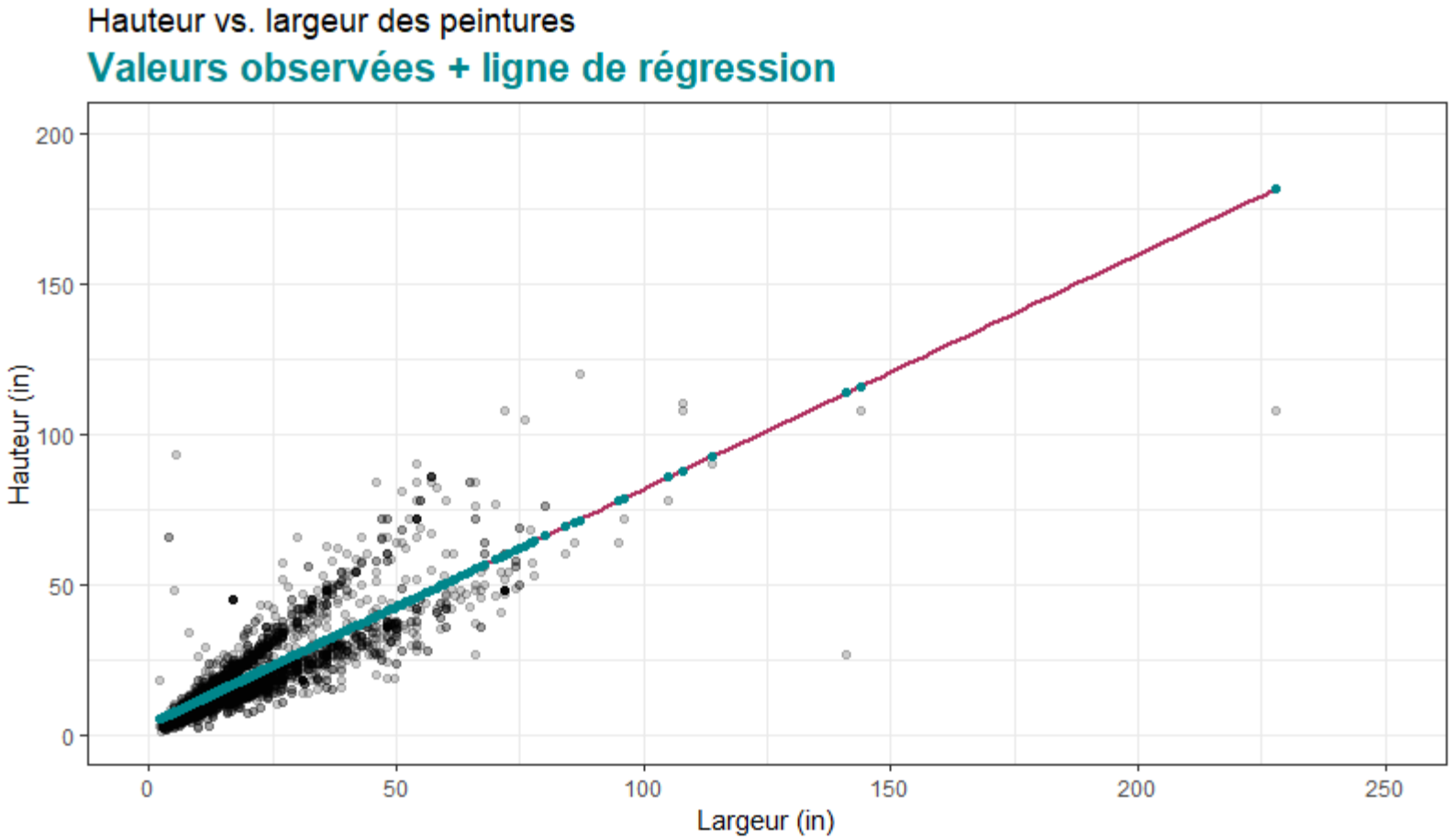
$$\hat{y} = \beta_0 + \beta_1 \times x_i$$

- La courbe de régression minimise la somme des carrés des résidus.
- Si $e_i = y_i - \hat{y}_i$, alors, la droite de régression minimise $\sum_{i=1}^{n} e_i^2$.

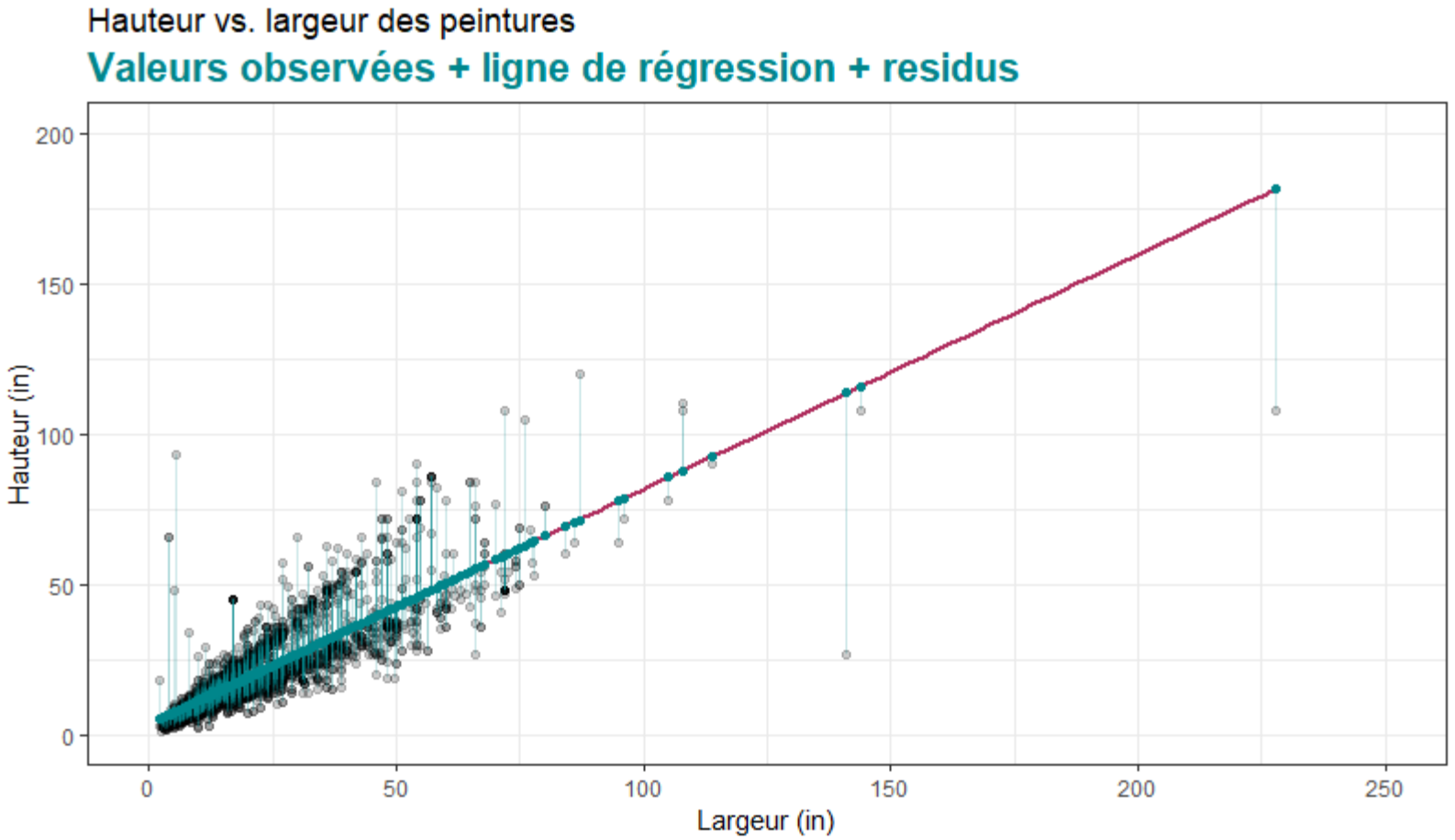
Visualisation des résidus



Visualisation des résidus



Visualisation des résidus



Propriétés de la régression par les moindres carrés

- La ligne de régression passe par le centre de masse, les coordonnées correspondant à la moyenne x et à la moyenne y, $(\bar{x}, \bar{y})$.

$$\bar{y} = b_0 + b_1\bar{x} \rightarrow b_0 = \bar{y} - b_1\bar{x}$$

- La pente a le même signe que le coefficient de corrélation : $b_1 = r\frac{s_y}{s_x}$
- La somme des résidus est zéro : $\sum_{i=1}^{n} e_i^2 = 0$
- Les résidus et les valeurs x ne sont pas corrélés.

Les modèles avec des variables explicatives catégorielles

Prédicteur catégorique avec 2 niveaux

```
## # A tibble: 3,393 × 3
## name Height_in landsALL
## <chr> <dbl> <dbl>
## 1 L1764-2 37 0
## 2 L1764-3 18 0
## 3 L1764-4 13 1
## 4 L1764-5a 14 1
## 5 L1764-5b 14 1
## 6 L1764-6 7 0
## 7 L1764-7a 6 0
## 8 L1764-7b 6 0
## 9 L1764-8 15 0
## 10 L1764-9a 9 0
## 11 L1764-9b 9 0
## 12 L1764-10a 16 1
## 13 L1764-10b 16 1
## 14 L1764-10c 16 1
## 15 L1764-11 20 0
## 16 L1764-12a 14 1
## 17 L1764-12b 14 1
## 18 L1764-13a 15 1
## 19 L1764-13b 15 1
## 20 L1764-14 37 0
## # … with 3,373 more rows
```

- **landsALL = 0** : Pas d'éléments paysagers
- **landsALL = 1** : Un où plus d'éléments du paysage

Hauteur et éléments du paysage

```
linear_reg() %>%
set_engine("lm") %>%
fit(Height_in ~ factor(landsALL), data = pp) %>%
tidy()
## # A tibble: 2 × 5
##   term              estimate std.error statistic p.value
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          22.7      0.328      69.1       0
## 2 factor(landsALL)1 -5.65       0.532     -10.6   7.97e-26
```

Hauteur et éléments du paysage

$$\widehat{hauteur} = 22.7 - 5.65 \times landsALL_i$$

- **La pente** : Les peintures contenant des éléments de paysage devraient, en moyenne, être plus courtes de 5,65 pouces que les peintures sans éléments de paysage.
  - Comparaison entre le niveau de base (landsALL = 0) et l'autre niveau (landsALL = 1)
- **L'ordonnée à l'origine** : On s'attend à ce que les peintures sans éléments de paysage mesurent, en moyenne, 22,7 pouces de haut.

Les relations entre hauteur et école

```
linear_reg() %>%
  set_engine("lm") %>%
  fit(Height_in ~ school_pntg, data = pp) %>%
  tidy()
```

```
## # A tibble: 7 × 5
##       term          estimate std.error statistic p.value
##      <chr>            <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)       14.0       10.0      1.40    0.162
## 2 school_pntgD/FL    2.33      10.0      0.232   0.816
## 3 school_pntgF      10.2       10.0      1.02    0.309
## 4 school_pntgG       1.65      11.9      0.139   0.889
## 5 school_pntgI      10.3       10.0      1.02    0.306
## 6 school_pntgS      30.4       11.4      2.68    0.00744
## 7 school_pntgX       2.87      10.3      0.279   0.780
```

Les relations entre hauteur et école

```
## # A tibble: 7 × 5
##      term         estimate std.error statistic p.value
##     <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      14.0      10.0      1.40     0.162
## 2 school_pntgD/FL  2.33      10.0      0.232    0.816
## 3 school_pntgF     10.2      10.0      1.02     0.309
## 4 school_pntgG     1.65      11.9      0.139    0.889
## 5 school_pntgI     10.3      10.0      1.02     0.306
## 6 school_pntgS     30.4      11.4      2.68     0.00744
## 7 school_pntgX     2.87      10.3      0.279    0.780
```

- Lorsque la variable explicative catégorielle comporte plusieurs niveaux, ceux-ci sont codés en **variables muettes**.
- Chaque coefficient décrit la différence attendue entre les hauteurs dans cette école particulière par rapport au niveau de référence.

Les relations entre hauteur et école

```
## # A tibble: 7 × 5
##      term         estimate std.error statistic p.value
##      <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      14.0      10.0      1.40     0.162
## 2 school_pntgD/FL  2.33      10.0      0.232    0.816
## 3 school_pntgF     10.2      10.0      1.02     0.309
## 4 school_pntgG     1.65      11.9      0.139    0.889
## 5 school_pntgI     10.3      10.0      1.02     0.306
## 6 school_pntgS     30.4      11.4      2.68     0.00744
## 7 school_pntgX     2.87      10.3      0.279    0.780
```

- Les peintures de **l'école autrichienne (A)** devraient, en moyenne, mesurer **14 pouces de** haut.
- Les peintures de **l'école hollandaise/flamande (D/FL)** devraient, en moyenne, **mesurer 2,33 pouces de plus** que les peintures de *l'école autrichienne*.
- Les peintures de **l'école française (F)** devraient, en moyenne, **mesurer 10,2 pouces de plus** que les peintures de *l'école autrichienne*.
- Les peintures de **l'école allemande (G)...**
- Les peintures de **l'école italienne (I)...**
- Les peintures de **l'école espagnole (S)...**
- Les peintures dont l'école est **inconnue (X)...**

Exercices – Demo

Compléter les exercices de 5-10.

Sources:

- https://datasciencebox.org/
- Wickham, H., Çetinkaya-Rundel, M. and Grolemund, G. (2023). R for Data Science (2nd ed.). O'Reilly Media, Inc.