Introducing linear models

Prof. Maria Tackett

Click for PDF of slides

The language of models

Modeling

- We use models to...
  - understand relationships
  - assess differences
  - make predictions
- We will focus on linear models but there are many other types.

Data: Paris Paintings

Paris Paintings

```r
paris_paintings <- read_csv("data/paris_paintings.csv",
                            na = c("n/a", "", "NA"))
```

Click here for Paris Paintings codebook

Sandra van Ginhoven

Hilary Coe Cronheim

PhD students in the Duke Art, Law, and Markets Initiative in 2013

Source: Printed catalogs of 28 auction sales in Paris, 1764- 1780 - 3,393 paintings, their prices, and descriptive details from sales catalogs over 60 variables

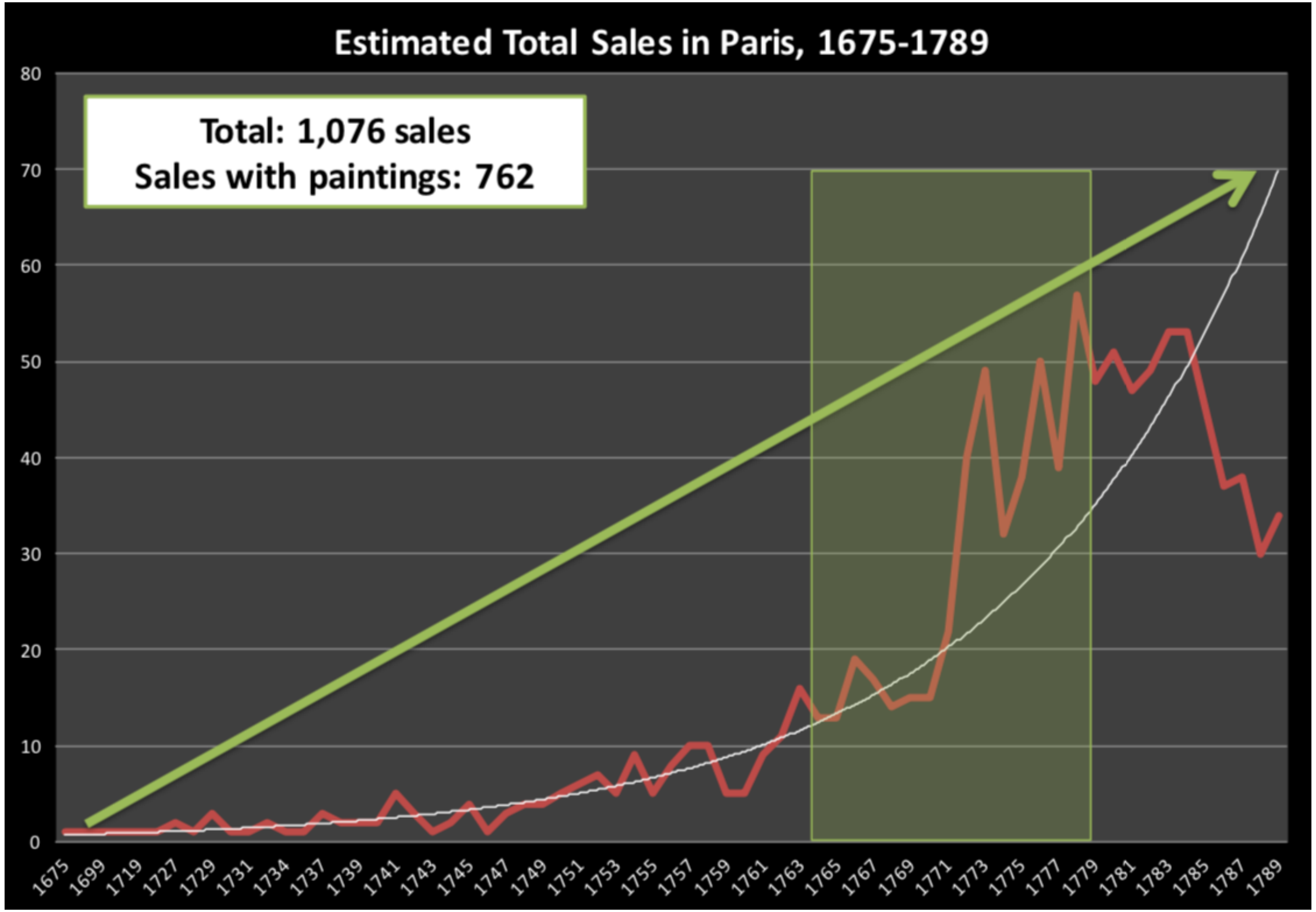
Auctions today

https://www.youtube.com/watch?v=apaE1Q7r4so

Auctions back in the day

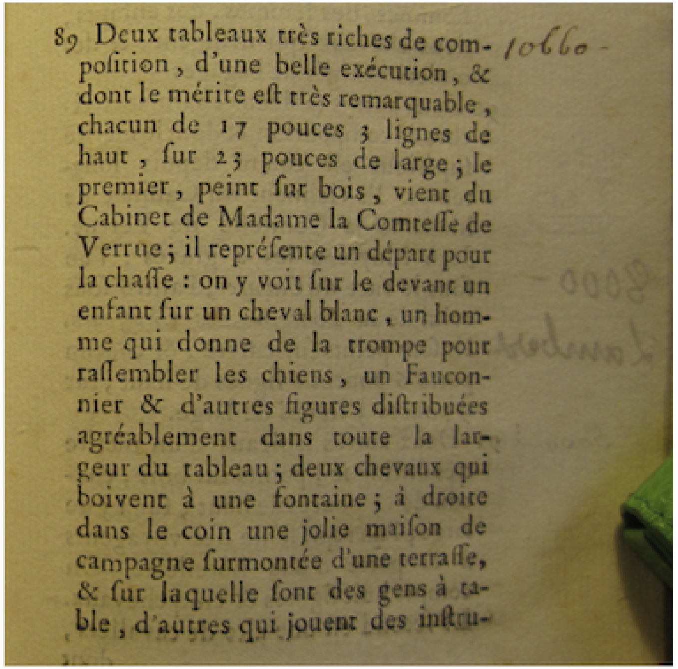
de Machy, Public Sale at the Hôtel Bullion, Musée Carnavalet, Paris (18th century)

Paris auction market

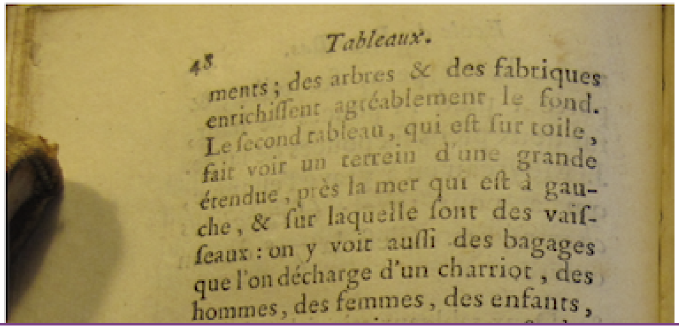


Depart pour la chasse

Auction catalogue text



89 Deux tableaux très riches de com- 10660
poſition, d'une belle exécution, &
dont le mérite eſt très remarquable,
chacun de 17 pouces 3 lignes de
haut, ſur 23 pouces de large; le
premier, peint ſur bois, vient du
Cabinet de Madame la Comteſſe de
Verrue; il repréſente un départ pour
la chaſſe : on y voit ſur le devant un
enfant ſur un cheval blanc, un hom-
me qui donne de la trompe pour
raſſembler les chiens, un Faucon-
nier & d'autres figures diſtribuées
agréablement dans toute la lar-
geur du tableau; deux chevaux qui
boivent à une fontaine; à droite
dans le coin une jolie maiſon de
campagne ſurmontée d'une terraſſe,
& ſur laquelle ſont des gens à ta-
ble, d'autres qui jouent des inſtru-



48 Tableaux.
ments; des arbres & des fabriques
enrichiſſent agréablement le fond.
Le ſecond tableau, qui eſt ſur toile,
fait voir un terrein d'une grande
étendue, près la mer qui eſt à gau-
che, & ſur laquelle ſont des vaiſ-
ſeaux : on y voit auſſi des bagages
que l'on décharge d'un charriot, des
hommes, des femmes, des enfans,

Two paintings very rich in composition, of a beautiful execution, and whose merit is very remarkable, each 17 inches 3 lines high, 23 inches wide; the first, painted on wood, comes from the Cabinet of Madame la Comtesse de Verrue; it represents a departure for the hunt: it shows in the front a child on a white horse, a man who gives the horn to gather the dogs, a falconer and other figures nicely distributed across the width of the painting; two horses drinking from a fountain; on the right in the corner a lovely country house topped by a terrace, on which people are at the table, others who play instruments; trees and fabriques pleasantly enrich the background.

```r
paris_paintings %>% filter(name == "R1777-89a") %>%
  select(name:endbuyer) %>% glimpse()
```

```
## Rows: 1
## Columns: 21
## $ 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\u008epart pour la chasse"
```

```r
paris_paintings %>% filter(name == "R1777-89a") %>%
  select(Interm:finished) %>% glimpse()
```

```
## Rows: 1
## Columns: 21
## $ 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
```

```r
paris_paintings %>% filter(name == "R1777-89a") %>%
  select(lrgfont:other) %>% glimpse()
```

```
## Rows: 1
## Columns: 19
## $ 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
```

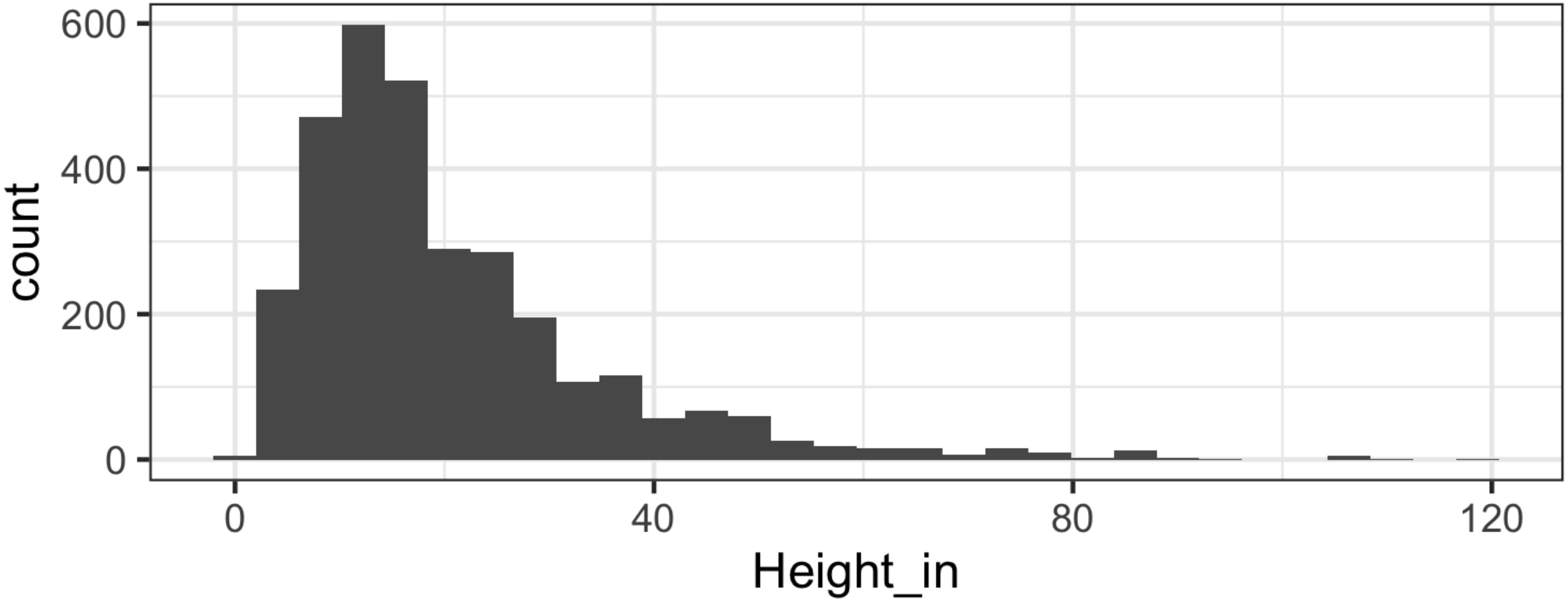
Modeling the relationship between variables

Describe the distribution of price

```r
ggplot(data = paris_paintings, aes(x = price)) +
  geom_histogram(binwidth = 1000) +
  labs(title="Distribution of Price (in Livres)")
```

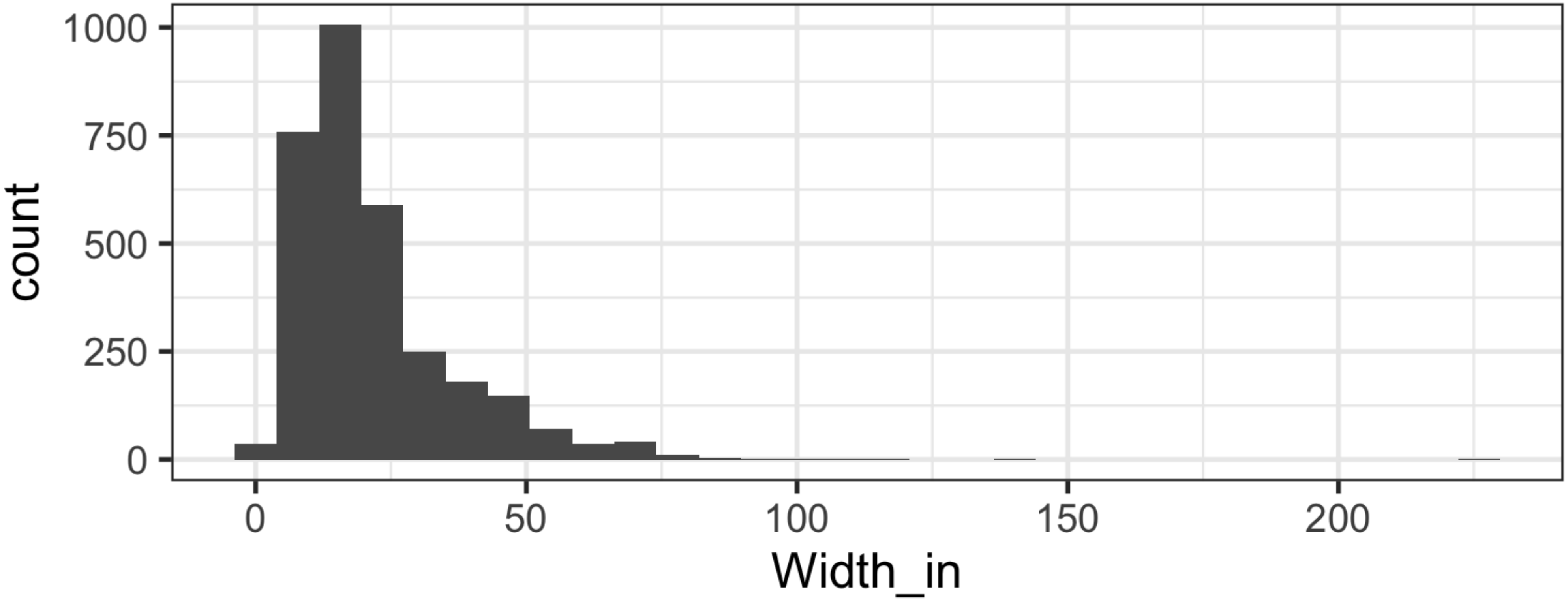
Height

```
ggplot(data = paris_paintings, aes(x = Height_in)) +
  geom_histogram()
```



Width

```r
ggplot(data = paris_paintings, aes(x = Width_in)) +
  geom_histogram()
```



Models as functions

- We can represent relationships between variables using **functions**
- A function in the *mathematical* sense is the relationship between one or more inputs and an output created from those inputs.
  - Plug in the inputs and receive back the output

Models as functions

- We can represent relationships between variables using **functions**
- A function in the *mathematical* sense is the relationship between one or more inputs and an output created from those inputs.
  - Plug in the inputs and receive back the output
- The formula $y = 3x + 7$ is a function with input x and output y.
  - When x is 5, the output y is 22

```
y = 3 * 5 + 7 = 22
```

Models as functions

- We can represent relationships between variables using **functions**
- A function in the *mathematical* sense is the relationship between one or more inputs and an output created from those inputs.
  - Plug in the inputs and receive back the output
- The formula $y = 3x + 7$ is a function with input x and output y.
  - When x is 5, the output y is 22

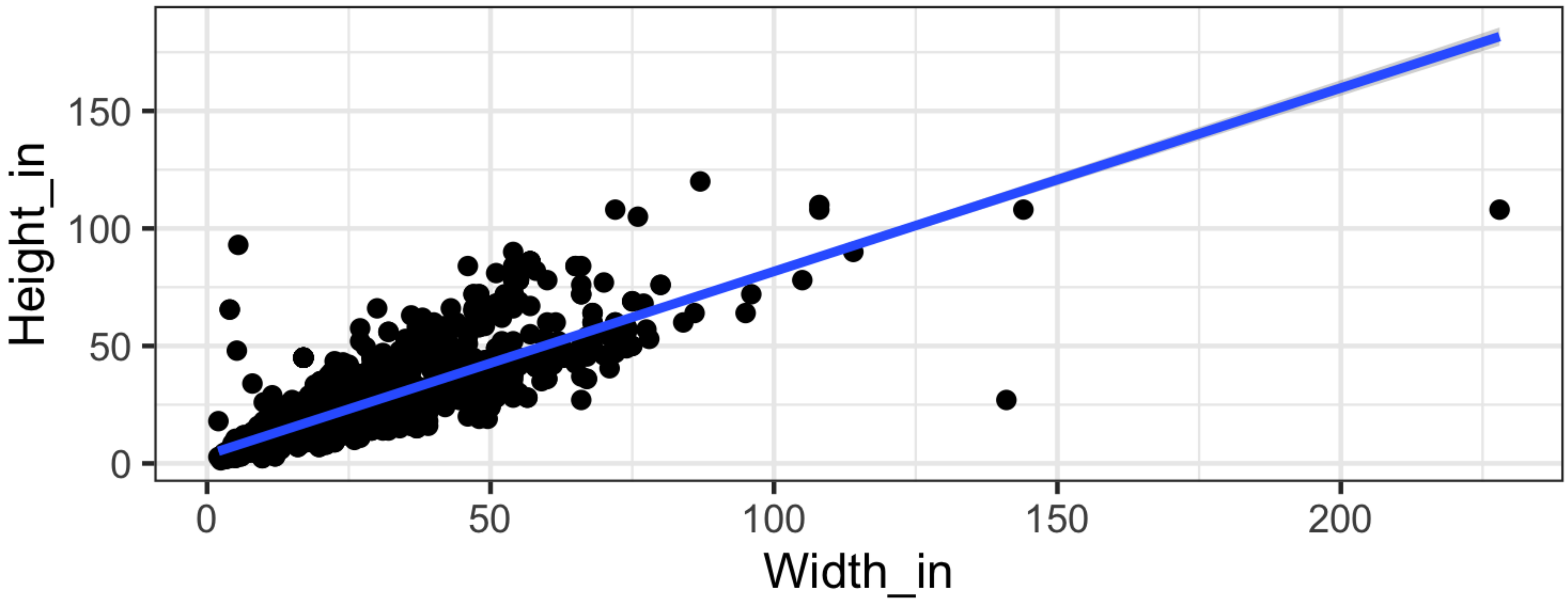
```
y = 3 * 5 + 7 = 22
```

  - When x is 12, the output of y is 43

```
y = 3 * 12 + 7 = 43
```

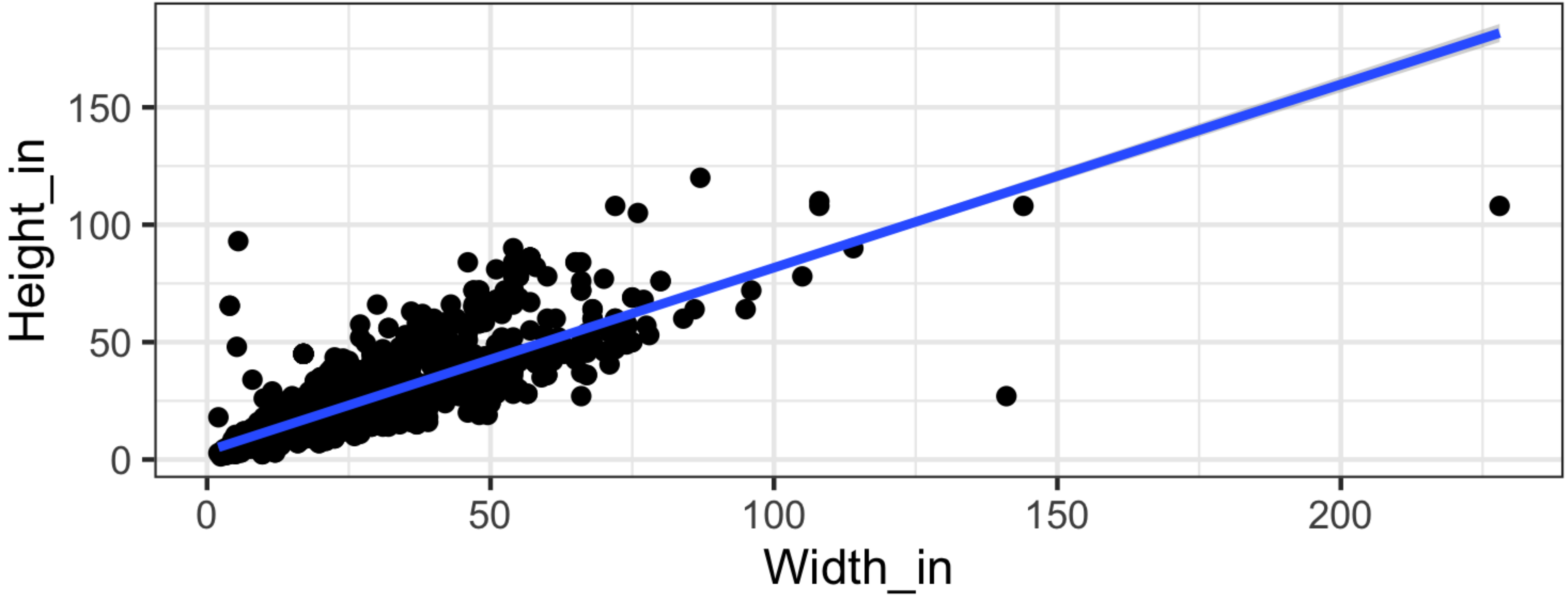
Visualizing the linear model

```
ggplot(data = paris_paintings, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm")
```



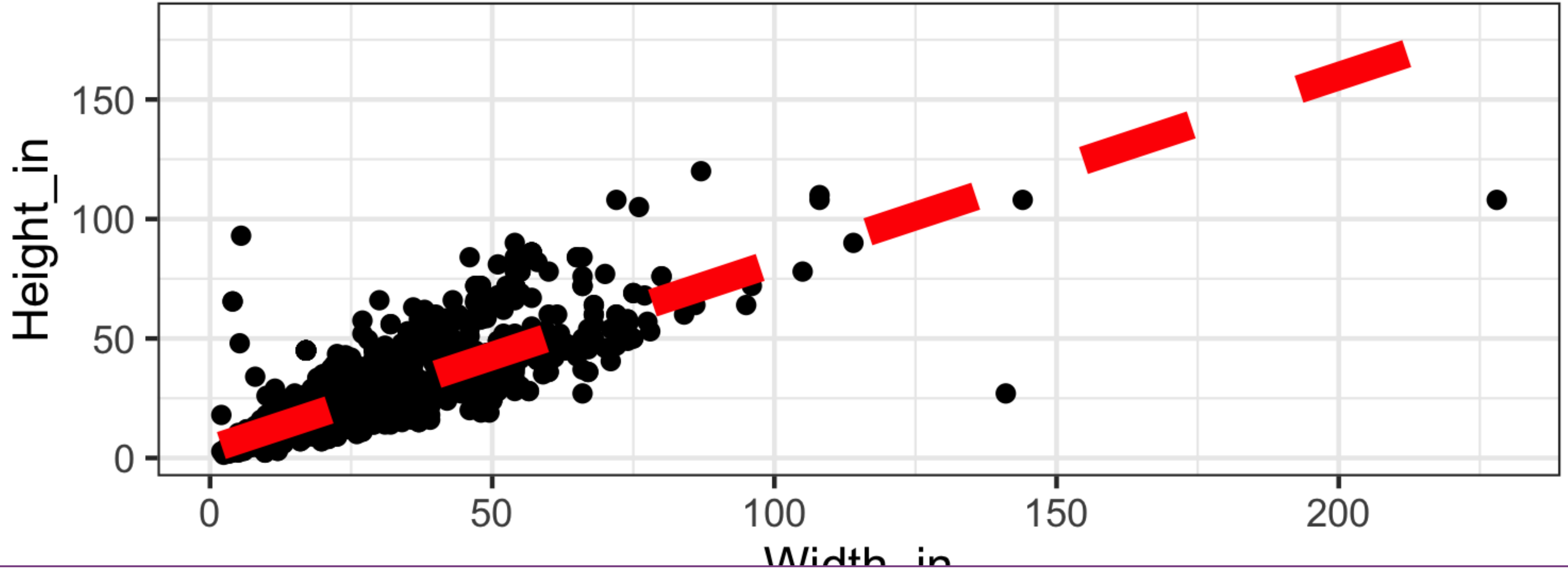
Visualizing the linear model

```r
ggplot(data = paris_paintings, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Visualizing the linear model

```
ggplot(data = paris_paintings, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE,
              col = "red", lty = 2, lwd = 3)
```



Vocabulary

- Response variable: Variable whose behavior or variation you are trying to understand, on the y-axis. Also called the **dependent variable**.

Vocabulary

- Response variable: Variable whose behavior or variation you are trying to understand, on the y-axis. Also called the **dependent variable**.

- Explanatory variables: Other variables that you want to use to explain the variation in the response, on the x-axis. Also called **independent variables**, **predictors**, or **features**.

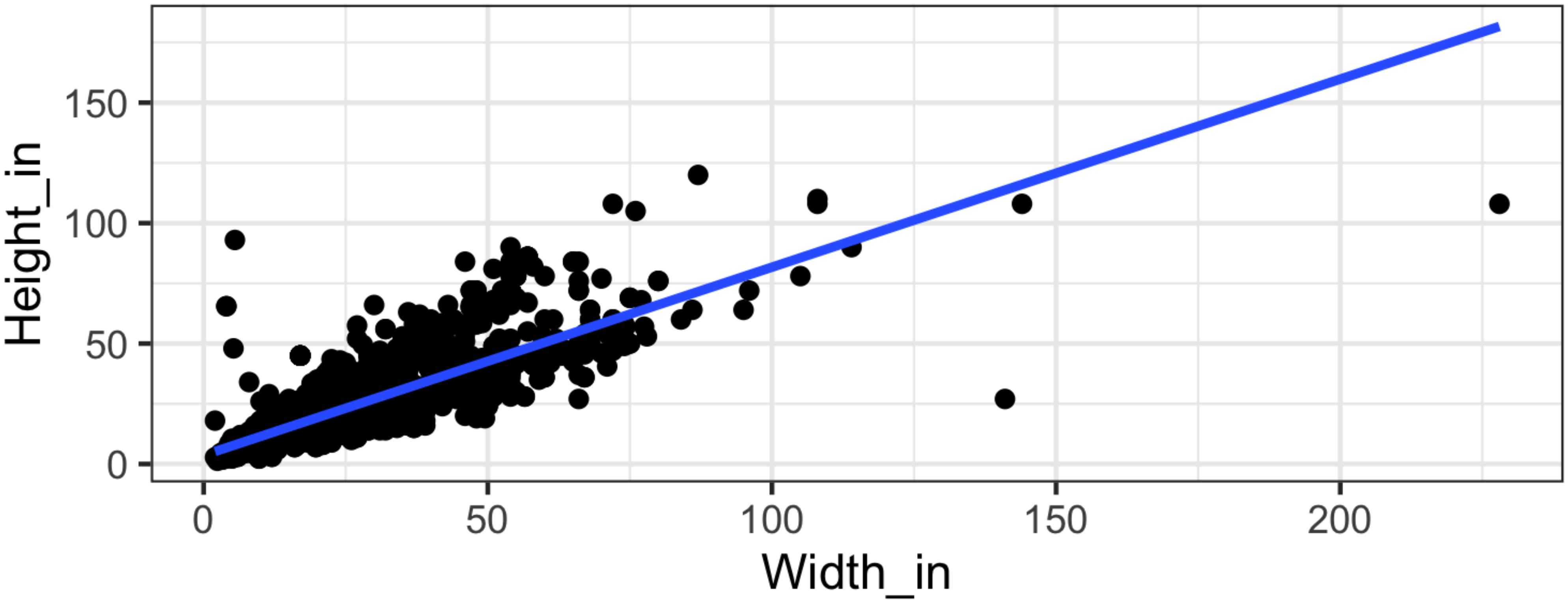
Vocabulary

- Response variable: Variable whose behavior or variation you are trying to understand, on the y-axis. Also called the **dependent variable**.
- Explanatory variables: Other variables that you want to use to explain the variation in the response, on the x-axis. Also called **independent variables**, **predictors**, or **features**.
- Predicted value: Output of the **model function**
  - The model function gives the typical value of the response variable *conditioning* on the explanatory variables (what does this mean?)

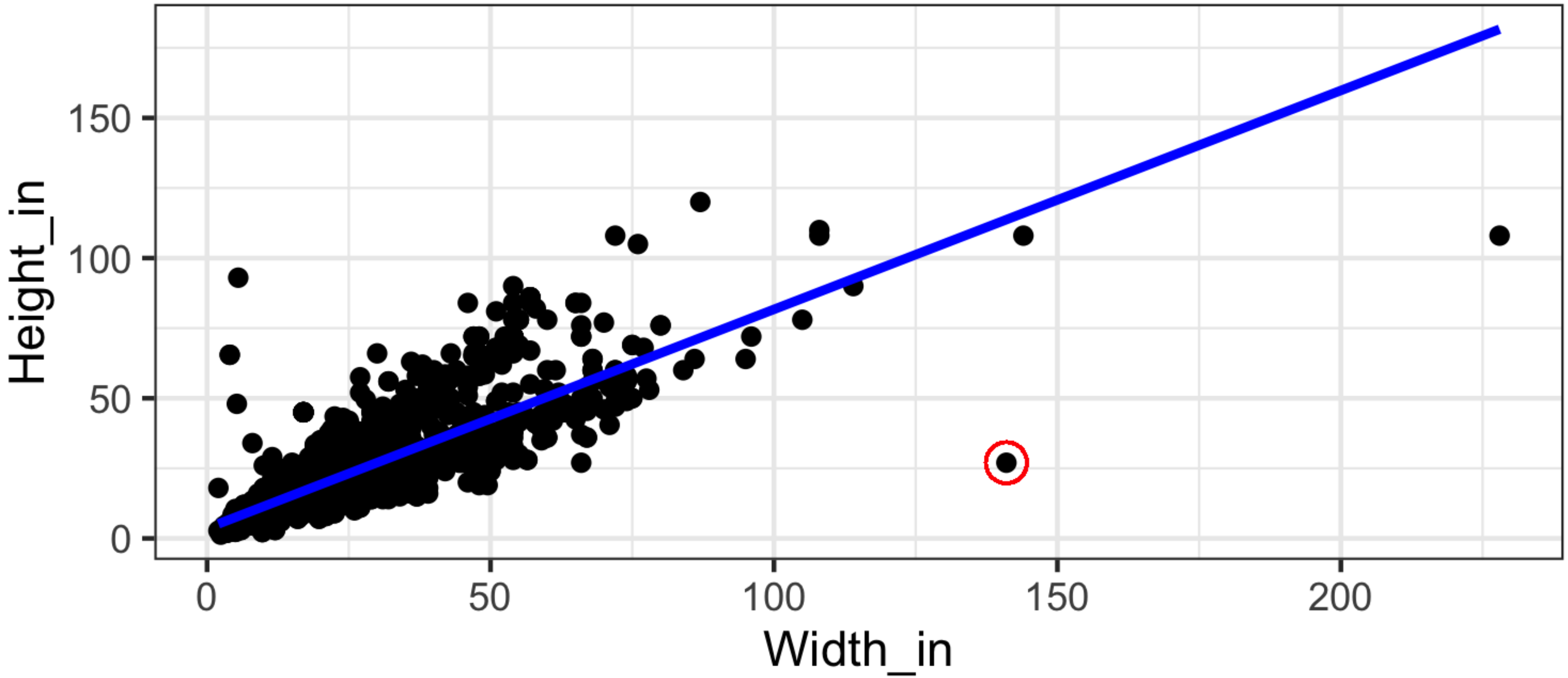
Vocabulary

- Residuals: Shows how far each case is from its predicted value
  - **Residual = Observed value - Predicted value**
  - Tells how far above/below the model function each case is

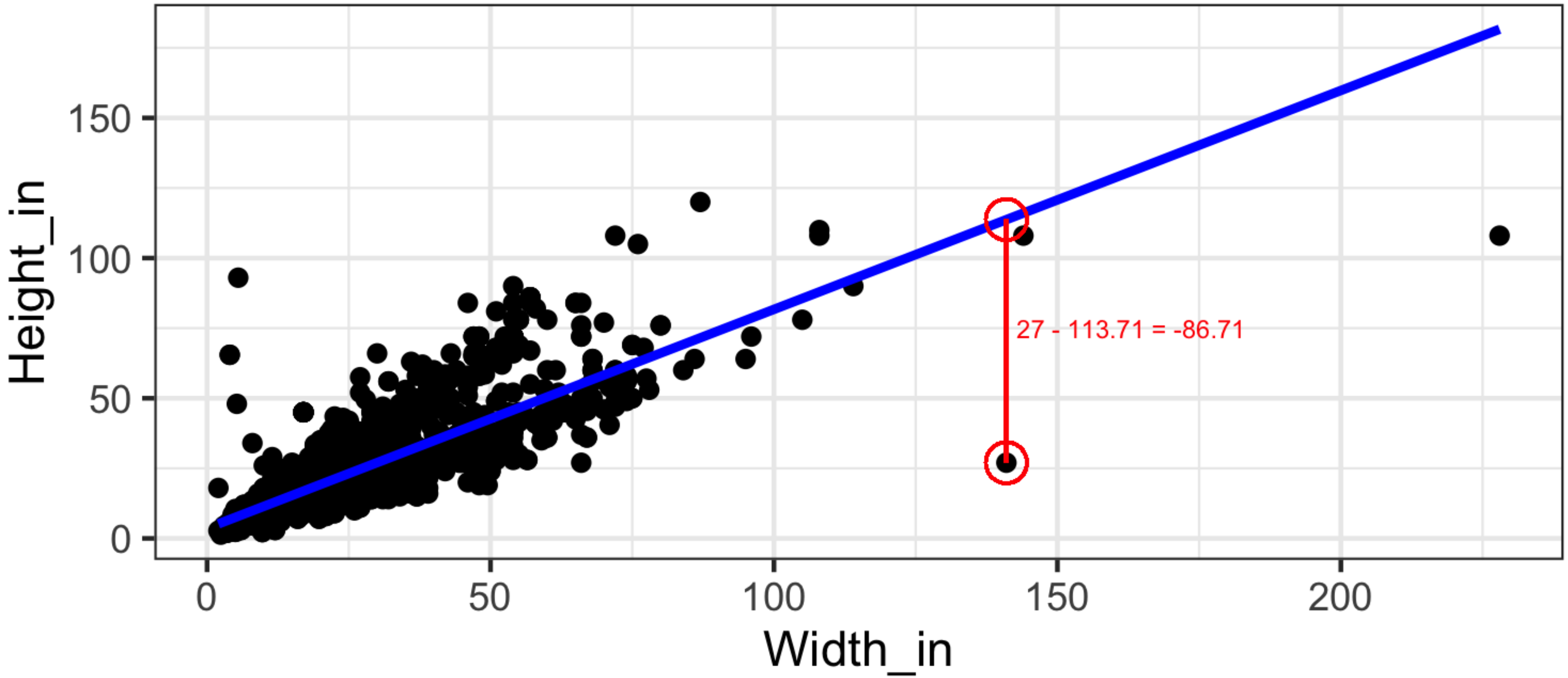
What does a negative residual mean? Which paintings on the plot have have negative residuals, those below or above the line?



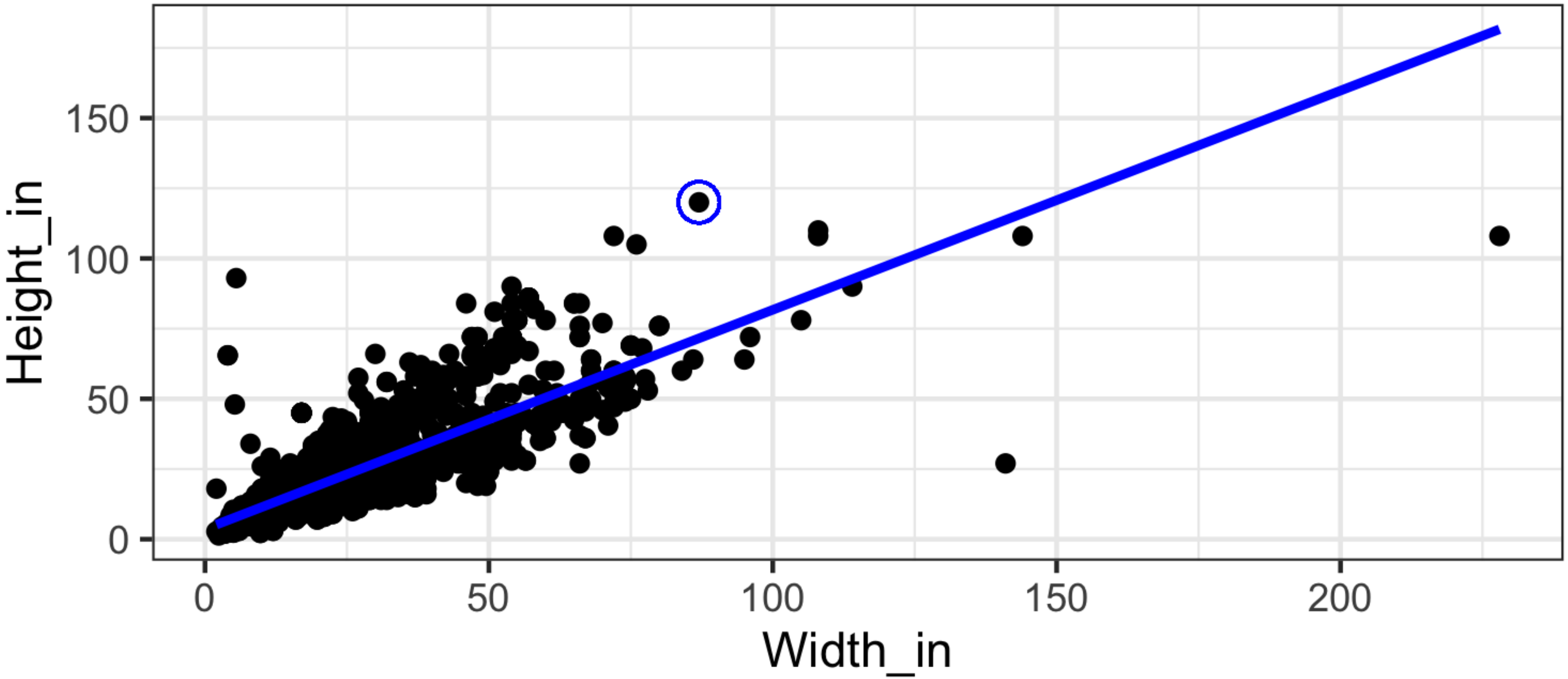
Residuals



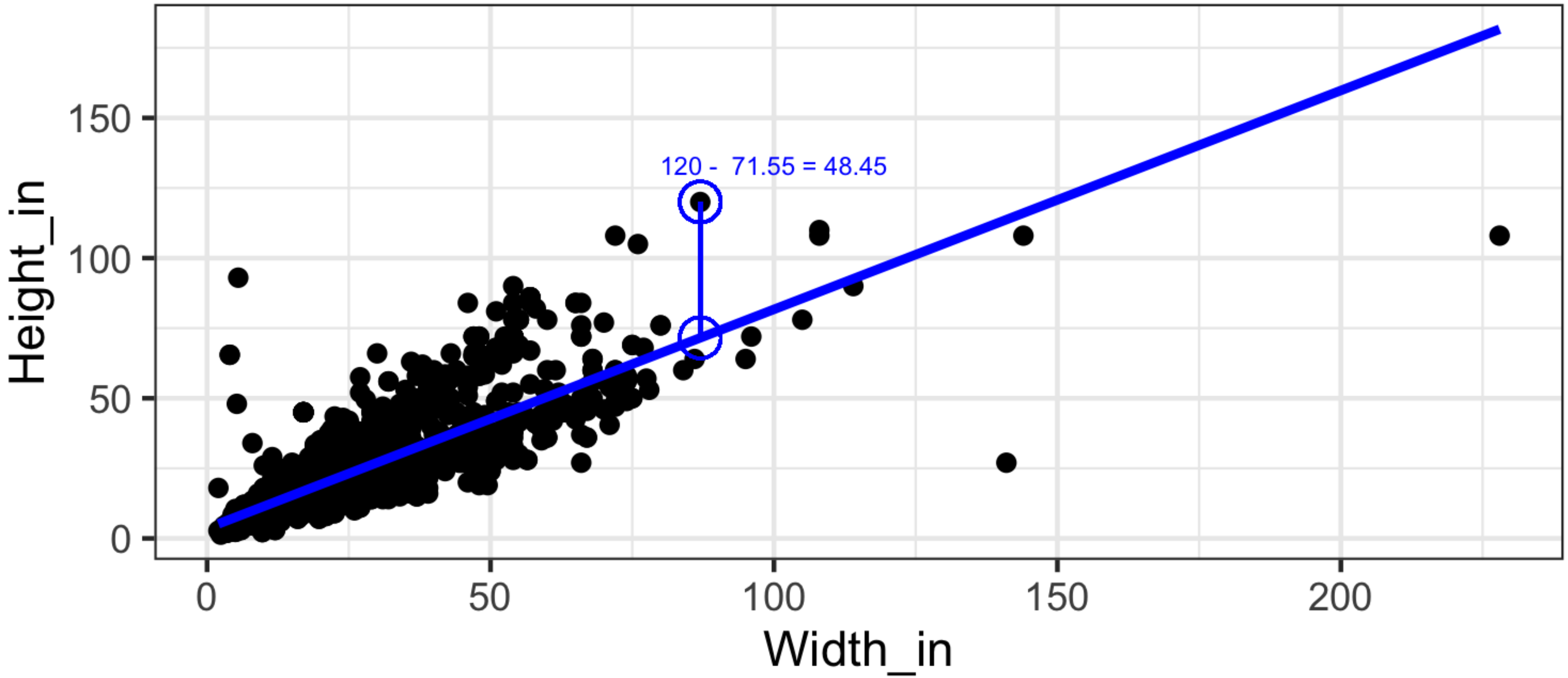
Residuals



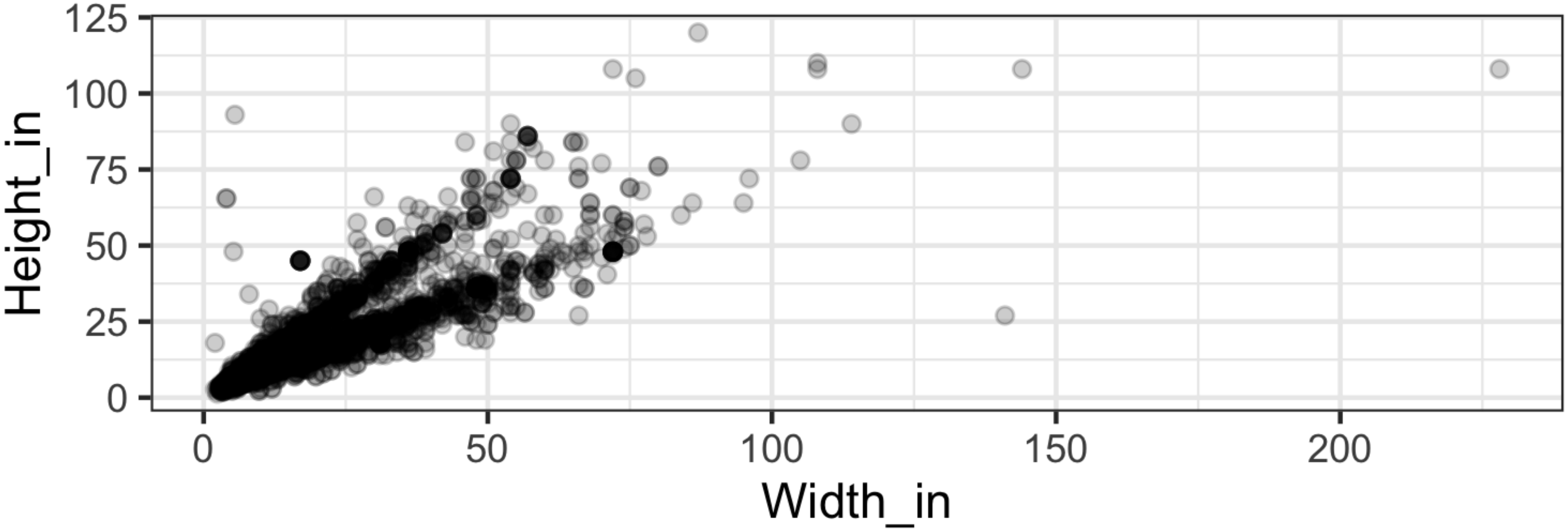
Residuals



Residuals



The plot below displays the relationship between height and width of paintings, but with a lower alpha level. What feature is apparent in this plot that was not (as) apparent in the previous plots? What might be the reason for this feature?

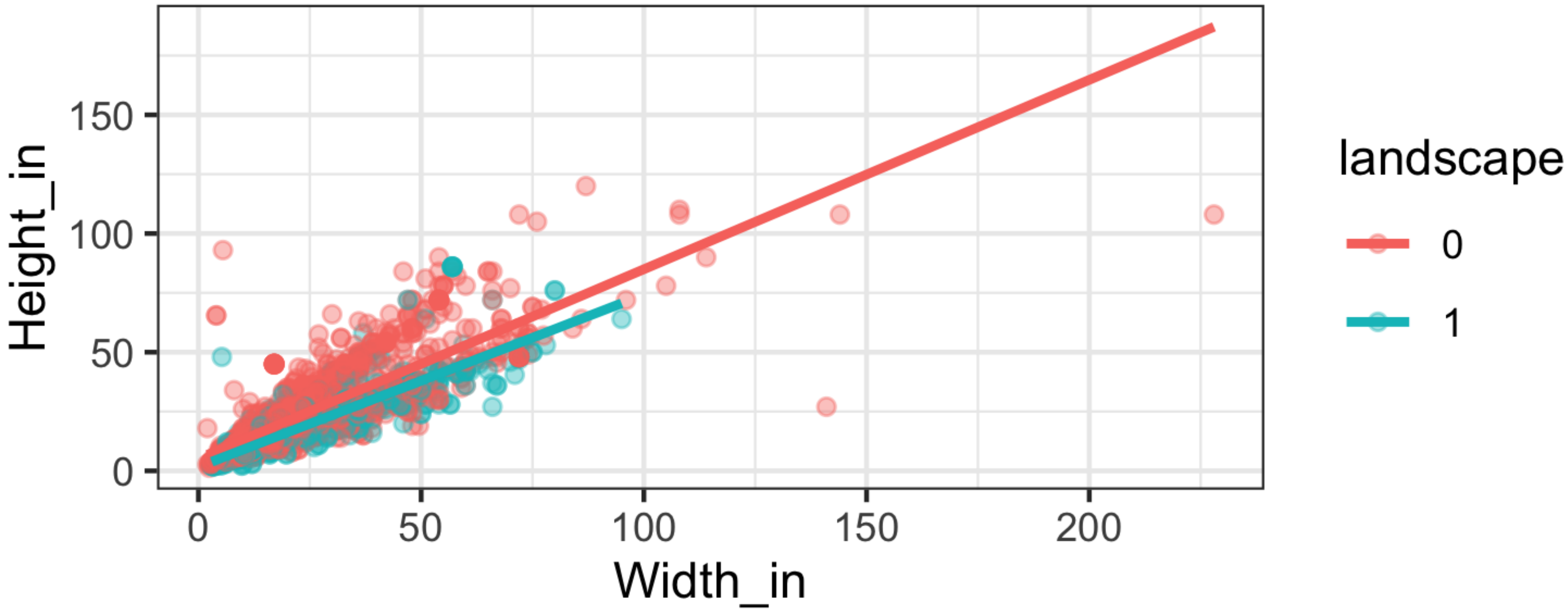


Landscape paintings

- Landscape painting is the depiction in art of landscapes – natural scenery such as mountains, valleys, trees, rivers, and forests, especially where the main subject is a wide view – with its elements arranged into a coherent composition.[1]
  - Landscape paintings tend to be wider than longer.
- Portrait painting is a genre in painting, where the intent is to depict a human subject.[2]
  - Portrait paintings tend to be longer than wider.

[1] Source: Wikipedia, Landscape painting [2] Source: Wikipedia, Portait painting

How, if at all, does the relationship between width and height of paintings vary by whether or not they have any landscape elements?



Models - upsides and downsides

- Models can sometimes reveal patterns that are not evident in a graph of the data. This is a great advantage of modeling over simple visual inspection of data.

- There is a real risk, however, that a model is imposing structure that is not really there on the scatter of data, just as people imagine animal shapes in the stars. A skeptical approach is always warranted.

Variation around the model...

is just as important as the model, if not more!

> *Statistics is the explanation of variation in the context of what remains unexplained.*

- The scatter suggests that there might be other factors that account for large parts of painting-to-painting variability, or perhaps just that randomness plays a big role.
- Adding more explanatory variables to a model can sometimes usefully reduce the size of the scatter around the model. (We'll talk more about this later.)

How do we use models?

1. **Explanation:** Characterize the relationship between y and x via *slopes* for numerical explanatory variables or *differences* for categorical explanatory variables

2. **Prediction:** Plug in x, get the predicted y

Interpreting Models

Packages

- You're familiar with the tidyverse:

```r
library(tidyverse)
```

- The broom package takes the messy output of built-in functions in R, such as **`lm`**, and turns them into tidy data frames.

```r
library(broom)
```

broom

- **broom** follows tidyverse principles and tidies up regression output
- **`tidy`**: Constructs a tidy data frame summarizing model's statistical findings
- **`glance`**: Constructs a concise one-row summary of the model
- **`augment`**: Adds columns (e.g. predictions, residuals) to the original data that was modeled

https://broom.tidyverse.org/

Height & width

```
m_ht_wt <- lm(Height_in ~ Width_in, data = paris_paintings)
tidy(m_ht_wt)
```

```
## # A tibble: 2 x 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.
```

Height & width

```
m_ht_wt <- lm(Height_in ~ Width_in, data = paris_paintings)
tidy(m_ht_wt)
```

```
## # A tibble: 2 x 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.
```

$$\widehat{Height_{in}} = 3.62 + 0.78\ Width_{in}$$

Height & width

```
m_ht_wt <- lm(Height_in ~ Width_in, data = paris_paintings)
tidy(m_ht_wt)
```

```
## # A tibble: 2 x 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.
```

$$\widehat{Height_{in}} = 3.62 + 0.78\ Width_{in}$$

- **Slope**: For each additional inch the painting is wider, the height is expected to be higher, on average, by 0.78 inches.

Height & width

```r
m_ht_wt <- lm(Height_in ~ Width_in, data = paris_paintings)
tidy(m_ht_wt)
```

```
## # A tibble: 2 x 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.
```

$$\widehat{Height_{in}} = 3.62 + 0.78\ Width_{in}$$

- **Slope**: For each additional inch the painting is wider, the height is expected to be higher, on average, by 0.78 inches.
- **Intercept**: Paintings that are 0 inches wide are expected to be 3.62 inches high, on average.

The linear model with a single predictor

- We're interested in the β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

$$\hat{y} = \beta_0 + \beta_1 \ x$$

The linear model with a single predictor

- We're interested in the β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

$$\hat{y} = \beta_0 + \beta_1 \; x$$

- Unfortunately, we can't get these values

The linear model with a single predictor

- We're interested in the β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

$$\hat{y} = \beta_0 + \beta_1~x$$

- Unfortunately, we can't get these values

- So we use sample statistics to estimate them:

$$\hat{y} = b_0 + b_1~x$$

Least squares regression

The regression line minimizes the sum of squared residuals.

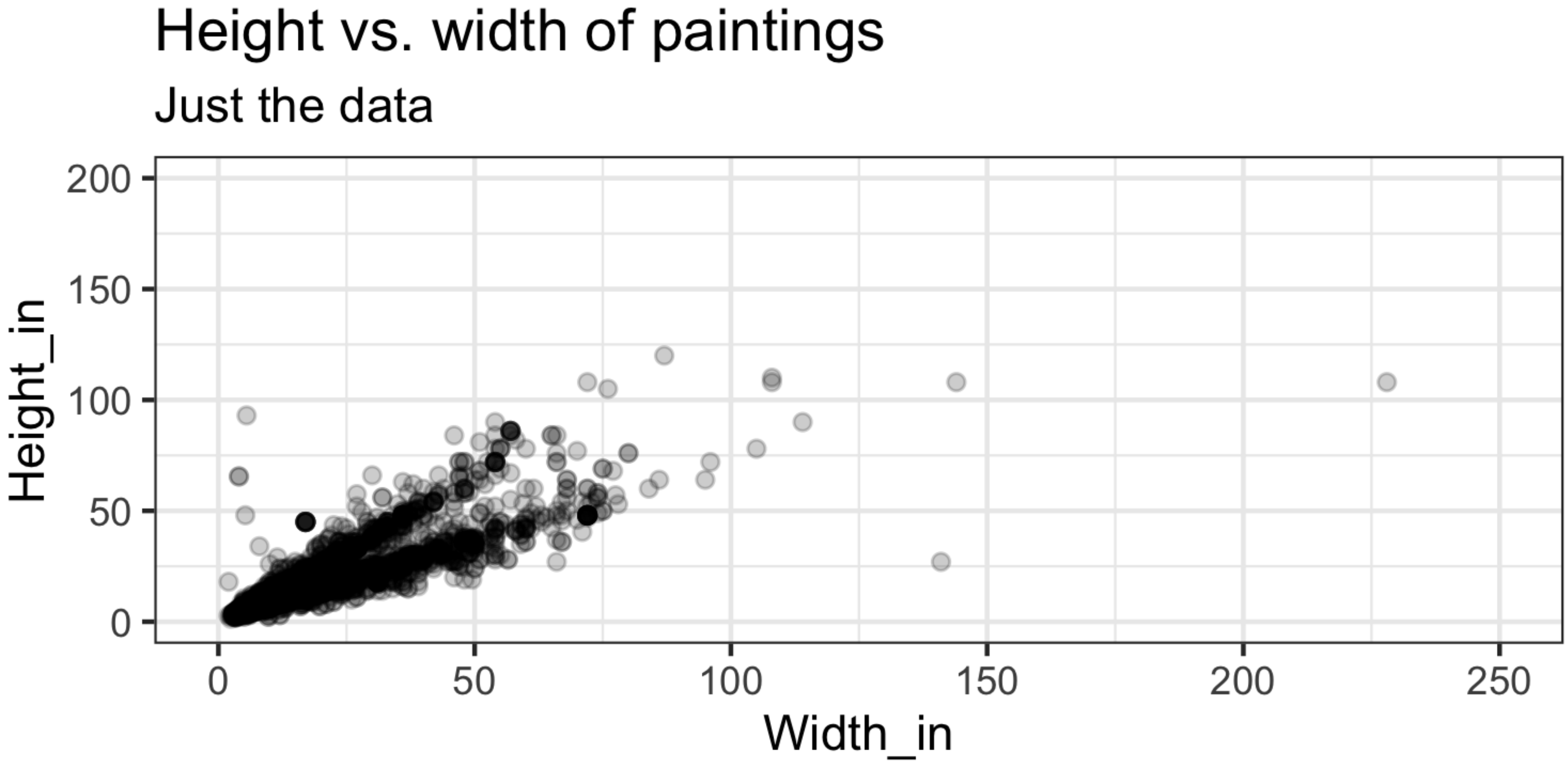
Least squares regression

The regression line minimizes the sum of squared residuals.

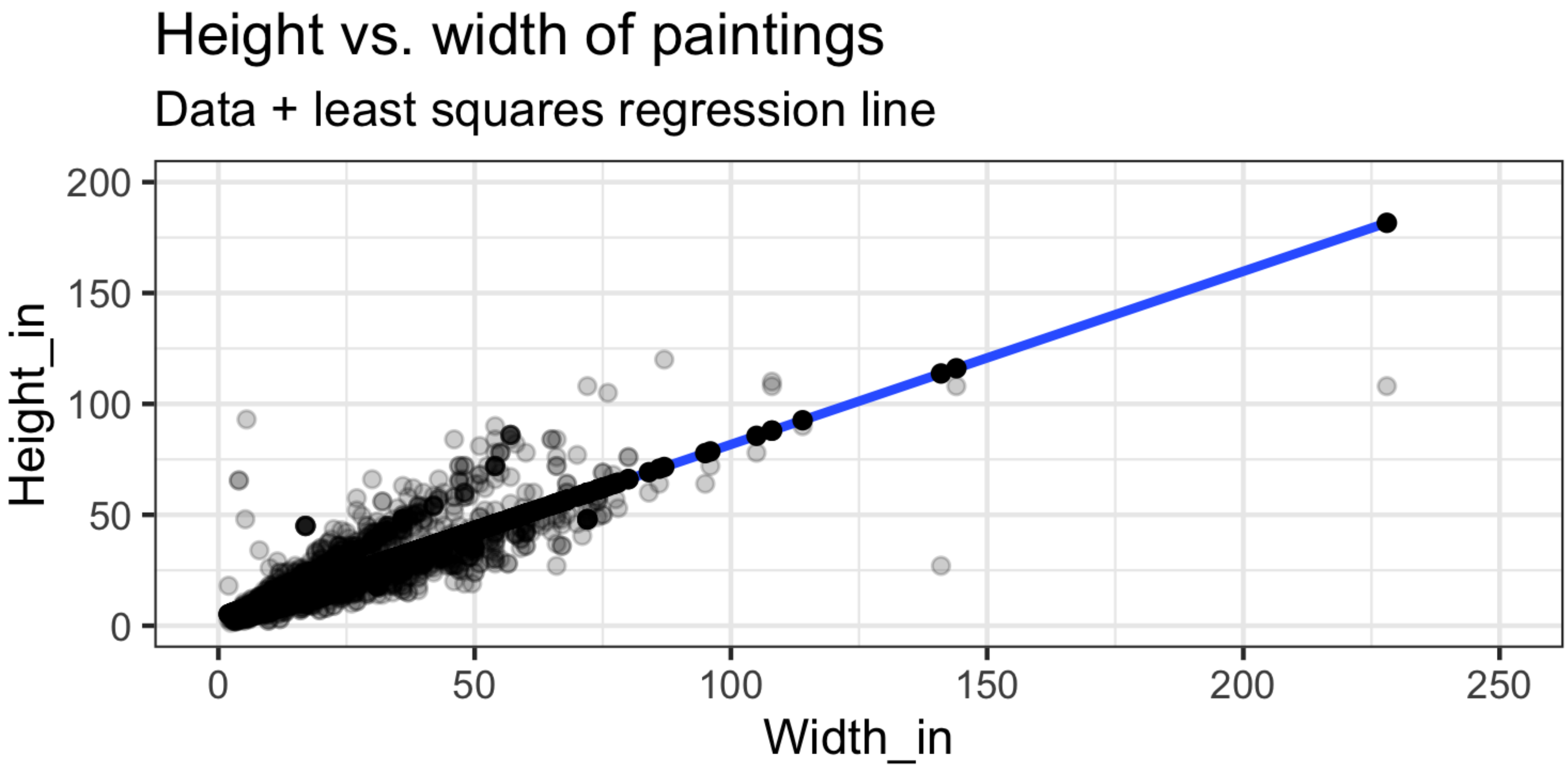
- **Residuals**: $e_i = y_i - \hat{y}_i$,
- The regression line minimizes $\sum_{i=1}^n e_i^2$.
- Equivalently, minimizing $\sum_{i=1}^n [y_i - (b_0 + b_1~x_i)]^2$

Why do we minimize the *squares* of the residuals?

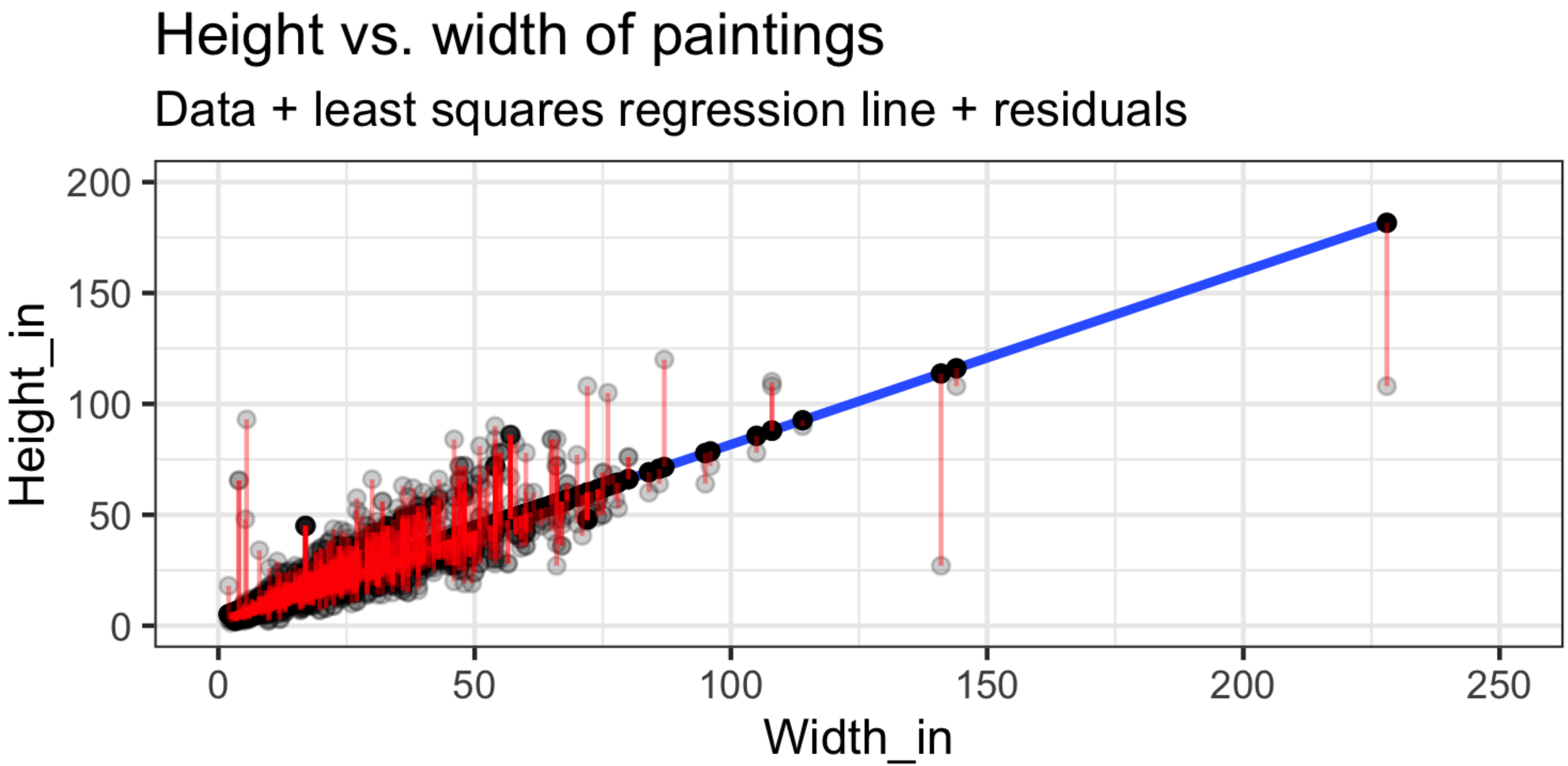
Visualizing residuals



Visualizing residuals (cont.)



Visualizing residuals (cont.)



Properties of the least squares regression line

- The estimate for the slope, b_1, has the same sign as the correlation between the two variables.

- The regression line goes through the center of mass point, the coordinates corresponding to average x and average y: $(\bar{x}, \bar{y})$

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0$$

- The residuals and x values are uncorrelated.

Categorical Predictors

What about non-continuous predictors?

Height & landscape features

```r
m_ht_lands <- lm(Height_in ~ factor(landsALL), data = paris_paintings)
tidy(m_ht_lands)
```

```
## # A tibble: 2 x 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
```

What about non-continuous predictors?

Height & landscape features

```
m_ht_lands <- lm(Height_in ~ factor(landsALL), data = paris_paintings)
tidy(m_ht_lands)
```

```
## # A tibble: 2 x 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
```

$$\widehat{Height_{in}} = 22.68 - 5.65\ landsALL$$

(cont.)

$$\widehat{Height_{in}} = 22.68 - 5.65\ landsALL$$

- **Slope:** Paintings with landscape features are expected, on average, to be 5.65 inches shorter than paintings that without landscape features.
    - Compares baseline level (`landsALL = 0`) to other level (`landsALL = 1`).
- **Intercept:** Paintings that don't have landscape features are expected, on average, to be 22.68 inches tall.

Categorical predictor with 2 levels

```
## # A tibble: 8 x 3
##   name      price landsALL
##   <chr>     <dbl>    <dbl>
## 1 L1764-2     360        0
## 2 L1764-3       6        0
## 3 L1764-4      12        1
## 4 L1764-5a      6        1
## 5 L1764-5b      6        1
## 6 L1764-6       9        0
## 7 L1764-7a     12        0
## 8 L1764-7b     12        0
```

Categorical predictors with more than 2 levels

```
m_ht_sch <- lm(Height_in ~ school_pntg, data = paris_paintings)
tidy(m_ht_sch)
```

```
## # A tibble: 7 x 5
##   term            estimate std.error statistic p.value
##   <chr>              <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)        14.        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
```

What do these rows correspond to? Why are there only six schools listed, but seven schools total (what happened to the Austrian school?)

Categorical predictors with more than 2 levels

```
## # A tibble: 7 x 5
##   term              estimate std.error statistic p.value
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          14.        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
```

- When the categorical explanatory variable has many levels, the levels are encoded to dummy variables

- Each coefficient describes the expected difference between heights in that particular school compared to the baseline level.

How dummy variables are made

```
## # A tibble: 7 x 7
## # Groups:   school_pntg [7]
##   school_pntg  D_FL     F     G     I     S     X
##   <chr>       <int> <int> <int> <int> <int> <int>
## 1 A               0     0     0     0     0     0
## 2 D/FL            1     0     0     0     0     0
## 3 F               0     1     0     0     0     0
## 4 G               0     0     1     0     0     0
## 5 I               0     0     0     1     0     0
## 6 S               0     0     0     0     1     0
## 7 X               0     0     0     0     0     1
```

Correlation does not imply causation!!

Remember this when interpreting model coefficients

Prediction with models

Predict height from width

On average, how tall are paintings that are 60 inches wide?

$$\widehat{Height_{in}} = 3.62 + 0.78\ Width_{in}$$

Predict height from width

On average, how tall are paintings that are 60 inches wide?

$$\widehat{Height_{in}} = 3.62 + 0.78\ Width_{in}$$

```
3.62 + 0.78 * 60
```

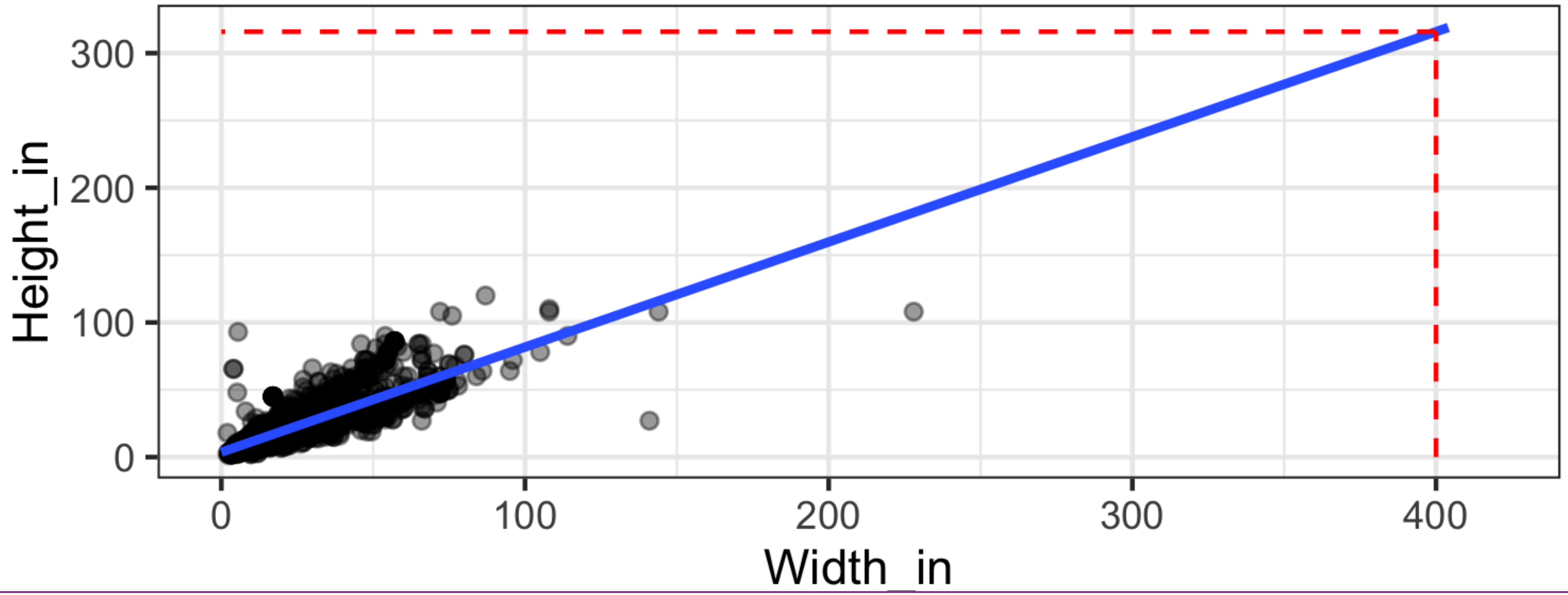
```
## [1] 50.42
```

"On average, we expect paintings that are 60 inches wide to be 50.42 inches high."

Warning: We "expect" this to happen, but there will be some variability.

Prediction vs. extrapolation

On average, how tall are paintings that are 400 inches wide?



Watch out for extrapolation!

> "When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on."[1]
> Stephen Colbert, April 6th, 2010

[1] OpenIntro Statistics. "Extrapolation is treacherous." OpenIntro Statistics.

Tidy vs. not-so-tidy regression output

Let's revisit the model predicting heights of paintings from their widths:

```r
m_ht_wt <- lm(Height_in ~ Width_in, data = paris_paintings)
```

❌ Not-so-tidy regression output

- You might come across these in your googling adventures, but we'll try to stay away from them
- Not because they are wrong, but because they don't result in tidy data frames as results.

❌ Not-so-tidy regression output (1)

Option 1:

```
m_ht_wt
```

```
##
## Call:
## lm(formula = Height_in ~ Width_in, data = paris_paintings)
##
## Coefficients:
## (Intercept)     Width_in
##      3.6214       0.7808
```

❌ Not-so-tidy regression output (2)

Option 2:

```
summary(m_ht_wt)
```

```
##
## Call:
## lm(formula = Height_in ~ Width_in, data = paris_paintings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.714  -4.384  -2.422   3.169  85.084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.621406   0.253860   14.27   <2e-16
## Width_in    0.780796   0.009505   82.15   <2e-16
```

Review

What makes a data frame tidy?

Review

What makes a data frame tidy?

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

✅ Tidy regression output

Achieved with functions from the broom package:

- `tidy`: Constructs a data frame that summarizes the model's statistical findings. We've talked about coefficient estimates and standard errors, but it also displays *test statistics and p-values* (more on these in a few weeks!).

- `augment`: Adds columns to the original data that was modeled. This includes predictions and residuals.

- `glance`: Constructs a concise one-row summary of the model, computed once for the entire model.

✅ Tidy your model's statistical findings

```
tidy(m_ht_wt)
```

```
## # A tibble: 2 x 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.
```

✅ Tidy your model's statistical findings

```
tidy(m_ht_wt) %>%
  select(term, estimate) %>%
  mutate(estimate = round(estimate, 3))
```

```
## # A tibble: 2 x 2
##   term        estimate
##   <chr>          <dbl>
## 1 (Intercept)    3.62
## 2 Width_in       0.781
```